



SDS
Statistica e
Data Science



Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia - Leonardo Egidì - Antonino
Abbruzzo

—

Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

Questo volume è rilasciato sotto licenza Creative Commons
Attribuzione - Non commerciale - Non opere derivate 4.0



© 2024 The Authors

Contents

1	Keynote Sessions	11
1.1	Classification with imbalanced data and the (eternal?) struggle between statistics and data science. <i>Nicola Torelli</i>	11
1.2	Deep residual networks and differential equations. <i>Gérard Biau</i>	13
2	Invited - Complex data: new methodologies and applications	15
2.1	Link selection in binary regression models with the Model Confidence Set. <i>Michele La Rocca and Marcella Niglio and Marialuisa Restaino</i>	15
2.2	A cluster-weighted model for COVID-19 hospital admissions. <i>Daniele Spinelli, Paolo Berta, Salvatore Ingrassia and Giorgio Vittadini</i>	23
2.3	Multi-class text classification of news data. <i>Maurizio Romano and Maria Paola Priola</i>	28
3	Invited - Data science and dataspace: challenges, results, and next steps	35
3.1	Data-Centric AI : A new Frontier emerging in Data Science. <i>Donato Malerba, Vincenzo Pasquadibisceglie, Vito Recchia and Annalisa Appice</i>	35
3.2	Data Spaces strategy to unleash agriculture data value: a concrete use case. <i>Nicola Masi, Delia Milazzo, Giulia Antonucci and Susanna Bonura</i>	42
3.3	Addressing Agricultural Data Management Challenges with the Enhanced TRUE Connector. <i>Sergio Comella, Delia Milazzo, Mattia Giuseppe Marzano, Giulia Antonucci, Susanna Bonura and Angelo Marguglio</i>	48
4	Solicited - Data Science for Official Statistics	55
4.1	Data science at Istat for urban green. <i>Fabrizio De Fausti, Marco Di Zio, Giuseppe Lancioni, Stefano Mugnoli, Alberto Sabbi and Francesco Sisti</i>	55

4.2	Twitter (X) as a Data Source for Official Statistics: Monitoring Italian Debate on Immigration through Text Analysis. <i>Elena Catanese, Gerarda Grippo, Francesco Ortame and Maria Clelia Romano</i>	62
5	Solicited - Sustainable Artificial Intelligence in Finance	69
5.1	Feature Dependence and Prediction Explanations in P2P Lending. <i>Paolo Pagnottoni and Thanh Thuy Do</i>	69
6	Solicited - Young SIS	77
6.1	Merging data and historical information via optimal power prior selection in Bayesian models. <i>Roberto Macrì Demartino, Leonardo Egidi, Nicola Torelli and Ioannis Ntzoufras</i>	77
6.2	Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables. <i>Dalila Failli, Bruno Arpino, and Maria Francesca Marino</i>	84
6.3	A Simultaneous Spectral Clustering for Three-Way Data. <i>Cinzia Di Nuzzo and Salvatore Ingrassia</i>	90
7	Solicited - From Data Analysis to Data Science	97
7.1	Optimal Scaling: New Insights Into an Old Problem. <i>Gilbert Saporta</i>	97
8	Solicited - Statistical methods for textual data	101
8.1	PROCSIMA: Probability Distribution Clustering Using Similarity Matrix Analysis. <i>Marco Ortu</i>	101
8.2	Exploring Anti-Migrant Rhetoric on Italian Social Media. <i>Lara Fontanella, Annalina Sarra, Emiliano del Gobbo, Alex Cucco and Sara Fontanella</i>	108
8.3	Causal inference from texts: a random-forest approach. <i>Chiara Di Maria, Alessandro Albano, Mariangela Sciandra and Antonella Plaia</i>	114
9	Solicited - Data analysis methods for data in non-Euclidean spaces	121
9.1	Riemannian Statistics for Any Type of Data. <i>Oldemar Rodriguez Rojas</i>	121
9.2	PAM clustering algorithm for ATR-FTIR spectral data selection: an application to multiple sclerosis. <i>Francesca Condino, Maria Caterina Crocco and Rita Guzzi</i>	128
9.3	Random Survival Forest for Censored Functional Data. <i>Giuseppe Loffredo, Elvira Romano and Fabrizio Maturo</i>	134
9.4	Advancing credit card fraud detection with innovative class partitioning and feature selection technique. <i>Mohammed Sabri, Antonio Balzanella and Rosanna Verde</i>	140
10	Solicited - Functional Data Analysis in Action	147
10.1	Functional Linear Discriminant Analysis for Misaligned Surfaces. <i>Tomas Masak</i>	147
10.2	Leveraging weighted functional data analysis to estimate earthquake-induced ground motion. <i>Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio</i>	155

10.3	Functional autoregressive processes on a spherical domain for global aircraft-based atmospheric measurements. <i>Almond Stöcker and Alessia Caponera</i>	161
11	Solicited - Bayesian Inference for Graphical Models	169
11.1	Log-likelihood approximation in Stochastic EM for Multilevel Latent Class Models. <i>Silvia Columbu, Nicola Piras and Jeroen K. Vermunt</i>	169
11.2	MCMC Sampling in Bayesian Gaussian Structure Learning. <i>Antonino Abbruzzo, Nicola Argentino, Reza Mohammadi, Maria De Iorio, Willem van den Boom and Alexandros Beskos</i>	176
12	Contributed - Promoting Equity: Statistical Insights into Tourism, Sustainability and Digital Divide	183
12.1	Lesson Learnt in the Data Science Worldview: New Dimension of Digital Divide. <i>Rita Lima</i>	183
12.2	An overview of Tourism Statistical Literacy. <i>Yasir Jehan, Giuseppina Lo Mascolo and Stefano De Cantis</i>	192
12.3	Scalable bootstrap inference via averaged Robbins-Monro approximations. <i>Giuseppe Alfonzetti and Ruggero Bellio</i>	198
12.4	The impact of sustainability on Initial Coin Offering: advantages in trading. <i>Alessandro Bitetto and Paola Cerchiello</i>	204
13	Contributed - High dimensional and functional data	211
13.1	Analysis of Brain-Body Physiological Rhythm Using Functional Graphical Models. <i>Rita Fici, Luigi Augugliaro and Ernst C. Wit</i>	211
13.2	A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects. <i>Ruggero Bellio and Cristiano Varin</i>	218
13.3	Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models. <i>Antonino Gagliano, Gianluca Sottile, Nicolina Sciaraffa, Claudia Coronello and Luigi Augugliaro</i>	224
13.4	Investigating the association between high school outcomes and university enrollment choice: a machine learning approach. <i>Andrea Priulla, Alessandro Albano, Nicoletta D'Angelo and Massimo Attanasio</i>	230
14	Contributed - Statistical Analysis in economic and market dynamics	237
14.1	A comparison of multi-factor stochastic models for commodity prices C3. <i>Luca Vincenzo Ballestra, Christian Tezza and Paolo Foschi</i>	237
14.2	Nonparametric ranking estimation with application to the propensity for Circular Economy of Italian economic sectors. <i>Stefano Bonnini, Michela Borghesi and Massimiliano Giacalone</i>	246
14.3	Impact of the Russian invasion of Ukraine on coal markets: Evidence from an event-study approach. <i>Yana Kostiuk, Paola Cerchiello and Arianna Agosto</i>	252
14.4	Labour market and time series: a forecast approach for European countries from 1995 to 2022. <i>Paolo Mariani, Andrea Marletta and Piero Quatto</i>	258

15	Contributed - Innovations in cluster and latent class models . . .	263
15.1	Biclustering of discrete data by extended finite mixtures of latent trait models. <i>Dalila Failli, Maria Francesca Marino and Francesca Martella</i>	263
15.2	Seismic events classification through latent class regression models for point processes. <i>Giada Lo Galbo, Giada Adelfio and Marcello Chiodi</i>	270
15.3	Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization. <i>Agostino Stavolo, Maria Gabriella Grassia, Marina Marino and Rocco Mazza</i>	276
16	Contributed - Modelling on spatial phenomena	283
16.1	Integrating computational and statistical algorithms in RT-GSCS for spatial survey administration. <i>Yuri Calleo, Simone Di Zio and Francesco Pilla</i>	283
16.2	Sensitivity mapping as a tool to support siting of offshore wind farms and increase citizens' acceptability. <i>Giovanna Cilluffo, Gianluca Soffile, Laura Ciriminna, Geraldina Signa, Agostino Tomasello and Salvatrice Vizzini</i>	290
16.3	Investigating hotel consumers' purchase intention on web analytics data through PLS-SEM. <i>Giuseppina Lo Mascolo, Chiara di Maria, Marcello Chiodi and Arabella Mocciano Li Destri</i>	296
16.4	Spatio-temporal analysis of lightning point process data in severe storms. <i>Nicoletta D'Angelo, Milind Sharma, Marco Tarantino and Giada Adelfio</i>	302
17	Contributed - Statistical machine learning for predictive modelling	309
17.1	Application of statistical techniques to predict the effective temperature of young stars. <i>Marco Tarantino, Loredana Prinszano and Giada Adelfio</i>	309
17.2	Topological Attention for Denoising Astronomical Images. <i>Riccardo Cecaroni and Pierpaolo Brutti</i>	316
17.3	LSTM-based Battery Life Prediction in IoT Systems: a proof of concept. <i>Vanessa Verrina, Andrea Vennera and Annarita Renda</i>	322
17.4	Predictive modeling of drivers' brake reaction time through machine learning methods. <i>Alessandro Albano, Giuseppe Salvo and Salvatore Rusotto</i>	328
18	Contributed - Ordinal and preference data analysis .	335
18.1	OSILA (Order Statistics In Large Arrays): an original algorithm for an efficient attainment of the order statistics. <i>Andrea Cerasa</i>	335
18.2	The Mallows model with respondents' covariates for the analysis of preference rankings. <i>Marta Crispino, Lucia Modugno and Cristina Mollica</i>	343
18.3	Value-Based Predictors of Voting Intentions: An Empirical and Explainable approach. <i>Luca Pennella and Amin Gino Fabbrucci Barbagli</i>	349
18.4	A dynamic version of the Massey's rating system with an application in basketball. <i>Paolo Vidoni and Enrico Bozzo</i>	355
19	Contributed - Advances in text mining	361
19.1	Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?. <i>Andrea Sciandra and Arjuna Tuzzi</i>	361

- 19.2 A Fuzzy Topic Modeling approach to legal corpora. *Antonio Calcagni and Arjuna Tuzzi* 368
- 19.3 EmurStat: a digital tool for statistical analysis of emur flow. *Simone Paesano, Maria Gabriella Grassia, Marina Marino, Dario Sacco and Rocco Mazza* 374
- 19.4 Graph Neural Networks for clustering medical documents. *Vittorio Torri and Francesca Ieva* 380

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data. Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science. In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized its 3rd international conference held in Palermo on the 11st and 12nd of April 2024, attended by over 100 researchers from different scientific fields. A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Palermo, April 11st and 12th, 2023

Antonella Plaia - Leonardo Egidi - Antonino Abbruzzo

Editors

1. Keynote Sessions

1.1 - Classification with imbalanced data and the (eternal?) struggle between statistics and data science

1.2 - Deep residual networks and differential equations

Classification with imbalanced data and the (eternal?) struggle between statistics and data science

Nicola Torelli

Abstract

The problem of classification with imbalanced data is recognized, among data science practitioners, as an issue of primary relevance since it occurs in many real situations and domains. When data on the dichotomous (or possibly polytomous) response variable are skewed, the classifiers typically used in statistical and machine learning could perform poorly. Several remedies to this problem have been proposed and among the most popular and successful methods are those that alter the class distribution by non-proportionally sampling data or by generating new data.

The analysis of some of the solutions will offer the opportunity to discuss how different scholars addressed the issue. First, it received much more attention from data scientists in the machine learning community than from statisticians who, nonetheless, gave important contributions. Also, the more popular solutions and methods for dealing with imbalanced data are rooted in the different methodological approaches of the two communities.

To illustrate the aforementioned points the algorithm ROSE, introduced in 2014 by Menardi and Torelli, will be reviewed along with some recent proposals for refining it, also borrowing some ideas from the machine learning community.

Key words: ROSE, SMOTE, over- and under-sampling, synthetic data generation, bootstrap.

Nicola Torelli

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Piazzale Europa 1, 34127, Trieste, Italy, e-mail: nicola.torelli@deams.units.it

Deep residual networks and differential equations

G rard Biau

Abstract Deep learning has become a prominent approach for many applications, such as computer vision or neural language processing. However, the mathematical understanding of these methods is still incomplete. A recent approach is to consider neural networks as discretized versions of differential equations. I will first give an overview of this emerging field and then discuss new results on residual neural networks, which are state-of-the-art deep learning models.

G rard Biau
Laboratoire de Probabilit s, Statistique et Mod lisation, Sorbonne Universit , Paris,
France, Bo te 158, tour 15-25, bureau 219, Paris. E-mail: gerard.biau@sorbonne-universite.fr

2. Invited - Complex data: new methodologies and applications

- 2.1 - Link selection in binary regression models with the Model Confidence Set
- 2.2 - A cluster-weighted model for COVID-19 hospital admissions
- 2.3 - Multi-class text classification of news data

Link selection in binary regression models with the Model Confidence Set

Michele La Rocca and Marcella Niglio and Marialuisa Restaino

Abstract The choice of the link function in Generalized Linear Regression models may be a key aspect when the dependent variable is binary and imbalanced. After the introduction of this class of models and a brief presentation of some asymmetric distribution functions that are used in the literature to obtain asymmetric links, the paper proposes the use of the Model Confidence Set as a tool to select, among candidate link functions, the “best set” for the data. A simulation study gives empirical evidence of the Model Confidence Set performance when asymmetric links are used to model imbalanced binary data.

Key words: GLM, symmetric and asymmetric links, Model Confidence Set.

1 Introduction

Generalized Linear Models (GLM) are regression models largely used in many empirical domains: economics, medicine, finance, education, etc. Their flexibility and attitude to treat data sets with the response variable distribution belonging to the one-parameter exponential family, make their use quite extensive.

Among them, binary response models, with Bernoulli dependent distribution, are very popular in the literature (see for example [5], [8]).

Michele La Rocca
Università degli Studi di Salerno Via Giovanni Paolo II, 132 - Fisciano (SA) Italy, e-mail:
larocca@unisa.it

Marcella Niglio
Università degli Studi di Salerno Via Giovanni Paolo II, 132 - Fisciano (SA) Italy, e-mail:
mniglio@unisa.it

Marialuisa Restaino
Università degli Studi di Salerno Via Giovanni Paolo II, 132 - Fisciano (SA) Italy, e-mail:
mlrestaino@unisa.it

Let Y be a binary response variable with Bernoulli distribution $P(Y = y) = \pi^y(1 - \pi)^{1-y}$, for $y = \{0, 1\}$. Let n be the size of the sample extracted from Y and $\mathbf{X} = (X_1, X_2, \dots, X_k)$ be a vector of k covariates, such that the expectation $E[Y_i | \mathbf{X}_i = \mathbf{x}_i] = \pi_i$, for $i = 1, 2, \dots, n$. The rationale of the GLMs is based on the selection of a monotone and differentiable function $g(\cdot)$ such that:

$$g(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta}, \quad (1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ is the $[(k+1) \times 1]$ parameters vector, with $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, and $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})'$ is the vector of covariates of unit i .

The function $g(\cdot)$, called *link function*, joins the response variable Y and the linear predictors included in $\mathbf{X}$:

$$P(Y_i = 1) \equiv \pi_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$$

where $g(\cdot)$ may be selected parametrically or estimated nonparametrically. Here, we consider the former case where the selection of $g(\cdot)$ is related to the probability π_i and the distribution function $F(\cdot)$:

$$\pi_i \equiv F(\mathbf{x}_i' \boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}).$$

We discuss how the proportion of zeroes and ones in the dependent variable (that defines its balance or imbalance) may be crucial to selecting the proper link function for the model (1).

In the presence of a balanced response variable (where the proportion of ones and zeroes is approximately the same), the Logit and Probit link functions are the most commonly used. On the contrary, when Y is imbalanced their use produces biased estimates of $\boldsymbol{\beta}$ [8] and then, from the practical point of view, inaccurate estimations of the probabilities π_i that may lead to inappropriate decisions.

Starting from these remarks, in Sect. 2 we briefly present some asymmetric link functions that may improve the accuracy of the estimations and the model fitting in the presence of an imbalanced response variable. Sect. 3 deals with the selection of the asymmetric link function and introduces the proposal of adopting the Model Confidence Set (MCS) introduced in [6], for reducing the number of candidate asymmetric functions. A simulation study in Sect. 4 shows the performance of the MCS, making more evident the advantages of using such an approach.

2 Asymmetric links

The introduction of asymmetric link functions has been differently faced in the literature and a large number of proposals have been presented in this domain. Among them, here the main attention is focused on the class of power distribution functions and the Generalized Extreme Values (GEV) distributions. For the latter class of models consolidated results are given in [2] and [9] whereas for the power distri-

bution class, the main results are presented in [1], [3], [4] and [7]. The issue faced with the asymmetric link functions is related to the inadequacy of the symmetric links (such as the logit link based on the logistic distribution) when the rates of ones and zeroes are not the same in the dependent variable. This imbalance requires the use of skewed distribution functions that can be characterized by positive skewness if the rate approaching 1 is faster than the rate to approach 0, and negatively skewed otherwise.

To give empirical evidence of the skewness of some distributions used in defining link functions to model binary imbalanced data, in Fig. 1 the Logistic distribution is compared to the GEV, the power Logistic, the power Normal, the power Cauchy, the power t and the Stukel distributions (see the references cited before for details on these distributions).

The common factor of all these distributions is that their asymmetry is related to a shape parameter that regulates the tail behavior and consequently the rate of their distribution functions to approach 0 or 1. Fig. 1 shows the different behavior, for the Logistic distribution, of the left tail of the GEV distribution, the Power Normal, and the Power t distributions with shape parameter 1.8. Furthermore, these last two distributions show the main differences, always with the Logistic case, on the right tail when the shape value is 0.7. It means that for these power distributions, the shape parameter < 1 or > 1 impacts the right and left tails respectively.

In the Stukel case, the asymmetric behavior in the tails is recognized when the two shape parameters have opposite signs.

The previous evaluations imply that the tails of the distribution functions affect the selection of the link function. This observation allows opening the discussion on an important issue related to the “link selection”. In this regard, in Sect. 3 we propose the use of the Model Confidence Set [6] that allows us to define a set of link functions that include the “best link” with a given probability.

3 Selection of the asymmetric link functions based on the MCS

Let $\mathcal{M}^0$ be a model set containing m_0 nested, nonnested, or also overlapping models and let L_j be the loss associated with each model j , for $j = 1, \dots, m_0$. Further, let:

$$d_{jr,i} = L_{j,i} - L_{r,i} \quad \text{for } j, r \in \mathcal{M}^0, \text{ and } i = 1, \dots, n, \quad (2)$$

be the variable obtained comparing two losses, with $\mu_{jr} = E[d_{jr,i}]$.

The set of superior models is defined as:

$$\mathcal{M}^* = \{j \in \mathcal{M}^0 : \mu_{jr} \leq 0, \forall r \in \mathcal{M}^0\}$$

where model j is superior to model r if $\mu_{jr} < 0$. In practice, the ranking of model j is higher than model r if its expected loss is the lowest.

The definition of $\mathcal{M}^*$ is based on a sequence of significance tests with the following hypothesis:

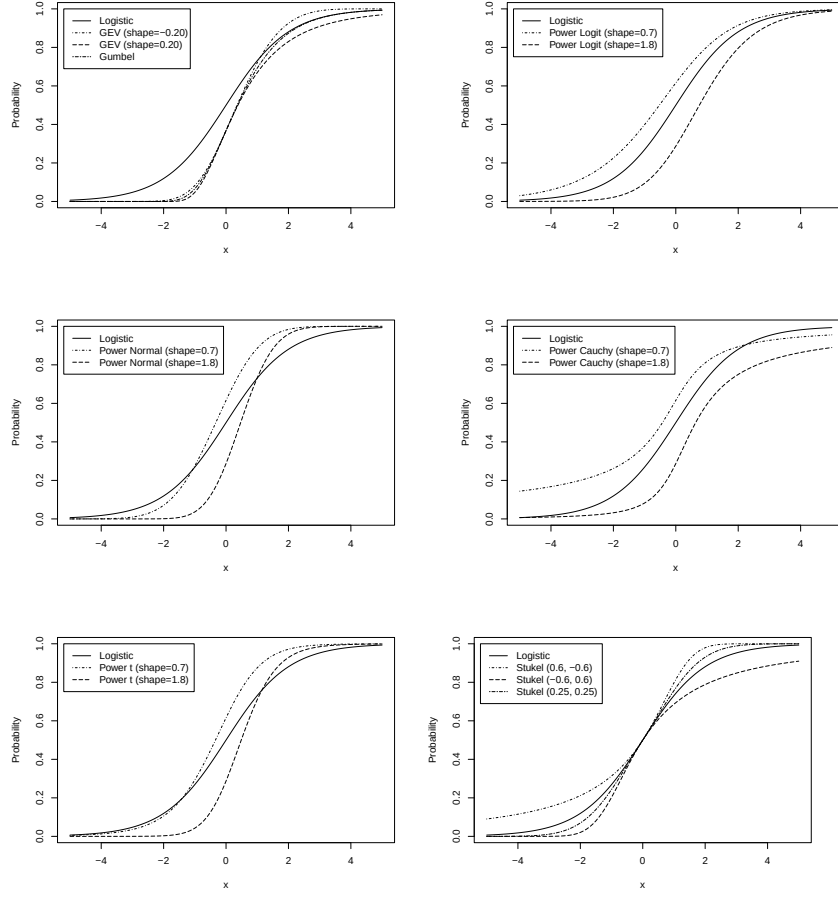


Fig. 1 GEV, Power Logit, Power Normal, Power Cauchy, Power T, and Stukel distribution functions compared to the Logistic distribution

$$H_0 : \mu_{jr} = 0 \quad \text{vs} \quad H_1 : \mu_{jr} \neq 0, \quad \forall j, r \in \mathcal{M}$$

with $\mathcal{M} \subset \mathcal{M}^0$. If H_0 is not rejected, the two models have equal expected loss. Finally, the MCS is defined by all models included in $\mathcal{M}^*$ with equal expected losses.

From (2) it is clear the key role of the loss function, which needs to be properly defined to discriminate among candidate link functions and to define $\mathcal{M}^*$.

In the following, we apply the MCS to select the set of link functions that can be properly used to estimate binary regression models with an imbalanced dependent variable. Given the novelty of the application, the use of the MCS is eval-

uated through a Monte Carlo study to assess its performance in this context. For this aim, the R package MCS, available on the CRAN repository, <http://cran.r-project.org/web/packages/MCS/index.html>, is employed.

4 Simulation study

To evaluate the performance of the MCS in selecting the set of candidate link functions for GLMs, we implement a Monte Carlo study with 1000 replicates where we generate the data from the GEV regression model with $k = 5$ covariates, obtained from independent standard Normal distributions. The binary variable, Y , is generated as independent Bernoulli distributions such that the proportion of ones is imbalanced compared to the zeroes, and it is fixed to $p = \{0.05, 0.10, 0.20\}$. The vector of parameters β has values $\beta_j = 1/k$, for $j = 1, \dots, k$, and β_0 is set at different levels to guarantee the proportion of zeroes and ones given before. We replicate the simulation considering sample sizes, n , varying in the set $n = \{250, 500, 1000\}$.

Then we estimate, on all the generated datasets, the GLMs with link functions obtained from Logistic, Power Logit, and GEV (which include the Gumbel case) distributions.

Further to perform at each iteration the significance tests on which the MCS is based, we consider the following steps:

1. Generate the data with a training set of n units (namely $\mathcal{S}$) and a test set of $n/2$ units (namely $\mathcal{T}$);
2. Estimate the parameters of the four candidate models (Logit, Power Logit, GEV and Gumbel) on the training set, $\mathcal{S}$;
3. Given the vector $\hat{\beta}$, estimate the probabilities $\hat{\pi}_{j,i}$ for all four models ($j = 1, 2, 3, 4$), by using the data included in the test set, with $i \in \mathcal{T}$;
4. Define the MCS ($\mathcal{M}^*$) by using the differences $d_{jr,i} = L_{j,i} - L_{r,i}$, for $j, r = 1, 2, 3, 4$, $j \neq r$, and $i \in \mathcal{T}$, where $L_{j,i} = (\hat{\pi}_{j,i} - \pi_{j,i})^2$ and with $\pi_{j,i}$ the “true” probabilities of model j obtained on the test set by using the parameters β defined, as described before, to generate the artificial data (similar interpretation is given to $L_{r,i}$, for model r).

The estimation of the parameters is obtained by maximizing the corresponding log-likelihood. Further, to simplify the estimation of the shape parameter in the asymmetric link functions, a grid of values is chosen for it and then the model with the highest log-likelihood is selected.

The results of the MCS are summarized in Table 1. The values are the proportion of cases for which the MCS includes the GEV, the Logit, the PLogit, and the Gumbel as a candidate model for the data at different sample sizes (n) and with different proportions of ones (p). In all cases, the significance level is $\alpha = 0.05$.

It can be noted that as n grows the frequency at which $\mathcal{M}^*$ includes the link function based on the GEV distribution tends to $1 - \alpha$, with empirical levels close to the nominal level when $n = 1000$. As the imbalance of the dependent variable

Table 1 Proportion of cases for which the MCS (with $\alpha = 0.05$) includes the GEV, Logit, PLogit, and Gumbel link functions.

p	n	GEV	Logit	PLogit	Gumbel
0.20	250	0.865	0.754	0.693	0.179
	500	0.892	0.725	0.598	0.000
	1000	0.930	0.679	0.547	0.000
0.10	250	0.893	0.740	0.749	0.015
	500	0.913	0.652	0.700	0.000
	1000	0.952	0.620	0.629	0.000
0.05	250	0.871	0.713	0.835	0.000
	500	0.901	0.625	0.767	0.000
	1000	0.945	0.611	0.702	0.001

is higher (and consequently p decreases) the performance of the MCS (evaluated as the proportion of cases where the MCS includes the true GEV model) improves. When p goes from 0.20 to 0.05 and n increases, the frequency of MCS's that include the Logit model decreases (as expected) whereas the proportion of PLogit increases. However, the proportion of Logit models included in the final MCS remains high. Therefore, when the sample size is small with imbalanced data sets, it may be difficult to make a clear distinction between symmetric and alternative asymmetric links.

These are the earliest results of using the MCS to select a set of link functions among m_0 candidates. More evaluations need to be done on the selection of the loss which is a critical issue that can further improve the MCS performance in selecting the asymmetric link.

References

1. Bazán, J. L., Torres-Avilés, F., Suzuki, A. K., Louzada, F.: Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Appl. Stochastic Models Bus. Ind.* **33**, 22–34 (2017)
2. Calabrese, R., Osmetti, S.: Modelling SME loan defaults as rare events: An application to credit defaults. *J. Appl. Stat.* **40**, 1172–1188 (2013).
3. Chen, M.H., Dey, D.K., Shao, Q.M.: A New Skewed Link Model for Dichotomous Quantal Response Data. *J. Am. Stat. Assoc.* **94**, 1172–1186 (1999)
4. de la Cruz Huayanay, A., Bazán, J. L., Cancho, V.G., Dey, D.K.: Performance of asymmetric links and correction methods for imbalanced data in binary regression. *J. Stat. Comput. Simul.* **89**, 1694–1714 (2019)
5. Dobson, A.J., Barnett, A.G.: *An Introduction to Generalized Linear Models* (3rd ed). CRC Press, Boca Raton (2008)
6. Hansen, P.R., Lunde, A., Nason, J.M.: The model confidence set. *Econometrica* **79**, 453–497 (2011)
7. Lemonte A.J., Bazán, J. L.: New links binary regression: an application to coca cultivation in Peru. *Test* **27**, 597–617 (2018)
8. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Chapman Hall, London (1989)

9. Wang, X., Dey, D.K.: Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *Ann. Appl. Stat.* **4**, 2000–2023 (2010).

A cluster-weighted model for COVID-19 hospital admissions

Daniele Spinelli, Paolo Berta, Salvatore Ingrassia and Giorgio Vittadini

Abstract We propose a cluster-weighted model to analyze the mortality and the latent heterogeneity of COVID-19 patients. We focus on administrative data collected during in the earliest phases of the COVID-19 pandemic. Results highlight that a model-based clustering approach is helpful to detect unobserved clusters of COVID-19 patients.

Key words: Cluster-Weighted Models, COVID-19, clustering, administrative data

1 Introduction

Several studies have focused on the characteristics underlying COVID-19 mortality. Contributing causes of death by COVID-19 have been investigated using different statistical approaches to highlight the effect on mortality of age, gender, comorbidities, time of onset of illness, and other characteristics. What seems to dominate from a statistical point of view, is the large heterogeneity of the populations affected by COVID-19 and the difficulty in identifying subpopulations when classical statistical

Daniele Spinelli

Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: daniele.spinelli@unimib.it

Paolo Berta

Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: paolo.bera@unimib.it

Salvatore Ingrassia

Department of Economics and Business , University of Catania e-mail: salvatore.ingrassia@unict.it

Giorgio Vittadini

Department of Statistics and Quantitative Methods, University of Milano-Bicocca e-mail: giorgio.vittadini@unimib.it

approaches, such as generalized linear models, are adopted. The most common empirical strategies assume that the population affected by COVID-19 is homogeneous. However, some studies have considered latent heterogeneity in relation to COVID-19 [3, 6, 8]. These studies have also used clustering methods to reveal unobserved subpopulations in COVID-19 patients. The first study [3] used unsupervised clustering techniques on clinical data to identify three subgroups of pediatric patients and identified the characteristics of a multisystem inflammatory syndrome in children phenotype. Moreover, [6] used unsupervised cluster analysis on Spanish hospital records to identify subgroups of patients exhibiting post-COVID symptoms. The obtained clusters are to be used to assign therapeutic interventions. Further, [8] used a two-step method combining clustering and generalized linear models. In the first step, they used a k -means algorithm to reveal clusters of COVID-19 patients on a longitudinal dataset of biomarkers measurements. In the second stage, they used the obtained clusters as covariates to predict mortality using logistic regression. Their purpose is similar to ours, as they relate the clustering to mortality. However, we employ a different modeling technique.

Specifically, we propose a clustering framework to study COVID-19 mortality adopting a cluster-weighted model (CWM; [4, 5]). Compared to the above mentioned unsupervised methods, this supervised method has the advantage to model the latent heterogeneity, the outcomes of interest and the covariates affecting the outcome simultaneously. This also means, in contrast to the estimating procedure used in [8], that the effects of the covariates on mortality is allowed to vary within each cluster instead of being assumed independent of the latent class structure. Indeed, the CWM does not rely on the assignment independence assumption which states that the assignment of the data points to the latent clusters is independent from the covariates distribution [7].

2 Data

We consider administrative data regarding 2,617 Covid-19 hospitalizations occurred in the period from January to June 2020 at the Spedali-Civili hospital in Brescia, Italy [1]. The data include demographic characteristics, clinical information (diagnoses according to ICD-9-CM classification) and admission characteristics such as admission date, discharge date, in-hospital mortality, unit of admission and procedures performed during the hospitalization.

3 Methods

We assume a sample $(x_1, y_1), \dots, (x_n, y_n)$ concerning a response variable Y and a set of covariates X . The sample is hypothesized to come from a heterogeneous population formed by K latent classes. The CWM models the density of (Y, X) as outlined

by Eq. 1

$$p(x, y, \theta) = \sum_{j=1}^K \pi_j d(y|x; \beta_j) q(x; \eta_j). \quad (1)$$

In Eq. 1, π_j is the mixing proportion of latent class j , $d(y|x; \beta_j)$ is the class j -specific conditional density of the response variable and $q(x; \eta_j)$ is the marginal density of X in class j . Parameter vectors β_j and η_j are included in the generic vector of parameters θ to be estimated for all $j = 1, \dots, k$. Specifically, β_j is a vector of regression coefficients, and η_j depends on distributional assumptions regarding the covariates. Assuming $d(y|x; \beta_j) = 1$ in Eq. 1 leads to a mixture of distributions, while $q(x; \eta_j) = 1$ leads to a finite mixture of regressions (FMR).

Our response variable Y is a binary mortality indicator. We assume that Y follows a Bernoulli distribution with probability function $f(x; \beta_j)$, where x is a set of categorical and numeric risk-adjustment covariates, $f(\cdot)$ is the logistic function and β_j is a set of latent class-specific regression coefficients to be estimated. Therefore, $d(y|x; \beta_j)$, the conditional density of Y for latent class j (Eq. 1) is as follows:

$$d(y|x; \beta_j) = \left[f(x; \beta_j) \right]^y \left[1 - f(x; \beta_j) \right]^{1-y}. \quad (2)$$

In Eq. 1, the conditional density $q(x; \eta_j)$ depends on distributional assumptions on the covariates. The vector of covariates X includes age (continuous), sex (dichotomous), number of comorbidities (discrete) and period of admission in weeks from the start of 2020 (continuous). Patient characteristics are included to control for clinical pre-existing conditions, and they can be considered as a risk factor for in-hospital mortality. The period of admission is included as a proxy for the stress experienced by the healthcare system.

Period of admission and age are modeled according to a multivariate Gaussian distribution with mean vector μ_j and variance-covariance matrix Σ_j , considering Gaussian parsimonious models [2]. As for the number of comorbidities, we assume a Poisson distribution with mixture component-specific mean λ_j , while sex is assumed to be Bernoulli distributed with probability ψ_j . Hence, in keeping with the notation in Eq. 1, the expression for the marginal density of the covariates is the following:

$$\begin{aligned} q(x; \eta_j) &= q(x; \mu_j, \Sigma_j, \lambda_j, \psi_j) = \\ &= \phi(\text{Age}, \text{Period}; \mu_j, \Sigma_j) \times \frac{\lambda_j^{\text{Comorb}} e^{-\lambda_j}}{\text{Comorb}!} \times (\psi_j)^{\text{Sex}} (1 - \psi_j)^{1-\text{Sex}} \end{aligned} \quad (3)$$

where, $\phi(\cdot)$ is the bivariate Gaussian probability density function. The eigenvalue decomposition of Σ_j leads to 14 possible parametrizations. The model obtained by combining Eqs. 1, 2 and 3 is estimated in Stata by the EM algorithm implemented in the `cwmglm` package [9]. We consider competing CWMs and FMRs with varying number of latent clusters and different parsimonious model specification on the multivariate Gaussian covariates and deselect them using the Akaike information

criterion (AIC) and Bayesian information criterion (BIC). We cluster observations according to their maximum a posteriori probability.

4 Results

Our preferred model, a CWM with $k = 3$, is used to identify different profiles in terms of covariates (age, sex, period of admission and comorbidities) and mortality risks. The first group ($\pi_1 = 0.07$) is composed mainly by the oldest people (average age 74 years), who have been hospitalized later, on average. Patients in group 2 ($\pi_2 = 0.79$) are on average aged 70 years old and have been admitted earlier. Group 3 ($\pi_3 = 0.14$) is characterized by the lowest mean age (46) and by the lowest mortality. The impact of covariates on mortality differs from what obtained in [8], where the chosen statistical method could only estimate a single odds ratio for age and sex after controlling for clustering assignment. Our modeling framework is more flexible as the risk related to the covariates is allowed to be cluster-dependent. Consistent with [8], our estimation reveals a positive effect of age on mortality in all of the clusters, but the magnitude is different. The odds ratio are 1.04, 1.10 and 1.24; their estimate is 1.08 after controlling for cluster membership.

References

- [1] Berta, P., Ingrassia, S., Vittadini, G., Spinelli, D.: Latent heterogeneity in covid-19 hospitalisations: a cluster-weighted approach to analyse mortality. *Aust NZ J Stat* (2024)
- [2] Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. *Pattern Recognit.* **28**(5), 781–793 (1995)
- [3] Geva, A., Patel, M.M., Newhams, M.M., Young, C.C., Son, M.B.F., Kong, M., Maddux, A.B., Hall, M.W., Riggs, B.J., Singh, A.R., et al.: Data-driven clustering identifies features distinguishing multisystem inflammatory syndrome from acute covid-19 in children and adolescents. *EClinicalMedicine* **40** (2021)
- [4] Ingrassia, S., Minotti, S., Vittadini, G.: Local statistical modeling via the cluster-weighted approach with elliptical distributions. *J. Classif.* **29**(3), 363–401 (2012)
- [5] Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S.: The generalized linear mixed cluster-weighted model. *J. Classif.* **32**(1), 85–113 (2015)
- [6] Fernández-de Las-Peñas, C., Martín-Guerrero, J.D., Florencio, L.L., Navarro-Pardo, E., Rodríguez-Jiménez, J., Torres-Macho, J., Pellicer-Valero, O.J.: Clustering analysis reveals different profiles associating long-term post-covid symptoms, covid-19 symptoms at hospital admission and previous medical comorbidities in previously hospitalized covid-19 survivors. *Infection* **51**(1), 61–69 (2023)

- [7] Punzo, A., Ingrassia, S., Maruotti, A.: Multivariate generalized hidden markov regression models with random covariates: physical exercise in an elderly population. *Stat. Med.* **37**(19), 2797–2808 (2018)
- [8] San-Cristobal, R., Martín-Hernández, R., Ramos-Lopez, O., Martinez-Urbistondo, D., Micó, V., Colmenarejo, G., Villares Fernandez, P., Daimiel, L., Martínez, J.A.: Longwise cluster analysis for the prediction of covid-19 severity within 72 h of admission: Covid-data-save-lives cohort. *J. Clin. Med.* **11**(12), 3327 (2022)
- [9] Spinelli, D., Ingrassia, S., Vittadini, G.: cwmglm: Stata module to estimate Cluster Weighted Models (CWM) (2022). URL <https://ideas.repec.org/c/boc/bocode/s459090.html>

Multi-class text classification of news data

Maurizio Romano and Maria Paola Priola

Abstract Several Multi-class text classification (MCC) strategies, namely One-Vs-Rest (OVA), One-Vs-One (OVO), Best-of-Best (BOB), and Error-Correcting-Output-Codes (ECOC), are compared in terms of accuracy and computational efficiency. Each strategy is implemented utilizing several classifiers such as Naïve Bayes, Random Forest, Logistic Regression, Neural Networks, Linear Discriminant Analysis, Support Vector Machine, and the recently-introduced Threshold-based Naïve Bayes (Tb-NB). We run a horse race involving the analysis of the 20News-Group dataset, well known in the literature for its complexity. Our results highlight the importance of choosing the right classifier whilst pairing it with an optimal strategy, providing valuable insights for optimizing classifier performance in MCC classification tasks considering both environmental implications and the need for accurate predictions.

Key words: Statistical Learning, One-Vs-Rest, One-Vs-All, Naïve Bayes, Tb-Nb, Accuracy, 20NewsGroup dataset.

1 Introduction

Multiclass classification (MCC) is known to be a complex challenge given the rising demand for such classification tasks in various domains. Most of the time selecting the right approach is crucial, particularly because it is often required to find the right balance between model complexity and computational time. Concerns about

Maurizio Romano
University of Cagliari, Dep. of Business and Economics, Cagliari, 09123, Italy e-mail: romano.maurizio@unica.it

Maria Paola Priola
University of Cagliari, Dep. of Business and Economics, Cagliari, 09123, Italy e-mail: mariap.priola@unica.it

climate change and the environmental impact of large-scale models emphasize this need.

This paper focuses on MCC, also known as multinomial classification in statistical learning. MCC is a classification task characterized by the presence of an outcome (dependent, or response, variable) that presents (often many) more than two classes. MCC is distinct from multilabel classification (MLC). As pointed out by [17], in MCC an instance is assigned to one class out of multiple mutually exclusive classes; on the other hand, MLC refers to a classification problem where each instance could be assigned to multiple class labels simultaneously, under the assumption that there can be more than one correct class label for a given instance or data point.

A common feature of all the proposed approaches for MCC is their computational burden. Given the ongoing hardware availability, practitioners have increasingly employed complex models thus researchers have raised concerns about the climate impact of training large-scale models. Notably, [16] focused on the environmental and economic costs associated with the computations required for deep learning or Artificial Intelligence (AI) research. They highlight that trainings have a significant carbon footprint and are not energy-efficient and advocate the shift from a model training with massive resources to obtain a state-of-art outcome, namely "Red AI", to a more climate-friendly AI procedure, the "Green AI" - essentially a shift from accuracy to efficiency performance metrics.

Considering the factors mentioned above, the objective of this study is to conduct a comparative evaluation of the performance of some statistical classifiers, including Naïve Bayes, Threshold-based Naïve Bayes, Random Forest, Logistic Regression, and an Artificial Neural Network on MCC problems involving text data. Specifically, we focus on an extremely complex MCC task by analyzing the well-known 20NewsGroups dataset available in sklearn¹.

Our study positions itself in the context of MCC research for text data. It involves a comparative analysis of various combinations of MCC strategies and classifiers, striking a balance between performance and computational efficiency. Importantly, it explores the suitability of combining different classifier-strategy combinations pairs in handling heterogeneous textual data. Our line of research aligns with the principles of Green AI, emphasizing the environmental impact of model training. The ultimate goal is to contribute to the ongoing debate concerning the use of computationally demanding MCC strategies for text data by potentially proposing efficient strategies or highlighting the strengths of specific classifiers, offering valuable insights for optimizing MCC tasks while prioritizing environmental sustainability.

¹ <https://scikit-learn.org/stable/>

2 Data and preprocessing

To conduct our investigation, we selected the 20 Newsgroup Dataset [9], a well-known dataset of news articles that is available from the Sklearn library. The 20 Newsgroup Dataset is widely used in the field of text classification and NLP for research purposes. It contains 18,846 unique news articles, which are already categorized into 20 pre-labeled categories. Each news varies from a minimum of 1 to a maximum of 5761 words, with an average of 85, a median of 38, and a standard deviation of 241. We cluster the labels into six main topics: IT (Information Technology) (27.07%), miscellaneous (5.40%), politics (10.24%), recreation (22.02%), religion (13.41%) and science (21.87%). A brief summary of this composition is provided in Table 1.

Table 1 Topic-Group Mapping Table

<i>IT</i>	<i>misc</i>	<i>recreation</i>	<i>politics</i>	<i>religion</i>	<i>science</i>
comp.graphics		rec.autos			sci.crypt
comp.os.ms-windows.misc		rec.motorcycles	talk.politics.guns	alt.atheism	sci.electronics
comp.sys.ibm.pc.hardware	misc.forsale	rec.sport.baseball	talk.politics.mideast	soc.religion.christian	sci.med
comp.sys.mac.hardware		rec.sport.hockey	talk.politics.misc	talk.religion.misc	sci.space
comp.windows.x					

In this work, the dataset is downloaded with the a priori exclusion of headers, footers, and quotes. This is because, generally, these segments of text contain highly specific or detailed information [15]. Next, we clean and process the documents by applying a set of text mining techniques popular in Natural Language Processing. Specifically, the news are converted in text-corpus following a methodology delineated by [7]. We clean the corpus by removing numbers, lowercasing the words, deleting extra spaces, punctuations, and standard stopwords downloaded with nltk package. Lastly, we also opt to delete suffixes and prefixes by applying lemmatization² and the [12] stemming³ algorithm. At the end of this cleaning, the goal of parsing data is to end up with tokens, which are text readable by the machine. In this case, we include unigrams⁴. As a result of this phase, we achieve a document-term matrix. Each row represents a document (observation), and each column represents a word (feature) contained in at least one of the documents. The matrix is filled with the standard practice of the term-frequency inverse-document-frequency (tf-idf) [1].

² Lemmatization refers to the process for which words are morphologically analyzed with a dictionary, so that the word "saw" can be converted in "see", its present tense.

³ Stemming is a pure heuristic process based on Porter's algorithm that works, without getting into the deep, in this way: it has five sequential steps of word cuts based on rules. For the first step, for instance, the algorithm may reduce the words based on doubled typical letters (caresses becomes caress", "economies" "economy" and so forth).

⁴ n-grams are series of n adjacent elements from a given string of tokens

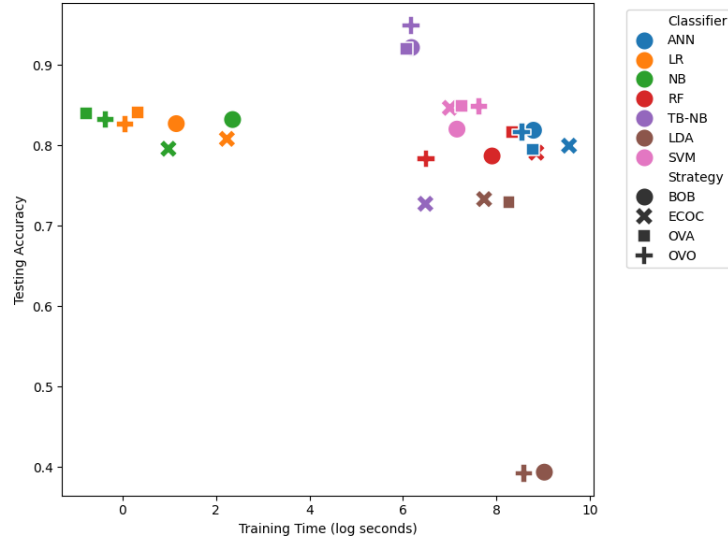
3 Methodology & Results

Our study focuses on comparing a set of classifiers (Logistic Regression (LR) [8], Naïve Bayes (NB) [11], Random Forest (RF) [2], Artificial Neural Network (ANN) [14], Linear Discriminant Analysis (LDA) [10], Support Vector Machine (SVM) [3], and Threshold-based Naïve Bayes (Tb-NB) [13]) with several Multi-class text classification strategies: One-Vs-Rest (OVA), One-Vs-One (OVO) [6], Best-of-Best (BOB) [4] and Error-Correcting-Output-Codes (ECOC) [5]. Thus, we focus on a widely popular dataset that has been employed in multiple research questions, namely the 20NewsGroup Dataset [9], a well-known dataset of news articles that is available from the Sklearn library. Table 2 report the ranks of the classifier-strategy pairs (strategy-classifier pairs in parenthesis) for Accuracy performance metric (i.e. the ratio of correctly classified points (prediction) to the total number of predictions). The best average rank for both pairs is highlighted in bold. Moreover, Figure 1 allows to also consider the Training Time (seconds) for each classifier-strategy. Those predictive results allow to say that:

- Tb-NB shows superior performance, especially when utilizing OVO. However, its performance notably declines under ECOC. This variability suggests a sensitivity to the chosen strategy-classifier combination and raises questions about Tb-NB ability to generalize effectively across different strategies. Training time is moderate and varies based on the strategy employed, with OVA being the quickest;
- NB demonstrates robust performance across various strategies, with its optimal performance occurring when paired with OVA. It shows relatively short training times, with its quickest being under a second with OVA;
- ANN, LDA, SVM, and RF exhibit increased variability between training and test sets, as this aspect could speak about the generalization power of the models. They necessitate longer training times, indicating a poor trade-off between complexity and time efficiency;
- LR is distinguished by its adaptability, maintaining consistent performance across all strategies. It consistently requires the shortest training times across all strategies, demonstrating exceptional efficiency.

Table 2 Accuracy average Rank for Classifier-Strategy Pair on the 20NewsGroup Dataset

Classifier	BOB	ECOC	OVA	OVO	avg. classifier Rank
ANN	5 (1)	3 (3)	6 (4)	5 (2)	4.75
LR	3 (2)	2 (4)	3 (1)	4 (3)	3
NB	2 (2)	4 (4)	4 (1)	3 (2)	3.25
RF	6 (3)	5 (2)	5 (1)	6 (4)	5.5
TB-NB	1 (2)	7 (4)	1 (3)	1 (1)	2.5
LDA	7 (3)	6 (1)	7 (2)	7 (4)	6.75
SVM	4 (4)	1 (3)	2 (1)	2 (2)	2.25
avg. strategy Rank	(2.4)	(3)	(1.9)	(2.6)	

Fig. 1 Trade-off between Training Time and Testing Accuracy

4 Concluding remarks

Our analysis shows that classifier choice significantly affects prediction quality and time impact. Environmental implications of model and strategy selection are significant and should be considered in the broader context of sustainable AI and computational efficiency. As we move forward, the integration of eco-friendly practices in machine learning will be a priority, necessitating a balance between model performance, computational demand, and environmental impact. However, the strategy's role is crucial and should not be underestimated. Optimal performance likely depends on the right combination of classifier and strategy, tailored to the data distribution.

References

- [1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval, vol. 463. ACM press New York (1999)
- [2] Breiman, L.: Random forests. *Machine learning* **45**, 5–32 (2001)
- [3] Chervonenkis, A.Y.: Early History of Support Vector Machines, pp. 13–20. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). DOI 10.1007/978-3-642-41136-6_3. URL https://doi.org/10.1007/978-3-642-41136-6_3

- [4] Conversano, C.: Interactive visualization in multiclass learning: integrating the sassc algorithm with klimt. *Computational Statistics* **26**, 711–731 (2011)
- [5] Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research* **2**, 263–286 (1994)
- [6] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition* **44**(8), 1761–1776 (2011). DOI <https://doi.org/10.1016/j.patcog.2011.01.017>. URL <https://www.sciencedirect.com/science/article/pii/S0031320311000458>
- [7] Grün, B., Hornik, K.: topicmodels: An r package for fitting topic models. *Journal of statistical software* **40**(1), 1–30 (2011)
- [8] Hosmer, D.W., Lemeshow, S.: *Applied Logistic Regression*. John Wiley & Sons, Ltd (2000). DOI <https://doi.org/10.1002/0471722146.ch1>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471722146.ch1>
- [9] Lang, K.: Newsweeder: Learning to filter netnews. In: *Machine learning proceedings 1995*, pp. 331–339. Elsevier (1995)
- [10] Mitteroecker, P., Bookstein, F.: Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology* **38**(1), 100–114 (2011). DOI [10.1007/s11692-011-9109-8](https://doi.org/10.1007/s11692-011-9109-8). URL <https://doi.org/10.1007/s11692-011-9109-8>
- [11] Murty, M.N., Devi, V.S.: *Bayes Classifier*, pp. 86–102. Springer London, London (2011). DOI [10.1007/978-0-85729-495-1_4](https://doi.org/10.1007/978-0-85729-495-1_4). URL https://doi.org/10.1007/978-0-85729-495-1_4
- [12] Porter, M.F.: An algorithm for suffix stripping. *Program* **40**, 211–218 (1980)
- [13] Romano, M., Contu, G., Mola, F., Conversano, C.: Threshold-based naïve bayes classifier. *Advances in Data Analysis and Classification* pp. 1–37 (2023)
- [14] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65**(6), 386 (1958)
- [15] Sarkar, D.: Text classification. In: *Text Analytics with Python: A Practitioners Guide to Natural Language Processing*, pp. 275–342. Springer (2019)
- [16] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green ai. *Communications of the ACM* **63**(12), 54–63 (2020)
- [17] Sharma, S., Mehrotra, D.: Comparative analysis of multi-label classification algorithms. In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pp. 35–38. IEEE (2018)

3. Invited - Data science and dataspace: challenges, results, and next steps

3.1 - Data-Centric AI : A new Frontier emerging in Data Science

3.2 - Data Spaces strategy to unleash agriculture data value: a concrete use case

3.3 - Addressing Agricultural Data Management Challenges with the Enhanced TRUE Connector

Data-Centric AI : A new Frontier emerging in Data Science

Donato Malerba, Vincenzo Pasquadibisceglie, Vito Recchia and Annalisa Appice

Abstract Artificial Intelligence (AI) has historically relied on two key elements: data and algorithms. However, the traditional Model-Centric AI paradigm has typically emphasized algorithms, often handling data as static entities. Data are typically gathered, pre-processed, and kept unchanged, with significant efforts focused on refining learned models. This conventional approach has led to the development of increasingly complex and opaque decision models, requiring substantial effort in data training. On the other hand, the emerging Data-Centric AI (DCAI) paradigm focuses on the systematic and algorithmic generation of optimal data to fuel Machine Learning (ML) and Deep Learning (DL) techniques. The primary aim of the DCAI paradigm is to improve data quality, thereby achieving model accuracy deemed unattainable levels through model-centric techniques alone. This paper investigates the transformative effects of recent advancements in the DCAI paradigm on the future use of AI, ML and DL in data science. The objective is to inspire further innovations in DCAI research, ultimately influencing the future landscape of in data science applications.

Key words: Data-Centric AI, Data collection and preparation, Model development, Applications

Donato Malerba
University of Bari Aldo Moro e-mail: donato.malerba@uniba.it

Vincenzo Pasquadibisceglie
University of Bari Aldo Moro e-mail: vincenzo.pasquadibisceglie@uniba.it

Vito Recchia
University of Bari Aldo Moro e-mail: vito.recchia@uniba.it

Annalisa Appice
University of Bari Aldo Moro, e-mail: annalisa.appice@uniba.it

1 Introduction

The big amount of data generated nowadays has led to the recent surge of AI techniques across various data science applications ranging from remote sensing, healthcare and industry. Despite the close relationship between AI and Big Data, traditional AI methods have primarily operated within a model-centric paradigm, prioritizing the search for highly accurate models trained with fixed input data during the training phase. Generally, this model-centric approach pays great attention to model design and training processes, such as algorithm design and hyperparameter optimization, while it often overlooks issues related to data quality, such as missing values, incorrect labels or outliers. Consequently, models developed under the model-centric AI paradigm are commonly specialized and tailored to specific tasks and datasets, making it challenging to transfer them across tasks or datasets within the same problem domain.

Neglecting data quality can lead to diminished accuracy, especially in scenarios involving changes in data distribution shifts or adversarial samples. In addition, it may introduce biases towards certain classes or sensitive attributes like gender [1]. Due to the limitations of the model-centric AI paradigm, growing importance has been recently recognized to improve data quality and stabilize models. This has definitively prompted a gradual shift towards a data-centric AI paradigm. In this new paradigm, the more emphasis is devoted to data preparation to develop more robust models that can easily adapt to varying datasets with minimal changes. However, transitioning from a model-centric to a data-centric AI paradigm poses significant challenges, including training data development, inference data development and data maintenance tasks.

The paper is organized as follows. Section 2 describes the key components of the DCAI paradigm. Section 3 illustrates some examples of applications of the DCAI paradigm in different data science domains. Finally, Section 4 summarises the goals of the DCAI, sketches conclusions and identifies possible future research directions.

2 Data-Centric AI paradigm

Several emerging AI enterprises have recently prioritized data as the central element for enhancing precision, expediting deployment, and standardizing their data science workflows. Open AI, exemplified due to its realization of Chat GPT and GPT4 [2], is a recent instance of an effective industrial implementation of data-centric AI principles. The company adheres to a uniform learning model architecture across various applications, with variations in number of parameters only. It emphasizes the acquisition of top-tier data and user input to conduct reinforcement learning [3] and deliver more precise responses from a human-centric standpoint.

In general, DCAI denotes a framework for developing, refining, and sustaining data for AI systems [4]. Recent studies have started the definition of tasks and methods for extracting smart data from raw information sources in order to build effec-

tive training data, design proper inference data, and maintain the data in the DCAI perspective. In particular, the DCAI paradigm is divided in three main steps: *training data development*, *inference data development*, and *data maintenance* [4]. These steps, that are all essential for developing a robust DCAI process, are interconnected one to each other.

Training data development. The main goal of the training data development step is to collect and produce rich and high-quality training data to support the training of decision models. Both quality and quantity of training data are achieved in the training data development through the steps of data creation and data processing. Data creation focuses on encoding human intentions into datasets. Data processing focuses on preparing the data for the learning stage. In particular, the training data development step includes the following steps:

- *Data collection* [11, 12] that aims to identify the most related and useful datasets from data lakes and data marketplaces. This step may often require data integration operations.
- *Data labeling* [13] that assigns one or more labels to data samples enabling the use of supervised learning algorithms. As this operation is a time-consuming and resource-intensive process, various techniques have been identified to improve efficiency and reduce the cost of data labeling (e.g., crowd source labeling, consensus learning, semi-supervised and active learning).
- *Data preparation* [14] that prepares raw data for the learning stage by handling noise, inconsistencies and any unnecessary information that may lead to inaccurate and biased results. Both feature extraction and feature transformation are two examples of transformation operations.
- *Data reduction* [15] that reduces the complexity of a given dataset, while keeping its representative information. This is achieved by reducing either the feature size (feature reduction) or the sample size (sampling).
- *Data augmentation* [16] that is a technique to increase the size and diversity of data.

Inference data development. The model-centric paradigm mainly evaluates a decision by limiting to use performance metrics (e.g. accuracy metrics, computation time, memory usage). However, this may lead to perform the evaluation phase neglecting some critical aspects of a decision model such as resilience, adaptability and the reasoning behind decision-making. The aim of the inference data development phase is to generate new evaluation sets that offer more detailed insights into the model or activate a particular capability of the model using engineered data inputs. Some sub-goals of this phase are:

- *In-distribution evaluation* [17] that involves generating samples that align with the training data. This is crucial to identify and calibrate underrepresented groups to prevent biases and errors, understand decision boundary and scrutinize ethical considerations.
- *out-of-distribution evaluation* [18] that aims to generate samples that differ significantly from training data. For example, adversarial samples can aid in understanding the robustness of models to out-of-distribution.

Data maintenance. In real-world applications, data are not created once, but they need to be continuously updated and curated. The data maintenance step aims at maintaining the quality and reliability of data in a dynamic environment. It involves three sub-goals:

- *Data understanding* [19] that may use visual summarization, clustering or data statistics to help to organize complex data and make human readable insights.
- *Data quality assurance* [21, 24] that is commonly performed in dynamic environment where continuous monitoring and qualitative improvement are mandatory. Quality assessment involves objective and subjective metrics. Objective assessment metrics measure inherent data attributes (accuracy, timeliness, consistency, completeness), while subjective assessments evaluate data quality from a human perspective.
- *Data storage and retrieval* [20, 23] that manages exponentially growing data through resource allocation strategies to optimize throughput and latency in data administration systems.

3 Data-Centric AI: Applications

DCAI promises to transform numerous industrial and data science domain applications by gaining decision-making efficiency and effectiveness through insights derived from data [6]. At present, the use of the DCAI paradigm has emerged as a new trend in multiple data domain applications such as remote sensing, healthcare and industry.

Data-centric AI in spatial data analysis. DCAI is gaining importance in remote sensing, where a big amount of Earth images (e.g. Sentinel-1 and Sentinel-2 images) is free available with the Copernicus Programme.¹ [9] has recently described the main principles of the DCAI paradigm in both remote sensing and geospatial data applications. This study shows that geospatial data acquisition and curation should receive as much attention as data engineering and model development and evaluation. In addition, it describes one of the first data-centric remote sensing pipelines experimented for land cover classification in satellite imagery. Contrastive learning has been recently used to handle lack of labels in land cover classification by resorting to semi-supervised learning [8]. In addition, [10] has recently explored data-driven approaches for deep feature extraction in Sentinel-2 data.

Data-centric AI in Healthcare. Healthcare providers can leverage DCAI to analyze patient data in real time and make more informed therapeutic decisions. Adopting this technique could involve adjusting medication dosages for specific patients, administering targeted therapies, and predicting diseases more accurately. Given its myriad capabilities and potential benefits, DCAI is increasingly embraced and integrated into precision healthcare, reshaping traditional healthcare into digitized, patient-centred healthcare [22]. Zahid *et al.* [5] have recently illustrated a systematic

¹ <https://www.copernicus.eu/en>.

review of emerging information technologies used for data modeling and analytics to achieve Data-Centric Health-Care (DCHC) for sustainable healthcare.

Data-centric AI in Industry. DCAI principles have decisively influenced not only academia but industrial research and development. In [7], the authors address the challenges encountered by small and medium-sized enterprises (SMEs) when used in AI-based data science pipelines. Although the prevailing model-centric AI paradigm recommend to amass large datasets for the data scientists' analysis, SMEs often encounter obstacles such as limited data, lack of labels, data drift and insufficient knowledge in ML and DL techniques which hinder their data science implementation efforts. The DCAI paradigm, however, prioritizes the systematic engineering of data used in constructing an AI system. In [7], the authors describe a tangible, adaptable implementation of a DCAI development process tailored for industrial applications, particularly in machining and manufacturing sectors.

4 Conclusion

The DCAI paradigm signifies a fundamental shift in AI system design within data science applications, prioritizing a data-centric approach over a model-centric one. This paper emphasizes the importance of leveraging data effectively to realize its full potential. In addition, it outlines key steps in the data science process that are aligned with the DCAI principles and highlights recent applications. Future research in DCAI is expected to focus on extracting smart data from raw sources for practical training, designing inference data, and managing data integrity over time. Methods will address challenges such as data imbalance, diverse data types, and maintaining model performance amidst data drifts and adversaries. Co-designing input data and models will be crucial for creating adaptable and transferable models, with implications for data monitoring and enhancing the synergy between AI, data science, and human decision-making processes. This approach aims to address critical challenges in data quality and model stability.

Acknowledgements Donato Malerba, Vincenzo Pasquadibisceglie and Vito Recchia are partially supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007), under the NRRP MUR program funded by the NextGenerationEU. Annalisa Appice is partially supported by project SERICS (PE00000014) under the NRRP MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

1. Sambasivan, N., Kapania, S., Highfill, H., et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings Of The 2021 CHI Conference On Human Factors In Computing Systems*. (2021)
2. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., et al. GPT-4 Technical Report. (2023)

3. Christiano, P., Leike, J., Brown, T., et al. Deep Reinforcement Learning from Human Preferences. *Advances In Neural Information Processing Systems*. **30** (2017)
4. Zha, D., Lai, K., Yang, F., et al. Data-centric AI: Techniques and Future Perspectives. *Proceedings Of The 29th ACM SIGKDD Conference On Knowledge Discovery And Data Mining*. pp. 5839-5840 (2023)
5. Zahid, A., Poulsen, J., Sharma, R., et al. A systematic review of emerging information technologies for sustainable data-centric health-care. *International Journal Of Medical Informatics*. **149** pp. 104420 (2021)
6. Kumar, S., Sharma, R., Singh, V., et al. Potential Impact of Data-Centric AI on Society. *IEEE Technology And Society Magazine*. **42**, 98-107 (2023)
7. Luley, P., Deriu, J., Yan, P., et al. From concept to implementation: The data-centric development process for AI in industry. *2023 10th IEEE Swiss Conference On Data Science (SDS)*. pp. 73-76 (2023)
8. Ienco, D., Gaetano, R., and Interdonato, R., A constrastive semi-supervised deep learning framework for land cover classification of satellite time series with limited labels, *Neurocomputing*, 2024.
9. Roscher, R., Rußwurm, M., Gevaert, C., et al., Data-Centric Machine Learning for Geospatial Remote Sensing Data, arXiv, 2023.
10. Phillips, J., Zhang, C., Williams, B., et al., "Data-Driven Sentinel-2 Based Deep Feature Extraction to Improve Insect Species Distribution Models," EGU General Assembly, 2022.
11. X. Chu, I. F. Ilyas, S. Krishnan, et al., "Data curation at scale: the data tamer system," *CIDR*, 2013.
12. Stonebraker, M. and Ilyas, I. F., "Data Integration: The Current Status and the Way Forward," *IEEE Data Eng. Bull.*, vol. 2, no. 41, pp. 3-9, 2018.
13. Dekel, O. and Shamir, O., "Vox Populi: Collecting high-quality labels from a crowd," *COLT*, 2009.
14. Wan, M., Zha, D., Liu, N., et al., "Modeling Techniques for Machine Learning Fairness: A Survey," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2022.
15. J. C. Riquelme, J. S. Aguilar-Ruiz and M. Toro, "Finding representative patterns with ordered projections," *Pattern recognition*, vol. 36, no. 4, pp. 1009-1018, 2003.
16. M. Frid-Adar, E. Klang, M. Amitai, et al., "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification," *ISBI*, 2018.
17. E. Ötleş, J. Oh, B. Li, et al., "Mind the Performance Gap: Examining Dataset Shift During Prospective Validation," *MLHC*, 2021.
18. Madry, A., Makelov, A., Schmidt, L., et al., A. Towards Deep Learning Models Resistant to Adversarial Attacks. *6th International Conference On Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. (2018)
19. Burch, M. and Weiskopf, D., "On the benefits and drawbacks of radial diagrams," *Handbook of human centric visualization*, pp. 429-451, 2014.
20. Aken, D. V., Pavlo, A., Gordon, G. J., et al., "Automatic Database Management System Tuning Through Large-scale Machine Learning," *SIGMOID*, 2017.
21. Pipino, L., Lee, Y. & Wang, R. Data quality assessment. *Commun. ACM*. **45**, 211-218 (2002,4).
22. Oberste, L., and Heinzl, A., "User-centric explainability in healthcare: A knowledge-level perspective of informed machine learning," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 840-857, Aug. 2023.
23. Ardagna, C. A., Bena, N., Hebert, C. , Krotsiani, M., Kloukinas, C. and Spanoudakis, G. (2023). Big Data Assurance: An Approach Based on Service-Level Agreements. *Big Data*, 11(3), pp. 239-254.
24. Anisetti M., Ardagna C., Bena N. and Damiani E., "Rethinking Certification for Trustworthy Machine-Learning-Based Applications" in *IEEE Internet Computing*, vol. 27, no. 06, pp. 22-28, 2023.

Data Spaces strategy to unleash agriculture data value: a concrete use case

Nicola Masi*, Delia Milazzo*, Giulia Antonucci*, Susanna Bonura*

*Engineering Ingegneria Informatica SpA—Piazzale dell'Agricoltura, Rome, Italy e-mail: nicola.masi@eng.it; delia.milazzo@eng.it; giulia.antonucci@eng.it; susanna.bonura@eng.it

Abstract In order to ensure that all actors in the European agricultural sector can improve the sustainability and profitability of their operations while addressing issues such as food security and climate change, the European Commission is committed to create an international Data Space aimed at creating a virtual environment in which public and private agricultural data converge, simplifying interconnections between stakeholders and promoting interoperability between different applications and stakeholders in the agricultural sector. In this paper the International Data Space Association's Reference Architecture Model is presented together with its concrete implementation within a running research initiative in the smart and sustainable agriculture context.

Abstract *Al fine di garantire che tutti gli attori del settore agricolo europeo possano migliorare la sostenibilità e la redditività delle loro operazioni, affrontando al contempo tematiche quali la sicurezza alimentare e il cambiamento climatico, la Commissione Europea è impegnata nel creare uno Spazio Dati internazionale volto a creare un ambiente virtuale in cui convergono dati agricoli pubblici e privati, semplificando le interconnessioni tra gli stakeholder e promuovendo*

l'interoperabilità tra diverse applicazioni e soggetti del settore agricolo. In questo articolo, viene presentato il Modello di Architettura di Riferimento dell'Associazione IDSA, insieme alla sua concreta implementazione all'interno di un'iniziativa di ricerca nel contesto dell'agricoltura intelligente e sostenibile.

Key words: Data Space, Digital Agriculture, Data Sovereignty.

Introduction

Within the agri-food supply chain, stakeholders utilize their data at their specific stage. However, for more efficiencies and added value in the agri-data economy, there is a need for further development and demonstration in sharing data collectively along the chain. Although some ad-hoc sharing occurs, each new group of potential data providers/consumers faces similar challenges. Additionally, the absence of data sharing precedents in agriculture hinders data owners from adopting a more exploratory perspective.

According to the last European Strategy for data [1], the key to incentivizing data sharing is the creation of Data Spaces, that are ecosystems where diverse actors can share and use data in a secure, reliable and trustworthy way, following common governance, organizational, regulatory and technical mechanisms, so to provide the basis for fruitful cooperation, lowered barriers to entry and limitless innovation in the data economy of the future.

In the next sections the International Data Space Association's Reference Architecture Model (IDSA RAM) and its use case within the Agritech Project are presented.

International Data Space Reference Architecture Model Overview

The IDSA Reference Architecture Model¹ enables data sharing through data spaces characterized by uniform rules, certified data providers and consumers and trust among partners guaranteed by certified components and secure interactions (Figure 1).

¹ IDS Reference Architecture Model 3.0 (internationaldataspaces.org)

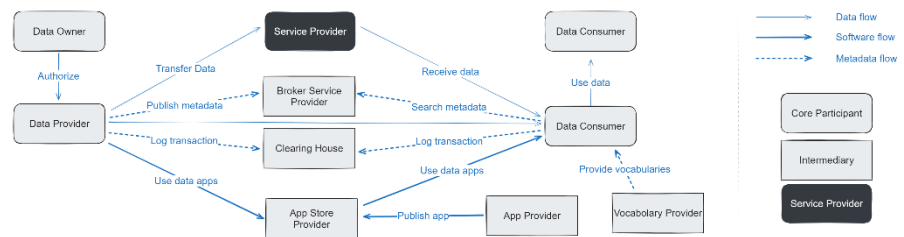


Figure 1: Roles and Interaction in the Industrial Data Space (source: IDS Reference Architecture Model 3.0 (internationaldataspaces.org))

Typically, a participant who serves as the Data Owner also automatically assumes the role of the Data Provider. However, there are instances where the Data Provider is not the same as the Data Owner. In cases where the Data Owner does not simultaneously act as the Data Provider, their sole responsibility is to authorize a separate Data Provider to make their data accessible for use by a Data Consumer. This authorization process should be well-documented through a contract, which includes information about data usage policies for the provided data.

Similar to the Data Owner, who is the legal entity with control over their data, the Data User is the legal entity with the right to use the data of a Data Owner as specified by the usage policy. In most cases, the Data User is the same as the Data Consumer.

Before establishing a connection with a Data Provider, the Data Consumer has two options. Firstly, they can search for existing datasets by making an inquiry at a Broker Service Provider. The Broker Service Provider then provides the necessary metadata for the Data Consumer to connect to the Data Provider. Alternatively, the Data Consumer can directly establish a connection with the Data Provider without involving a Broker Service Provider. In cases where the Data Consumer already possesses information to connect with the Data Provider, they may request the data (along with the corresponding metadata) directly from the Data Provider.

The Clearing House records all activities during data exchanges. After a data exchange (or parts of it) is completed, both the Data Provider and the Data Consumer confirm the data transfer by logging transaction details with the Clearing House. This logging information serves as the basis for billing. Furthermore, the Clearing House provides reports on the executed (logged) transactions, aiding in billing and conflict resolution.

The App Store offers Data Apps, which are applications that can be deployed within the Connector—the essential technical component for participants to join the International Data Spaces. These Data Apps play a crucial role in facilitating data processing workflows.

The Vocabulary Provider oversees and provides access to vocabularies, including ontologies, reference data models, and metadata elements. These vocabularies serve to annotate and describe datasets.

In the next section an IDSA-compliant Data Space concrete architecture is shown that enables the secure, reliable and trustworthy data sharing in smart

agriculture. Such architecture design is the basis for the developments in the Agritech project.

Agritech Data Space

The Agritech project is a large Italian project funded with the National Recovery and Resilience Plan to contribute, among other objectives, to the digitization of Agriculture in Italy and to the implementation of an information technology platform and its web-based abstraction (called Web of Agriculture - WoA) that will support research, production and commercial activities in the agrifood sector in Italy, by leveraging on innovative artificial intelligence tools [2]. We present here the architectural design of the Agritech Data Space.

The Figure 2 below represents the components forming the platform based on the logical references to which they belong, highlighting interactions with Agritech actors.

The storage layer shows different types of platform storage, aiming to provide adequate support for the types of data to be saved and, most importantly, the type of usage these data are intended to fulfil: *(i)* document storage, to store, organize, and manage documents and digital files efficiently; *(ii)* in-memory data store, intended to offer some platform components with a temporary data storage for facilitating or speeding up data retrieval; *(iii)* graphDb, which provides platform components with the ability to manage complex relationships between entities; *(iv)* message broker, in case some components need to handle streaming data flows; *(v)* object storage, useful for storing large volumes of data, even unstructured.

This lower layer supports the services within the Components layer, which contains modules that collaborate to deliver the promised services to Agritech users. Specifically:

- "Harmonization/Standardization": The platform aims to provide a data sharing service, both within and outside the platform. For this purpose, data standardization plays a key role in enabling effective interoperability. This component offers the possibility of defining standards based on input data.
- "Traceability Manager": This component provides a traceability module to retrieve information useful for product traceability and tracking.
- "Data Analytics": This module enables the execution of Big Data Analytics (both stream and batch data) created through workflows composed of microservices. It offers a service not only to data scientists (who have the ability to design algorithms and inject them into their workflows) but also to less experienced users through a No-Code interface referring to a catalogue of ready-to-use algorithms. It is based on ALIDA, a well known asset by ENGINEERING coming from past and current research and innovation initiatives [3].
- "Decision Support System" (DSS): it highlights a series of elements useful for providing decision support services based on the type of knowledge on which it is based and how it is intended to be offered. Indeed, modules have

been inserted for serving ML models, modules for storage and retrieval of data intended for LLM use, components offering a remote development environment capable of retrieving datasets owned or shared by third parties for the development of DSS applications directly on the platform, and finally components allowing exposure to the outside through APIs.

Many of the applications described above require an interface through which users can interact. For this purpose, in the "Application" layer, there are components of User Interface (UI) or generic APIs that users can utilize according to their purposes.

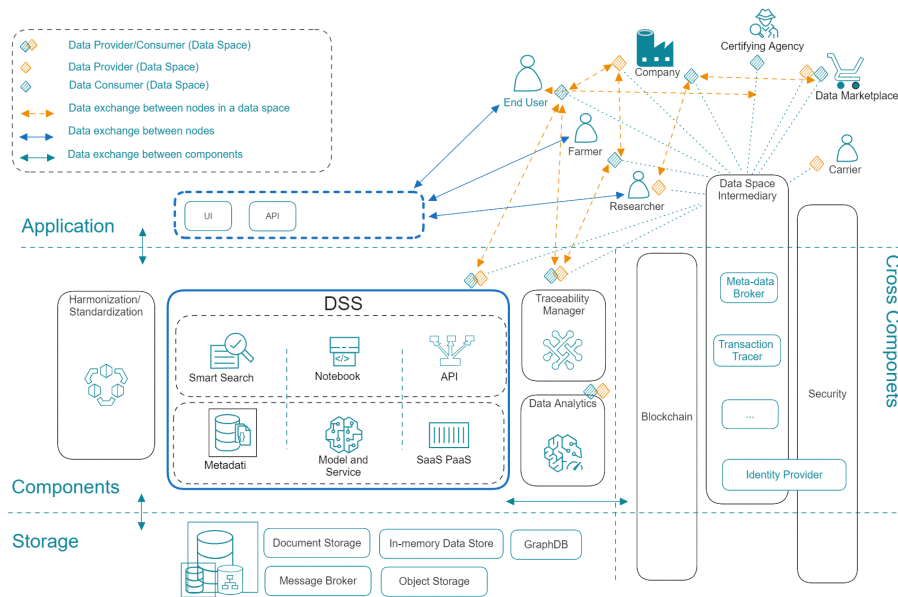


Figure 2: Data Space within Agritech

On the right, three macro blocks that intersect with some blocks play the role of Cross components: blockchain, "Data Space Intermediary," and Security. In particular: the blockchain module is intended to provide a data storage service with the prerogative of being immutable; the security module aims to offer authentication and authorization services for users interacting with the platform. Finally, the Data Space Intermediary module aims to offer all those components useful for data exchange between participants (Data Provider and Data Consumer) within a Data Space according to International Data Space Association (IDSA) specifications. Actors in the supply chain or users interested in data exchange via Agritech Data Space can request or provide data while maintaining data sovereignty.

Several components and external actors have a diamond shape next to their graphical representation, a symbol that represents the ability for that actor or component to provide and/or request data through Data Space connectors (Orange as Provider, Green as Consumer, and double diamond for both roles). The arrows show

three different types of data exchange: dashed orange for data transmission relying on the system described on Data Space, continuous blue arrow for data exchange between users and the platform through a graphical interface or API, and finally continuous green arrow for exchange between internal components of the platform.

Conclusions

In the agriculture domain Data Spaces play a pivotal role in ensuring a surge in data availability. These spaces facilitate secure and dependable access and sharing through digital platforms that adhere to transparent and shared standards and governance. The Common European Data Spaces aim to unlock more data for utilization in both the economy and society, all while maintaining control for the companies and individuals who contribute to this data. This paper provided an overview of the IDSA-based Data Space and its benefits within the Agritech project.

References

1. Summary Report on the open public consultation on the European strategy for data ([Summary report of the public consultation on the European strategy for data | Shaping Europe's digital future \(europa.eu\)](#))
2. Chessa, S., Dimitri, G.M., Gori, M., Kocian, A. (2023). WoA: An Infrastructural, Web-Based Approach to Digital Agriculture. In: Novais, P., et al. Ambient Intelligence – Software and Applications – 14th International Symposium on Ambient Intelligence. ISAmI 2023. Lecture Notes in Networks and Systems, vol 770. Springer, Cham. https://doi.org/10.1007/978-3-031-43461-7_12
3. D. Profeta, N. Masi, D. Messina, D. Dalle Carbonare, S. Bonura and V. Morreale, "A Novel Micro-Service Based Platform for Composition, Deployment and Execution of BDA Applications," 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Kallithea, Greece, 2019, pp. 182-185, doi: 10.1109/SEAA.2019.00037. keywords: {Data visualization;Big Data;Sparks;Connectors;Pipelines;Business;Graphical user interfaces;Big Data analytics;Big Data systems;Spring Cloud Dataflow;Microservices;Data visualization;Visual analytics;Workflow management system;Optimization;BDA Application}

Acknowledgements

This work is funded within the project AGRITECH Spoke 9 - Codice progetto MUR: AGRITECH "National Research Centre for Agricultural Technologies" - CUP CN00000022, of the National Recovery and Resilience Plan (PNRR) financed by the European Union "Next Generation EU".

Addressing Agricultural Data Management Challenges with the Enhanced TRUE Connector

Sergio Comella*, Delia Milazzo*, Mattia Giuseppe Marzano*, Giulia Antonucci*, Susanna Bonura*, Angelo Marguglio*¹

Abstract This paper addresses the challenge of interoperability in agricultural data management and proposes the TRUE Connector as an enhanced solution to enable agricultural data spaces ecosystem. Interoperability barriers resulting from incompatible digital platforms and privacy concerns hinder effective data sharing in agriculture. The study presents the TRUE Connector, supported by adherence to the Data Space Protocol, as a transformative solution. By this integration, the agricultural sector can achieve seamless data exchange while maintaining data sovereignty and complying with European regulations. This innovative approach promises to improve economic and environmental performance, paving the way for a more efficient and sustainable agricultural future.

Abstract Il presente lavoro affronta la sfida dell'interoperabilità nella gestione dei dati agricoli introducendo il TRUE Connector come possibile opportunità per una più fluida implementazione di ecosistema di data spaces. Gli ostacoli all'interoperabilità derivanti dalla presenza di piattaforme digitali spesso incompatibili tra loro, congiuntamente alle tematiche attinenti agli aspetti di privacy, impediscono una fluida condivisione dei dati nel settore agricolo. Il paper in oggetto presenta il TRUE Connector, supportato dall'adesione dello stesso al Data Space Protocol, come possibilità di ottenere uno scambio di dati mantenendo la sovranità degli stessi nel rispetto delle normative europee, garantendo dunque maggiore efficacia e sostenibilità nel settore agricolo.

Key words: data space, interoperability, data sovereignty, agriculture, TRUE Connector, Data Space Protocol

¹ Sergio Comella, Engineering Ingegneria Informatica S.p.A., sergio.comella@eng.it
Delia Milazzo, Engineering Ingegneria Informatica S.p.A., delia.milazzo@eng.it
Mattia Giuseppe Marzano, Engineering Ingegneria Informatica S.p.A., mattiagiuseppe.marzano@eng.it
Giulia Antonucci, Engineering Ingegneria Informatica S.p.A., giulia.antonucci@eng.it
Susanna Bonura, Engineering Ingegneria Informatica S.p.A., susanna.bonura@eng.it
Angelo Marguglio, Engineering Ingegneria Informatica S.p.A., angelo.marguglio@eng.it

1 Challenges in exploiting European agricultural data

The European Commission is committed to adopt strategies that promote efficiency, sustainability, and competitiveness in the European agricultural sector. It is undeniable that digital technologies and data have the potential to revolutionise agriculture, assisting farmers in working more precisely, efficiently, and sustainably.

However, the digitalisation of the sector presents urgent challenges, among which compatibility issues - different digital platforms may not be compatible making data sharing and integration of diverse applications difficult - and privacy and data ownership concerns - privacy and data ownership may hinder data sharing among agricultural stakeholders.

In this context, a significant change can be achieved by the implementation of data spaces, where data can flow between countries and sectors, and be accessed and used easily, respecting European values and regulations.

While significant progress has been made in promoting interoperability, persistent technological and operational challenges remain. The current landscape is characterised by diversity with a lack of standardised solutions making interoperability more difficult. In addition, the lack of best practices or comprehensive cost-benefit analyses for data sharing further complicates efforts to encourage collaboration among stakeholders. To address the above-mentioned challenges and to facilitate the free flow of data across borders and sectors, DIVINE (Demonstrating Value of agri data sharing for boosting data Economy in agriculture) project, founded by European Commission, aims to establish a unified gateway through which agricultural data spaces - both public and private - can be smoothly accessed by interested data consumers.

In this paper it is introduced an innovative architecture for data spaces ecosystem for agricultural domain which enables transparent data sharing between existing data spaces. Aligned with the IDSA (International Data Spaces Association) Reference Architecture Model¹, it prioritises data sovereignty and provides robust methods and tools for implementation.

In the following section it will be introduced the concept of data space and DIVINE project, highlighting its innovative approach to addressing the challenges of interoperability in the agricultural sector. In the third section, it will be proposed the TRUE Connector² as a key component in solving the interoperability challenges; by aligning with the IDSA Data Space Protocol, TRUE Connector offers a promising solution to improve data interoperability and facilitate seamless data exchange between different stakeholders.

¹ <https://internationaldataspaces.org/offers/reference-architecture/>

² <https://www.eng.it/case-studies/true-connector-per-facilitare-la-condivisione-di-dati-in-gaiax>

2 Mitigating Agricultural Data Challenges

In the era of digital transformation, the concept of data spaces has emerged as a fundamental framework for enabling secure and standardised data exchange across different sectors. A data space serves as a secure digital infrastructure that facilitates the trusted exchange of data and the provision of data-based services between different stakeholders. Within the IDSA paradigm, a data space is defined as a virtual environment that provides a standardised framework for data exchange. This framework is based on common protocols and formats, supported by robust mechanisms for secure and trusted data exchange. Central to the concept of data spaces is the principle of data sovereignty which refers to the pillar that data owners retain control over their data, determining who can access it and under what conditions. This principle ensures that data is protected and used according to the preferences and permissions of its owners.

In this context, the DIVINE project emerges as a pioneering initiative within the agricultural sector and represents a concerted effort to influence data practices in agriculture aiming to address the challenges of digitisation in agriculture by creating an ecosystem for sharing and analysing agricultural datasets.

Leveraging the principles of data spaces, DIVINE aims to create a seamless platform for data sharing, interoperability, and collaboration between stakeholders.

The architecture designed envisages the use of a smart connector that enables the straightforward federation of data spaces within the ecosystem (see Figure 1). It encompasses data security, data analytics, decision support, data sharing, open data sources, enablers, applications, and the commitment to open data principles. The principal components that make up the architecture are as follows: Dashboard, Collaboration Space, Knowledge Tools, Enablers App, Data Management, Security and Agora Management. The Dashboard is the access point for agriculture stakeholders and allows them to access the different services available. The Stakeholder Open Collaboration Space (SOCS) is an online platform dedicated to all stakeholders, where providers and consumers of digital technologies can match resources and needs and work together towards common innovations. Knowledge tools include assets as AI Models ready to be used for specific use case, recommendations, and content-based resources. Security matters are managed by Identity Management, Identity Service Provider and Identity wallet to Authentication and Authorisation aspects. The platform allows the identification of members during data exchange or use of its functionalities. Within the Agora Management it is placed a Federated Catalogue which is the repository enabling participants to find other participants' information and service offerings in the shape of the Self-Descriptions. The Data Management includes all functionalities for data preparation, harmonisation, semantic interoperability, and data fusion.

The architecture described has a strong focus on enablers and the innovative connector. The following section will present the mentioned smart connector which is the result of the TRUE Connector alignment with the guidelines embedded in Data Space Protocol that not only addresses interoperability challenges but also promotes collaboration and innovation within the agricultural sector.

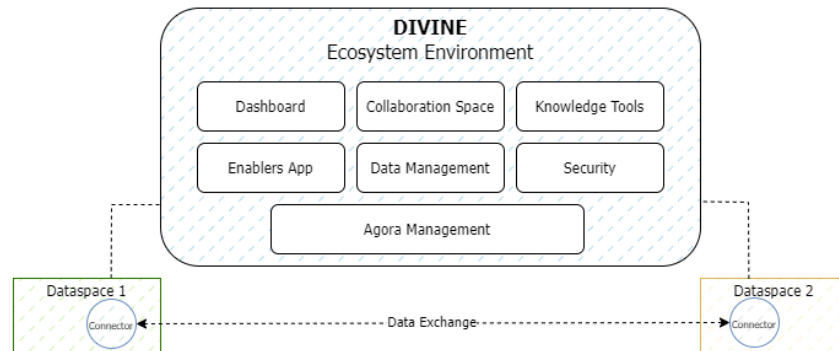


Figure 1: Main components of DIVINE's architecture

3 TRUE Connector meets Data Space Protocol

TRUE Connector is the open source International Data Spaces (IDS) Connector by Engineering which enables interoperability thanks to its adherence to Data Space Protocol [2]. The design decisions of the Protocol are directed towards aligning with the principles of IDSA [3]. The foremost principle centres on prioritising data sovereignty for all participants. The Protocol's design aims to empower participants with the ability to exercise choice, agency, and autonomy within the data space, positioning it as the unequivocal guiding tenet. While individual connectors may introduce partial modifications to these principles, ensuring the data sovereignty of participants remains the highest priority within the Protocol itself.

The second principle highlights the conception of a data space as a context of trust, distinct from a physical environment or marketplace. Within the Protocol, a data space is primarily envisioned as a realm fostering trust between two sovereign participants. The third principle draws an analogy between the rules governing a data space and a government-provided lawbook. Certain components of the Protocol may need central implementation by a Dataspace Authority, acting as a governance body. This authority is tasked with defining the laws and rules of the data space without compromising the data sovereignty of the participants, ensuring a harmonious and effective operational framework.

The Figure 2 represents the schema of a data space, with the minimum viable data space (MVDS) thanks to the two connectors (provider and consumer) and the Identity Provider composed by Dynamic Attribute Provisioning Service (DAPS) and a Certificate Authority (CA) [4]. Moreover, the schema shows other components that enrich the dataspace such as the Metadata Broker and the Clearing House, those aren't mandatory components but enable useful functionalities for data discovery and data exchange.

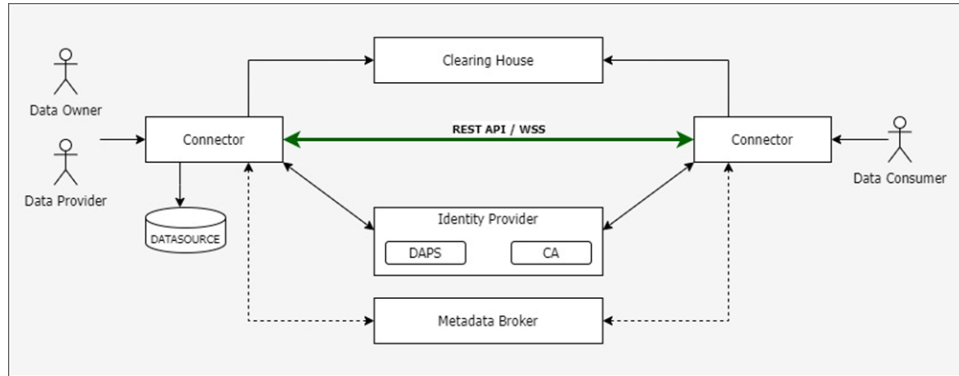


Figure 2: IDSA Minimum Viable Data Space

Figure 3 shows the difference of an interaction between a data provider and a data consumer within a data space, compared to the interaction between dataspace. The latter scenario introduces the added value of the Data Space Protocol [2]. The Protocol, implemented in the TRUE Connector, is a set of specifications designed to promote interoperable data sharing among entities subject to usage control and using Web technologies. These specifications articulate the schemas and protocols necessary for entities to disseminate data, negotiate agreements, and access data within a federation of technical systems known as a data space.



Figure 3: IDSA Data Space interoperability

4 Conclusions

By integrating Data Space Protocol guidelines, TRUE Connector plays a key role in enabling interoperability between different existing data spaces, thus facilitating the establishment of an ecosystem that aggregates private and public agri data that interconnects existing data spaces.

In summary, its application will enable smooth interoperability between different data spaces, with a positive impact on economy and sustainable aspects.

References

1. Center, D. S. (2023, 10 12). Control plane vs. Data plane. Tratto da Dataspace Support Center: <https://dssc.eu/space/BBE/178422298/Control+plane+vs.+Data+plane>
2. IDSA. (2023, 10). Dataspace Protocol Overview. Tratto da Dataspace Protocol: <https://docs.internationaldataspaces.org/ids-knowledgebase/v/dataspace-protocol/overview/readme>
3. IDSA. (2023, 10). Ensuring Dataspace interoperability. Tratto da Dataspace Protocol: <https://internationaldataspaces.org/dataspace-protocol-ensuring-data-space-interoperability/>
4. IDSA. (2023). Minimum viable Dataspace. Tratto da IDS Testbed: <https://github.com/International-Data-Spaces-Association/IDS-testbed/blob/master/minimum-viable-data-space/MVDS.md>

Acknowledgements

The research presented in this paper is partially based on work carried out under the Horizon Europe DIVINE project (Grant agreement ID: 101060884) that is funded by the European Commission under HORIZON.2.6.3 - Agriculture, Forestry and Rural Areas (HORIZON-CL6-2021-GOVERNANCE-01-20).

4. Solicited - Data Science for Official Statistics

4.1 - Data science at Istat for urban green

4.2 - Twitter (X) as a Data Source for Official Statistics: Monitoring Italian Debate on Immigration through Text Analysis

Data science at Istat for urban green

Fabrizio De Fausti¹, Marco Di Zio², Giuseppe Lancioni³, Stefano Mugnoli⁴, Alberto Sabbi⁵, Francesco Sisti⁶

Abstract This paper discusses a machine learning-based approach for urban green estimation that exploits information from orthophotos. The techniques discussed aim to automatically determine a threshold of the Normalized Digital Vegetation Index, a metric used for quantifying density of vegetation based on the spectral reflectance measurement, for the classification of urban greenery taking into account issues related to the variability naturally inherent remote sensing data.

Abstract *In questo lavoro viene discusso un approccio basato su tecniche di machine learning per la stima del verde urbano tramite ortofoto. Le tecniche discusse mirano a determinare automaticamente una soglia per il Normalized Digital Vegetation Index, una metrica utilizzata per quantificare la densità della vegetazione basata sulla misurazione della riflettanza spettrale, utile per la classificazione del verde urbano tenendo conto delle problematiche legate alla variabilità naturalmente insita nelle immagini telerilevate.*

Key words: Machine learning, clustering, remote sensing data.

1 Introduction

Data science is intended as the science of learning from data. This definition is so broad that it is sometimes difficult to design the boundaries with other disciplines, as for instance statistics. An interesting discussion on this topic is in Donoho (2017). One context that is frequently associated to data science is the analysis and use of big data. In recent years, National Statistical Institutes (NSI) have started investing in the analysis of such data for the production of official statistics. Istat has launched various

¹ Fabrizio De Fausti, Istituto Nazionale di Statistica, Istat, Via Cesare Balbo 39, Roma; defausti@istat.it

² Marco Di Zio, Istituto Nazionale di Statistica, Istat, Via Cesare Balbo 39, Roma; dizio@istat.it

³ Giuseppe Lancioni, Istituto Nazionale di Statistica, Istat, Via Cesare Battisti, 18, Trieste; lancioni@istat.it

⁴ Stefano Mugnoli, Istituto Nazionale di Statistica, Istat, Piazza Marconi, 24, Roma; mugnoli@istat.it

⁵ Alberto Sabbi, Istituto Nazionale di Statistica, Istat, Piazza Marconi, 24, Roma; sabbi@istat.it

⁶ Francesco Sisti, Istituto Nazionale di Statistica, Istat, Via Cesare Balbo 39, Roma; sisti@istat.it

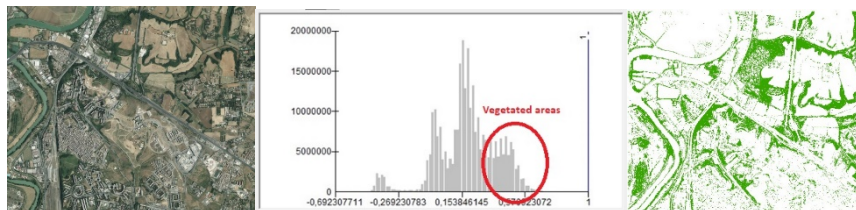
projects that range from the use of AIS data for maritime statistics, to social network analysis, from web scraping to the use of mobile phone data (Bruno and Scannapieco, 2022). In this context, the use of remote sensing for official statistics is certainly one of the most important field of study. The advantages of using remote sensed images are considerable, starting with acquiring territorial information very easily compared to other ways; furthermore, the possibility of having images continuously, allows the study of phenomena that would be impossible to investigate in other ways. In this paper, we report methodologies and some experimental results for the use orthophotos for urban green estimation.

The paper is structured as follows. Section 2 describes the problem, Section 3 illustrates the methodological proposals, and Section 4 reports some preliminary experimental results and some final remarks.

2 Urban Green

Measuring the greenery in urban areas is important for the analysis and the development of many indicators related to various aspects of the cities life (Jabbar M. et al 2021). Several papers report studies on the use of remote sensing for vegetation analysis. They are based on the chlorophyll behavior in relation to its ability to absorb light radiation at various wavelengths. Some indicators for classifying the vegetation are produced in literature (Xue and Boafeng, 2017). We decided to work with the Normalized Difference Vegetation Index (NDVI). We use the high resolution remote sensed images (AGEA Orthophotos with 20 and 50 cm pixel resolution). These images cover, over a three years period, the entire Italian territory and starting from 2012, have been released to Istat at four spectral bands. Pixel of the images must be re-classified to extract the ones related to the areas covered by vegetation. To this aim, the histogram of the images' NDVI is analysed to select the threshold value beyond which there is a high probability that the pixel is 'green'. Figure 1 illustrate how the pixels from the vegetation index image are grouped within the image histogram.

Figure 1 – Image R,G,B of a portion of a urban areas of Rome (Castel Giubileo, Villa Spada, Serpentara, Colle Salaria, the NDVI histogram, and the image related to the 'green pixels'



3 Machine learning for determining a vegetation threshold

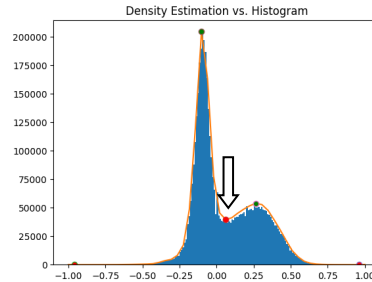
In general, we expect four classes of NDVI identifying different objects (water, buildings, arid soil, vegetation) and a threshold value around 0.2 (Arval, J. et al- 2022) delimiting the vegetation class. In practice, we observe a various number of clusters and a variability of the NDVI threshold value for the vegetation. For this reason, we study different methods to automatically detect the threshold in order to adapt the search to different situations.

At first, data are pre-processed as follows:

- orthophotos in ecw format are converted in GeoTiff;
- images are cropped with a shape file to select urban areas;
- NDVI is computed;
- multiple images are combined to form a mosaic (the city image).

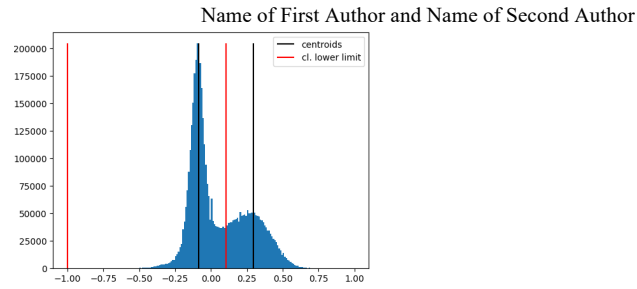
Then, three group of methods (hierarchically nested) are studied. The first one is based on kernel density estimation (KDE), the second is based on clustering techniques, and the third is referred to a refined detection of vegetation, henceforth RDV. KDE works by estimating the probability distribution through a Gaussian kernel with bandwidth related to the standard deviation of data points. The local maxima of KDE are computed. We use them for indicating the likely number of clusters, and the threshold for the vegetation is set to the rightmost local minimum. In figure 2, that is referred to the municipality of Ravenna, the NDVI threshold for identifying the ‘greenery’ is the red point addressed by the vertical arrow.

Figure 2: NDVI histogram, KDE (orange curve) with local maxima (green dots) and minima (red dots)



In the second approach, we used K-means and K-medians clustering algorithm. The number of clusters are set to the number of local maxima identified by KDE, these values are also set as the initial centroids for the algorithms. The urban green threshold is the lower limit of the rightmost cluster, i.e., the cluster related to the centroid with the highest NDVI value, see figure 3.

Figure 3: clusters centroids (black lines) and clusters lower limits (red lines) found by K-means



The third approach - RDV - refines the analysis obtained via the clustering methods. The motivation of this approach is to deal with two challenging aspects that are commonly encountered in this field. First, the varying green nuances across different images; a significant factor contributing to this variation is related to the seasonal effects, which can alter the type of green detected in each image. For instance, the lushness of vegetation in spring presents a different shade of green compared to those of a later period (Eastman JR et al., 2013). Second, the excessive presence of non-green areas in urban landscapes, which often overshadows the green clusters in our analysis. We adopt the approach of Donchyts et al., (2018). This was originally applied for water-index detection and offer valuable insights for our context. A customization of that methodology was needed for the vegetation analysis, that is indeed our target.

The method consists in the following pipeline:

- the KDE threshold is set as a preliminary threshold for green pixels in the image. Pixels below the KDE threshold are turned black, while those above are kept unchanged;
- the Canny Edge Detection algorithm is applied on the masked image. It performs a segmentation of the input image based on the variation of the NDVI values. This generates segments along the KDE threshold since there is a steep variation between green and non-green areas, and in the full green areas above KDE threshold, where it can separate faint-green by strong-green pixels.
- A buffering technique is applied around the detected edges. This step narrows the focus to the immediate areas surrounding the vegetation, providing a more targeted region for analysis. Buffered image is used as a mask on the original image to resample pixels in the regions of high index variation. This reduces noise from irrelevant regions.
- The Otsu clustering is applied to the resampled histogram, to determine an optimal threshold for binary separation. It works on single tone images (such as the NDVI images are) and split the values in two classes with a threshold. It efficiently distinguishes non-green from green areas.

4 Preliminary results and some final remarks

We applied the methods to Ravenna, a municipality in Northern Italy. Table 1 shows the estimated NDVI thresholds, and the green area in terms of hectares and percentage of covered urban territory. There is a general agreement of the results for KDE, K-means and K-medians. KDE and K-medians are closer to each other, while K-means is a bit more different, this is due to the less robustness of the algorithm to outliers. RDV, as expected, provides lower estimates, in fact it works refining the set of green pixel of the clustering algorithms. In this case, there is a strong segmentation inside the green area, meaning a great variety of vegetal structures with highly varying NDVI, and the Otsu clustering determines a shift in the threshold to discard areas of less intense green vegetation. RDV makes us able to differentiate soft-green from strong-green, refining the previous threshold determinations and enhancing the overall quality of the approach.

Table 1: Thresholds and green area in hectares (Ha) and percentages (%) for Ravenna

<i>Method</i>	<i>Threshold</i>	<i>Green area (Ha)</i>	<i>Green Area (%)</i>
KDE	0.06	729,29	39,86
K-means	0.11	663,32	36,25
K-medians	0.08	698,79	38,19
RVD	0.18	544,24	29,74

Figure 4 is an example of the difference between K-medians and RDV algorithms, as expected RDV selects stronger green areas.

Figure 4: Orthophoto of Ravenna (left), identification of green areas shown in black for K-medians (center) and RVD (right).



All elaboration are executed using python public libraries.

These first experiments show promising methodological approaches based on machine learning that are applicable to the orthophotos and are useful to identify the urban green. Those methods automatically deal with the variability of such data, in an adaptive manner. The results are encouraging and future works will extend the application of the procedures to the largest Italian municipalities.

References

1. Aryal, J., Sitaula, C., Aryal, S: NDVI threshold-based urban green space mapping from sentinel-2a at the Local Governmental Area (LGA) level of Victoria, Australia. Land 11.3, 351, (2022).
2. Bruno M., Scannapieco M.: Methodologies for big data at Istat: state of the art and open challenges. Proceedings of the 1st Workshop on Methodologies for official Statistics, Istat, Rome, Italy 5-6 December, 2022, available at <https://www.istat.it/en/archivio/290538>
3. Donchyts, G., Schellekens, J., Winsemius, H., Eisemann, E., van de Giesen, N.: A 30 m resolution surface water mask including estimation of positional and thematic differences using Landsat 8, SRTM and OpenStreetMap: A case study in the Murray-Darling Basin, Australia. Remote Sensing, 8(5), 386, (2016).
4. Donoho, D.,; 50 years of data science. Journal of Computational and Graphical Statistics; 26.4: 745-766 (2017).
5. Jabbar, Muhammad, Mariney Mohd Yusoff, and Aziz Shafie.: Assessing the role of urban green spaces for human well-being: A systematic review. GeoJournal, 1-19 (2021).
6. Xue, J., Boafeng, S.: Significant Remote Sensing Vegetation Indices: a Review of Developments and Applications; Journal of Sensors, pp. 1-17, (2017).

Twitter (X) as a Data Source for Official Statistics: Monitoring Italian Debate on Immigration through Text Analysis

Elena Catanese, Gerarda Grippo, Francesco Ortame, Maria Clelia Romano

Abstract Official Statistics (OS) is more and more interested to investigate the possible use of new sources, including social media website to enrich its production. Aim of the work is to study public opinions towards immigrants in Italy by exploiting Twitter data. In this study, a large dataset of Italian tweets between 2018 and 2022 containing a set of keywords related to immigration is analyzed using text mining techniques such as topic modelling. This kind of analysis provides valuable insights about the debate on immigration, serving as a preliminary step for the construction of sentiment-based synthetic indices and as a continuous process to enhance their interpretability. Emerging patterns about immigration are discussed along with possible generalizations according to the chosen methodology.

Key words: topic modelling, NLP, twitter, immigration, LDA, opinion mining

1. Introduction

Social networks have become very popular communication tools among Internet users, with millions of messages exchanged daily through trendy platforms such as Facebook, Twitter (X), Instagram, LinkedIn, etc. Hence, these web sites and platforms represent valuable sources to study phenomena related to people's opinions and sentiments. Official Statistics (OS) recognizes the added value the new

¹

Elena Catanese, Italian National Institute of Statistics (Istat), catanese@istat.it

Gerarda Grippo, Italian National Institute of Statistics (Istat), grippo@istat.it

Francesco Ortame, Italian National Institute of Statistics (Istat), francesco.ortame@istat.it

Maria Clelia Romano, Italian National Institute of Statistics (Istat), romano@istat.it

data sources can offer for social analysis, especially thanks to their timeliness and huge dimensionality, features that traditional sources typically lack. However, using social media sources for OS poses many challenges to be faced both on the methodological and thematic points of view. For example, likewise most Big Data sources, X's data generating mechanism does not fall under the statistician's control and is unknown. Indeed, X users are not a representative sample of the population. However, it is still important for OS to invest in the direction of these new sources, strengthening the methodological tools necessary to analyze them. This paper is part of a broader experimental project aimed at studying attitudes toward immigration using data from the X platform. Immigration is an increasingly central topic in public debate, as new migration waves are recorded at a global level due to new conflicts. These waves often lead to an increase in hostile sentiments toward immigration. Despite being frequently based on misinformation and prejudice, such attitudes risk compromising immigrants' ability to integrate with local communities. Monitoring and understanding public opinion regarding immigrants is therefore essential to prevent discrimination and strengthen social cohesion in an increasingly globalized society. As many studies have highlighted [4, 5, 7, 8], social network platforms, especially X, where citizens can freely and publicly express their opinions can be a useful complement to information from traditional sources by allowing real-time observation and quantification of users' attitude towards immigration and the interest aroused by the multiple issues related to it.

This work is structured as follows: section 2 describes data and methods; section 3 presents the results; conclusions are drawn in section 4.

2. Data and methods

2.1 Data

Istat opened a download stream, since 2016, to collect samples of tweets of interest for OS through a filter of 278 words. To extract tweets related to immigration, we used a further filter of topic-specific lemmas. This filter has undergone a two-step verification: top-down (expert choice); bottom-up (data driven analysis). The final filter and the relevance of the content of this sample of tweets has been validated both by word embedding techniques as well as by topic modelling. In this work, we analyse the tweets collected in real-time in the period between January 1st, 2018, and December 31st, 2022. The average number of tweets selected in this period is about 7,000 per day; the total number of tweets in the period is 20 million.

2.2 Topic Modelling

Topic Modelling is a famous unsupervised machine learning technique used to 'detect hidden topics' from a corpus, in a similar way to clustering algorithms which partition data. The algorithms and models used to extract patterns from texts arise

from native Bayesian generative probabilistic models, linear algebra models based on matrix decompositions and clustering algorithms on Word Embedding spaces.

LDA [2] is an extension of Probabilistic Latent Semantic Analysis, which introduces sparse Dirichlet prior distributions over document-topic (parameter α) and topic-word (parameter β) distributions. LDA is a more complete probabilistic generative model with a three-level hierarchical Bayesian structure, for its components, documents, topics and words. The assumption is that documents are represented as random mixtures over latent topics, each characterized by a distribution of words. The words with the highest probability on each topic are used to identify the subject. The distribution of topics across all documents shares a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution of words, with the word distributions of topics sharing a common Dirichlet prior. The theoretical underpinnings of LDA rely on exploiting the concept of exchangeability, i.e. a major simplifying assumption of text processing that allows for computationally efficient methods. The bag-of-words representation makes LDA an adequate model for learning hidden themes, but does not account for a document's deeper semantic structure. However, LDA is considered state of the art in topic modeling and is a powerful textual analysis technique to find and quantify underlying topics [6].

Practically, an assessment of the LDA model reveals a few weaknesses, namely the necessity of a fixed k value (number of topics), the inability of Dirichlet distributions in capturing correlations, the static nature does not show the evolution of topics over time, and lastly the simplifying "bag-of-words" exchangeability assumption. Out of these limitations, none is sufficient to abandon this topic modeling method, but awareness is necessary to understand the boundaries of results.

3. Results

3.1. Model choice and evaluation

Texts were cleaned by removing URLs, numbers, and non-alphabetic characters (excluding #, hashtags) and a user-defined list of stop-words. Then texts containing at least three words were included. We then restricted our list of terms to attain an 86% coverage of the total word occurrences. This threshold was computed in order to exclude on average no more than one or two words per tweet (tweet median and average length is 14, bi-modal at 8 and 16). As a result, the number of features was set to 12,109 from a total size of more than 500 thousands.

When implementing LDA it is necessary to fine-tune the hyper-parameters, specifically the number of topics (k), the number of features (V , i.e. fixed vocabulary size), and α and β Dirichlet prior parameters. For the choice of α and β , we used heuristic reasoning based on the knowledge of topics and words over time. We hypothesized that both the distributions of topics in each document and the

distribution of words in each topic are sparse and constant over time. For these reasons, we chose fixed values for α and β , both equal to $1/k$.

Although there is little consensus about topic model evaluation, the performance is often evaluated using quantitative and qualitative measures, i.e., perplexity and coherence scores as well as interpretability and usefulness of topics. The performance of the topic models in this paper is computed through two widely used metrics, namely topic coherence and topic diversity. In order to evaluate the optimal number of k , topic coherence was evaluated using normalized pointwise mutual information (NPMI) [5], obtaining 55.

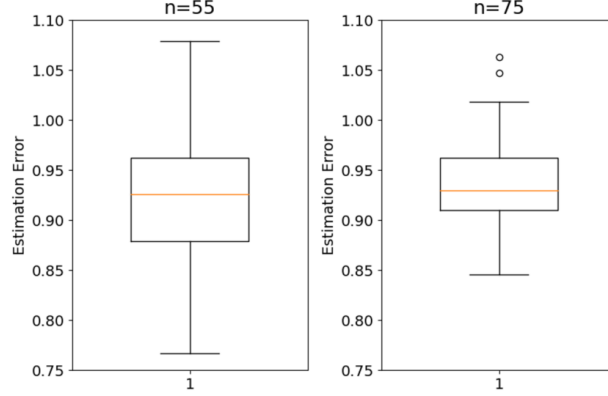


Figure 1 – Box-plot of the ratio between estimated and real distributions

Topic diversity, as defined by [3], is the percentage of unique words for all topics. We here evaluate the differences between 55 and the adopted choice of 75 clusters. The latter shows a better separation degree 53% compared to 49%. Moreover, we added a qualitative measure of the distortion between the estimated frequency and the real frequency of the most 50 frequent words of the Document Term Matrix.

In this study, we prioritize as choosing criteria the lack of distortion in the estimation of frequencies, the separation degree and human judgement.

3.2. Main results

The results highlight a great heterogeneity of topics that are the subject of conversations on immigration, which reproduce the complexity of public debate. However, semantic analysis allows us to trace them back to some macro-themes and to identify their different relevance/interest in the debate based on the number of tweets that fall into the various topics. Topics have undergone a human hierarchical reduction to summarize results by macro-area, as shown in table 1.

In the following, we briefly describe the macro areas.

Table 1 – Main macro-themes, number of topics and percentages of related tweets

Macro-area	No. of clusters	%
Internal and European political debate	12	16.5
Illegal landings, NGOs	9	14.3
Crime, violence	9	13.1
Work, pensions, welfare	6	9.7
Covid-19 pandemic	6	8.8
Citizenship rights	5	5.9
Islam	2	3.8
Hate	3	3.3
Other/Miscellaneous	23	24.6

Internal and European political debate (16.5%): clusters characterized by internal and European political debate on immigration. The most prominent terms refer to the main political players and measures adopted in the field of immigration, and concerns caused by the arrival of illegal migrants.

Illegal landings, NGOs (14.3%): illegal immigration and NGOs' activity in the Mediterranean, relative to several distinct events, represent the main theme. Words such as *NGOs*, *illegal immigrants*, *sea*, and *ships* are among the most frequent.

Crime, violence (13.1%): discussion about crime and violence in their various declinations, from *rape* to *drug trafficking* to the *Nigerian mafia* to aiding and abetting illegal immigration. The overall most numerous cluster (3.2% of tweets) belongs to this macro-area and is characterized by words that refer to episodes of sexual violence against women.

Work, pensions, welfare (9.7%): topics concerning means to support the full integration of migrants as well as the impact of immigration on the economic and social situation of the country, in a context where the scarcity of resources (*work*, *social housing*, etc.) also creates problems for Italians.

Covid-19 pandemic (8.8%): focus on the role of *China*, both as the country from which the *virus* originated and as the main producer of *masks*; on the regulation of *vaccines*, *green pass* and risks, in the absence of possible controls, related to illegal immigration for the spread of *Covid-19*.

Citizenship rights (5.9%): clusters related to specific regulatory aspects associated to migration, such as *residence permits*, *political asylum* applications) and *jus soli* as well as the importance of respecting laws and rules, the protests to claim rights, and economic subsidies (as *citizenship income*).

Islam (3.8%): two topics fall in this category. The first one focuses on political issues (*regime*, *Iran*, *death*, *Turkey*, *police*, *terrorism*, *freedom*, *attack*, etc.), the second one on cultural and religious aspects containing words such as *religion*, *culture*, *family*, *Christian*, *love*, *rights*, *values*, *peace*, *civilization*, etc.

Hate and intolerance (3.3%): topics characterized by the word *hate* associated with words such as *racism*, *gypsies* and verbal expressions that refer to criminal acts (such as *rape*, *kill*). The discussions also contain clearly hostile forms of expression such as *#primagliitaliani* (*#italiansfirst*), *#stopinvasione* (*#stoptheinvasion*) as well as references to the risk of ethnic substitution.

Other/miscellaneous (24.6%): topics related to migration that could not be easily assigned to other categories or that have a strong semantic heterogeneity that makes them difficult to interpret.

4. Conclusions and future perspectives

Topic coherence measures the degree of semantic similarity between the words in each topic, while thematic value is based on human interpretation of topic quality and measures the usefulness of the topics for achieving specific study goals, as in the present study. Coherence scores do not tell the whole story about the quality of a model. However, it might still be necessary to perform either different topic modelling algorithms, or LDA analysis with different parameters, to provide a more accurate estimation of topics, reducing both the subjectivity in topic classification and the fluctuations induced by the parameters' choice.

Results confirm the strong semantic heterogeneity of the debate about immigration, and allow quantifying the relative weight of each topic. In such a way, it is possible to identify both the issues of greatest interest and those that, despite being minor, deserve particular attention due to the content of hatred, that risks fostering a climate of indiscriminate hostility toward immigrants.

This analysis, as the preliminary step in a larger project, provides a comprehensive but static picture. It is advisable to continuously monitor topics, study their semantic evolution over time, and explore the immigration theme through a wider set of tools.

References

1. Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453
2. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.
3. Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.
4. Calderón, C.A.; de la Vega, G.; Herrero, D.B. Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain. *Soc. Sci.* 2020, 9, 188. <https://doi.org/10.3390/socsci9110188>.
5. Freire-Vidal, Y., Graells-Garrido, E., Rowe, F., A Framework to Understand Attitudes towards Immigration through Twitter. *Appl. Sci.* 2021, 11, 9689. <https://doi.org/10.3390/app11209689>
6. Hamed Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *CoRR*, abs/1711.04305, 2017.
7. Menshikova A., Frank van Tubergen, What Drives Anti-Immigrant Sentiments Online? A Novel Approach Using Twitter, *European Sociological Review*, Volume 38, Issue 5, October 2022, Pages 694–706, <https://doi.org/10.1093/esr/jcac006>
8. Rowe F., Mahony M., Graells-Garrido E., Rango M., Sievers N., Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. *Data & Policy*. 2021;3. e36. doi:10.1017/dap.2021.38

5. Solicited - Sustainable Artificial Intelligence in Finance

5.1 - Feature Dependence and Prediction Explanations in P2P Lending

Feature Dependence and Prediction Explanations in P2P Lending

Paolo Pagnottoni and Thanh Thuy Do

Abstract Complex Machine Learning (ML) models used to support decision-making in peer-to-peer (P2P) lending often lack clear, accurate and interpretable explanations. While the game-theoretic concept of Shapley values and its computationally efficient variant Kernel SHAP may be employed to this aim, similarly to other existing methods, the latter makes the assumption that the features are independent. The assumption of uncorrelated features in credit risk management is fairly restrictive and, thus, prediction explanations coming from correlated features might result in highly misleading Shapley values, even when considering simple models. We therefore propose an evaluation of different dependent-feature estimation methods of Kernel SHAP for classification purposes in credit risk management. We show that dependent-feature estimation of Shapley values can improve the understanding of true prediction explanations, their robustness, and is essential for better identifying the most relevant variables to default predictions coming from black-box ML models.

Key words: Feature Dependence; Shapley Values; Machine Learning; Explainability

1 Introduction

Machine Learning (ML) model explainability and, thereby, eXplainable Artificial Intelligence (XAI) have gained significant importance over recent times for sup-

Paolo Pagnottoni
University of Insubria, Via Monte Generoso 71, Varese, 21100, Italy e-mail:
paolo.pagnottoni@uninsubria.it

Thanh Thuy Do
University of Insubria, Via Monte Generoso 71, Varese, 21100, Italy e-mail:
ttido@studenti.uninsubria.it

porting decision-making in various domains such as medicine, healthcare, credit risk, and fraud detection. However, one of the major challenges in using these models is the lack of clear, accurate, and interpretable explanations for their predictions. Model explanation is crucial and features in two main aspects: a) a practical one, which raises the question as to whether the model is trustworthy enough to rely on its predictions; b) a legal one, which pertains the compliance with regulations on automated decisions and the General Data Protection Regulation. This is of particular relevance in the context of P2P lending, where the decisions of lenders are based on the creditworthiness of borrowers, and transparent and trustworthy explanations are necessary for establishing trust, and rendering transparent and explainable decisions to lenders and borrowers.

In this regard, credit risk measurement and management have been experiencing spectacular growth over the recent past, resulting in significant improvements in the accuracy of prediction models. However, while many of these algorithms are highly accurate, they are often complex and difficult to understand, making it challenging to interpret the results they produce. To cope with this, researchers keep on developing methods for making ML models more interpretable and transparent. By making it possible to understand how a model performs its predictions, XAI can help to build trust in these algorithms and increase their adoption in credit risk management and other fields

The game-theoretic concept of Shapley values has been proposed as a potential solution. Recently, approaches exploiting model-agnostic explanations paved the way to a better understanding of the contribution of each predictor as a proportion of the overall prediction Lundberg and Lee (2017) proposed the so-called Shapley Additive exPlanations (SHAP) method, a model-independent attribution technique stemming from the Shapley values developed in the field of game theory. The Shapley values measure the feature-specific average contribution to all potential coalitions when it is subtracted from the feature of interest. In the feature attribution literature, SHAP has been a widely used method, and numerous improvements have been made. Covert and Lee (2021) elaborates on the computational costs associated with the determination of Shapley values and discusses possible solutions.

In this brief paper, we render a comparison of methods for prediction explanations through Kernel SHAP in credit risk management when features are dependent. We consider three predictive models, i.e. logistic regression, eXtreme Gradient Boosting (XGBoost), and Generalized Additive Models (GAM) and four feature dependence structures, i.e. independent, multivariate Normal distribution, copula and empirical conditional distribution.

We report in Section 2 the application of the different methods to the P2P lending database. Section 3 concludes and contains some remarks for future research avenues.

2 Empirical Findings

2.1 Results

We first train our models splitting the observations into training and test data as conventionally done in empirical research. In our setting, the dataset is randomly split into 80% of the observations which enter the training dataset, and the remaining ones which compose the test dataset. Firstly, to illustrate how SHAP values represent the contribution of observations and make a comparison between the significance of variables in both classes, fully-paid and default classes, we provide Figure 1 which depicts SHAP values found based on the predictions given by the logistic regression model.

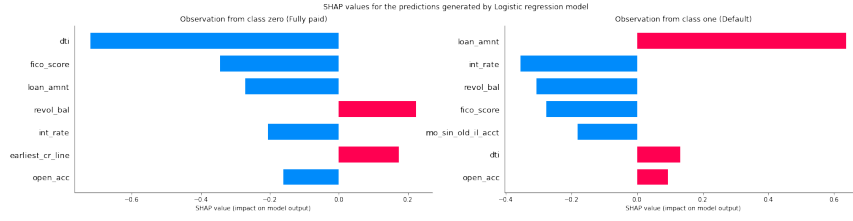


Fig. 1 SHAP values of different class observations. The figure shows the SHAP values for two exemplifying observations from the two relevant classes, namely "Fully-paid" and "Default" loans.

It should be noticed that the sum of SHAP values yield the difference of actual and average prediction; in other words, it evaluates how much has each feature value contributed to the predicted value, compared to the average predicted value. It can be seen that almost all seven significant variables, in terms of explanation, are the same for both observations; except for one variable. However, this cannot lead to a global conclusion since these two observations are selected randomly from a huge amount of data dataset. Focusing on left panel, representing fully-paid class, we can see that "dti" has the highest contribution and its SHAP value is -0.719765. This value shows the contribution of this variable that in fact is the difference that debt-to-income can cause to the final score. On the other hand, SHAP value for "dti" for the observation with label one, default loan, is only 0.131998. In addition, for the default observation, "loan amount" is the variable which has concurred at most to the increase in the probability of default.

For robustness purposes, we perform a random subsampling procedure for the test dataset. This allows the evaluation of variations in the distributions of the model performances and of the Shapley values according to variations in the sample drawn from the test dataset. In other words, we randomly sample the test data $n = 100$ times, iteratively assess the predictive performance of the considered models, and derive the corresponding series of Shapley values. Therefore, after training the classification models using the train data, the KernelExplainer is created based on the

trained model and the train data. Then, this explainer is used to find explanations for the predictions of the observations extracted from test data through random subsampling. Table 1 represents the average of the accuracy measure, AUROC of all the trained models. Considering the mean AUROC values, XGBoost is the most accurate model over all the subsets of the test data.

Table 1 Accuracy comparison of different classification methods. The table shows the averages and standard deviations of the AUROC values obtained via the sampling procedure with $n = 100$ for each predictive model (XGBoost, logistic regression, GAM).

Classification Model	Mean AUROC	STD AUROC
XGBoost	0.626952	0.002872
Logistic	0.617635	0.000365
GAM	0.612613	0.001585

We report in Table 2 the results regarding the average of Shapley values across all dependence methods and the distribution obtained from random subsampling related to the 13 variables analyzed. It is obvious that "Loan amnt" has the highest average SHAP value.

Table 2 Average Shapley values from random subsampling for each dependence method. The table shows the average Shapley values, across dependence methods, of the distribution obtained from random subsampling related to the 13 variables analyzed for each predictive model.

Variable	Classic	Gaussian	Empirical	Copula	Average
Loan amnt	0.002494	0.003688	0.003102	0.006304	0.003897
Mo sin old il acct	0.000275	0.003289	0.000690	-0.000020	0.001058
Int rate	0.000320	0.001708	0.000356	0.001745	0.001032
Mort acc	-0.000336	0.001224	0.000770	-0.000477	0.000295
Dti	-0.000127	0.001103	-0.000077	-0.000005	0.000224
Revol bal	-0.000664	0.000679	-0.000543	0.001288	0.000190
Earliest cr line	0.002168	-0.002995	0.002008	-0.000493	0.000172
Pub rec	0.000420	0.000157	-0.000582	0.000257	0.000063
Open acc	-0.000267	0.001514	-0.000389	-0.000747	0.000028
Revol util	-0.000822	0.000786	0.000011	-0.000462	-0.000122
Emp length	-0.000342	-0.000221	0.000982	-0.001583	-0.000291
Fico score	-0.001487	-0.001750	0.000255	-0.001353	-0.001084
Annual inc	0.000792	-0.007051	-0.000753	0.000124	-0.001722

We report the entire output of our repeated random subsampling procedure in Figure 2. We illustrate the estimated SHAP values across the $n = 100$ iterative estimations for each variable, across the considered predictive models and dependence structures. Overall, feature importance as measured by SHAP values varies across models, dependence methods, and - for some variables - even iterations. Despite that, some patterns are evident. The figure suggests that the variability of the estimated SHAP values associated with Loan amount and Dti (first two) and, to a lesser extent, Fico score and Annual Income (last two) across samples is sensibly higher

than those of the rest of the features. This appears a relevant finding towards the importance of such variables for prediction explanations.

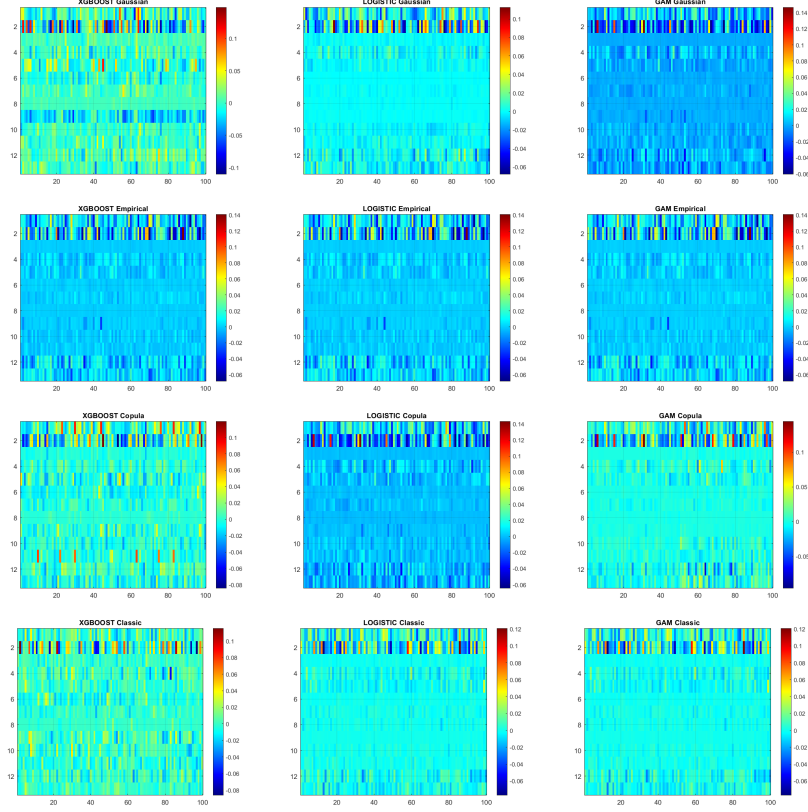


Fig. 2 Repeated random subsampling estimation of feature SHAP values. The figure shows the output of the $n = 100$ repeated random subsampling estimation of feature SHAP values for each predictive model and dependence method.

3 Concluding Remarks

In this short paper, we have worked on an extension of the Kernel SHAP method which explicitly models dependence structures when modeling P2P lending default risk. We demonstrate the effectiveness of their approach by using four predictive ML models: logistic regression (LR), generalized additive models (GAMs) and XG-Boost. We show that our method can enhance the understanding and estimation

robustness of true prediction explanations coming from complex ML models, especially in the case where highly correlated predictors are in place. As a result, our proposed method has practical applications in improving the interpretability and transparency of decision-making ML models for credit risk management in P2P lending.

Future research might extend the present work in several directions. From a methodological viewpoint, Shapley value estimation for dependent features should facilitate the introduction of Kernel SHAP computation for more complex types of data, such as time series and panel data. The adoption and comparison of classical and ML multivariate time series models to explain and forecast time series and panel data can enhance the explainability of time series predictions - see, for instance, Pagnottoni and Spelta (2022); Pagnottoni (2023) - also in the context of explaining investment and portfolio management decision support systems - see Spelta et al. (2022). This enables the analysis of multivariate time series shock transmission across assets of a portfolio, possibly enhancing forecasting performances and explainability of multivariate dependence structures. Further, extending to background data selection can improve the robustness of prediction explanations. From an empirical viewpoint, the proposed method is also applicable to any kind of classification and prediction task, and the inclusion of non-conventional, alternative data on credit risk could sensibly improve prediction explanations, other than accuracy, in P2P lending default classification.

References

- Covert, I. and Lee, S.-I. (2021). Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Pagnottoni, P. (2023). Superhighways and roads of multivariate time series shock transmission: Application to cryptocurrency, carbon emission and energy prices. *Physica A: Statistical Mechanics and its Applications*, 615:128581.
- Pagnottoni, P. and Spelta, A. (2022). The motifs of risk transmission in multivariate time series: Application to commodity prices. *Socio-Economic Planning Sciences*, page 101459.
- Spelta, A., Pecora, N., and Pagnottoni, P. (2022). Chaos based portfolio selection: A nonlinear dynamics approach. *Expert Systems with Applications*, 188:116055.

6. Solicited - Young SIS

- 6.1 - Merging data and historical information via optimal power prior selection in Bayesian models
- 6.2 - Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables
- 6.3 - A Simultaneous Spectral Clustering for Three-Way Data

Merging data and historical information via optimal power prior selection in Bayesian models

Roberto Macrì Demartino, Leonardo Egidi, Nicola Torelli and Ioannis Ntzoufras

Abstract In Bayesian analysis power prior distributions represent a key tool for integrating past information in clinical trials and similar studies: a crucial role is played by the weight parameter, which regulates the influence of the historical information with respect to the current sample. This parameter can be either fixed or random, and various strategies exist for its determination. We introduce a novel fully Bayesian method employing a calibrated Bayes factor to fine-tune the weight parameter's initial distribution. This new method ensures that historical data is incorporated in a way that is proportional to its relevance and accuracy compared to current data. We apply our approach on some real melanoma clinical trial data, demonstrating its effectiveness in achieving a more balanced and accurate integration of historical information according to a modern machine learning perspective.

Key words: Bayes factor, Clinical trials, Historical information, Machine learning, Strength of evidence.

Roberto Macrì Demartino
Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, e-mail: roberto.macridemartino@phd.unipd.it,

Leonardo Egidi
DEAMS, University of Trieste, Via Valerio 4/1, Trieste, e-mail: legidi@units.it

Nicola Torelli
DEAMS, University of Trieste, Via Valerio 4/1, Trieste, e-mail: nicola.torelli@deams.units.it

Ioannis Ntzoufras
Department of Statistics, Athens University of Economics and Business, Trias 2, Athens, e-mail: ntzoufras@aueb.gr

1 Introduction

Integrative data analyses have recently been shown to be an effective tool for researchers interested in synthesizing datasets from multiple studies in order to enhance statistical conclusions. This is particularly critical in clinical trials, where recruiting sufficient participants can be difficult. Bayesian approaches are particularly suited for integrating efficiently past data into new analyses. This is possible by incorporating historical information into current analyses through informative priors on model parameters. However, the elicitation of these type of priors is widely recognized as a complex task.

In this context, power priors [1] allow the historical data to influence the prior distribution in a flexible and controlled way by specifying a weight parameter δ , that determines the degree to which historical data influences the prior distribution. Multiple strategies exist to specify the parameter δ . The initial approach involves fixing it on a predetermined and reasonable value. Particularly, let consider an historical dataset D_0 with corresponding likelihood $L(\boldsymbol{\theta} | D_0)$, the basic power prior formulation is

$$\pi(\boldsymbol{\theta} | D_0, \delta) \propto L(\boldsymbol{\theta} | D_0)^\delta \pi_0(\boldsymbol{\theta}), \quad (1)$$

where $0 \leq \delta \leq 1$ is the scalar weight parameter and $\pi_0(\boldsymbol{\theta})$ is the initial prior for $\boldsymbol{\theta}$. When the data scientist fixes the weight parameter, sensitivity analysis should be used to identify an appropriate level of borrowing information while accounting for prior-data conflict and several statistical methods have been proposed [2]. A natural extension of the power prior in (1) can be achieved by treating the weight parameter as a random variable with its own prior distribution. Hence, the joint normalized power prior [3] for $(\delta, \boldsymbol{\theta})$ is

$$\pi(\boldsymbol{\theta}, \delta | D_0) = \frac{L(\boldsymbol{\theta} | D_0)^\delta \pi_0(\boldsymbol{\theta}) \pi_0(\delta)}{\int L(\boldsymbol{\theta} | D_0)^\delta \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \quad (2)$$

where $\pi_0(\delta)$ is an initial prior distribution for δ .

However, eliciting an appropriate initial prior distribution has proven to be challenging. Specifically, the challenge lies in crafting a Beta prior that fulfill two purposes: it should enable data borrowing when historical and current data align and restricting it when these data sources present conflicts. Notably, Shen et al. [4] developed two techniques to fine-tune the shape parameters of a Beta initial prior for δ , using two optimization criteria: Kullback-Leibler (KL) divergence and mean squared error (MSE).

This paper introduces a novel Bayesian algorithmic method to enhance an optimal hyperparameter tuning of the initial prior distribution of δ using a calibrated Bayes factor approach. Specifically, the proposed algorithmic strategy is based on the use of hypothetical replications generated from the posterior predictive distribution. This would enable the selection of candidate models based on a more accurate and reliable assessment of the available evidence, thereby enhancing the validity and robustness of statistical inference. We present some preliminary results using a well-known

example involving two clinical trials aimed at assessing the efficacy of interferon treatment for patients affected by melanoma [5], demonstrating the practical application and potential of our approach in real-world data, with a particular focus on biostatistical applications.

2 Calibrated Bayes Factor

Within the Bayesian framework, the Bayes factor [6, 7, 8] plays a pivotal role as a tool for model comparison. It serves as a quantitative measure, assessing the strength of evidence in favor of one statistical model over another. It is defined as the ratio of the marginal likelihood of two competing hypotheses, typically denoted as $\mathcal{H}_0$ (null hypothesis) and $\mathcal{H}_1$ (alternative hypothesis). Considering the prior formulation in (2), then the Bayes factor is

$$\text{BF}_{0,1}(D) = \frac{m(D | \mathcal{H}_1)}{m(D | \mathcal{H}_0)} = \frac{\int_0^1 \int \frac{L(\theta|D) L(\theta|D_0)^\delta \pi_0(\theta) \pi_0(\delta|\mathcal{H}_1)}{\int L(\theta|D_0)^\delta \pi_0(\theta) d\theta} d\theta d\delta}{\int_0^1 \int \frac{L(\theta|D) L(\theta|D_0)^\delta \pi_0(\theta) \pi_0(\delta|\mathcal{H}_0)}{\int L(\theta|D_0)^\delta \pi_0(\theta) d\theta} d\theta d\delta}, \quad (3)$$

where $\pi_0(\delta|\mathcal{H}_0)$ and $\pi_0(\delta|\mathcal{H}_1)$ represent the initial Beta prior distribution for δ under the two competing hypothesis.

However, decisions based solely on the observed Bayes factor may lead to potentially misleading conclusions due to the considerable noise often present in the data. To effectively and responsibly employ the Bayes factor approach, it is crucial to measure and account for this variability, adjusting and calibrating the method according to the specified model [9].

2.1 Simulation-based calibration

Before observing the sample, the Bayes factor can be viewed as a random variable [10]. Hence, it is crucial to calibrate the Bayes factor to take account of this randomness, ensuring that it accurately represents the variability inherent in the data. However, the challenge arises when the analytic form of the Bayes factor distribution is not available. To overcome this, we employed a simulation-based strategy, utilizing hypothetical replications generated from the posterior predictive distribution to approximate the distribution of the Bayes factor. A main advantage of this procedure is that the replicated Bayes factor $\text{BF}_{0,1}(D^{\text{rep}})$ mirrors the behavior of the observed Bayes factor $\text{BF}_{0,1}(D)$. Furthermore, by utilizing the knowledge from the current data D , we are concentrating our interest on a significant region within the parameter space [11], improving the precision and relevance of our analysis.

To enhance a more comprehensive data-driven decision rule, it is essential to define a dual-criterion approach. This method not only considers the distributional characteristics of the Bayes factor but also integrates the observed Bayes factor $\text{BF}_{0,1}(D)$ as an index of surprise. Specifically, our proposed criterion centers around the survival function of Bayes factor, $S_{\text{BF}_{0,1}}(\cdot)$, and the inclusion of the observed Bayes factor, $\text{BF}_{0,1}(D)$, within a defined Highest Posterior Density Interval (HPDI).

In refining our methodology, we focus on the logarithmic transformation of the Bayes factor, termed Log-BF, to enhance the interpretability. Furthermore, we adopt a reference model computing the Bayes factor between each alternative model and the reference reducing the number of comparisons. Specifically, as reference model we choose the Beta(1, 1) prior for δ , that is a default non-informative prior for the weight parameter [2]. Consequently, the optimal prior specification is determined using the following criteria

$$\mathcal{H}_{\text{Opt}} = \left\{ \arg \max_{i=1, \dots, M} \left\{ \mathbb{1}_{\{S_{\log \text{BF}_{0,i}}(0) > 0.5\}} S_{\log \text{BF}_{0,i}}(0) \times \mathbb{1}_{\{\log \text{BF}_{0,i}^{\text{obs}} \in (\cdot) \% \text{HPDI}\}} \log \text{BF}_{0,i}^{\text{obs}} \right\} \right\}, \quad (4)$$

where the first indicator function focuses on distributions in which the alternative hypothesis is more likely to be true than the null hypothesis. This means selecting distributions where the survival function of the Log-BF at zero is greater than 0.5. The second indicator function evaluates the presence of the observed Log-BF within a specific HPDI functioning, from a Bayesian perspective, as a measure of surprise.

The computational step of the procedure are summarized in Algorithm 1.

Algorithm 1 Calibrated Bayes Factor Procedure

- 1: Define a grid of size M of potential competing hypotheses regarding the Beta prior for the weight parameter $\mathcal{H} = \{\mathcal{H}_1, \dots, \mathcal{H}_i, \dots, \mathcal{H}_M\}$, with $\mathcal{H}_i : \delta \sim \text{Beta}(\eta_i, \nu_i)$, $i = 1, \dots, M$.
 - 2: For $i = 1, \dots, M$ compute the Observed Bayes factor $\text{BF}_{0,i}(D)$ between hypothesis $\mathcal{H}_i$ and $\mathcal{H}_0$ as in (3).
 - 3: For $i = 1, \dots, M$ generate K hypothetical samples from the posterior predictive distribution under the alternative hypothesis $\mathcal{H}_i$, that is $\mathbf{D}_{\mathcal{H}_i}^{\text{rep}} = (D_{\mathcal{H}_i,1}^{\text{rep}}, \dots, D_{\mathcal{H}_i,K}^{\text{rep}})$.
 - 4: For $i = 1, \dots, M$ and for $k = 1, \dots, K$ compute the replicated Bayes factor $\text{BF}_{0,i}(D_{\mathcal{H}_i,k}^{\text{rep}})$ between hypothesis $\mathcal{H}_i$ and $\mathcal{H}_0$ as in (3) using the hypothetical posterior predictive samples generated in Step 3.
 - 5: Select the optimal model using the criteria in (4).
-

3 Melanoma trials

To illustrate the application of the calibrated Bayes factor in the context of prior elicitation for the weight parameter, we consider two well-known clinical trials on phase III melanoma, a two-arm and a three-arm trial conducted by the Eastern

Cooperative Oncology Group (ECOG), see Kirkwood et al. [5], Ibrahim et al. [2] for further details. Specifically, we examine data from the historical trial, E1684, and the current trial, E1690, which involved 284 and 427 patients, respectively. The clinical trials under consideration assigned patients to either receive interferon treatment (IFN) or a placebo.

We focus on the patients' probability of mortality, denoted as θ , using a Binomial model. We assigned as initial prior for θ a Beta(1, 1). As optimal initial prior of δ , a Beta(0.5, 6) was chosen. This selection was derived from the calibrated Bayes factor procedure, adhering to the selection criteria in (4). This process involved a comprehensive evaluation of competing initial priors, systematically analyzed across a range of different Beta hyperparameters from 0.5 to 6 in increments of 0.5.

The top panel of Figure 1 presents a comparative analysis of the marginal posterior for θ in two distinct scenarios. The blue line represents results using the default Beta(1, 1) initial prior for the weight parameter, while the red line illustrates a more peaked posterior, derived from the optimally selected prior via the calibrated Bayes factor method, leading to a lower variance compared to the default prior. The bottom panel shows the initial priors for δ , depicted with dotted lines, contrasting them with the corresponding marginal posteriors $\pi(\delta \mid \theta, D_0, D)$, shown as continuous lines. The initial prior for δ , selected through the calibrated Bayes factor, is significantly more focused around zero. This indicates a limited integration of data from the historical E1684 trial. This finding aligns with the outcomes observed in the E1690 trial compared to E1684, particularly regarding the efficacy of interferon treatment, which demonstrated contrasting results across the two studies. Moreover, to measure the similarity between posterior distributions, we computed the overlap index η [12, 13] using the R package `overlapping`. This resulted in overlap estimates of $\hat{\eta}_\theta = 0.88$ and $\hat{\eta}_\delta = 0.04$.

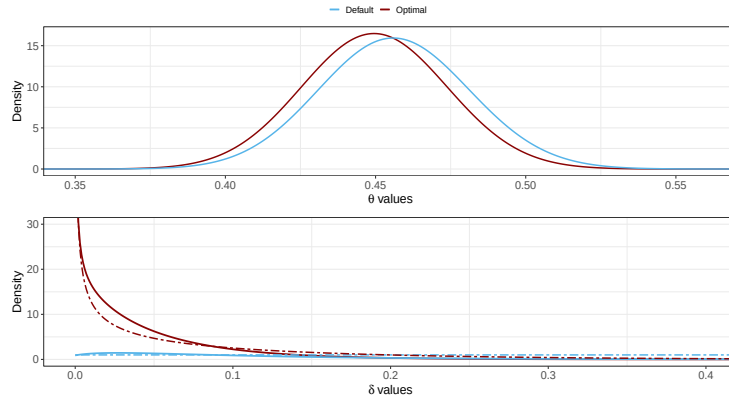


Fig. 1: Marginal posterior distribution for δ and θ under the default (blue continuous lines) and optimal (red continuous lines) initial prior for δ . The dashed lines represents the initial prior for δ under the two scenarios.

4 Discussion

This paper provides a preliminary and concise assessment of applying the calibrated Bayes factor algorithmic approach for power prior elicitation within a fully Bayesian framework, particularly when a Beta initial prior is applied to the weight parameter δ . Our preliminary application suggests that the proposed methodology is a valuable and powerful measure of evidence for discriminating between some competing prior specifications. Moreover, we feel our method could help clinicians to take decisions, or, at least, use past information with more awareness.

However, there are several directions for future research. We plan to conduct simulation studies to deepen our understanding on how the calibrated Bayes factor influences the selection of an optimal initial prior for δ . Additionally, we aim at undertaking a more comprehensive analysis to refine and enhance the criterion used for determining a somehow defined optimal prior.

References

1. Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1):46 – 60, 2000. doi: 10.1214/ss/1009212673.
2. Joseph G. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015. doi: <https://doi.org/10.1002/sim.6728>.
3. Yuyan Duan, Keying Ye, and Eric P. Smith. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106, 2006. doi: <https://doi.org/10.1002/env.752>.
4. Yueqi Shen, Luiz M. Carvalho, Matthew A. Psioda, and Joseph G. Ibrahim. Optimal priors for the discounting parameter of the normalized power prior, 2023.
5. John M Kirkwood, Myla H Strawderman, Marc S Ernstoff, Thomas J Smith, Ernest C Borden, and Robert H Blum. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial est 1684. *Journal of Clinical Oncology*, 14(1):7–17, Jan 1996. doi: 10.1200/JCO.1996.14.1.7.
6. Harold Jeffreys. *The theory of probability*. Oxford University Press, 1961.
7. Robert E. Kass. Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42(5):551–560, 1993. ISSN 00390526, 14679884.
8. Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 01621459.
9. Daniel J Schad, Bruno Nicenboim, Paul-Christian Bürkner, Michael Betancourt, and Shravan Vasishth. Workflow techniques for the robust use of bayes factors. *Psychological Methods*, 28(6):1404–1426, 2023. doi: 10.1037/met0000472.
10. Gonzalo Garcia-Donato and Ming-Hui Chen. Calibrating Bayes factor under prior predictive distributions. *Statistica Sinica*, 15(2):359–380, 2005. ISSN 10170405, 19968507.
11. Christian P. Robert. 50 shades of bayesian testing of hypotheses. In Arni S.R. Srinivasa Rao, G. Alastair Young, and C.R. Rao, editors, *Advancements in Bayesian Methods and Implementation*, volume 47 of *Handbook of Statistics*, pages 103–120. Elsevier, 2022. doi: <https://doi.org/10.1016/bs.host.2022.06.003>.
12. Massimiliano Pastore. Overlapping: a R package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, 3(32):1023, 2018. doi: 10.21105/joss.01023.
13. Massimiliano Pastore and Antonio Calcagni. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01089.

Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables

Dalila Failli, Bruno Arpino, and Maria Francesca Marino

Abstract We extend the Mixture of Latent Trait Analyzers (MLTA) with concomitant variables in a multilevel framework to perform a hierarchical clustering of first- and second-level units. The use of Mixture of Latent Trait Analyzers allows us to cluster first-level units, while also accounting for the residual variability of items in the data matrix. Furthermore, a multilevel approach allows us to account for the hierarchical structure of the data. Last, the inclusion of concomitant variables enables us understanding how first-level units' characteristics influence clustering formation.

Key words: Model-based clustering, Multilevel analysis, Variational Inference, Non-Parametric Maximum Likelihood

1 Introduction

The Mixture of Latent Trait Analyzers (MLTA) model as introduced in the literature by [6, 7] allows us to identify latent groups of units sharing unobserved characteristics, while also capturing the residual dependence between items in a binary data matrix. [5] extended this proposal by allowing for unit features to have an impact on the prior probability to belong to a given mixture component, in a concomitant variable perspective. In this work, we focus on a three-way binary data structure,

Dalila Failli
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze, e-mail: dalila.failli@unifi.it

Bruno Arpino
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Battisti 241 - 35121 Padova, e-mail: bruno.arpino@unipd.it

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze, e-mail: mariafrancesca.marino@unifi.it

where rows correspond to first-level units, columns to items, and layers to second-level units. Given the three-way structure of the data, we further extend the model in a multilevel fashion, by including second-level random effects on the multinomial logit model for the prior component probabilities. These are assumed to have an un-specific distribution which is directly estimated from the data in a Non-Parametric Maximum Likelihood (NPML; [9]) framework. This leads to an estimated discrete distribution for the random effects that, as a by-product, allows for a clustering of second-level units. Overall, the proposed model allows for a hierarchical clustering of first- and second-level units.

2 Method

Let Y_{hik} be the k -th binary item associated to the i -th first-level unit nested within the h -th second-level unit, with $h = 1, \dots, H$, $i = 1, \dots, n_h$, and $k = 1, \dots, R$. The whole set of data can be formally thought of as a three-way structure, where rows correspond to first-level units, columns to items, and layers to second-level units. The proposed model assumes that first-level units can be grouped into homogeneous classes identified by the latent random variable $\mathbf{z}_{hi} = (z_{hi1}, \dots, z_{hiG})' \stackrel{iid}{\sim} \text{Multinomial}(1, (\eta_{hi1}, \dots, \eta_{hiG})')$, with elements

$$z_{hig} = \begin{cases} 1 & \text{if first-level unit } i \text{ within second-level unit } h \text{ belongs to group } g, \\ 0 & \text{otherwise.} \end{cases}$$

Here, η_{hig} denotes the prior probability for the i -th first-level unit, within the h -th second-level unit, to belong to group g , with $g = 1, \dots, G$. Further, as in the latent trait analysis, a D -dimensional continuous latent trait, $\mathbf{u}_{hi} \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I})$, is assumed to capture residual dependence between the multiple binary items recorded from the same first-level unit belonging to the g -th group. Such a parametric assumption is mathematically convenient and computationally efficient. Moreover, it has been shown in the literature to be a reasonable approximation for many real phenomena, and this makes the model flexible and applicable in a wide range of contexts. The model assumes that, conditional on $\mathbf{z}_{hi}$ and $\mathbf{u}_{hi}$, the elements in the vector $\mathbf{Y}_{hi} = (Y_{hi1}, \dots, Y_{hiR})'$ are independent Bernoulli random variables with success probability $\pi_{gk}(\mathbf{u}_{ig})$, $k = 1, \dots, R$, modeled via the following logistic function:

$$\pi_{gk}(\mathbf{u}_{hi}) = \Pr(y_{hik} = 1 \mid \mathbf{u}_{hi}, z_{hig} = 1) = \frac{1}{1 + \exp[-(b_{gk} + \mathbf{w}'_{gk} \mathbf{u}_{hi})]}, \quad 0 \leq \pi_{gk}(\mathbf{u}_{hi}) \leq 1.$$

Here, the model intercept b_{gk} represents the attractiveness of the k -th item for first-level units belonging to the g -th group, while the slopes, $\mathbf{w}_{gk}$, capture the influence of the latent trait on the probability of a response between item k and the first-level units belonging to the g -th group.

To account for the effect that covariates may have on the clustering of first-level

units, we follow the approach proposed by [4] and we assume that probabilities η_{hig} depend on a J -dimensional vector of (component-specific) parameters, $\boldsymbol{\beta}_g$. These are meant to measure the impact of covariates $\mathbf{x}_{hi}$ on clustering formation. Further, with the aim of capturing sources of unobserved heterogeneity determining dependence between responses coming from the same second-level unit, we let component probabilities η_{hig} also depend on a second-level-specific random effect, ε_h . That is,

$$\eta_{hig} = \Pr(z_{hig} = 1 \mid \mathbf{x}_{hi}, \varepsilon_h, \boldsymbol{\beta}_g) = \frac{\exp\{\mathbf{x}_{hi}' \boldsymbol{\beta}_g + \varepsilon_h\}}{1 + \sum_{g'=2}^G \exp\{\mathbf{x}_{hi}' \boldsymbol{\beta}_{g'} + \varepsilon_h\}}, \quad g = 2, \dots, G.$$

To obtain a partitioning of second level units, we leave the random effects distribution unspecified and estimate it from the data via a Non-Parametric Maximum Likelihood approach (NPML; [9]), as suggested by [1, 2]. This leads to the estimation of a discrete distribution defined over a finite number of support points $\{\gamma_1, \dots, \gamma_Q\}$ with corresponding weights $\{\rho_1, \dots, \rho_Q\}$. In detail, each second-level unit is assumed to belong to one out of Q unobserved groups identified by the Q -dimensional, discrete, latent variable $\mathbf{v}_h = (v_{h1}, \dots, v_{hQ})'$. Its generic element is

$$v_{hq} = \begin{cases} 1 & \text{if second-level unit } h \text{ belongs to group } q, \\ 0 & \text{otherwise,} \end{cases}$$

with $q = 1, \dots, Q$. Here, $\rho_q = \Pr(v_{hq} = 1)$ is the prior probability for the h -th second-level unit to belong to group q , under the constraints that $\sum_{q=1}^Q \rho_q = 1$ and $\rho_q \geq 0, \forall q = 1, \dots, Q$.

2.1 Parameter estimation

Let $\boldsymbol{\theta}$ be the vector of all free model parameters. Given the assumptions described in the previous section, the log-likelihood function of the model can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^H \log \left(\sum_{q=1}^Q \rho_q \prod_{i=1}^{n_h} \sum_{g=1}^G \eta_{higq} \int f(\mathbf{y}_{hi} \mid \mathbf{u}_{hi}, z_{hig} = 1) f(\mathbf{u}_{hi}) d\mathbf{u}_{hi} \right). \quad (1)$$

where $f(\mathbf{y}_{hi} \mid \mathbf{u}_{hi}, z_{hig} = 1) = \prod_{k=1}^R f(y_{hik} \mid \mathbf{u}_{hi}, z_{hig} = 1)$, with $f(y_{hik} \mid \mathbf{u}_{hi}, z_{hig} = 1) = (\pi_{gk}(\mathbf{u}_{hi}))^{y_{hik}} (1 - \pi_{gk}(\mathbf{u}_{hi}))^{1-y_{hik}}$. Further, $f(\mathbf{u}_i)$ is a standard Gaussian density, and $\eta_{higq} = \Pr(z_{hig} = 1 \mid \mathbf{x}_{hi}, \varepsilon_h = \gamma_q; \boldsymbol{\beta}_g)$. From Equation (1), we can easily recognize the log-likelihood function of a hierarchical mixture model [11].

Maximum likelihood estimates of model parameters can be obtained by maximizing this function via an EM algorithm [3]. However, the integral appearing in Equation (1) does not admit a closed-form expression; to solve the issue, we consider its variational approximation, following the approach suggested by [8]. Compared to

other methods, such as the Gauss-Hermite quadrature, estimation using variational approximation is faster and straightforward to implement. Overall, model parameter estimates are derived by alternating the following steps until convergence.

- E-step: compute the expectation of the complete data log-likelihood, conditional on the observed data and the current parameter estimates $\boldsymbol{\theta}^{(t)}$; this consists in computing the posterior expectation of z_{hiq} , v_{hiq} , and the posterior expectation of z_{hiq} , conditional on $v_{hiq} = 1$.
- M-step: update $\boldsymbol{\beta}_g$, $\boldsymbol{\gamma}_q$, and $\boldsymbol{\rho}_q$ by maximizing the expected complete data log-likelihood function with respect to $\boldsymbol{\theta}$. A second EM algorithm, nested within the first, is required to update the posterior mean vectors and covariance matrices of $\mathbf{u}_{hi}$, as well as the variational parameters, and the latent trait parameters, b_{gk} and $\mathbf{w}_{gk}$.

These steps are iterated until convergence is reached. Furthermore, an initialization strategy based on multiple starting points may help the estimation algorithm avoid being trapped in local maxima solutions.

3 Simulation study

The ability of the proposal in correctly estimating model parameters and clustering first- and second-level units is evaluated through a simulation study based on $B = 100$ replicates, each based on 100 random starting values. We consider different scenarios based on a varying number of first-level units ($N = 500, 1000, 2000$), while the number of second-level units is set equal to $H = 20$. Further, we consider $G = 3$ first-level groups, and $Q = 2$ second-level groups. Lastly, the latent trait size and the number of items are set equal to $D = 1$ and $R = 7$, respectively. Class membership is defined in terms of a single nodal attribute (i.e., $J = 1$) drawn from a Gaussian distribution with mean and variance both equal to 1.

The ability of the proposal in correctly classifying first- and second-level units is evaluated via the Adjusted Rand Index (ARI; [10]) reported in Table 1. Looking at this table, we note that, as the number of first-level units N increases, the classification improves for both first- and second-level units.

Table 1 Adjusted Rand Index: mean (median) across samples for first- and second-level partitions, with varying N .

	First-level units	Second-level units
$N=500$	0.791 (0.788)	0.960 (1.000)
$N=1000$	0.793 (0.781)	0.980 (1.000)
$N=2000$	0.796 (0.785)	1.000 (1.000)

Table 2 shows the Mean Squared Error (MSE) values across samples for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and

ρ . Looking at this table, it is evident that, as the number of first-level units increases, we are more and more able to correctly estimate model parameters.

Table 2 MSE values across samples for β , γ , and ρ , with varying N .

	β	γ	ρ
$N=500$	[0.01 0.02]	[0.00 0.07]	[0.01 0.01]
$N=1000$	[0.01 0.02]	[0.00 0.04]	[0.01 0.01]
$N=2000$	[0.01 0.00]	[0.00 0.04]	[0.01 0.01]

4 Conclusion

The proposed model provides a hierarchical clustering of first- and second-level units in a multilevel framework. This is achieved by considering a hierarchical mixture of latent trait models. In detail, the success probability for a first-level unit to a given item is modeled as a function of attractiveness and influence parameters. The former are assumed to depend on the mixture component the first-level unit comes from; the latter allows us to explain residual heterogeneity between first-level units in terms of the way they respond to the different items. Observed covariates are included in the model with the aim of analyzing the effect they have on clustering formation. Further, the inclusion of discrete, second-level-specific random variables, whose distribution is estimated via a NPML approach, allows for the clustering of second-level units too. The simulation study highlights the effectiveness of the proposal in correctly classifying first- and second-level units and identifying model parameters.

From a methodological point of view, the proposed approach can be extended to frameworks with nominal and ordinal response variables. We may also consider more than two levels of clustering, as well as concomitant variables at each level.

References

1. Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalized linear models. *Stat. Comput.* **6**, 251–262 (1996)
2. Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128 (1999)
3. Dempster, A.P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc., Ser. B*, **39**, 1—38 (1977)
4. Failli, D., Marino, M.F., Martella, F.: Extending finite mixtures of latent trait analyzers for bipartite networks. In: Balzanella A., Bini M., Cavicchia C., Verde R. (eds.) *Book of short Paper SIS 2022*, Pearson (2022)

5. Failli, D., Marino, M.F., Martella, F.: An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks. In: Chelli, F.M., Ciommi, M., Ingrassia, S., Mariani, F., Recchioni, M.C. (eds.) *Book of short Papers SIS 2023*, Pearson (2023)
6. Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* **24**, 569–588 (2014)
7. Gollini, I.: A mixture model approach for clustering bipartite networks. In: *Challenges in Social Network Research Volume in the Lecture Notes in Social Networks (LNSN - Series of Springer)* (2020)
8. Jaakkola, T. S., Jordan, M. I.: Bayesian logistic regression: A variational approach. In: Madigan, D., Smyth, P. (eds.) *Proceedings of the 1997 Conference on Artificial Intelligence and Statistics*. Ft. Lauderdale, FL, (1997)
9. Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. *JASA* **73**, 805–811 (1978)
10. Rand, W. M.: Objective Criteria for the Evaluation of Clustering Methods. *JASA* **66**, 846–850 (1971)
11. Vermunt, J. K., Magidson, J.: Hierarchical mixture models for nested data structures. In: Weihs, C., Gaul, W. (eds.), *Classification: The Ubiquitous Challenge*, 176–183, Springer (2005)

A Simultaneous Spectral Clustering for Three-Way Data

Cinzia Di Nuzzo and Salvatore Ingrassia

Abstract We introduce a novel approach to spectral clustering for three-way data, which integrates simultaneous dimensionality reduction and clustering. While conventional spectral clustering methods focus on two-way data and treat dimensionality reduction and clustering separately, our proposed method extends to handle three-way data, capturing temporal repetition and multivariate interactions. This is the first method, which tackles this challenge purely through statistical techniques.

Key words: Spectral Clustering, Dimensionality reduction, Three-way data

1 Introduction and Background Theory

The Spectral Clustering algorithm is a clustering technique that leverages the spectral structure of data to partition them into homogeneous groups, see [11], [10]. Unlike traditional approaches such as K -Means, Spectral Clustering does not require assuming a specific cluster shape and can handle non-linearly separable data. The Spectral Clustering procedure can be divided into several phases. Initially, a similarity matrix capturing the relationships between data instances is constructed. Subsequently, the Laplacian matrix of the similarity matrix is computed, and the eigenvectors corresponding to the smallest eigenvalues of the Laplacian matrix are determined. These eigenvectors are used to represent the data in a reduced-dimensional space, where the cluster structure is more evident. Finally, a clustering algorithm (such as K -Means, but also mixture models have been taken into account [2]) is applied to this reduced representation to assign instances to clusters. Spectral Clus-

Cinzia Di Nuzzo

Department of Economics and Business, University of Catania, Corso Italia, 55, Catania, 95129, Italy, e-mail: cinzia.dinuzzo@unict.it

Salvatore Ingrassia

Department of Economics and Business, University of Catania, Corso Italia, 55, Catania, 95129, Italy, e-mail: salvatore.ingrassia@unict.it

tering is particularly useful when data exhibit complex or non-linear structures, and when assuming a specific cluster shape is inappropriate. It is effective even when clusters have irregular shapes or when there are clusters of vastly different sizes. Moreover, Spectral Clustering can efficiently handle high-dimensional data and can be adapted to address clustering problems on graphs or non-Euclidean/categorical data [5].

The simultaneous integration of the spectral clustering method as a dimensionality reduction and clustering algorithm is crucial for maximizing efficiency and coherence in data representation. A simultaneous approach allows capturing the intrinsic relationships among the data, reducing computational complexity, and enhancing precision in creating homogeneous groups. A first method to combine the dimensionality reduction and the clustering step in this context has been proposed by [9]. Moreover, extending the spectral clustering method simultaneously for three-way data is essential for addressing situations where the temporal repetition of data is relevant and interactions between three variables can influence results over time. An example can be data concerning user behaviour and interactions on an online social media platform over a certain time interval. Some developments of spectral clustering for three-way data have already been proposed in [4], [3]. By using the spectral clustering method simultaneously for dimensionality reduction and clustering on these three-way data, complex patterns in user behaviour over time can be identified. For example, discovering groups of users with similar characteristics interacting similarly on the platform during specific periods of the day or in response to certain types of content. This would enable better customization of the user experience, adapting the type of content shown or the timing of notifications based on the different user segments identified.

This paper proposes a novel approach to spectral clustering that integrates simultaneous dimensionality reduction and clustering for three-way data. The basic idea is to extend the results presented in [9]. By simultaneously conducting dimensionality reduction and clustering, our approach captures intricate patterns in data representation, thereby enhancing computational efficiency and coherence. We demonstrate the utility of our method through first empirical evaluations on synthetic and real-world datasets. The results showcase the effectiveness of our proposed approach in uncovering complex patterns and facilitating meaningful insights in diverse application domains. Overall, our contribution advances the field of spectral clustering by addressing the challenges posed by three-way data structures and providing a versatile framework for clustering in various contexts.

The rest of the work is structured as follows: Section 2 introduces the foundational concept of the model, while Section 3 provides an illustration using one artificial data and one real-world data.

2 The Model

A novel spectral clustering method for three-way data is presented. Spectral clustering is a two-step sequential procedure involving the reduction of data dimensionality

through Laplacian embedding, followed by the application of a clustering algorithm to partition the data into K clusters. Although spectral clustering methods have conventionally been applied to two-way data, we recently proposed a spectral clustering approach for three-way data in [3] and [4]. The objective is to partition a three-way dataset comprising N units, M variables, and H occasions into K clusters. The innovation of the proposed method lies in modelling a set of H Laplacian matrices $\mathbf{L}_h$ (for $h = 1, \dots, H$) of size $N \times N$ as follows:

$$\mathbf{L}_h = \mathbf{A}\mathbf{C}_h\mathbf{A}' + \mathbf{E}_h, \quad (1)$$

such that $\mathbf{A}\mathbf{A}' = \mathbf{I}$, for $h = 1, \dots, H$; where $\mathbf{L}_h$ denotes the Laplacian matrix, $\mathbf{A}$ represents the $N \times K$ matrix for Laplacian embedding, $\mathbf{C}_h$ is a $K \times K$ diagonal matrix with non-negative elements, and $\mathbf{E}_h$ is the $N \times N$ error term. We adopt a least squares approach to estimate the part of dimensionality reduction of the model, based on the following minimisation under orthogonality constraints

$$\underset{\mathbf{A}, \mathbf{C}_h}{\operatorname{argmin}} \sum_{h=1}^H \|\mathbf{L}_h - \mathbf{A}\mathbf{C}_h\mathbf{A}'\|^2. \quad (2)$$

According to [1], because the direct minimisation of (2) seems difficult; a Candecomp method can be developed to minimize (2), therefore the minimisation problem can be solved by

$$\underset{\mathbf{A}, \mathbf{B}, \mathbf{C}_h}{\operatorname{argmin}} \sum_{h=1}^H \|\mathbf{L}_h - \mathbf{A}\mathbf{C}_h\mathbf{B}'\|^2, \quad (3)$$

where $\mathbf{B}$ is an $N \times K$ matrix. In this context, a notable contribution arises from the findings presented in [6] and the associated algorithm known as Indort. Particularly, [6] demonstrates that if the matrices $\mathbf{L}_h$ are positive definite and the elements of $\mathbf{C}_h$ are non-negative, then the resulting matrices $\mathbf{A}$ and $\mathbf{B}$ in (3) obtained from the Indort algorithm are identical, thus (3) solves the least squares problem of (2). These results facilitate the derivation of a unique configuration of Laplacian embedding shared among all H Laplacian matrices $\mathbf{L}_h$, rendering the matrix $\mathbf{A}$ representative of the entire dataset. Within this framework, to conduct dimensionality reduction, an algorithm for clustering is also modelled concurrently, employing K -means and essentially extending the model outlined in [9] to the three-way case.

Finally, it should be noted that the computational burden of the algorithm proposed by [9] is of the order $O(N^2K)$; in the model proposed here, an additional step is merely appended wherein individual occasions sums are computed within the specific SVD calculation in the estimation process of matrices $\mathbf{A}$ and $\mathbf{B}$. Another difference from the method proposed in [9] is the estimation of matrix $\mathbf{C}_h$, which is introduced in this model to represent the weights of the occasion h in each cluster. It is also important to note that the latter part of the method, concerning the clustering process, remains unaltered compared to the method already proposed in [9], as in the model we are proposing here, we opt to reduce the original data dimensionality of size $N \times M \times H$ to a common space of dimension $N \times K$ shared across all occasions. Hence, in terms of computational burden and utility of the method in the dimension-

ality reduction process, this model facilitates resolving situations with initially high dimensions, which are subsequently condensed into a much smaller common space containing the structure of clusters from the original data.

3 Some results on synthetic and real-word datasets

Here, we provide some examples illustrating the effectiveness of the method. The first example concerns an artificial dataset originating from $K = 3$ normal distributions each comprising $N = 20$ units, with means $\mu_1^{(1)} = (1.5, 1, 2)$, $\mu_2^{(1)} = (0, 0, 0)$, and $\mu_3^{(1)} = (-1, -2.5, -2)$, respectively, and, covariance matrices equal to $\Sigma^{(1)} = \mathbf{I}$. We repeat the random generation of such data $H = 3$ times, specifically, we generate randomly the second occasion from 3 normal distribution with $\mu_1^{(2)} = (2.5, 1.5, 2.5)$, $\mu_2^{(2)} = (0.2, 0.3, 0.4)$, and $\mu_3^{(2)} = (-2.5, -2.4, -2)$ and $\Sigma^{(2)} = 2\mathbf{I}$. Finally we generate the third occasion from normal distribution with $\mu_1^{(3)} = (6, 5, 5)$, $\mu_2^{(3)} = (5, 4, 3.7)$, and $\mu_3^{(3)} = (3.7, 2.8, 1.8)$ and $\Sigma^{(3)} = 0.5\mathbf{I}$, respectively.

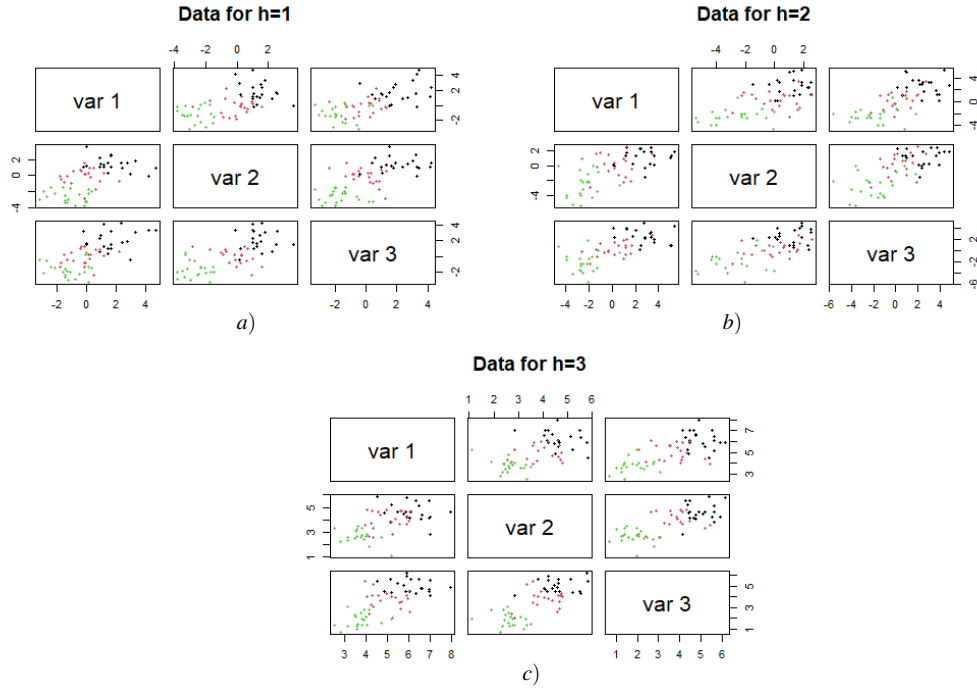


Fig. 1 Synthetic data. a) $\mathbf{X}_1$. b) $\mathbf{X}_2$. c) $\mathbf{X}_3$.

Denoting the dataset of $N = 60$ units and $M = 3$ variables across $H = 3$ occasions as $\mathbf{X}$, the original data are depicted in Figure 1, it should be noted that the other settings exhibit similar patterns, differing primarily in the magnitude of the data. Upon constructing the original dataset $\mathbf{X}$, three Laplacian matrices, namely $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$, are generated such that $\mathbf{L}_h = \mathbf{D}_h^{1/2} \mathbf{W}_h \mathbf{D}_h^{-1/2}$, for $h = 1, 2, 3$, where $\mathbf{W}_h$ represents the similarity matrix constructed using a kernel function (refer to [12]):

$$w_{ij}^{(h)} = \exp \left(-\frac{\|\mathbf{x}_i^{(h)} - \mathbf{x}_j^{(h)}\|^2}{\varepsilon_i^{(h)} \varepsilon_j^{(h)}} \right), \quad \text{for } h = 1, \dots, 3 \quad (4)$$

with $\varepsilon_i^{(h)} = \|\mathbf{x}_i^{(h)} - \mathbf{x}_m^{(h)}\|$, where $\mathbf{x}_m^{(h)}$ is the m -neighbor of the point $\mathbf{x}_i^{(h)}$, for $i, j = 1, \dots, N$, and $m = 2$. Additionally, the matrix $\mathbf{D}_h$ signifies the degree matrix associated with $\mathbf{W}_h$ for $h = 1, \dots, 3$. Employing the proposed algorithm outlined in (3) to process the set of matrices $\mathbf{L}_h$, with the number of groups predetermined, the resulting reduced-dimensional space, as determined by the clustering outcome, is depicted in Figure 2 (ARI=1).

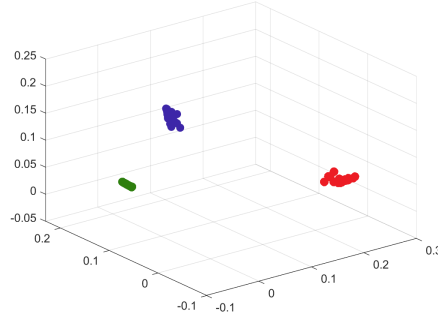


Fig. 2 *Synthetic data*. Clustering result on the embedded space.

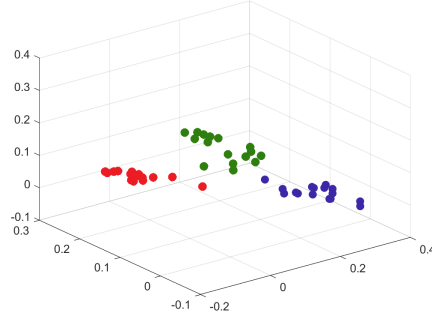


Fig. 3 *Blue Crabs data*. Clustering result on the embedded space.

The second example pertains to a commonly used dataset called Blue Crabs data. Blue crabs data can be accessed at <https://three-mode.leidenuniv.nl/> (see [7] and [8]). The dataset comprises gill, hepatopancreas, and muscle tissue samples extracted from blue crabs, categorized into three groups based on their origin and health condition: Albermarle Sound (Healthy), Pamlico River (Healthy), and Pamlico River (Diseased). The dataset consists of $N = 48$ units (tissue samples) by $M = 3$ variables (tissue types) by $H = 25$ occasions (trace elements). Our model was applied to this dataset, and the result in the embedded space is depicted in Figure 3. In this case as well, our method succeeded in achieving correct classification (ARI=1).

Acknowledgement

This study was funded by the GRINS project PE00000018-CUP E63C22002120006 (*Di Nuzzo Cinzia*), and by PNRR MUR project PE0000013-FAIR (*Salvatore Ingrassia*).

References

1. Carroll, J.D., and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition, *Psychometrika*, **35**, 283–319.
2. Di Nuzzo, C., Ingrassia, S. (2022). A mixture model approach to spectral clustering and application to textual data, *Statistical Methods & Applications*, **31**, 1071–1097.
3. Di Nuzzo, C., Ingrassia, S., Vicari, D. (2022). An INDSCAL-Type Approach for Three-Way Spectral Clustering. In “García-Escudero L.A., Gordaliza A., Mayo-Iscar A., Lubiano Gomez M.A., Gil M.A., Grzegorzewski P., Hryniewicz O. (Eds.) *Building Bridges between Soft and Statistical Methodologies for Data Science*”, *Advances in Intelligent Systems and Computing*, Springer, **1433**, 128–135.
4. Di Nuzzo C., Ingrassia S. (2023). Three-way Spectral Clustering. In “Brito P., Dias J.G., Lausen B., Montanari A., Nugent R. (Eds.) *Classification and Data Science in the Digital Age*”, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 111–118.
5. Di Nuzzo, C. (2024). Advancing Spectral Clustering for Categorical and Mixed-Type Data: Insights and Applications. *Mathematics*, **12**(4): 508.
6. Dosse, M.B., ten Berge, J.M. & Tendeiro, J.N. (2011). Some New Results on Orthogonally Constrained Candecomp. *Journal of Classification*, **28**(2), 144–155.
7. Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., & Bray, J. T. (1992). Principal component analysis, trace elements, and blue crab shell disease. *Analytical Chemistry*, **64**, 523–531.
8. Kroonenberg, P. M., Basford, K. E., & Gemperline, P. J. (2004). Grouping three-mode data with mixture methods: the case of the diseased blue crabs. *Journal of Chemometrics*, **18**, 508–518.
9. Labiod, L., Nadif, M. (2021). Efficient regularized spectral data embedding. *Advances Data Analysis Classification*, **15**, 99–119.
10. Ng, A.Y., Jordan, M., Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In “Dietterich T., Becker S. and Ghahramani Z. (Eds.) *Advances in neural information processing systems*”, MIT Press, **14**.
11. von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing* **17**, 4, 395–416.
12. Zelnik-Manor, L., Perona, P. (2004). Self-tuning spectral clustering. In “L. Saul and Y. Weiss and L. Bottou (Eds.) *Advances in Neural Information Processing Systems*”, MIT Press, **17**.

7. Solicited - From Data Analysis to Data Science

7.1 - Optimal Scaling: New Insights Into an Old Problem

Optimal Scaling: New Insights Into an Old Problem

Gilbert Saporta

Abstract Processing qualitative variables with a very large number of categories in Machine Learning is an opportunity to revisit the theory of optimal scaling and its applications.

Key words: optimal scaling, encoding, correspondence analysis.

1 A Brief History

Coding (or scoring) a qualitative variable consists in assigning numerical values to its modalities, thus transforming it into a discrete numerical variable allowing the use of methods designed for numerical data. Scoring qualitative variables has a long history going back to Hirschfeld [8], Fisher [4], Williams [17], Guttman [6], Hayashi [7,16], among others. It was the origin of correspondence analysis [11,13].

It is interesting to note, as did Kendall & Stuart [9], that Lancaster's theorem [10] implies that the search for separate scoring systems for two categorised variables such as to maximise their correlation, comes down to trying to produce a bivariate normal distribution by operations upon the marginal distributions.

The 1970s and early 1980s were a particularly fertile period for the development of optimal scaling (ie scoring) in supervised and unsupervised contexts, performed with alternating least squares between model parameters and data parameters (the codings) [5,18].

Then, for almost 40 years, the topic did not generate much research; applications became routine, such as risk scores in banking and insurance [2,15].

¹

Gilbert Saporta, Cedric-Cnam, Paris, France; gilbert.saporta@cnam.fr

2 Machine Learning and Variable Encoding

With the availability of massive data, machine-learning researchers and practitioners were confronted with categorical data, ill-suited to neural networks with moreover a large number of categories (eg zip codes).

Generally ignoring the old works of statisticians, dozens of encoding methods [14], some quite arbitrary or rediscovered, have flourished in ML literature, like Hash encoding, methods where the encoding only depends on the response variable (conditional average), as well as the One-Hot Encoding (OHE). OHE is nothing else than the well-known disjunctive form of categorical variables with as many indicators as categories. It should be noted that OHE is more a representation of categories than an encoding, since it is not a transformation into a single numerical variable. A. Di Ciaccio [3] proposes a review of encoding methods from the point of view of a statistician.

Machine Learning also offers methods adapted to massive data. Linear methods can be modelled by neural networks, which in this case are not very efficient computationally, but the use of networks such as auto-encoders with nonlinear links and minimising cross-entropy may give better results than the ALS methods [1].

One of the essential contributions of Machine Learning lies in the learning-validation approach to avoid overfitting. A large number of categories for some variables raises problems of stability and overfitting. These issues were neglected in usual statistical applications where the number of modalities is small. [12] shows how regularization may be applied in the context of regression with optimal scaling features.

3 Conclusions and Perspectives

The confrontation of Statistics and Machine Learning worlds allows us to consider a renewal of the coding methods, from both a theoretical and a practical point of view. It must be emphasised that there is no optimal coding per se: it depends on the problem and on the criterion to be optimised.

References

1. Abdi, H., Di Ciaccio, A., Saporta, G.: Old and New Perspectives on Optimal Scaling. In: Beh, E.J., Lombardo, R., Clavel, J.G. (eds.) *Analysis of Categorical Data from Historical Perspectives. Behaviormetrics: Quantitative Approaches to Human Behavior*, vol 17, pp. 131-154. Springer, Singapore. (2023)
2. Bouroche, J. M., Saporta, G.: Les méthodes et les applications du credit-scoring. In: 34^e Riunione Scientifica della Società Italiana di Statistica, Siena, pp. 19-26. (1988)
3. Di Ciaccio, A.: Optimal Coding of categorical data in machine learning. In: L. Grilli, M. Lupporelli, E. Rocco, C. Rampichini, M. Vichi (eds.), *Statistical Models and Methods for Data Science. Studies in Classification, Data Analysis and Knowledge Organization*, pp. 39-51. Springer, Cham (2023)

4. Fisher, R.A.: The precision of discriminant functions. *Ann. Eugenics* **10**, 422–429 (1940)
5. Gifi, A.: *Nonlinear Multivariate Analysis*. Wiley, Chichester (1990)
6. Guttman, L.: The quantification of a class of attributes: a theory and method of a scale construction. In: Horst, P., Wallin, P., Guttman, L. (eds.) *The Prediction of Personal Adjustment*, 321–348. Social Science Research Council, New York (1941)
7. Hayashi, C.: On the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Stat. Math.* **2**, 35–47 (1950)
8. Hirschfeld, H.O.: A connection between correlation and contingency. *Math. Proc. Camb. Philos. Soc.* **31**, 520–524 (1935)
9. Kendall, M.G., Stuart, A.: *The Advanced Theory of Statistics*, vol. II. Charles Griffin, London (1961)
10. Lancaster, H.O.: Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **44**, 289–292 (1957)
11. Lebart, L., Saporta, G.: Historical elements of correspondence analysis and multiple correspondence analysis. In: Blasius, J., Greenacre, M. (eds.) *Visualization and Verbalization of Data*, 73–86. Chapman & Hall/CRC, Boca Raton (2014)
12. Meulman, J.J., van der Kooij, A.J., Duisters, K.L.: ROS regression: integrating regularization with optimal scaling regression. *Stat. Sci.* **34**, 361–390 (2019)
13. Nishisato, S.: *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press, Toronto (1980)
14. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. *Int. J. Comput. Appl.* **175**(4), 7–9 (2017)
15. Saporta, G., Niang-Keita, N.: Correspondence analysis and classification. In: Greenacre, M., Blasius, J. (eds.) *Multiple Correspondence Analysis and Related Methods*, pp. 371–392. Chapman and Hall/CRC, Boca Raton (2006)
16. Tanaka, Y.: Review of the methods of quantification. *Environ. Health Perspect.* **32**, 113–123 (1979)
17. Williams, E.J.: Use of scores for the analysis of association in contingency tables. *Biometrika* **39**, 274–289 (1952)
18. Young, F.W.: Quantitative analysis of qualitative data. *Psychometrika* **46**, 357–388 (1981)

8. Solicited - Statistical methods for textual data

- 8.1 - PROCSIMA: Probability Distribution Clustering Using Similarity Matrix Analysis
- 8.2 - Exploring Anti-Migrant Rhetoric on Italian Social Media
- 8.3 - Causal inference from texts: a random-forest approach

PROCSIMA: Probability Distribution Clustering Using Similarity Matrix Analysis

Marco Ortu

Abstract This study presents PROCSIMA, a methodological approach to document clustering, that defines a similarity metric derived from Jensen-Shannon divergence, for measuring similarities between topic probability distributions obtained from Topic Modeling techniques, such as Latent Dirichlet Allocation (LDA). Unlike conventional approaches that allocate documents to a singular, most pertinent topic, PROCSIMA allocates the clustering of documents by considering their comprehensive topic distribution. By transforming the similarity matrix into an adjacency matrix and subsequently applying community detection algorithms it defines document clusters. Empirical validation on both synthetic and real-world datasets is performed by PROCSIMA by bootstrapping the optimal number of network communities to outperform traditional clustering methods.

Key words: Document Clustering, Topic Modeling, Textual Similarity Metric

1 Introduction

The task of organizing documents into meaningful clusters is a fundamental problem in Information Retrieval (IR) and Natural Language Processing (NLP). Specifically, when documents are represented as probability distributions over topics, typically the output of topic modeling techniques such as Latent Dirichlet Allocation (LDA) [1], the challenge becomes finding an optimal partitioning that groups documents into clusters based on their topical similarity. Identifying the optimal number of clusters k , is critical for achieving meaningful and interpretable clustering results. Several works such as [3, 4] have studied the application of probability distribution metrics, such as Hellinger distance, and network representation to better interpret and visualize topics respectively. This work introduces an algorithm to estimate the

Marco Ortu
University of Cagliari, Cagliari (IT), e-mail: marco.ortu@unica.it

optimal k by maximizing the silhouette score, a measure of how similar an object is to its own cluster compared to other clusters. We define a *Similarity* metric based on Jensen-Shannon diversity, which measures the divergence of two probability distributions. Then we use this metric to compute the similarity matrix of textual documents represented by their topic distributions. Finally, we translate the similarity matrix into an adjacency matrix, and we designed an algorithm to determine the optimal cluster through the partition of the similarity network. Section 2 presents the formal method along with the required background. Section 3 details the findings with both synthetic and real world data and finally Section 4 draws the conclusions and paves the way for future works.

2 Methodology And Background

Let $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ denote a set of documents, where each document d_i is represented as a probability distribution over a fixed set of topics. The objective is to partition $\mathcal{D}$ into k clusters $\{C_1, C_2, \dots, C_k\}$, such that documents within the same cluster are more similar to each other than to those in different clusters. The similarity between documents is derived using a metric appropriate for probability distributions, such as the Jensen-Shannon divergence shown in Equation 1. Such a measure is called Jensen-Shannon Divergence (JSD) and is a particular form of the more general Kullback-Leiber divergence shown in Equation 2, where M is the *mixture distribution*. The root square of JSD is a distance measure as it respects the properties of non-negativity, symmetry, and triangular inequality [5]. We define the **Similarity** between two topic distributions shown in Equation 3.

$$\text{JS}(P, Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M) \quad (1)$$

$$D(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad M = \frac{1}{2}(P + Q) \quad (2)$$

$$\text{Similarity}(P, Q) = 1 - e^{\left(1 - \frac{1}{\sqrt{\text{JS}(P, Q)}}\right)} \quad (3)$$

The exponential transformation $1 - \exp\left(1 - \frac{1}{f(x, y)}\right)$ maps the distance to a range $[0; 1]$, where values closer to 1 indicate low distance (high similarity), and values closer to 0 indicate high distance (low similarity). This transformation is useful for converting distances into similarities and tends to converge to 0 and 1 with the exponential.

To estimate the optimal number of clusters, we employ a bootstrap resampling approach combined with silhouette score analysis. This PROCSSIMA systematically explores different cluster configurations and identifies the number of clusters that, on average, yield the highest silhouette scores across multiple bootstrap samples. In our implementation of PROCSSIMA, the clustering function is defined by the Lou-

vain community detection algorithm [2], and the number of clusters k is determined by bootstrapping the network parameter τ which represents a threshold to retain the network’s edges. This parameter avoids fully connected networks and guarantees the sparsity of the adjacency matrix.

3 Results

We validated our methodology using both synthetic and real word data. For synthetic data, we created random topic distributions across a set of documents, mimicking real-world datasets’ complexity and variability with control parameters. For a real-world case study, we used the 20 Newsgroups dataset¹. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

Synthetic Data. Each document is associated with a distribution across a specified number of topics, and these documents are grouped into a predetermined number of clusters.

Two parameters, cluster diversity and sample diversity are used to generate topic distributions. Cluster Diversity controls the variability among cluster centers. A higher value leads to greater differences between the topic distributions of each cluster’s center. Sample diversity controls the extent of variation among the documents belonging to the same cluster. A higher in-sample diversity value allows for a broader range of topic distributions within a single cluster. We run the experiments for all combinations of cluster and sample diversity from 0.05 to 1 with 0.05 increase. We setup PROCSIMA Algorithm with 1000 bootstraps. Figure 1 reveals significant insights into the relationship between cluster diversity, sample diversity, and clustering performance (measured by ARI scores). As one might expect, best ARI values are concentrated in the lower left part of Figure 1, with higher values of cluster diversity and lower values of sample diversity. We can see that our algorithm works well in a range $[0, 1]$ for cluster diversity and approximately $[0, 0.65]$ for sample diversity, with an average ARI score of 0.47.

Real World Data. We pre-computed the topic distribution using LDA setting 20 topics (the real number of categories), we then sampled 1000 documents and computed the similarity matrix shown in Figures 2, and we apply our algorithm to obtain the optimal clusters.

We compared the ARI score obtained by PROCSIMA to the one obtained by LDA (using the dominant topic as cluster id). We considered two metrics definition (JSD and Hellinger similarity), and two community detection algorithms (Louvain and Spectral Clustering). In all cases, PROCSIMA achieved better results than the LDA, as shown in Table 1. We can see that the defined JSD similarity achieves similar results compared to the well-known Hellinger distance, with slightly higher results in terms of ARI score, except for the Reuters dataset with Spectral Clustering.

¹ <http://qwone.com/~jason/20Newsgroups/>

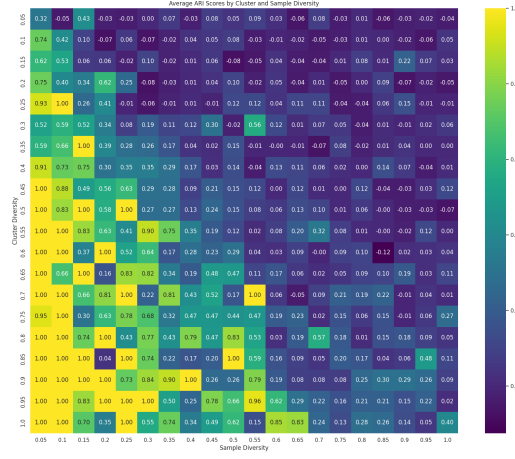


Fig. 1: ARI vs Cluster and Sample Diversity.

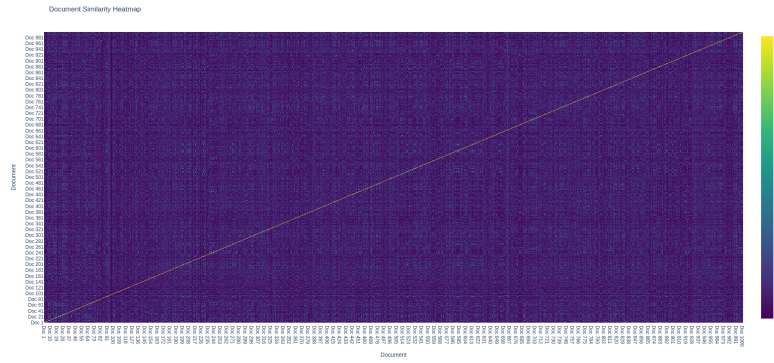


Fig. 2: Similarity Matrix.

In Figures 3 and 4, selected the 20 Newsgroup dataset along with the JSD similarity to visualize the clustering results. The similarity matrix is reordered using cluster ids, clusters visible as blocks along the diagonal where higher similarity within these blocks suggests that the data points are more closely related to each other than to points outside the block. The existence of such blocks can be indicative of a successful clustering process, revealing intrinsic groupings within the data. The variability in cluster sizes suggests heterogeneity in the dataset, with some clusters being more densely populated than others. This could reflect a natural variation in the number of data points related to certain real-world topics or due to the sample considered.

Dataset	Topics	Metric	Algorithm	Clusters	ARI LDA	ARI PROCSIMA
Newsgroup	20	Jensen-Shannon	Louvain	5	0.004	0.110
Newsgroup	20	Jensen-Shannon	Spectral Clustering	30	0.004	0.133
Newsgroup	20	Hellinger	Louvain	5	0.004	0.082
Newsgroup	20	Hellinger	Spectral Clustering	30	0.004	0.128
Reuters	7	Jensen-Shannon	Louvain	4	0.145	0.399
Reuters	7	Jensen-Shannon	Spectral Clustering	16	0.145	0.283
Reuters	7	Hellinger	Louvain	4	0.145	0.381
Reuters	7	Hellinger	Spectral Clustering	17	0.145	0.305

Table 1: Evaluation of Clustering Algorithms on Different Datasets using Jensen-Shannon and Hellinger Similarities.

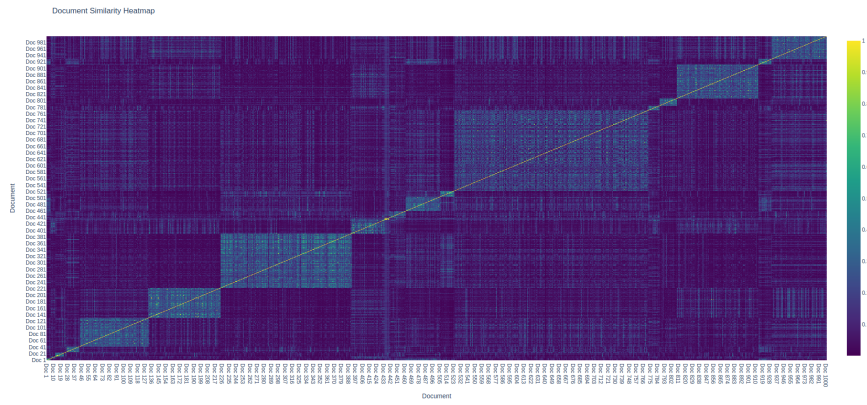


Fig. 3: Similarity Matrix Reordered Using Louvain.

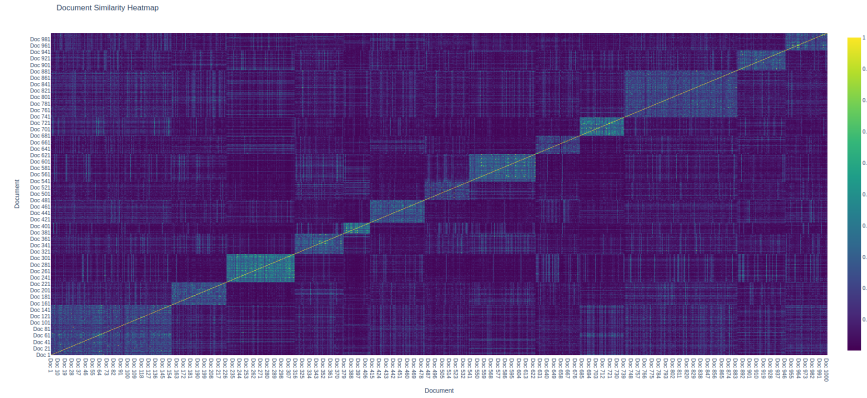


Fig. 4: Similarity Matrix Reordered Using Spectral Clustering.

4 Conclusions

In this study we present PROCSSIMA, a methodological approach to clustering document topic probability distributions based on the definition of *Similarity* measure and similarity network analysis. We validated our methodology with both synthetic and real-world data. The analysis of clustering performance in relation to cluster and sample diversity, indicates a nuanced interaction between diversity levels and clustering outcomes. Higher cluster diversity consistently correlates with improved Adjusted Rand Index (ARI) scores, suggesting its importance for optimal clustering. The real-world examples and comparison with well-established document metrics, such as Hellinger similarity, show the feasibility of the PROCSSIMA algorithm and the JSD similarity. These findings suggest that the proposed methodology achieves better results for a wide range of cluster diversity and lower values of sample diversity. The findings with real-world data tend to confirm the potential of our methodology in document clustering.

References

- [1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.
- [3] Arun S Maiya and Robert M Rolfe. Topic similarity networks: Visual analytics for large document sets. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 364–372. IEEE, 2014.
- [4] Giorgia Minello, Carlo RMA Santagiustina, and Massimo Warglien. Lda2net: Digging under the surface of covid-19 topics in scientific literature. *arXiv preprint arXiv:2112.01181*, 2021.
- [5] Frank Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2019.

Exploring Anti-Migrant Rhetoric on Italian Social Media

Lara Fontanella, Annalina Sarra, Emiliano del Gobbo, Alex Cucco, Sara Fontanella

Abstract In our digital era, the pervasive expansion of social media is leading to a significant transformation in the dynamics of communication. Social media platforms have evolved into arenas where both individuals and communities articulate their sentiments and convictions, be they positive or negative, regarding relevant topics. In the wake of the refugee crisis of 2015, anti-immigration sentiments are permeating public discourse across the global North. Digital environments unleash and amplify everyday racism, facilitated by the ability to remain anonymous and the widespread availability of these messages. The primary objective of this study is to systematically analyze how social media platforms have portrayed immigrants, migrants, asylum seekers, and refugees in recent years. Over 185,000 comments were collected and analyzed using a seededLDA technique.

Key words: social media platforms, racism, ambient racism, LDA, seededLDA

Lara Fontanella

Department of Legal and Social Sciences, University G. d'Annunzio of Chieti-Pescara (Italy)
e-mail: lara.fontanella@unich.it

Annalina Sarra

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio of Chieti-Pescara (Italy) e-mail: annalina.sarra@unich.it

Emiliano del Gobbo

Department of Economics, Management and Territory, University of Foggia (Italy) e-mail: emiliano.delgobbo@unifg.it

Alex Cucco

National Heart and Lung Institute, Imperial College London, UK e-mail: a.cucco20@imperial.ac.uk

Sara Fontanella

National Heart and Lung Institute, Imperial College London, UK e-mail: s.fontanella@imperial.ac.uk

1 Introduction

Social media has profoundly reshaped how people connect, communicate, and share information in the digital age. It serves as one of the most crucial channels for the public to express opinions on global events. However, the anonymity offered by social media platforms empowers individuals to freely express controversial opinions without revealing their identities, often leading to offensive behaviors. Users can release emotions without restraint and are absolved from the responsibility for hurtful messages.

Behavioral researchers are actively investigating how online conversations shape public opinion and convey prejudiced viewpoints on different contentious subjects. One such controversial topic in global politics and society that emerged during the 21st century is international migration. The continuous dissemination of “negative news” regarding immigrants and refugees persists through circulation on expansive social media platforms, notably Facebook and Twitter Social networking sites, with varying technical abilities, are utilized by racist and far-right actors to manipulate public online discussions. A popular anti-immigration strategy involves reframing and reinterpreting mainstream media coverage of immigration and migrants, focusing on criminal and negative issues such as public disorder, cultural ignorance, social problems, and financial costs. Rather than focusing solely on explicit manifestations of racism and tactics employed by extremist figures, it is essential to recognize the concept of ambient racism, as outlined by [4]. This subtle form of racism emerges in online communication, including user comments, tweets, and emotional responses, spanning various scales in virtual public spaces. Social media platforms facilitate a continuous flow of communication, blurring the lines between intentional messaging from far-right figures and genuine emotional expressions of ordinary users. Consequently, anti-immigrant actors capitalize on the constant information flow in social networks, manipulating narratives and shaping public perception by leveraging emotional reactions, continuous news streams, and shared experiences on online platforms.

In this paper, we aim to investigate the various expressions of rhetoric and hostility directed towards immigrants on widely used social media platforms, specifically Facebook, Instagram and YouTube within the context of Italy. To accomplish this goal, we gathered and scrutinized more than 185,000 comments. Our approach involves categorizing social media posts into two distinct groups—specifically, a racist narrative and a neutral narrative—through the application of seeded Latent Dirichlet Allocation [3]. This serves as an initial phase of a broader research initiative intended to develop lexical resources. In future work, lexicon-based features, which include blacklists of racist hateful and offensive terms, will be utilized to construct an ontology for racist and xenophobic speech. The paper is structured as follows. Section 2 describes the data. Section 3 provides background knowledge on LDA and seededLDA models while Section 4 reports the main results. Finally, Section 5 draws conclusions and emphasizes avenues for future research.

2 Data

To assess public sentiments on migration in Italy, we collected data from Facebook, Instagram, and YouTube. The choice of these platforms was based on their accessibility, diversity and widespread user engagement. We analyzed comments posted across the last decade, gathered through *exportcomments.com*, compiling a corpus of 185,734 documents. To enhance the overall quality and coherence of the data for subsequent analysis and modeling tasks, we conducted thorough data pre-processing. This involved refining textual inputs through techniques such as tokenization, elimination of stopwords, and noise removal. Our data pre-processing utilized the R *quanteda* package [1], resulting in a final vocabulary of 9,060 tokens.

3 Semi-supervised topic modeling

One of the most popular methods among researchers looking to leverage social media data for studying a phenomenon of interest is topic modeling, which offers a useful representation of the associated corpus. The predominant technique for topic modeling is Latent Dirichlet Allocation (LDA), initially introduced by [2]. LDA postulates a generative model that explains how a topic influences the generation of words in a document. Essentially, it assumes that each topic is a distribution over words and each document is a distribution of topics, i.e. every word in a document is generated by first sampling a topic and then sampling from the selected topic word distribution. However, a notable drawback of LDA is its unsupervised nature.

In recent times, text-mining research has explored innovative strategies derived from topic models to address significant substantive questions. Our approach to examining the anti-migrant rhetoric is based on a variant of LDA known as seededLDA [3]. A seeded topic model allows researchers to input a set of keywords into the model, defining the desired themes before the estimation process. In respect to vanilla LDA, seededLDA involves two model enhancements. Firstly, it introduces seed-topic distributions, restricting word generation to specific seed sets. Each word is randomly chosen from either the standard LDA distribution or the seed set. Secondly, it incorporates a second Dirichlet distribution before determining the topic composition in a document. This second distribution depends on a binary vector indicating the word sets present in the document and influences the mean value of the Dirichlet distribution. While vanilla LDA uncovers structures without prior knowledge, seededLDA is beneficial for substantial external influence. Besides, vanilla LDA allows topics to emerge based solely on the statistical patterns in the data. In contrast, seededLDA provides more control over the topics by incorporating prior information through seed words, leading to more interpretable and predefined topic structures. SeededLDA involves extra steps for seed word incorporation and initialization. The generative story of seededLDA is detailed in [3] and can be summarized as follows.

For each topic $k = 1 \dots T$:

- choose a regular topic $\phi_k^r \sim \text{Dir}(\beta_r)$
- choose seed topic $\phi_k^s \sim \text{Dir}(\beta_s)$
- choose $\pi_k \sim \beta(1, 1)$.

For each seed set $s = 1 \dots S$,

- choose group-topic distribution $\psi \sim \text{Dir}(\alpha)$.

For each document d ,

- choose a binary vector $\mathbf{b}$ of length S .
- choose a document-group distribution $\zeta^d \sim \text{Dir}(\tau \mathbf{b})$
- choose a group variable $g \sim \text{Mult}(\zeta^d)$
- choose $\theta_d \sim \text{Dir}(\psi_g)$. // of length T
- For each token $i = 1 \dots N_d$:
 - select a topic $z_i \sim \text{Mult}(\theta_d)$
 - select an indicator $x_i \sim \text{Bern}(\pi_{z_i})$
 - if x_i is 0
 - select a word $w_i \sim \text{Mult}(\phi_{z_i}^r)$
 - if x_i is 1
 - select a word $w_i \sim \text{Mult}(\phi_{z_i}^s)$.

The selection of seed words is guided by two main criteria. The first criterion is theoretical relevance, aiming to operationalize and represent abstract concepts within the model. The second criterion is based on empirical frequency, focusing on incorporating words with higher occurrences in the dataset, thereby enhancing the capabilities of machine learning. In our analysis, to guide the discovery of a topic linked to an anti-immigrant narrative, we consider only seed terms related to a racist perspective and allow for a residual topic. The seed terms are derived from the revised HurtLex lexicon [5] considering the category “negative stereotypes ethnic slurs”.

4 Results

The outcomes of the seededLDA model were obtained using the R package *quanteda* [1]. They include probabilities of words associated with topics and probabilities of topics within documents. The word clouds presented in Figure 1 visually highlight a clear distinction between terms linked to comments identified as “anti-immigrants” or “neutral” comments. The size of each word within the word clouds corresponds to the probability to be in the retrieved topic. Taking an overall racist perspective, the terms depicted in the left panel of Figure 1 provide valuable insights into the extensively debated issue of immigration and the socio-political concerns associated with it in Italy. Focusing on anti-immigrant narratives, the inclusion of “loro” (they/them) implies a distancing of immigrants, setting them apart from the native population. By linguistically separating immigrants as “they/them” instead of

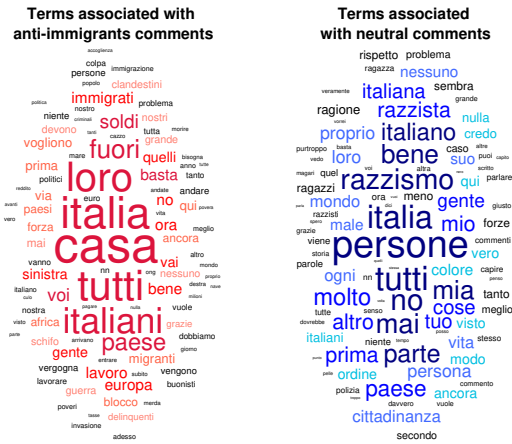


Fig. 1 Wordcloud of seededLDA word-topic probabilities

including them as part of the broader community (“we/us”), it reinforces the division between the perceived in-group and out-group. This linguistic choice is central to the “othering” process, meaning the simultaneous construction of the superior “self” or in-group and the inferior “other” or out-group. The linguistic choices, including the use of terms like “Italians”, “immigrants,” and “migrants”, contribute to this process by delineating clear boundaries between who is considered part of the favored group and who is considered an outsider. Othering revolves around the conscious or unconscious belief that a specific identified group poses a threat to another favored group. In a racist narrative, words employed to promote an “us versus them” mentality will portray immigrants as a threat to the nation’s identity (term “Italia”, “Paese” - country), security, and prosperity. More specifically, the inclusion of “soldi” (money), “accoglienza” (hospitality), “tasse” (taxes), and “pagare” (to pay) suggests concerns about the economic impact of immigration on public resources and welfare systems. Along the same lines, “lavoro” (work) encapsulates concerns about job availability and competition in the workforce. “Invasione” (invasion) fuels fears of demographic change and cultural dilution due to immigration, framing immigrants as invaders.

Terms like “blocco” (blockade) and “navale” (naval) are likely associated with sentiments about preventing immigrants from entering, encapsulating discussions around border control and immigration policies. On the other hand, “fuori” (out), “via” (go away), and “home” further emphasize the necessity of enforcing spatial boundaries and maintaining the perceived purity and dominance of the in-group.

It is worth noting, the juxtaposition of the term “gente” (people) in the anti-immigrant content with “persone” (individuals or persons) in the neutral comments. “Gente” can be used in a more general and detached manner, often carrying neutral or even negative connotations, to refer to groups of individuals in an impersonal or generalized way. This can contribute to a more stereotyped or dehumanized view of

people. On the other hand, “persone” tends to be more individualized and respectful, suggesting a recognition of the dignity and humanity of each individual.

5 Conclusions and future work

In this study, we have introduced a text-mining strategy aimed at conducting an extensive analysis of social media comments. Our approach incorporates a semi-supervised topic modeling methodology, serving as a crucial initial step in the development and enhancement of lexical resources. Looking ahead to future research endeavors, our primary objective is to expand the lexicon derived from the revised HurtLex [5], intending to broaden its scope and create a more updated, comprehensive lexical foundation. Furthermore, we aspire to construct an ontology—a structured representation of the relationships between various terms and concepts. This ontological framework will play a pivotal role in the creation of an annotated corpus, a crucial component for the subsequent phase of training supervised classification algorithms. These algorithms will be specifically tailored to identify and classify anti-racist social media comments, contributing significantly to a more accurate understanding of digital discussions surrounding this contentious topic.

Acknowledgements This work is part of the research project PRIN-2022 PNRR “Identification and Critical Analysis of Online Racism and Xenophobia against (Im)migrants and Roma people” (Project Code: P2022APKJL), funded by the European Union – Next Generation EU.

References

- [1] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. quanteda: An R package for the quantitative analysis of textual data. *J. Open Source Softw.*, 3(30):774, 2018.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(1):993–1022, 2003.
- [3] J. Jagarlamudi, H. Daumé, and R. Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.
- [4] C. Sharma. Affect and the attention economy of online racism. In *Proceedings of the 67th annual conference of the international communication association, 24-28 May Prague, Czech Republic*, 2018.
- [5] A. Tontodimamma, L. Fontanella, S. Anzani, and V. Basile. An Italian lexical resource for incivility detection in online discourses. *Qual. Quant.*, 57:3019–3037, 2023.

Causal inference from texts: a random-forest approach

Chiara Di Maria, Alessandro Albano, Mariangela Sciandra and Antonella Plaia

Abstract This paper employs causal random forests to analyse textual reviews in an e-commerce context, specifically investigating the causal impact of sentiment on the Positive Feedback Count (PFC). The PFC denotes the number of users who found the review helpful. The results uncover a negative causal effect, indicating that transitioning from negative to positive sentiment reduces the count of users perceiving a review as helpful. The analysis further explores heterogeneity, highlighting the nuanced influence of specific words and variations in treatment effects. This research underscores the efficacy of causal inference in elucidating the intricate dynamics between sentiment and the perceived utility of reviews.

Key words: Causal inference, text analysis, sentiment analysis, causal forest

1 Introduction

In recent years, texts have experienced unprecedented popularity as a source of information. Indeed, many approaches and techniques have been developed to manipulate and analyse written corpora to achieve different aims: understanding sentiment, detecting patterns in words or sentences, or finding topics [10]. The peculiarity of texts as “data structures”, given their unstructuredness and high dimensionality, highlighted the need for fast and computationally efficient methods to handle them. Indeed, text mining encompasses a spectrum of methodologies. Machine and Deep learning involve algorithms that learn patterns from textual data to make predictions or classifications [12]. On the other hand, Vector Space Models [7] (VSM) remain a cornerstone in text mining methodologies for representing words or doc-

Chiara Di Maria, Alessandro Albano, Mariangela Sciandra, Antonella Plaia
Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze,
Building 13, 90128 Palermo e-mail: chiara.dimaria@unipa.it; alessandro.albano@unipa.it; mari-
angela.sciandra@unipa.it; antonella.plaia@unipa.it

uments as vectors in a high-dimensional space to capture semantic relationships. Finally, Large Language Models (LLMs) [11] have revolutionised natural language processing with their extensive pre-trained knowledge to understand and generate human-like text.

Recently, scholars have started to integrate textual analysis within other statistical methods, using, for example, pieces of information extracted from texts as variables in regressions or more advanced models. Specifically, some authors proposed to combine text mining and causal inference, either to find causal relationships in texts or to use some features of written documents as variables of interest in a causal framework (treatment, confounding variables, etc.). This work can be considered part of this research strand, since we estimate the causal effect of some textual features on a response of interest using random forests [1].

2 Methods

Let W denote a binary treatment, Y a response variable of interest, and X a set of observed covariates explaining Y and possibly confounding its relationship with W . Assuming that X is sufficient to make such a relationship unconfounded, the Average Treatment Effect (ATE) of W on Y is defined as $\tau = \mathbb{E}[Y(1) - Y(0)]$ where $Y(w)$ is the counterfactual value of Y if W were set to a specific value w . The ATE is a unique measure that summarises the causal effect for all subjects in the population. However, the causal effect may differ according to individual characteristics, leading to the so-called effect heterogeneity. The Conditional Average Treatment Effect (CATE) is defined as $\tau(X_i) = \mathbb{E}[Y(1) - Y(0)|X]$, which, as can be seen, depends on the covariates. In general, we can write

$$Y_i = \tau(X_i)W_i + f(X_i) + \varepsilon_i, \quad (1)$$

where f is an unknown function describing the dependence of the outcome on the covariates, and ε is an error term having zero mean conditional on W and X . Let $e(x) = \mathbb{E}[W_i|X_i]$ be the propensity score and $m(x) = \mathbb{E}[Y_i|X_i]$ the conditional mean of Y , Equation (1) can be rewritten as

$$Y_i - m(x) = \tau(X_i)(W_i - e(x)) + \varepsilon_i. \quad (2)$$

The causal effects, $\tau(X_i)$, can then be obtained as a weighted residual-on-residual regression; but since f is unknown, the problem is how to estimate $e(x)$ and $m(x)$. However, as we are only interested in their predicted values, they can be obtained with any machine-learning method, such as random forests. Athey et al. [1, 2] proposed an approach to estimate causal effects using random forests, which can be considered a combination of the approaches proposed by [3] and [6], where the objective function to maximise is the difference of treatment effects in the subgroups. The CATE is subsequently obtained with a weighted regression as shown in

Equation (2), possibly using cross-validation, where the weights correspond to the adaptive neighbourhood weights (see [5]).

Two measures that can be helpful in interpreting the CATE are the best linear projection (BLP) [8] and the Rank-Weighted Average Treatment Effect (RATE)[13]. The former provides doubly-robust estimates for the linear model

$$\tau(X_i) = \beta_0 + \beta_1 A_i,$$

where A_i is a set of covariates. This gives information on which variables are more influential in explaining the variance in causal effects.

The RATE is a metric that assesses how well a CATE estimator ranks units according to the estimated treatment benefit. It can be considered as an area under the curve, called Targeting Operator Curve (TOC), defined as

$$TOC(q) = \mathbb{E}[Y(1) - Y(0) | \hat{\tau} \geq F^{-1}(1 - q)] - \mathbb{E}[Y(1) - Y(0)] \quad (3)$$

where F is the distribution function of a prioritizing rule $S(X)$, i.e. a rule to assign treatment to specific subjects having certain characteristics. The area under the TOC (AUTOC) quantifies the benefit of treating only $q \times 100\%$ subjects with higher values according to the prioritizing rule, over the ATE.

3 Application to real data

We analysed data¹ about a female clothes e-commerce containing more than 20 thousand reviews written by users, for which several variables are available, like the age of the reviewer (Age), the rating assigned to the product on a scale from 1 to 5 (Rating), the number of people who positively evaluated the review (Positive Feedback Count, PFC) and some others related to the kind of product. The aim of our analysis is to understand if reviews' text characteristics have a causal impact on PFC.

The traditional setting of causal inference involves a binary treatment, although addressing continuous treatments is straightforward. Texts can be summarised in different ways, but one of the easiest binary variables to extract from text is the sentiment, positive or negative. This is our treatment. We retrieved sentiment by using the Valence Aware Dictionary and sEntiment Reasoner (VADER, [4]) implemented in the NLTK Python library. VADER is a model for sentiment analysis able to take into account both text polarity (positive/negative) and intensity. It returns a score of how positive, negative, and neutral a text is and a compound score. We used a dichotomic version of the latter, obtained by assigning 0 to all scores below or equal to 0 and 1 to scores larger than 0. The former is interpreted as negative, and the latter as positive sentiment.

¹ Publicly available at <https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews>

To perform causal inference, we employed the `grf` package [9] in R, which allows us to run causal forests as described in [1, 2, 13]. First of all, we estimated the average treatment effect of sentiment on PFC, adjusting for Age, Rating and the words used in the reviews. We created a word count matrix for the latter to include word frequencies in the model. The results show that the causal effect is negative and significant (estimate equal to -0.600, with standard deviation 0.177), meaning that moving from a negative to a positive sentiment decreases the number of people leaving a positive feedback on the review. In other words, negative reviews receive more positive feedbacks.

The next step was to deepen the investigation of this causal effect and evaluate the presence of heterogeneity. We estimated the best linear projection of the causal effect using the frequencies of words classified among the 100 most influential variables as predictors. The results are shown in Table 1, where we reported only the significant words.

Table 1: Best Linear Projection of the Conditional Average Treatment Effect

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.844	0.419	-2.016	0.044*
black	0.780	0.303	2.576	0.010*
dress	-0.545	0.257	-2.117	0.034*
neck	2.947	0.684	4.307	<1e-05***
return	3.099	0.991	3.126	0.002**
top	-0.798	0.334	-2.387	0.017*
...	...	...	...	...

It can be noticed that words have different polarities in terms of their meaning; some are neutral, like ‘neck’, and others are negative, like ‘return’. This means that the presence of these words in the review influences the causal effect, and, as a consequence, they can be a source of heterogeneity. To check this aspect, we represented the TOC curve. The aim is to understand if assigning the treatment only to a fraction of the observations leads to a significant difference in the causal effect with respect to the ATE. In this setting, assigning the treatment to an observation means devising an intervention to make that review positive. The two graphs in Figure 1 confirm the results in Table 1, showing that treating just a fraction q of the observations, ranked according to the number of occurrences of the words ‘neck’ and ‘return’, leads to different causal effects compared to the scenario where all observations are treated (ATE), i.e., when $q = 1$.

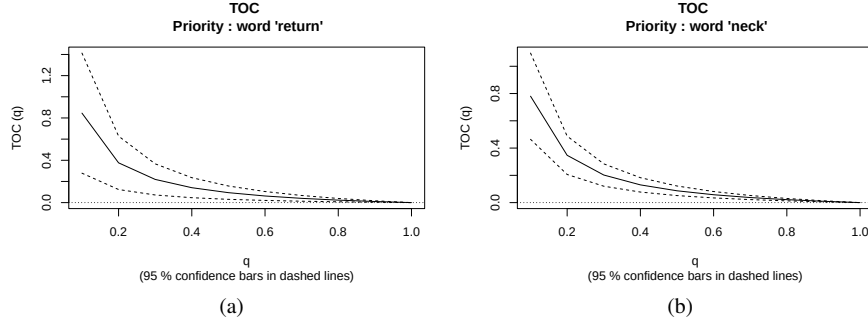


Fig. 1: Targeting Operator Characteristic curves for the words ‘return’ (a) and ‘neck’ (b)

Finally, we test for another source of heterogeneity: the hypothesis is that some reviews, conditional on their set of covariates, may benefit the most from the treatment. The prioritisation rule is based on CATE, in particular, we ranked by the most negative CATE. The results are shown in Figure 2. It can be noticed that the TOC is higher for smaller fractions of the sample, denoting heterogeneity of causal effects. This means that treating only a part of subjects showing the most negative causal effects would produce a higher benefit.

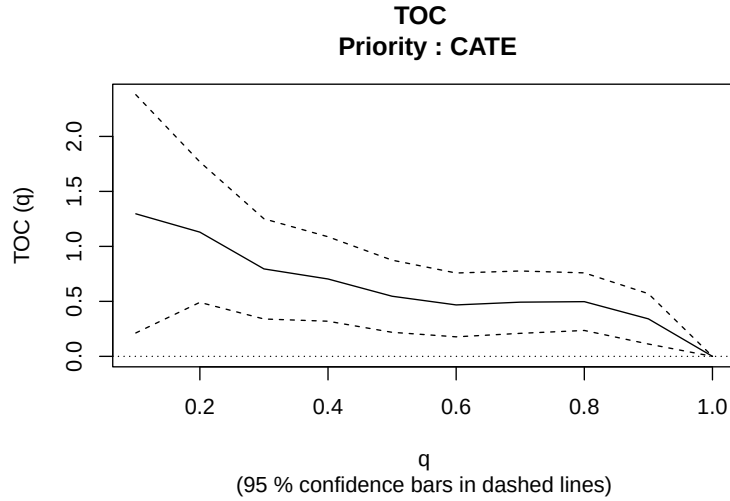


Fig. 2: Targeting Operator Characteristic curve prioritising for CATE.

4 Conclusions

In this paper, we analysed a set of reviews from an online clothing e-commerce site using causal random forests. We found that the review’s sentiment causally influences the number of positive feedback each review receives. We also tested for the presence of effect heterogeneity, focusing on some words classified as highly important in the splitting process of the forest and on the CATE. The TOC confirmed the presence of heterogeneity. The aim of this work was to show how it is possible to draw causal information from texts, exploiting the large amount of information that it is possible to extract from them.

Acknowledgements This research was partially funded by European Union – Next Generation EU. PRIN 2022 PNRR Project “A unified italian oral medicine and orthodontic language system: a prototype of Natural language processing application in healthcare” n. P202299ZWN CUP B53D23026050001.



References

1. Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Ann. Stat.*, 47(2):1148 – 1178, 2019.
2. Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Obs. Stud.*, 5(2):37–51, 2019.
3. Leo Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.
4. Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.
5. Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *J. Am. Stat. Assoc.*, 101(474), 2006.
6. Peter M Robinson. Root-n-consistent semiparametric regression. *Econom.*, pages 931–954, 1988.
7. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
8. V. Semenova and V. Chernozhukov. Estimation and inference about conditional average treatment effect and other structural functions. *The Econometrics Journal*, 24(2):264–289, 2021.
9. Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. *grf: Generalized Random Forests*, 2023. R package version 2.3.1.
10. Antonio Usai, Marco Pironti, Monika Mital, and Chiraz Aouina Mejri. Knowledge discovery out of text data: a systematic review via text mining. *J. Knowl. Manag.*, 22(7):1471–1488, 2018.
11. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
12. NI Widiastuti. Deep learning—now and next in text mining and natural language processing. In *IOP Conference Series: Materials Science and Engineering*, volume 407, page 012114. IOP Publishing, 2018.
13. Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

9. Solicited - Data analysis methods for data in non-Euclidean spaces

9.1 - Riemannian Statistics for Any Type of Data

9.2 - PAM clustering algorithm for ATR-FTIR spectral data selection: an application to multiple sclerosis

9.3 - Random Survival Forest for Censored Functional Data

9.4 - Advancing credit card fraud detection with innovative class partitioning and feature selection technique

Riemannian Statistics for Any Type of Data

Oldemar Rodríguez Rojas

Abstract This paper introduces a novel approach to statistics and data analysis, departing from the conventional assumption of data residing in Euclidean space to consider a Riemannian Manifold. The challenge lies in the absence of vector space operations on such manifolds. Pennec X. et al. in their book Riemannian Geometric Statistics in Medical Image Analysis proposed analyzing data on Riemannian manifolds through geometry, this approach is effective with structured data like medical images, where the intrinsic manifold structure is apparent. Yet, its applicability to general data lacking implicit local distance notions is limited. We propose a solution to generalize Riemannian statistics for any type of data.

Key words: Riemannian manifold; simplicial complexes; UMAP; homeomorphism.

1 Introduction

Each point in a data table can be imagined as a star or planet in the universe, especially when dealing with big data issues. In the universe, due to the infinitely different sizes of constellations, there are vastly different perceptions of distances between celestial bodies. For example, two constellations or galaxies that appear to be the same size from a distance (from Earth, for example) could be infinitely different, and one could even fit inside the other in a very small portion or empty space within it. Similarly, in data, there are local notions of distance corresponding to different regions of the data, and this should be considered when calculating statistical indices or models. For this reason, especially in problems involving Big Data, *thinking that the data is in Euclidean space is just as wrong as thinking that the earth is flat*. To address this, we propose considering that the data exists within a

Oldemar Rodríguez Rojas
School of Mathematics, CIMPA, University of Costa Rica, San José, Costa Rica, e-mail: olde-
mar.rodriguez@ucr.ac.cr

Riemannian manifold, where these local notions of distance can be effectively taken into account.

In [6] *Pennec et al.*, the authors had proposed the idea of analyzing data on Riemannian manifolds through the use of geometry. However, this idea is not readily applicable to general data where there are no implicit notions of local distance. The idea that we propose goes beyond, our main idea is to endow a Riemannian manifold structure to any given set of data. This approach leads to a significant improvement in the results of various statistical analyses as well as their interpretability.

In another significant contribution, McInnes et al. [4] introduce UMAP (Uniform Manifold Approximation and Projection), a novel technique for manifold learning and dimension reduction. Utilizing simplicial complexes, Čech complexes, and the Nerve theorem, UMAP gains additional benefits from this Riemannian metric-based approach. It generates a local metric space associated with each point, allowing for meaningful distance measurements. Consequently, the algorithm can assign weights to edges in a graph (simplicial complex), signifying the local metric-based separation between the original points. So the idea that we proposed in this paper is to use the local notions of distance that the UMAP algorithm generates in any data table to provide it with local distances.

UMAP, as a successor to t -SNE method, inherits a controversy associated with the t -SNE method. A comprehensive discussion on this topic can be found at <https://umap-learn.readthedocs.io/en/latest/clustering.html>. Despite these concerns, there are still valid reasons to utilize UMAP as a preprocessing step for clustering. As highlighted in the discussion, when applied to real high-dimensional datasets such as MNIST data [1] or cell RNA-seq data [2], and with appropriate parameterization, both t -SNE and UMAP yield significantly better clustering results than other algorithms.

2 Providing to a classical data table with a Riemannian manifold structure

UMAP method was designed to improve the main limitations of the t -SNE method. t -SNE means t -distributed Stochastic Neighbor Embedding and it was proposed by Laurens van der Maaten, see all the detail of this method in [3]. UMAP algorithm is competitive with t -SNE for visualization quality and it improves t -SNE limitations. UMAP (Uniform Manifold Approximation and Projection) is an algorithm for dimension reduction based on algebraic topology, topological data analysis and Riemannian geometry. It was proposed by the Mathematician Leland McInnes in [4]. UMAP works in a similar way to t -SNE, it finds distances in a space with many variables and then tries to reproduce these distances in a low-dimensional space. But UMAP does it very differently because more than distances it tries to reproduce the topology, not necessarily the geometry. UMAP assumes that data is distributed along a Riemannian manifold. A manifold is a uniform n -dimensional geometric shape in which, for each point of this manifold, there is a neighborhood around that

point that looks like a flat two-dimensional plane. Riemannian manifolds admit local notions of distances, area and angles. To explain the UMAP method we need to define the notion of k -simplex and simplicial complexes.

Let $\{x_0, \dots, x_k\}$ be points in $\mathbb{R}^n$. We will assume that these points satisfy the condition that the set of vectors in $\mathbb{R}^n$ represented by the differences with respect to x_0 , that is $\{x_1 - x_0, x_2 - x_0, \dots, x_k - x_0\}$ are linearly independent.

Definition 1. The k -simplex generated by the points $\{x_0, \dots, x_k\}$ is the set of all points $z = \sum_{i=0}^k a_i x_i$, where $\sum_{i=0}^k a_i = 1$. For a given z , we refer to a_i as the i -th barycentric coordinate.

Simplicial complexes are generalizations of graphs. A simplicial complex S in $\mathbb{R}^n$ is a set of simplices such that every face of a simplex in S is also a simplex in S . The intersection of two simplices in S is a face of each of them. Given data set presented as a finite metric space, we need to produce a simplicial complex such that the algebraic invariants of the simplicial complex reflect the shape of the data. To do that, we need to make the connection between clustering and components precise, via single-linkage clustering, which works as follows.

1. Choose a parameter ε .
2. Assign two points x and y to the same group if they are connected by a path of points (for some k) $x = x_0, x_1, x_2, \dots, x_{k-1}, x_k = y$ such that each point x_i is at a distance ε from x_{i+1} . See the Figure 1.

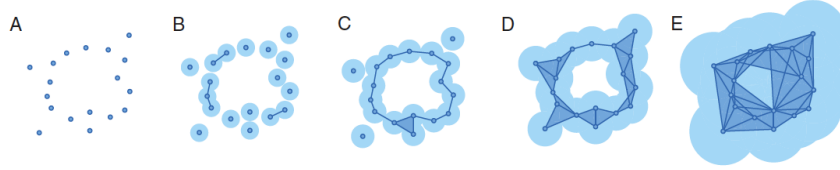


Fig. 1 As ε increases, more and more simplices are added to the simplicial complex and topological features emerge. In panels C and D, a circle can be detected.

The Nerve Theorem and its corollary [5] are the fundamental theoretical basis that allows to go from topological spaces to simplicial complexes and then to data. The Čech complex allows to demonstrate that there exists a homeomorphism between the union of balls (determined by the parameter ε) and the nerve and therefore we will have a bijection between the data and the simplicial complexes.

Definition 2. Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\varepsilon > 0$. The Čech complex $C_\varepsilon(X, \partial_X)$ is the simplicial complex with vertices the points of X , and a k -simplex $[v_0, v_1, \dots, v_k]$ when a set of points $\{v_0, v_1, \dots, v_k\} \subset X$ satisfies $\bigcap_i B_\varepsilon(v_i) \neq \emptyset$.

Theorem 1 (Nerve Theorem corollary). Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\varepsilon > 0$. There exists a homeomorphism: $\bigcup_{x \in X} B_\varepsilon(x) \cong |C_\varepsilon(X, \partial_X)|$ between the union of balls and the nerve $N(\mathcal{U})$ (the geometric realization) of the Čech complex.

The above theorem guarantees that there exists a homeomorphism between the union of balls and the nerve, so, there is relation one-to-one (bijection) between data and Čech complex. To apply these ideas, UMAP choose a radius from each point, connecting points when those radii overlap, then we can create a simplicial complex using 0, 1, and 2 simplexes as points, lines, and triangles. Choosing this radius is critical, too small choice will lead to small, isolated clusters, while too large choice will connect everything together. UMAP overcomes this challenge by choosing a radius locally, based on the local distance to each point to the k -th nearest neighbor. To do that, Riemannian metric is used, see the definition in [7]. The choice of k determines how locally we wish to estimate the Riemannian metric. A small choice of k means we want a very local interpretation, while, choosing a large k means our estimates will be based on larger regions. *This is very important, because it means that the UMAP algorithm provides the data table with local distance notions.* To do that, UMAP minimized the Cross Entropy, see [4].

3 Riemannian statistics for any type of data

Defining statistical methods on Riemannian manifolds poses a unique challenge due to the absence of fundamental vector space operations like addition and scalar product. A critical error would arise from employing statistical indices grounded in the Euclidean space structure of $\mathbb{R}^n$. To illustrate, consider the scenario where one intends to furnish the UMAP method with a correlation circle. To illustrate, we will utilize the following data table 1, which includes the school grades of ten students.

Table 1 Students data

	Math	Science	Spanish	History	Phys. Ed.
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

In Principal Component Analysis the coordinate in the correlation circle of variable X^j on axis r is given by $R(X^j, C^r)$ which is the correlation coefficient between the j -th variable and the r -th principal component. Using this idea, if we plot UMAP correlation circle using Pearson correlation index, the result is shown in left panel of Figure 2. Clearly, the correlation circle on the left in Figure 2 exhibits a sig-

nificant error, namely, certain variable arrows extend beyond the sphere of radius 1. That is to say, we don't have the property: $R^2(X^j, C^s) + R^2(X^j, C^r) \leq 1$. This discrepancy arises from computing correlations as if the data were in a Euclidean space, employing the classical index of correlation. However, the data resides on a Riemannian manifold, with local distances generated by UMAP. Consequently, there is a necessity to define Riemannian correlation, Riemannian mean, and, more broadly, necessitating the development of *Riemannian Statistics*. In the subsequent definition, we generalize Fréchet's mean:

Definition 3. Let $X \in M_{n \times p}$ the data table. We denote by $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ the rows of X and by $\mathbf{y}_1, \dots, \mathbf{y}_p \in \mathbb{R}^n$ the columns of X . Each vector $\mathbf{x}_i$ can be also considered a point in the Riemannian manifold M induced by the simplicial complex. Each pair of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ has associated a local distances $d_{\text{UMAP}}(\mathbf{x}_i, \mathbf{x}_j)$ generated by the UMAP algorithm with its k nearest neighbors¹. The *Riemannian mean* is the minimizer of the sum-of-squared distances to the data $\mathbf{g} = \arg \min_{\mathbf{x} \in M} \sum_{i=1}^n d_{\text{UMAP}}(\mathbf{x}, \mathbf{x}_i)^2$.

By leveraging the one-to-one relationship given by the Nerve Theorem corollary in 1 we define the the Riemannian correlation as follows.

Definition 4. Let $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$ rows of X , we define the subtraction induced by the UMAP algorithm as $\mathbf{x}_\alpha \ominus \mathbf{x}_\beta = \rho_{\alpha\beta}(\mathbf{x}_\alpha - \mathbf{x}_\beta)$, where $\rho_{\alpha\beta}$ is computed as follows²

$$\rho_{\alpha\beta} = \begin{cases} \frac{d_{\text{UMAP}}(\mathbf{x}_\alpha, \mathbf{x}_\beta)}{d(\mathbf{x}_\alpha, \mathbf{x}_\beta)} & \text{if } d(\mathbf{x}_\alpha, \mathbf{x}_\beta) \neq 0 \\ 1 & \text{if } d(\mathbf{x}_\alpha, \mathbf{x}_\beta) = 0 \end{cases}$$

with d the Euclidean distance in $\mathbb{R}^p$. We defined the variance-covariance matrix $S \in M_{p \times p}$ of X as $S = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i \ominus \mathbf{g})(\mathbf{x}_i \ominus \mathbf{g})^t$, where $(\mathbf{x}_i \ominus \mathbf{g}) = \rho_{i\lambda} (x_{i1} - \mathbf{g}_1, \dots, x_{ip} - \mathbf{g}_p)^t$. so, we define *Riemannian correlation* between y_i and y_j columns of X , that are in $\mathbb{R}^n$, as follows $R(\mathbf{y}_i, \mathbf{y}_j) = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$.

Now we have the property $R^2(X^j, C^s) + R^2(X^j, C^r) \leq 1$ and therefore if we plot again the UMAP correlation circle using the Riemannian correlation index, the result shown on right panel on Figure 2 is now correct.

4 Conclusions and Future Work

In this paper, we successfully extend the ideas proposed by *Pennec et al.* in [6], broadening the scope to compute Riemannian statistical indices and Riemannian data analysis models to any data table. Unlike previous approaches, our methodology is not restricted to data with an intrinsic Riemannian manifold structure. This

¹ If the vectors are not in the same k nearest neighbors then a merge distance is used.

² Note that $\rho_{\alpha\beta}$ is not a parameter of the model, it can be calculated thanks to the one-to-one relationship given by the Nerve Theorem in 1.

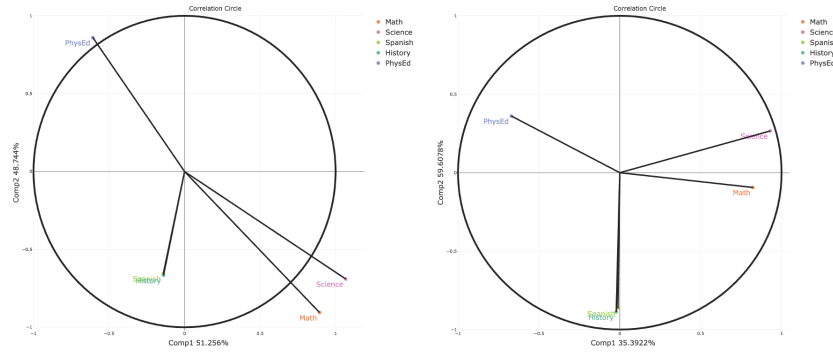


Fig. 2 UMAP circle of correlation with Pearson correlation on left and with Riemannian correlation on right.

advancement opens up a new field of research, where diverse methods like regression, k -means, and more, can be generalized for broader applicability.

Currently, we are actively engaged in implementing these novel ideas in both **R** and **Python**, ensuring practical adoption and seamless integration across different computational platforms.

References

1. Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 141–142.
2. Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, Joshua R. Sanes. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. In *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics*, Volume 166, Issue 5, Pages 1308-1323.e30, ISSN 0092-8674. <https://doi.org/10.1016/j.cell.2016.07.054>.
3. Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9 (Nov), 2579-2605. Submitted 5/08; Revised 9/08; Published 11/08.
4. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *cite arxiv: 1802.03426 Comment: Reference implementation available at <http://github.com/lmcinnes/umap/>* 6, 352–357. <http://arxiv.org/abs/1802.03426>
5. Oudot, S. Y. (2016). Persistence Theory: From Quiver Representations to Data Analysis (Mathematical Surveys and Monographs, Vol. 209). American Mathematical Society.
6. Pennec X. Sommer S. and Fletcher T. (Eds). (2020). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press, Elsevier.
7. Rabadán, R., and Blumberg, A. J. (2020). Topological Data Analysis for Genomics and Evolution: Topology in Biology. Columbia University, New York, y University of Texas, Austin. Cambridge University Press.

PAM clustering algorithm for ATR-FTIR spectral data selection: an application to multiple sclerosis

Francesca Condino, Maria Caterina Crocco and Rita Guzzi

Abstract The analysis of ATR-FTIR spectral data is more and more frequent in medical literature, since it often provides useful support for disease discrimination and early diagnosis for a number of pathologies, such as Multiple Sclerosis (MS). One of the main problem in this field is to identify the more informative features among a great number of possible candidates. Here, we propose a novel approach to perform feature selection based on redundant information among spectral data. In particular, we consider the Partition Around Medoids (PAM) algorithm to clusterize correlation matrix, in order to obtain groups of variables (wavenumbers) having similar pattern of pairwise dependence. Then, only the obtained medoids will be considered in the subsequent statistical methods for disease prediction.

Key words: feature selection, entropy, dependence

1 Introduction

In recent years, the use of Attenuated Total Reflectance Fourier-Transform Infrared (ATR-FTIR) spectroscopy applied on to biofluids, tissues, cells., has become crucial in the analysis of chemical and biological samples (see, for example, [1, 9, 5]). The big amount of data obtained are often characterized by high complexity that needs

Francesca Condino

Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Italy,
e-mail: francesca.condino@unical.it

Maria Caterina Crocco

Department of Physics, Molecular Biophysics Laboratory, University of Calabria, Italy; STAR
Research Infrastructure, University of Calabria, Italy, e-mail: mariacaterina.crocco@unical.it

Rita Guzzi

Department of Physics, Molecular Biophysics Laboratory, University of Calabria, Italy; CNR-
Nanotec Rende, Italy, e-mail: rita.guzzi@unical.it

a deep analysis for discerning signal from noise and enhancing the interpretability of spectroscopic results. This requires some statistical strategy for selection and feature extraction that also considers the link among data. Indeed, in the spectral data analysis it is fundamental to take into account the relation between features, since they are potentially involved in complex interplay of molecular vibrations. Therefore, a key concept is the statistical dependence among ATR-FTIR spectral data and a multivariate approach is often preferred, as it allows to handle dependence among covariates. In this paper, a simple measure of pairwise dependence, such as Spearman's rank correlation, is obtained and the clusterization of corresponding matrix is considered.

The concept of clusterizing the correlation matrix in biological context is not new. It is particularly frequent in gene expression data, where the aim is to find groups of genes characterized by similar pattern of dependence (see, for example, [4, 7, 8]). In the same way, in the context of ATR-FTIR spectroscopy, by applying clustering algorithms to correlation matrix, it becomes possible to identify groups of spectral features exhibiting similar behaviour and representing the underlying molecular structures. This can lead to synthesize spectral data information through the representative object of the cluster. Differently to the classical dimensionality reduction techniques, such as principal component analysis (PCA), this approach offers a better interpretability of the results, supporting researchers in identifying the most important molecular signatures within complex spectra.

The ATR-FTIR spectroscopy has emerged as a promising tool for unveiling molecular signatures associated with neurological disorders, as Multiple Sclerosis (MS) ([10, 3, 2]). Indeed, the diagnosis of MS poses significant challenges due to its heterogeneous clinical form and the absence of a specific diagnostic test. Complementary diagnostic tools that offer molecular insights into the pathological processes associated with MS are needed, allowing for earlier and more accurate diagnosis. In this sense, ATR-FTIR spectroscopy on biofluids such as blood and blood components can unveil specific molecular signatures associated with MS pathology. Alterations in biomolecular components including lipid profiles, protein secondary structures, and nucleic acid compositions can be identified in the spectra of biological samples from MS patients compared to healthy controls. These changes could be indicative of underlying pathological processes, including inflammation, de-myelination, and neurodegeneration, offering valuable insights into the molecular basis of the disease.

2 Data

To show the potentiality of the proposed methodology, in this work we consider the data already reported in [3]. They consist of spectral features registered on the plasma of 85 subjects, including 45 MS patients and 40 healthy controls, recruited from April 2017 to July 2018 in the MS Center of the Annunziata Hospital in Cosenza, Italy. The FTIR absorption spectra were analyzed both in the fingerprint

region ($1800 - 900 \text{ cm}^{-1}$) and in the high region ($3050 - 2800 \text{ cm}^{-1}$) and properly pre-processed before applying different univariate and multivariate statistical techniques aimed to discriminate patients from control subjects. Detailed information on patients and data are reported in [3]. Here, we consider the spectral variables (wavenumbers) of the two considered regions.

3 Synthesizing spectra: the PAM algorithm

In this work, to measure the dependence among spectral wavenumbers, we used the Spearman's rank correlation coefficient. This is a very simple measure, based on ranks, that allows us to assess the presence and the degree of a monotonic relationship among variables, differently from Pearson correlation that just measures the linear dependence. Moreover, Spearman correlation is more suitable when the data contain outliers.

As it is expected, the observed pattern of correlation for the available data seems quite peculiar, as it can be seen by the heatmap reported in the Fig. 1. Specifically, the presence of possible groups of ATR-FTIR spectral data are evident, since different block of wavenumbers are characterized by high correlation, while others exhibit lower dependence. This evidence suggests that some exploratory method, such as clustering analysis, can help bringing to light this grouping data structure and identify a representative element for each cluster that could synthesize block information.

In this paper, we consider the Partition Around Medoids (PAM) method, proposed by [6], to find groups of variables having similar pairwise pattern of dependence and to select the most relevant wavenumbers, reducing redundancy. One of the main advantage of PAM algorithm is that the centre of each cluster, called medoid, corresponds to an observed member of dataset. In our context, the specific wavenumber selected as group medoid can be considered as representative for each cluster. For this reason, only the resulting medoids will be chosen as covariates in the subsequent applications of methods for disease prediction. It is worth to note that PAM algorithm, being a non-hierarchical method, requires to establish a priori the number of clusters in the final solution. Here, we fix a number of groups equal to 16 and 18 for the fingerprint region and the high region, respectively.

4 Using medoids as covariates

After applying PAM clustering algorithm, a specific representative set of wavenumbers has been selected in the two selected spectral regions. These wavenumbers can be used as covariates in models for classification task. In particular, as an example, logistic regression and linear discriminant analysis has been applied to discriminates between SM and control subjects. Depending on the considered method, a measure

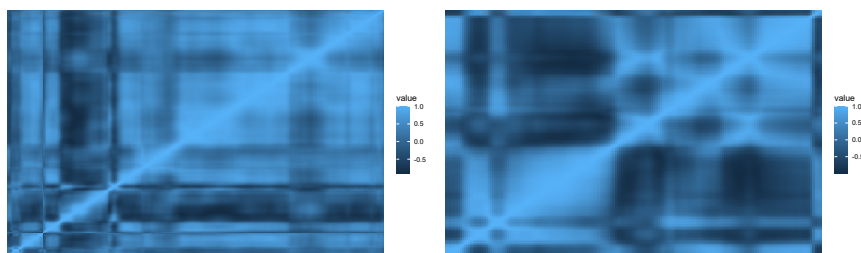


Fig. 1 Heatmap of the Spearman correlation matrix for the fingerprint region (left panel) and the high region (right panel)

of the variable importance is obtained to evaluate the effect of each wavenumber on the model. The obtained results are showed in Fig. 2.

Both the methods show good performance in separating the two groups of individuals, with an accuracy higher than 88% and 80% in the fingerprint and in the high region, respectively. Some usual measures of diagnostic performance are reported in Table 4. Moreover, the Receiver Operating Characteristic (ROC) curves depicted in Fig. 3 show high values of the area under the curves (AUC), suggesting a good ability of the two models in distinguishing between MS and HC.

1800 – 900 cm^{-1}				
Predicted (Logit model)				
Observed	SM	Ctrl		
SM	40	5	Sensitivity	88.89%
Ctrl	5	35	Specificity	87.50%
Tot	45	40	Accuracy	88.24%
Predicted (LDA)				
SM	38	2	Sensitivity	84.44%
Ctrl	7	38	Specificity	95.00%
Tot	45	40	Accuracy	89.41%
3050 – 2800 cm^{-1}				
Predicted (Logit model)				
Observed	SM	Ctrl		
SM	37	7	Sensitivity	82.22%
Ctrl	8	33	Specificity	82.50%
Tot	45	40	Accuracy	82.35%
Predicted (LDA)				
SM	37	8	Sensitivity	82.22%
Ctrl	8	32	Specificity	80.00%
Tot	45	40	Accuracy	81.18%

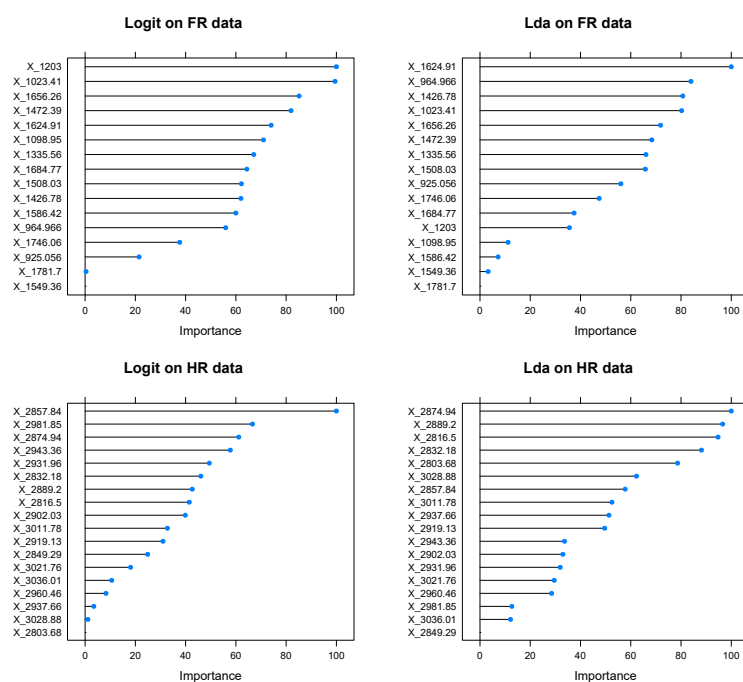


Fig. 2 Variable importance for the fingerprint region (left panel) and the high region (right panel)

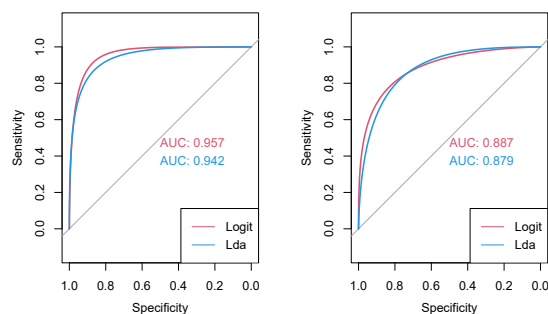


Fig. 3 Receiver Operating Curve (ROC) for logit regression and linear discriminant analysis for fingerprint region (left panel) and high region data (right panel)

5 Conclusion and future perspective

In the present work, a new approach to deal with variable selection in disease discrimination problems is proposed. The results are promising, although very preliminary, as the limited number of observations inhibits a suitable evaluation of validity and reproducibility of this new method. Moreover, for the future, particular aspects needs to be explored and analysed more in depth, such as the possibility to start from alternative dependence measures and to consider the number of clusters like a tuning parameter in a grid search context.

References

1. Baker, M. J. et al.: Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771–1791 (2014)
2. Condino, F., Crocco, M.C., Pirritano, D., Petrone, A., Del Giudice, F.; Guzzi, R.: A Linear Predictor Based on FTIR Spectral Biomarkers Improves Disease Diagnosis Classification: An Application to Multiple Sclerosis. *J. Pers. Med.* **13**, 1596 (2023)
3. Crocco, M.C., Moyano, M.F.H., Annesi, F. et al.: ATR-FTIR spectroscopy of plasma supported by multivariate analysis discriminates multiple sclerosis disease. *Sci Rep* **13**, 2565 (2023)
4. Fang, J., Ma, Q., Chu, C. et al.: PIRCh-seq: functional classification of non-coding RNAs associated with distinct histone modifications. *Genome Biol* **20**, 292 (2019)
5. Güler, G., Guven, U. & Oktem, G.: Characterization of CD133+/CD44+ human prostate cancer stem cells with ATR-FTIR spectroscopy. *Analyst* **144**, 2138–2149 (2019)
6. Kaufman, L. and Rousseeuw, P.J.: Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*. Eds L. Kaufman and P.J. Rousseeuw (1990)
7. Ménard, C., Dulong, et al.: Integrated transcriptomic, phenotypic, and functional study reveals tissue-specific immune properties of mesenchymal stromal cells, *Stem Cells* **38**(1), 146–159 (2020)
8. Pancaldi, V., Schubert, F., Bähler, J.: Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast. *Mol Biosyst.* **6**(3), 543–52 (2010)
9. Theakstone, A. G. et al.: Fourier-transform infrared spectroscopy of biofluids: A practical approach. *Transl. Biophoton.* **3**, e202000025, (2021).
10. Yonar, D., Ocek, L., Tiftikcioglu, B. I., Zorlu, Y. & Severcan, F.: Relapsing-remitting multiple sclerosis diagnosis from cerebrospinal fluids via Fourier transform infrared spectroscopy coupled with multivariate analysis. *Sci. Rep.* **8**, 1025 (2018)

Random Survival Forest for Censored Functional Data

Giuseppe Loffredo, Elvira Romano and Fabrizio Maturo

Abstract Functional Random Survival Forest (FRSF) is an extended version of the Survival Random Forest (SRF) algorithm specifically developed to incorporate functional data as predictors in survival analysis. Inspired by a recent study, where a supervised classification method via a combined use of functional data analysis and tree-based methods is proposed, innovative functional splitting rules are introduced within FRSF by introducing censored functional data, enabling the generation of functional predictions even in the presence of complex or unknown relationships in the data. These novel splitting rules are carefully designed to capture the essential features and patterns inherent in the functional predictors. By leveraging these functional indices, the predictive capabilities of the SRF algorithm are significantly enhanced, resulting in more accurate and reliable predictions. To generate the final prediction, the individual predictions are aggregated from all the trees in the forest. This ensemble approach leverages the collective knowledge of forecasts trees and incorporates the unique aspects of functional data, leading to improved performance in the prediction process.

Key words: Functional data analysis, survival analysis, functional random survival forest.

Giuseppe Loffredo
Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Caserta, Italy,
e-mail: giuseppe.loffredo@unicampania.it

Elvira Romano
Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Caserta, Italy,
e-mail: elvira.romano@unicampania.it

Fabrizio Maturo
Faculty of Technological and Innovation Sciences, Universitas Mercatorum, Rome, Italy, e-mail:
fabrizio.maturo@unimercatorum.it

1 Introduction

In the realm of statistical methodologies, survival analysis stands out as an indispensable tool, providing profound insights into the intricate interplay between temporal dynamics and significant events. Initially developed within the domain of medical research to analyze the duration until specific occurrences, such as disease or mortality, survival analysis has transcended disciplinary boundaries, finding utility across different domains [3], [9], [12], [17]. Within the framework of survival analysis, different techniques can be broadly classified into parametric, semi-parametric, and non-parametric methods, delineating the approaches to constructing survival models from data. Among these methodologies, supervised learning techniques assume a crucial role, where the goal is to predict the time until an event of interest occurs while considering censored observations. Tree-based methods have a great importance due to their intuitive interpretation and ability to handle complex datasets. Different popular versions of tree-based techniques have been developed specifically for survival analysis tasks. These include Survival tree (ST) [5], Survival Bagging (SB) [7], and Random Survival Forest (RSF) [6]. These methods offer different approaches to modeling survival data, each with unique strengths and characteristics. In this article, the traditional survival models are extended into functional context. Functional data, which represent curves observed over a continuum time-space, present unique challenges and opportunities in survival analysis. In order to capture the information in each instant time we introduce Censored Functional Data (CFD) and propose Functional Random Survival Forest (FRSF), aimed at enhancing prediction techniques in survival analysis. The FRSF methodology seeks to use the rich information contained in functional data while accounting for censoring mechanisms inherent in survival analysis. Through real-world examples, we demonstrate the utility of FRSF in addressing the challenges of predictive modeling techniques for time-to-event outcomes. The structure of our work is outlined as follows: In Section 2, we provide a comprehensive overview of Functional Data Analysis (FDA) alongside supervised methods, offering readers a foundational understanding of the methodologies underpinning our research. Section 3 is dedicated to the introduction of our novel approach, where we present our methodological framework, elucidate its theoretical bases, and detail its implementation. Finally, in Section 4, we draw our conclusions, reflecting on the findings derived from our proposed method and offering insights into potential future directions for research and application.

2 Functional Data Analysis

Functional Data Analysis (FDA) is a statistical field focused on analyzing data that is represented in functional form. Functional data objects are represented as random variables $\mathcal{X} = \{X(t); t \in \mathcal{T}\}$, where $X(t)$ is the value of the function at time t . The L^2 space allows us to define the distance between functional data objects as follows:

$$d_{L^2}(f, g) = \|f - g\|_{L^2} = \left(\int_{\mathcal{T}} |f(t) - g(t)|^2 dt \right)^{\frac{1}{2}}. \quad (1)$$

FDA techniques include smoothing, interpolation, and regression to convert discrete data into functional form. Functional data $X(t)$ can be approximated by a finite linear combination of basis functions:

$$X(t) = \sum_{j=1}^K c_j \phi_j(t) \quad (2)$$

where coefficients c_j are obtained minimizing the sum of squared errors. Mean and covariance functions in FDA are estimated as follows:

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i=1}^N X_i(t) \quad (3)$$

$$\hat{C}(t, s) = \frac{1}{N} \sum_{i=1}^N [(X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s))]. \quad (4)$$

FDA provides tools like Functional Principal Component Analysis (FPCA) [4], [11] for dimensionality reduction and Functional Classification Trees (FCTs) [10] for making predictions. The FPCA decomposes functional data into principal component scores v_s and eigenfunctions $\xi_s(t)$, allowing finite representation by linear combination:

$$X(t) = \sum_{s=1}^p v_s \xi_s(t) \quad (5)$$

Functional Classification Trees (FCTs) extend traditional classification trees to functional data. The FCT - with principal components (FCT-FPCS) is an extension of the classical classification tree in which the features are given by the scores obtained, for a sample of N subjects, from (5). Finally, the scores-matrix V is given by:

$$V = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1p} \\ v_{21} & v_{22} & \dots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N1} & v_{N2} & \dots & v_{Np} \end{pmatrix} \quad (6)$$

where splitting criteria like the Gini index determine nodes separation for classification purposes.

3 The Functional Survival Tree (FST)

In supervised learning, the Survival Tree (ST) is a special version of the classic Classification Tree (CT) designed for survival time data. Instead of classifying consid-

ering features collected into study, ST focuses on predicting the time until an event of interest occurs for a subject. Handling censored times in survival data makes ST a bit trickier to use compared to regular decision trees.

In survival studies, data collection involves monitoring subjects over a defined period, typically denoted as $\mathcal{T} = [a, b]$. Each subject, indexed by $i = 1, \dots, N$, is observed until either an event occurs (recorded as the survival time T_i) or until censoring happens (recorded as the censoring time C_i). The occurrence or non-occurrence of an event is captured by an indicator variable δ_i , which is set to 1 if the event occurs within the study period and 0 otherwise. During the observation period $\mathcal{T}$, multiple measurements are collected for each subject. These measurements, denoted as $Y_i \in R^{J_i}$, are recorded until the event or censoring time $T_i^* = \min(C_i, T_i)$. Thus, for each subject i , the available data includes the observation vector Y_i of J_i measurements, where J_i may vary across subjects. For each subject, we have an observation vector $Y_i = (Y_{i1}, \dots, Y_{iJ_i})^T$ corresponding to data collected in J_i values until the event time $T_i^* = \min(C_i, T_i)$ occurs for the i -th unit, and the response vector $Z = (Z_1, Z_2, \dots, Z_N)^T$ that can have binary values given by $Z_i = 0$ or $Z_i = 1$ for $i = 1, \dots, N$. Finally, we observe for the i -th subject the survival data given by $(Y_i, Z_i, T_i^*, \delta_i)$, with δ_i as the censoring indicator. Traditional CT is built to identify homogeneous terminal nodes using a splitting criterion. However, in ST, the splitting criterion used in node h , denoted by T , differs from classical methods when dealing with censored data. The procedure iteratively grows a tree until each node defines, at the very least, a singular distinct event, using survival splitting to maximize between-node survival differences defined through parametric (Cox regression) and non-parametric (Kaplan-Meier) approaches [1], [16].

The challenge in survival study lies in the fact that classical CT methods consider feature vectors $X \in R^J$, with the same size of values recorded for each unit. Starting from these data and defining the Censored Functional Data (CFD), the functional approach in ST develops new features using the values estimated on CFD by applying the FPCA decomposition. To reconstruct functional data from these discrete observations, even for unobserved intervals, we introduce the concept of CFD. The objective is to represent each subject's data as a continuous function over the interval $[a, T_i^*]$, incorporating both observed and unobserved periods.

Definition 1 *The censored functional data $\mathcal{X}_i$ for subject i is defined as a function $X_i(t)$ over the interval $[a, T_i^*]$, where T_i^* represents the minimum of the censoring time C_i and the survival time T_i . The indicator variable δ_i distinguishes between observed and censored data points, with $\delta_i = 1$ indicating observed data points*

This approach allows us to effectively handle censored data and reconstruct functional representations that capture the underlying patterns over time. CFD construction relies on classical representation methods to effectively model and interpret observational data. To reconstruct functional data, classical representation techniques are employed [15]. The observed values may be subject to random noise, captured by ε_{ij} for i subject in recorded value at time $t_{ij} \in \mathcal{T}$, necessitating an adjustment in the formulation to accommodate this noise component. Least squares criterion is then applied to estimate the coefficients, providing a best-fit representa-

tion of the functional data. By leveraging classical representation methods such as basis approximation, CFD allows for the continuous reconstruction of information, enabling robust modeling. The method, called Functional Survival Classification Tree with Principal Components (FSCTs-FPCs), considers new data $(\hat{v}_i, Z_i, T_i^*, \delta_i)$, where $\hat{v}_i = (\hat{v}_{i1}, \hat{v}_{i2}, \dots, \hat{v}_{ip})^T$ corresponds to the i -th score vector obtained from the decomposition. In this approach, the features correspond to the columns of the score matrix, given by the decomposition, where $\hat{V} \in R^{N \times p}$ and $\hat{v}_{im}$ correspond to functional principal component score of the i -th approximated curve relative to the m -th eigenfunction $\hat{\xi}_m$. Typically, considering node h of the tree T , given a score predictor $\hat{v}_j$, the method finds a value c and selects the best $\hat{v}_0$ such that survival differences between the two inequalities, $\hat{v}_0 \leq c$ and $\hat{v}_0 > c$, are maximized to predict the response Z . The aim is to define in node h two regions $R_1(j, c)$ and $R_2(j, c)$ to identify times for the two groups. One of the survival criteria used within node h is the log-rank splitting method, which can maximize a splitting value c for a predictor variable $\hat{v}$. The log-rank test formula is given by:

$$|L(\hat{v}, c)| = \frac{\sum_{k=1}^K (d_{k1} - r_{k1} \frac{d_k}{r_k})}{\sqrt{\sum_{k=1}^K \frac{r_{k1}}{r_k} \left(1 - \frac{r_{k1}}{r_k}\right) \left(\frac{r_k - d_k}{r_k - 1}\right) d_k}} \quad (7)$$

where d_{kj} counts events in daughter nodes j , r_{kj} counts individuals alive, and K denotes observation times between nodes. The FSCTs-FPCs use different methods for constructing predictive models, including log-rank splitting and the Nelson-Aalen estimator. Once the tree T is built, the survival predictor in terminal node h is defined in terms of the predictor within each terminal node. The Cumulative Hazard Function (CHF) and the Survival Function (SF) are given by using the Nelson-Aalen estimator and Kaplan-Meier estimator. To evaluate the importance of predicting survival outcomes after tree construction, metrics such as the Concordance Index (C-index) can be used. The extension of Survival Tree (ST) is referred to as Functional Random Survival Forest (FRSF), where multiple STs are grown for each bootstrap sample generated from a sample set N .

4 Discussion

The proposed FRSF algorithm is a significant advancement in survival analysis, specifically for datasets with functional predictors. FRSF introduces innovative splitting rules to capture patterns in functional data, handles censored data effectively, and utilizes an ensemble approach for improved prediction performance. By integrating functional data and leveraging ensemble learning, FRSF enhances the accuracy and reliability of survival predictions, making it a valuable tool in various fields.

Acknowledgements

We gratefully acknowledge the funding provided by the FEEDING project under the "Giovani Ricercatori" program of the University of Campania "Luigi Vanvitelli".

References

1. Abd ElHafeez S, D'Arrigo G, Leonardi D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxid Med Cell Longev*. 2021 Nov 30;2021:1302811. doi: 10.1155/2021/1302811. PMID: 34887996; PMCID: PMC8651375.
2. Balakrishnan S, Madigan D (2006) Decision trees for functional variables. In: Sixth International Conference on Data Mining (ICDM'06), IEEE, <https://doi.org/10.1109/icdm.2006.49>
3. Christodoulou, S.E. Water Network Assessment and Reliability Analysis by Use of Survival Analysis. *Water Resour. Manag.* 2011, 25, 1229–1238.
4. Ferraty F (2011) Recent Advances in Functional Data Analysis and Related Topics. Physica-Verlag HD.
5. Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treatment Reports*. 1985 Oct;69(10):1065-1069. PMID: 4042086.
6. Hemant Ishwaran. Udaya B. Kogalur. Eugene H. Blackstone. Michael S. Lauer. "Random survival forests." *Ann. Appl. Stat.* 2 (3) 841 - 860, September 2008.
7. Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M. (2004), Bagging survival trees. *Statist. Med.*, 23: 77-91.
8. Lin J, Li K, Luo S (2021). Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of Alzheimer's disease progression. *Statistical Methods in Medical Research*.
9. Martins Filho, S.; Duarte, M.L.; Venzon, M. Survival Analysis of the Green Lacewing, *Chrysoperla externa* (Hagen) Exposed to Neem-Based Products. *Agriculture* 2023, 13, 292.
10. Maturo, F., Verde, R. Supervised classification of curves via a combined use of functional data analysis and tree-based methods. *Comput Stat* 38, 419–459 (2023).
11. Ocana F, Aguilera A, Escabias M (2007) Computational considerations in functional principal component analysis. *Computational Statistics* 22(3):449–465.
12. Ozturk, S.; Fthenakis, V.; Faulstich, S. Assessing the Factors Impacting on the Reliability of Wind Turbines via Survival Analysis—A Case Study. *Energies* 2018, 11, 3034.
13. Preda C, Saporta G, Lévéder C (2007) PLS classification of functional data. *Computational Statistics* 22(2):223–235.
14. Quinlan JR (1986) Induction of Decision Trees. *Machine Learning*.
15. Ramsay J, Silverman B (2005) *Functional Data Analysis*, 2nd edn. Springer, New York.
16. Roshani D, Moradi G, Rasouli MA. Survival Analysis of Patients with Colorectal Cancer Undergoing Combined Treatment: A Retrospective Cohort Study. *J Res Health Sci.* 2023 Mar;23(1):e00572. doi: 10.34172/jrhs.2023.107. PMID: 37571943; PMCID: PMC10422145.
17. Royston, P., Parmar, M. K., Qian, W. (2002). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine*, 21(17), 2631-2642.

Advancing credit card fraud detection with innovative class partitioning and feature selection technique

Mohammed Sabri, Antonio Balzanella and Rosanna Verde

Abstract In the domain of credit card fraud detection, the application of supervised learning methodologies has been a prevalent approach. These methods rely on the analysis of historical transaction data to identify patterns of fraudulent activity. However, this approach is met with challenges due to the dynamic nature of consumer behavior and the continuous evolution of fraudster strategies. To overcome these challenges, this study proposes a novel model that incorporates both supervised and unsupervised learning techniques, along with an innovative feature selection method, to improve fraud detection capabilities. The integration of unsupervised K-means clustering allows for the identification of new, potentially fraudulent patterns within the data that were previously unrecognized. This phase is followed by a strategic feature selection process, which enhances the development of an advanced K-nearest neighbor algorithm. This algorithm is improved through informations obtained from the initial unsupervised learning phase. The empirical evaluation of this approach demonstrates its superiority in detecting fraudulent transactions, indicating a significant enhancement in detection accuracy.

Key words: Credit card fraud detection, K-nearest neighbor, K-means, Feature selection

Mohammed Sabri

Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy.
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, e-mail: sabri.mohammed@unicampania.it

Antonio Balzanella, Rosanna Verde

Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy,
e-mail: antonio.balzanella@unicampania.it, rosanna.verde@unicampania.it

1 Introduction

Credit cards are a popular and conventional method for making online payments. Credit card transactions, although widely accepted, pose significant challenges, especially in terms of security. The primary issue facing the online payment system is the detection and prevention of fraud [2, 4].

Supervised methods for fraud detection [6, 8], valued for their simplicity and capacity to generate clear results, have two major obstacles [3]. The initial challenge is their struggle to identify novel fraud tactics created by criminals to avoid detection systems. Another problem occurs when there are deviations in client behavior from the usual pattern, resulting in mistakes like false warnings. To overcome these challenges, combining unsupervised techniques with supervised methods in fraud detection systems can be advantageous. This hybrid method helps identify new patterns, enhancing the system's effectiveness in detecting and preventing fraud.

The main goal of this work is to improve fraud detection by developing a model that can identify both current fraud patterns and potential vulnerabilities before they become known types of fraud. To successfully tackle modern financial crime, it is essential to develop a proactive strategy that directly addresses the complex and constantly changing situations. Scammers always improve their strategies to avoid detection. Our algorithm utilizes a vast amount of historical data to search for hidden fraud indicators that, when revealed, greatly enhance the program's capacity to detect fraud. Our strategy involves not only identifying fraudsters in the act but also predicting their future actions by learning from previous fraud cases. This involves evaluating data to identify growing behavioral patterns that indicate either legitimate activities or new misleading fraud patterns. Identifying these noticeable differences enables us to categorize data in unique ways, which eventually improves our model's accuracy in distinguishing between positive and negative outcomes. This strategy addresses a significant gap in fraud detection by taking on the constantly changing issues presented by fraud, ensuring our defenses remain effective against both established and emerging threats.

2 The tri-phase credit card fraud detection framework

This study presents a new algorithm for partitioning inspired by the K-means method [1]. It uses a newly developed objective function to detect developing patterns within two preset classes of credit card transactions: Fraudulent and Legitimate. The strategy prioritizes optimizing two main aspects: enhancing cohesion between subgroups of the same class to ensure close relationships among subgroup members. Secondly, it seeks to maximize the dissimilarity between the prototype of one subgroup and the prototypes of subgroups from the opposite class. This dual optimization method improves the algorithm's capability to differentiate between legitimate and fraudulent transactions more efficiently.

The technique utilizes an iterative algorithm similar to K-means, focusing on minimizing a new objective function. An optimization technique was employed to determine the parameters of the weights and subgroup prototypes in the proposed objective function. We provide the variable solutions for partitioning the fraud group in the training set into subgroups, denoted by the variable F .

Let $T = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be the training set of n transactions. The vector $\mathbf{u}_i^F = (u_{i1}^F, \dots, u_{is}^F, \dots, u_{in_F}^F)$ represents the membership of the i -th transaction in the fraud class for the n_F subgroup. Similarly, $\mathbf{u}_i^L = (u_{i1}^L, \dots, u_{is}^L, \dots, u_{in_L}^L)$ represents the membership of the i -th transaction in the legitimate class for the n_L subgroup.

The solutions proposed for the subgroup's prototype variables are as follows:

$$z_{pj}^F = \frac{\sum_{i=1}^n u_{ip}^F x_{ij} (x_{ij} - z_{Lj})}{\sum_{i=1}^n u_{ip}^F (x_{ij} - z_{Lj})} \quad (1)$$

The variable z_{Lj} denotes the j^{th} feature value of the global centroid z_L of the legitimate group L , which is the average of the prototypes of the legitimate subgroups. It is represented as follows:

$$z_{Lj} = \frac{\sum_{q=1}^{n_L} \sum_{i=1}^n u_{iq}^L z_{qj}^L}{\sum_{q=1}^{n_L} \sum_{i=1}^n u_{iq}^L}; \quad (2)$$

The solutions proposed for the subgroup's weight variables are as follows:

$$w_{pj}^F = \frac{1}{\sum_{l=1}^m \left(\frac{D_{pj}^F}{D_{pl}^F} \right)^{\frac{1}{\beta-1}}} \quad (3)$$

where

$$D_{pj}^F = \sum_{i=1}^n u_{ip} \frac{(x_{ij} - z_{pj}^F)^2}{n_F (z_{pj}^F - z_{Lj})^2} \quad (4)$$

Follow the same process to determine the variables z_{pj}^L and w_{pj}^L for the legitimate class.

A new method for selecting features is introduced after the initial clustering step. A weighted method is incorporated into the clustering process at this stage to assess the significance of variables within subgroups. This phase improves the variable selection process by using a criterion based on the weights given to subgroups. We use the coefficient of variation (CV) of the weights as a key criterion in our method. We then apply a forward feature selection strategy based on this criterion to discover important characteristics [7]. The weight matrix W contains row vectors that represent the weights assigned to variables associated with different subgroups. The coefficient of variation (CV) for each variable is outlined as follows:

$$CV(W^j) = \frac{Std(W^j)}{\mu(W^j)} \quad j = 1, \dots, m \quad (5)$$

where W^j indicates the j -th feature column of the matrix W .

We then arrange the characteristics of W according to the criterion Eq. (5). We use forward feature selection on the sorted set of features, evaluating them based on their $F1 - score$ to determine if a feature negatively impacts the classifier and has to be removed. The algorithm finally returns the significant features.

This contribution presents a CCFD model that integrates a novel supervised classification approach. This model utilizes an advanced version of the KNN algorithm [5], drawing on insights from clustering and feature selection stages. It employs an Euclidean weighted distance metric, assigning weight vectors dynamically to distinguish fraud and legitimate subclasses within the training dataset. Each subgroup receives a unique weight vector. This improved KNN classifier employs a weighted distance methodology, enabling variable distances between query transactions and the closest points in the training set, based on the transaction's subgroup prototypes. The weighted distance used to compute the distance between a new query transaction x'_l and the subgroup's prototypes is described as follows:

$$d_w(x'_l, z_p^g) = \sum_{j=1}^m w_{pj}^g (x'_{lj} - z_{pj}^g)^2 \quad g \in \{F, L\} \quad (6)$$

Transactions within the same subgroups have identical vector weights. The CCFD model incorporates predictive attributes from the feature selection's second phase and uses the clustering phase's weights for feature differentiation. The refined KNN classifier is trained with labels updated through the K-means method, and new transactions are classified into predefined categories indicative of legitimate or fraudulent activities. The model aims to precisely classify transactions as fraudulent or legitimate within specific classes using the K-Nearest Neighbors (KNN) algorithm, with subgroup prototypes acting as nearest neighbors.

The proposed algorithm including the three steps using KNN with the subgroup's prototypes is represented in Algorithm 1.

Algorithm 1 Three step CCFD algorithm

-
- 1: **Input:** $T = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$: training dataset. n_F, n_L : the optimal number of subgroups within the fraud and legitimate groups, respectively. $\mathbf{x}'_l$: new transaction. k : number of nearest neighbors prototypes to $\mathbf{x}'_l$.
 - 2: **Output:** $y_{\mathbf{x}'_l}$ the label of $\mathbf{x}'_l$.
 - 3: **Step 1-** Partition of the fraud and legitimate classes of T using K-means clustering algorithm into n_F, n_L subgroup by updating the parameters Z and W using Eq. (1) and Eq. (3) respectively until the convergence.
 - 4: **Step 2-** Feature selection step: identify significant features using FFS algorithm.
 - 5: **Step 3-** KNN classifier using subgroups prototypes as NN
 - 6: By using Eq. (6) calculate the distances $d_w(\mathbf{x}'_l, \mathbf{z}_i^F), d_w(\mathbf{x}'_l, \mathbf{z}_i^L)$ ($i = 1, \dots, n_F$ and $r = 1, \dots, n_L$), by using the important features from Step 2.
 - 7: Choose the subset N_k that includes prototypes representing the k nearest neighbor subgroups of the query $\mathbf{x}'_l$. This selection is determined by the minimum distance between $\mathbf{x}'_l$ and the prototypes of the subgroups from both classes.

$$N_k(\mathbf{x}'_l) = \{\mathbf{z}_i^g : g \in F, L, i = 1, \dots, k\} \quad (7)$$

- 8: Allocate $\mathbf{x}'_l$ to its original class using majority voting.
 - 9: **return:** $y_{\mathbf{x}'_l}$
-

3 Application

The methodology described in the previous section is used on a simulated dataset and a real dataset provided by our European financial partner, consisting of 75 attributes. We balanced the training set of both datasets by applying the Synthetic Minority Over-sampling Technique (SMOTE) algorithm before applying the CCFD method to ensure balance in the dataset.

We compare our tri-phase model with current state-of-the-art Composite Credit Fraud Detection (CCFD) frameworks in our investigation. This comparative analysis using the simulated and the real dataset in question clearly demonstrates that our classifier strategy significantly improves the performance metrics of our model. Our analysis found that combining a three-step process involving K-means with a unique objective function, Forward Feature Selection (FFS) with a specific criterion for feature selection, and KNN using prototypes as nearest neighbors along with a data balancing strategy produced the best results in terms of various classification performance metrics like Accuracy, Recall, and Precision when compared to other classifiers. This result indicates that our model is effective in increasing the true positive rate and reducing the false negative rate, hence enhancing the identification of fraud cases. However, this method may mistakenly classify some valid transactions as fraudulent, highlighting the need for our model to discover more patterns within the group of legitimate transactions.

After discussing the outcomes with our financial partner, we have recognized a substantial interest in our tri-phase fraud detection technique. The interest in our

model is mainly because of its improved Recall, allowing it to identify more fraudulent transactions compared to other models. However, it is important to recognize that this approach also raises the likelihood of false positives. False positives occur when the model incorrectly labels a valid transaction as fraudulent, suggesting that the model may misclassify some legitimate transactions based on recognized patterns. This analysis is the first examination of the dataset, prompting us to propose various enhancements. The bank's current fraud detection software may not be proficient at recognizing the complex patterns created by fraudsters. Hence, additional preprocessing of the dataset is crucial before training the detection model.

References

1. Ahmed, M., Seraj, R., Islam, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, **9**(8), 1295 (2020)
2. Barker, K. J., D'amato, J., Sheridan, P. . Credit card fraud: awareness and prevention. *Journal of financial crime*, **15**(4), 398-410 (2008)
3. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F. and Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, **557**, 317–331 (2021)
4. Ileberi, E., Sun, Y., Wang, Z.: A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, **9**(1), 1-17 (2022)
5. Kataria, A., Singh, M. D. A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering*, **3**(6), 354-360 (2013)
6. Khatri, S., Arora, A. and Agrawal, A.P. Supervised machine learning algorithms for credit card fraud detection: a comparison. In 2020 10th international conference on cloud computing, data science & engineering (confluence) (pp. 680-683). IEEE. (2020).
7. Macedo, F., Oliveira, M. R., Pacheco, A., Valadas, R. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, **325**, 67-89 (2019)
8. More, R., Awati, C., Shirgave, S., Deshmukh, R., Patil, S. Credit card fraud detection using supervised learning approach. *International journal of scientific & technology research*, **9**(10), 216-219 (2021)

10. Solicited - Functional Data Analysis in Action

10.1 - Functional Linear Discriminant Analysis for Misaligned Surfaces

10.2 - Leveraging weighted functional data analysis to estimate earthquake-induced ground motion

10.3 - Functional autoregressive processes on a spherical domain for global aircraft-based atmospheric measurements

Functional Linear Discriminant Analysis for Misaligned Surfaces

Tomas Masak

Abstract The task of speech command classification is studied in the functional data analysis context. The raw audio signals are transformed into log-spectrograms, and the resulting bivariate random surfaces are smoothed and aligned by a functional registration procedure in order to adjust for different speech velocities among the speakers. The aligned and smoothed log-spectrograms are considered a random sample and classified using the functional linear discriminant analysis. A comparison is made against two corresponding approaches: one equivalent but not utilizing any functional registration steps, and another utilizing classical audio pre-processing steps in order to extract a set of features to be used for classification in a non-functional context.

Key words: classification, speech data, two-dimensional functional data, covariance separability, registration.

1 Introduction

The problem of classification of spoken commands is considered from the functional data analysis (FDA) perspective [15, 7]. FDA concerns with statistical inference on random samples of continuously indexed random processes. The available samples $X_1, \dots, X_N$, i.e. independent realizations of the underlying process $X : \mathcal{D} \rightarrow \mathbb{R}$, where the domain $\mathcal{D}$ is uncountable, are to be treated as elements of a Hilbert space $\mathcal{H}$ of functions on $\mathcal{D}$, e.g. the set of square-integrable functions $\mathcal{L}^2(\mathcal{D})$.

The full data set in consideration is provided in [21, 22] and consists of over 100,000 audio signals. We consider a subset corresponding to 10 English command words (*down, go, left, no, off, on, right, stop, up, yes*), and transform the audio signals into log-spectrograms [13]. Then we smooth the resulting surfaces, and perform alignment/registration using the elastic analysis [18] in a particular way. Finally, we explore the classification capabilities of functional linear discriminant analysis (fLDA) with a separable covariance [1]. Theoretically, fLDA was shown to be able to achieve zero misclassification probability under certain assumptions [4, 8] and is known to perform well with functional data.

Tomas Masak
Institute for Statistics and Mathematics, WU, Welthandelsplatz 1, 1020 Vienna, Austria
e-mail: tmasak@wu.ac.at

Since the considered data set was crowdsourced, i.e. recorded by various speakers in uncontrolled environments, a registration/alignment step is necessary in order to shift the surfaces and adjust for the different velocities in speech among the samples. Misalignment of this sort is a common complication when working with functional data, inducing a variability in phase. Failing to recognize the phase variability and wrongly attributing it to amplitude can seriously distort even the most basic features of the data [12, 17]

The same data set (or probably a subset of it) was used recently by Hoff et al. [6], who however used different pre-processing steps involving further discretization. They then tackled the data from the multivariate perspective, using maximum likelihood, as opposed to the functional perspective, where likelihood is not well defined. They utilized quadratic discriminant analysis, but did not take the alignment/registration problem into account, since their pre-processing approach (described in the following section) naturally flattened the data, making the alignment step inessential. Since those pre-processing steps correspond to those typically used in the signal processing community to define a set of features specifically for audio classification, this paper demonstrates how similar results can be achieved with rather general but purely functional considerations, which has the advantage of a much easier interpretation.

2 Methods

2.1 Pre-processing

The speech data considered come in the form of 1 second long `.wav` signals (we discard shorter instances), which can be interpreted as time series of integers capturing the air pressure sampled in a given time point and quantized to a specific bit depth [16]. A high air pressure is due to a short-period sound wave and vice versa. Since an audio signal is a combination of waves that is changing in time, the usage of the short-time Fourier transform is a reasonable first step.

From an audio signal $x_n = x_n(t)$, $n = 1, \dots, N$ and $t \in [0, 1]$, we calculate the short-time Fourier transform $Y_n(t, \omega)$ at time t frequency and ω as

$$Y_n(t, \omega) = \int_{-\infty}^{\infty} x_n(t)w(s-t)\exp(-i\omega s)ds,$$

where i is the imaginary unit and w is a window function (e.g. the triangular kernel). Then, the log-spectrogram is defined as $\tilde{X}_n(t, \omega) = 10 \log_{10}(|Y_n(t, \omega)|^2)$.

We use the R [14] package `tuneR` [9] to perform the calculation. The Hamming window function of length 12 ms is used for a discrete set of time points of $\{8k/\tau | k = 1, \dots, 124\}$, where $\tau = 16,000$ is the original sampling rate (i.e. 16 kHz). This results in an overlap and the maximum detectable frequency of $\tau \cdot 0.012 = 192$,

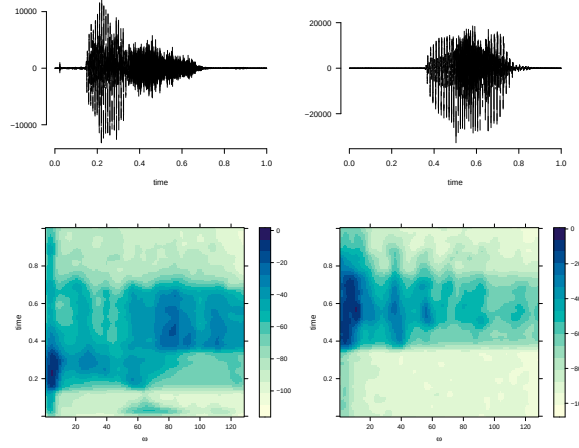


Fig. 1 Raw signals (top) and the corresponding smoothed log-spectrograms (bottom) for the templates of labels *yes* (left) and *no* (right).

with the nearest lower power of 2 being 128. Thus we end up with a discretization of the log-spectrograms of size 124×128 . See [19] for details.

At this point, our FDA approach diverges from the standard signal processing approach. While we smooth the log-spectrograms mildly (using the `loess` function with `span=0.01` in R [3]) and align them (see the next section), Hoff et al. [6] instead utilize the mel-frequency cepstral coefficients (MFCCs), a standard set of features for audio classification [19]. The MFCCs can be computed in a single step using `tuneR::melfcc`, which performs many intermediate steps. In principle, the MFCCs arise as filtered discrete cosine transforms of the log-spectrograms converted to a mel-frequency scale. See e.g. [19] for details.

2.2 Alignment

The log-spectrograms are expected to be misaligned in the temporal domain. In Figure 1, we clearly see that the *yes* was spoken earlier within the 1 second window than the *no* was in these particular instances. Moreover, speakers (note that we do not utilize the fact that some speakers provided recordings of multiple commands or even multiple recordings of the same command) can have different velocities, and these can change over the temporal domain.

There are various methodologies to perform functional alignment, c.f. the numerous references in [10, 2]. We opt for the elastic methodology, which was presented in [18] for multidimensional curves (paths from $[0, 1]$ to R^d , $d \in \mathbb{N}$), where only the

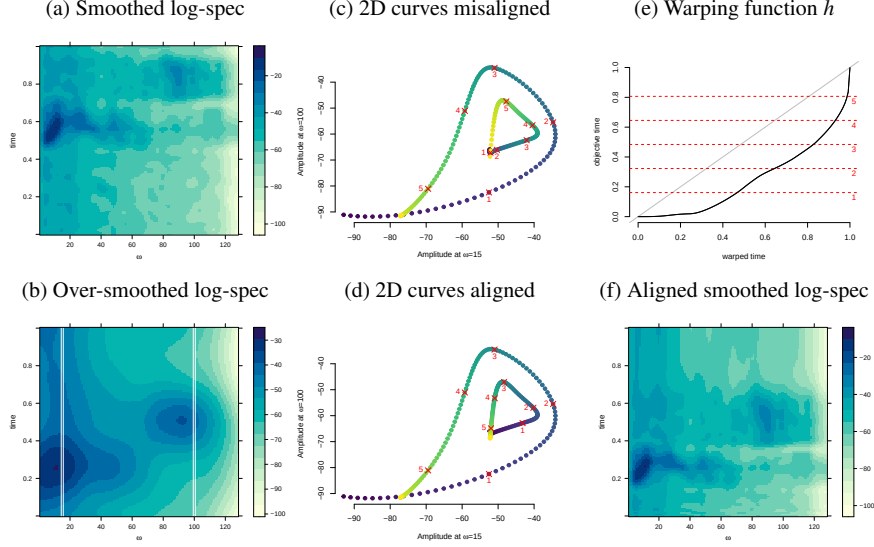


Fig. 2 The alignment procedure for a single instance of *yes* towards the template depicted in Figure 1, left. Firstly, the smoothed log-spectrogram (a) is over-smoothed (b). The amplitudes of frequencies $\omega = 15$ and $\omega = 100$, highlighted by white vertical lines in (b), are taken to form the inner 2D curve in (c). The outer 2D curve in (c) is obtained analogously from the template. Functional registration is applied to obtain aligned 2D curves (d). For (c) and (d), since the flow of time is only depicted by the color gradient, we also plot 5 checkpoints in red. The warping function that changes the aligned flow of time in (d) to the misaligned one in (c) is shown in (e). This is then applied to every frequency ω of the smoothed log-spectrogram (a) to obtain the aligned surface in (f).

images represent the objects of interest (i.e. the parametrization of $[0, 1]$ should not play a role), and the sparse sampling regime [23], where the observations are sampled irregularly on each curve. However, it can be utilized for our densely observed multivariate functional data as well, when those are interpreted as multidimensional images parametrized by a single variable, in our case the time. Such an interpretation is suitable for us, since we only want to register our log-spectrograms in time.

The alignment methodology is naturally sensitive to twists and turns, but we expect the changes in velocity to be smooth. Thus for the purposes of alignment only, we over-smooth the data, using `span=0.2` inside the `loess()` call in R before alignment. This over-smoothing can also be seen as biasing the alignment problem to avoid degenerate solutions, c.f. [20]. After over-smoothing, it does not make sense to consider a single surface as a 128-dimensional curve (recall that 128 is the no. of available frequencies), this would be both computationally prohibitive and overly variable. Instead, we constrain the problem further, focusing on only two frequencies $\omega = 15$ and $\omega = 100$, where the peaks in amplitude commonly arise. Hence an over-smoothed log-spectrogram $\tilde{X}$ is recast as a two-dimensional curve $\gamma = (\tilde{X}(t, 15), \tilde{X}(t, 100))$.

These two-dimensional curves are then aligned in a pair-wise manner. Since we naturally want to remove the phase variability inside a class, as it brings no useful information to the classification task, we simply choose a template for every label (as an arbitrary instance of that label, which clearly contains the full sound of the given word) and register data towards that template. The results of a single pairwise registration of a curve $\gamma: [0, 1] \rightarrow \mathbb{R}^2$ towards a template $\gamma': [0, 1] \rightarrow \mathbb{R}^2$ is a warping function $h: [0, 1] \rightarrow [0, 1]$ such that $\gamma \circ h^{-1}$ and γ' share the same parametrization. That is, we define the log-spectrogram $\tilde{X}$ registered towards the template as

$$X(t, \omega) = \tilde{X}(h^{-1}(t), \omega), \quad \text{i.e.} \quad X = \tilde{X} \circ (h^{-1} \otimes \text{id}).$$

See Figure 2 for an example of the whole process.

2.3 Classification

If X belongs to class $j \in \{1, \dots, G\}$, assume it follows the Gaussian distribution $\mathcal{N}(\mu_j, C)$ with a label-specific mean μ_j and a covariance operator C that is common across the labels. We calculate the scores

$$s_j(X) = \langle X - \mu_j, C^{-1}(X - \mu_j) \rangle, \quad j = 1, \dots, G,$$

and assign X to class with the minimum score. The means and the covariance need to be estimated, and the covariance needs to be regularized to facilitate a solution to the underlying inverse problem, see [8, 5].

The covariance C , i.e. a trace-class operator on $\mathcal{L}^2([0, T] \times [0, S])$, is called separable if $C = A \otimes B$, where A , resp. B , is a trace-class operator on $\mathcal{L}^2([0, T])$, resp. $\mathcal{L}^2([0, S])$. This can be understood as an alternative form of low-rankness [11].

We estimate the unstructured empirical covariance from the pooled sample, and then we use partial tracing [1] to obtain the estimates for the separable constituents of the covariance A and B . Crucially, when working on a discrete grid, the estimators $\hat{A}$ and $\hat{B}$ can be obtained efficiently, without the need to ever construct the unstructured empirical estimator explicitly (which would be computationally prohibitive). Finally, we produce the regularized separable covariance estimator to be plugged into the fLDA score calculations: $\hat{C} = \hat{A} \otimes \hat{B} + \varepsilon I$. Since the sample size is large, even small values of penalty parameters will work here, and we choose $\varepsilon = 10^{-3}$ without further investigation.

This estimator can be manipulated with ease. For example, an inverse problem of the form $\hat{C}U = V$ has the solution

$$U = \sum_i \sum_j \frac{1}{\alpha_i \beta_j + \varepsilon} \langle a_i \otimes b_j, V \rangle a_i \otimes b_j,$$

where $\hat{A} = \sum_i \alpha_i a_i \otimes a_i$ and $\hat{B} = \sum_j \beta_j b_j \otimes b_j$ are eigendecompositions of $\hat{A}$ and $\hat{B}$.

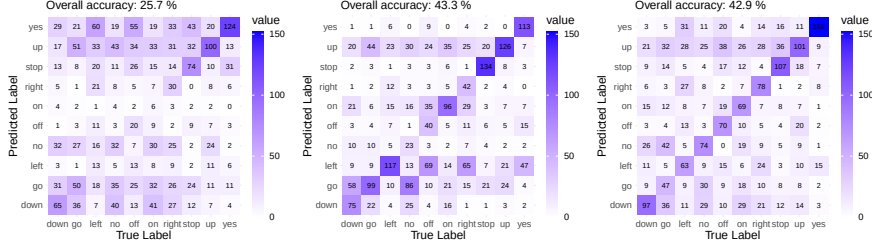


Fig. 3 Confusion matrices for the three considered algorithms: fLDA based on the log-spectrograms without alignment (left); with alignment, i.e. the proposed method (middle); and fLDA with the separable MLE estimator of the covariance based on the MFCCs (right).

3 Results

Firstly, the data set is randomly split into the training sample (2000 samples per class, so we can simplify the exposition working only with equal prior probabilities) and the test sample (200 samples per class). We register all the training samples towards the class-specific templates and estimate the class-specific means empirically and the common covariance as described in the previous section. Then, the classification of a new unlabeled datum X_* consists of the following steps.

1. Calculate the warping functions $h_1, \dots, h_G$ aligning X_* with the class-specific templates as described in Section 2.2.
2. Calculate the scores

$$s_j(X_*) = \left\langle X_* \circ (h_j^{-1} \otimes \text{id}) - \hat{\mu}_j, \hat{C}^{-1}(X_* \circ (h_j^{-1} \otimes \text{id}) - \hat{\mu}_j) \right\rangle, \quad j = 1, \dots, G.$$

3. Assign X_* to the class with the minimum score.

The previously described methodology is compared against the corresponding one without the alignment step, as well as the separable MLE approach of [6]. Of the various approaches presented in [6], we chose the one utilizing a separable covariance due to its correspondence to the classification methodology presented in this paper. There are essentially two differences. The MFCCs are used instead of the smoothed log-spectrograms, and the separable covariance is estimated via Gaussian MLE instead of partial tracing.

The results are reported in Figure 3. There are two key messages to be taken away. Firstly, we see that the algorithm without alignment performs quite poorly. This is due to the fact that the class means are estimated poorly from the log-spectrograms without alignment. Thus the alignment step is clearly important when working with log-spectrograms. Secondly, the presented algorithm outperforms the separable MLE approach, though just slightly (accuracy of 43.3 % vs 42.9 %). This is interesting because, despite MLE not being defined for functional data in general, it is in our experience often able to outperform moment estimators (such as partial tracing) even for functional data as long as the grid size vs. sample size ratio is relatively low, which it is the case of MFCCs here.

Acknowledgements I am grateful to Almond Stoecker for introducing me to the `elasdics` R package and to Davide la Manna, who performed the MFCC pre-processing of the data.

References

1. Aston, J. A., Pigoli, D., Tavakoli, S.: Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics* (2017)
2. Chakraborty, A., Panaretos, V. M.: Functional registration and local variations: Identifiability, rank, and tuning. *Bernoulli* Vol. 27, Issue 2 (2021)
3. Cleveland, W. S., Grosse, E., Shyu, W. M.: Local regression models. In *Statistical models in S*. Routledge (2017)
4. Delaigle, A., Hall, P.: Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74.2 (2012)
5. Delaigle, A., Hall, P.: Classification using censored functional data. *Journal of the American Statistical Association* 108.504 (2013)
6. Hoff, P., McCormack, A., Zhang, A.R.: Core shrinkage covariance estimation for matrix-variate data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2023)
7. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. Vol. 997. John Wiley & Sons (2015)
8. Kraus, D., Stefanucci, M.: Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika* 106.1 (2019)
9. Ligges, U., Krey, S., Mersmann, O., Schnackenberg, S.: `tuneR`: Analysis of Music and Speech. R package (2018)
10. Marron, J. S., Ramsay, J. O., Sangalli, L. M., Srivastava, A.: Functional data analysis of amplitude and phase variation. *Statistical Science* (2015)
11. Masak, T., Sarkar, S., Panaretos, V. M.: Separable expansions for covariance estimation via the partial inner product. *Biometrika*, 110(1) (2023)
12. Panaretos, V. M., Zemel, Y.: Amplitude and phase variation of point processes. *The Annals of Statistics* (2016)
13. Pigoli, D., Hadjipantelis, P. Z., Coleman, J. S., Aston, J. A.: The statistical analysis of acoustic phonetic data: Exploring differences between spoken Romance languages. *Journal of the Royal Statistical Society Series C: Applied Statistics* 67.5 (2018)
14. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria (2023) <https://www.R-project.org/>
15. Ramsay, J.O., Silverman, B.: *Functional Data Analysis*. Springer, New York (2005)
16. Spanias, A., Painter, T., Atti, V.: *Audio signal processing and coding*. John Wiley & Sons (2006)
17. Srivastava, A., Wu, W., Kurtsek, S., Klassen, E., Marron, J. S.: Registration of functional data using Fisher-Rao metric. *arXiv:1103.3817* (2011)
18. Steyer, L., Stöcker A., Greven, S.: Elastic analysis of irregularly or sparsely sampled curves. *Biometrics* 79.3 (2023)
19. Sueur, J.: *Sound analysis and synthesis with R*. Cham: Springer (2018)
20. Tang, R., Müller, H.-G.: Pairwise curve synchronization for functional data. *Biometrika*, 95(4) (2008)
21. Warden, P.: *Speech commands: A public dataset for single-word speech recognition*. (2017) Dataset available from http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz
22. Warden, P.: *Speech commands: A dataset for limited-vocabulary speech recognition*. *arXiv:1804.03209* (2018)
23. Zhang, X., Wang, J.-L.: From sparse to dense functional data and beyond. *The Annals of Statistics* (2016)

Leveraging weighted functional data analysis to estimate earthquake-induced ground motion

T. Bortolotti, R. Peli, G. Lanzano, S. Sgobba and A. Menafoglio

Abstract Ground motion models are fundamental tools for seismic hazard assessment, providing estimates of earthquake-induced ground motion based on seismic variables. A novel approach grounding on weighted functional data analysis is employed to extend a scalar ground motion model for Italy to the functional context. By incorporating observation-specific functional weights in the estimation routine, we aim to improve the accuracy and stability of model calibration in the presence of incomplete functional data. Through a simulation study, we show the effectiveness of the proposed methodology in enhancing the estimation accuracy and reducing variability compared to the traditional approach. Our findings highlight the potential of weighted functional data analysis to enhance the ground motion estimates and seismic hazard assessment, offering valuable insights for civil protection planning.

Key words: Functional data analysis, functional weights, partially observed data, ground motion model

Teresa Bortolotti

MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: teresa.bortolotti@polimi.it

Riccardo Peli

MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: riccardo.peli@polimi.it

Giovanni Lanzano

Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Milan, Italy, e-mail: giovanni.lanzano@ingv.it

Sara Sgobba

Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Milan, Italy, e-mail: sara.sgobba@ingv.it

Alessandra Menafoglio

MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: alessandra.menafoglio@polimi.it

1 Introduction

Ground motion models play a critical role in seismic hazard assessment by estimating earthquake-induced ground motion based on seismic variables. ITA18 is a scalar ground motion model for Italy proposed in [3] which separately estimates the mean of a longitudinal sequence of ground motion intensity measures defined over a set of vibration periods. By leveraging a novel weighted approach in the field of functional data analysis, in [1] we extend the ITA18 model to the functional framework. The inherent incompleteness of seismic response profiles across oscillation periods poses a significant challenge, prompting the need for advanced methodologies that can enhance the accuracy and stability of the estimated functional ground motion model. We employ state-of-the-art methods to reconstruct the incomplete response profiles and incorporate observation-specific functional weights in the fitting of the model. In doing so, we aim to address the uncertainty associated with partially observed seismic data, seeking a solution that is somewhere between reducing the domain of analysis to the part common to all curves and considering the full domain by analysing solely the fully observed curves. Through a comprehensive simulation study, we evaluate the performance of our proposed weighted functional methodology under various scenarios. By assessing the impact of different reconstruction methods and weight definitions on estimation outcomes, we show the effectiveness of our approach in reducing the estimation uncertainty and enhancing predictive accuracy. Eventually, the proposed methodology proves effective in improving the estimation precision in the calibration of the ground motion model, offering a more reliable framework for seismic risk assessment and civil protection planning in regions susceptible to seismic activity.

2 Method

Let $y_1, \dots, y_n$ be reconstructed curves belonging to $L^2(T)$, where T is an open subset of $\mathbb{R}$. The weighted approach couples each y_i to a functional weight $w_i : T \rightarrow [0, 1]$, taking value 1 where the curve is observed and decreasing to zero the more the reconstruction becomes uncertain. Analytically, this translates into formulating a new optimization criteria for concurrent regression, including the curve-specific functional weights. In doing so, full weight is given to errors made on the observed parts of the curves, and less weight to errors made on the reconstructed parts.

2.1 Weighted concurrent regression

The reconstructed functional observations are assumed to follow a concurrent regression model, namely

$$y_i = \sum_{j=1}^q \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $x_{i1}, \dots, x_{iq}$ are independent functional covariates, and $\beta_1, \dots, \beta_q$ are the functional coefficients defined on T . The errors $\varepsilon_1, \dots, \varepsilon_n$ are assumed to be realizations of a zero-mean stochastic process. We define a penalized fitting criterion which minimizes

$$\sum_{i=1}^n \int_T w_i (y_i - \beta_j x_{ij})^2 + \sum_{j=1}^q \int_T \lambda_j (D^2 \beta_j)^2. \quad (2)$$

The second sum in (2) is a roughness penalty regularizing the estimates, and $\lambda_1, \dots, \lambda_q$ are penalization parameters tuned via generalized cross-validation. Each coefficient is associated to a specific penalization parameter, meaning that the estimates of the coefficients are allowed to have diverse levels of smoothness.

The dimensionality of problem (2) is reduced by assuming that each β_j belongs to a finite dimensional space spanned by suitable basis functions $\theta_{j1}, \dots, \theta_{jL_j}$. Under this assumption, the vector $\beta := (\beta_1, \dots, \beta_q)$ can be expressed as $\beta = \Theta b$ and is uniquely identified by $b \in \mathbb{R}^L$, with $L = \sum_{j=1}^q L_j$. We show in [1] that the solution to problem (2) can be found in closed form for b , and reads

$$[J + R] b = \int_T \Theta(s)^T X(s)^T W(s) y(s) ds, \quad (3)$$

where $J := \int_T \Theta(s)^T X(s)^T W(s) X(s) \Theta(s) ds$, and R accounts for the ridge regularization.

2.2 Definition of the weights

We consider two alternative systems of weights, namely logistic weights and reconstruction driven weights.

Logistic weights Let the reconstructed curve y_i be originally observed up to t_i . Then the logistic weight is defined as

$$w_i(t) = \begin{cases} 1, & t \leq t_i \\ \frac{1}{1 + e^{(t - \mu_i)\alpha_i}} + d_i, & t > t_i \end{cases}, \quad (4)$$

where $\mu_i = (t_i + t_N)/2$ and $\alpha_i = a\hat{\sigma}_{t_i}$, $\hat{\sigma}_{t_i}$ being the empirical standard deviation of the observed values at t_i . Although continuity is not required, d_i is a corrective term guaranteeing continuity of the weight at t_i . Parameter $a > 0$ controls the rate of decay of the logistic functions. The advantage of using logistic weights lies in their interpretability and in the possibility of controlling the downweighting of the missing trajectory.

Reconstruction driven weights An alternative definition of the weights allows for fragmented patterns of missing information and relies on the covariance operator of the reconstruction error. Let y_i be a curve observed on O_i and reconstructed on M_i , $O_i \cup M_i = T$, $O_i \cap M_i = \emptyset$. Let $\mathcal{C}$ denote the covariance operator of y_i , and $\mathcal{V}_i$ the covariance operator of the reconstruction error. Additionally, let $c(t)$ be the diagonal of the kernel of $\mathcal{C}$ and $v_i(t)$ the diagonal of the kernel of $\mathcal{V}_i$. We define the reconstruction driven functional weight associated to y_i as

$$w_i(t) = 1 - \sqrt{\frac{\hat{v}_i(t)}{\hat{c}(t)}}, \quad \forall t \in M_i, \quad (5)$$

where $\hat{c}(t)$ and $\hat{v}_i(t)$ are the sample versions of $c(t)$ and $v_i(t)$, respectively. Notice that $v_i(t)$ quantifies the amount of uncertainty associated to the predicted trajectory in t , while $c(t)$ quantifies the uncertainty that there would be about $y_i(t)$ if we ignored the observed part. Then, $w_i(t)$ quantifies the reduction in uncertainty on $y_i(t)$ achieved through the reconstruction.

3 Simulation study

A simulation study is conducted to validate the proposed methodology and assess its robustness under different scenarios. We investigate the effectiveness of the weights in reducing the impact of the adopted reconstruction methods on the estimates, evaluate the accuracy of estimates with varying weight definitions, and analyze the performance of the weighted methodology with increasing fractions of partially observed data. Through simulations, we demonstrate the effectiveness of weighted functional analysis in reducing the variability of the estimates compared to the unweighted functional approach. The predictive performance of the methodology is further analyzed through leave-one-out cross-validation (LOO CV) on synthetic data, evaluating the functional prediction error under different weight definitions. By examining the empirical distributions of prediction errors, we show in [1] the existence of a trade-off between fully relying on observed data and neglecting information from missing trajectories, highlighting the importance of finely-tuned weighting systems in achieving optimal estimation outcomes. Such finding is in agreement with the work of [5], where a classification problem on partially observed functional data is tackled by considering an intermediate domain extension between the common and the full domain.

4 Case study

The analysed dataset includes 5607 seismological records, relative to 146 earthquakes and 1657 stations [4]. The spectral acceleration profiles of each record are

considered as response of the ground motion model. Each profile is identified by 37 longitudinal ground motion intensity measures, namely the peak ground acceleration and the spectral acceleration recorded at 36 vibration periods in the interval $[0.04 \text{ s}, 10 \text{ s}]$. In the dataset under analysis, less than 75 % of the records are observed up to the largest period of 10 s, while the remaining profiles are only partially observed up to a certain period $T_i < 10 \text{ s}$.

The functional ground motion model that we propose is a functional extension of the model proposed in [3], which separately fits a scalar linear regression to the peak ground acceleration and the 36 longitudinal values of spectral acceleration. The embedding of the scalar model into a functional framework reads

$$\begin{aligned} \log_{10} SA = & a + b_1(M_w - M_h)1_{(M_w \leq M_h)} + b_2(M_w - M_h)1_{(M_w > M_h)} + f_1 SoF_1 + f_2 SoF_2 \\ & + c_1(M_w - M_{\text{ref}})\log_{10} R + c_2 \log_{10} R + c_3 R + k \log_{10} \frac{V_0}{800} + e. \end{aligned} \quad (6)$$

In (6), SA is a random variable with values in the space of square integrable functions, a , M_h , M_{ref} and R are known functions with domain T and e is assumed to be generated by a zero mean stochastic process.

Before fitting (6), we carry out the three critical steps of calibrating the ridge penalization parameters entering (2), selecting the optimal system of weights, and selecting the optimal reconstruction method. In particular, the calibration of the ridge penalization, implying the search of an optimum in a nine-dimensional space, requires the use of an evolutionary algorithm for parameter selection (Centofanti et al. 2023). In all three steps, optimality is by means of a global measure of error in predicting the observed parts of the spectral acceleration profiles. To highlight the impact of the weighted functional methodology on the model calibration, we here comment on the estimates of coefficient c_1 , obtained with the scalar and with the weighted functional approaches (see Figure 1). The results relating to the other model coefficients are in [1]. The functional boxplot associated to the point estimate is obtained via bootstrap sampling. The functional estimate follows the trend of the scalar estimate while displaying a smoother behavior. However, we notice a significant difference in the right half of the range due to how the data is weighted. When we extend the scalar model to longer periods, we overlook some information from partially observed SA profiles, relying only on 75 % of the data used for functional estimation. Conversely, by leveraging the correlation among SA ordinates for profile reconstruction and using a weighted functional approach, our method prevents this loss of information. This allows us to obtain dependable estimates throughout the relevant oscillation period range.

Acknowledgements Teresa Bortolotti and Alessandra Menafoglio acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023–2027.

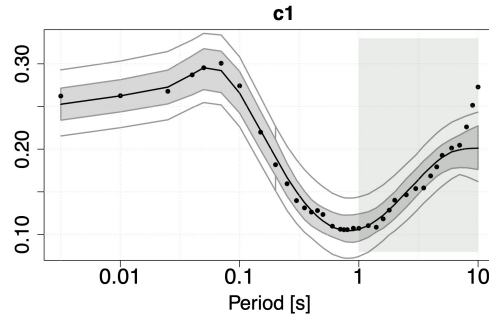


Fig. 1 Functional boxplot of the bootstrap sample of the estimated functional regression coefficient c_1 and comparison with its ITA18 point estimates (black dots). The black line represents the functional point estimate of the coefficient. The bootstrap sample has size 1000.

References

1. Bortolotti T., Peli R., Lanzano G., Sgobba S., Menafoglio A.: Weighted Functional Data Analysis for the Calibration of a Ground Motion Model in Italy. *Journal of the American Statistical Association*. 1-19 (2024)
2. Centofanti F. et al.: Adaptive smoothing spline estimator for the function-on-function linear regression model. *Computational Statistics*. **38**, 191-216 (2023)
3. Lanzano G. et al.: A Revised Ground Motion Prediction Model for Shallow Crustal Earthquakes in Italy. *Bulletin of the Seismological Society of America*. **109**(2), 525-540 (2019)
4. Lanzano G. et al.: Parametric table of the ITA18 GMM for PGA, PGV and Spectral Acceleration ordinates. Istituto Nazionale di Geofisica e Vulcanologia (INGV) (2022)
https://doi.org/10.13127/ita18/sa_flatfile/
5. Stefanucci M., Sangalli L., Brutti P.: PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. *Statistica Neerlandica*. **72**(3), 246-264 (2018)

Functional autoregressive processes on a spherical domain for global aircraft-based atmospheric measurements

Almond Stöcker and Alessia Caponera

Abstract Motivated by global analysis of aircraft-based measurements of air pollutants and climate variables, and specifically the COVID-19 pandemic’s possible impact on ozone concentrations, a functional autoregressive model is proposed to capture global spatio-temporal variability, incorporating solar radiation cycles. Efficient estimation techniques are developed and means of suitable visualization demonstrated, paving the way for similar analyses in the future.

Key words: functional autoregressive processes, spherical domain, ozone concentrations, aircraft-based, measurements, COVID-19

1 Introduction

Social and political reactions to the COVID-19 pandemic from late 2019 onward, causing a worldwide economic downturn of unprecedented extend, have lead to a reduction in pollutant emission that was soon reported in various studies [9]. The unusual abruptness of the change compared to typical economic developments has been considered a “unique scientific opportunity” [9], promising considerable new insights into large scale temporal volatility in atmospheric composition and the human’s impact on it. Many early studies have focused on its immediate impact on primary pollutants, such as carbon monoxide (CO) or nitrogen dioxide (NO₂), at the surface, say in an urban context, [9] or in the troposphere [6], often with a local, regional focus. Low level ozone (O₃) in the troposphere presents a secondary pollutant, where it is driven by CO and NO₂ among others [5]. [6] analyze the effect of COVID-19 reactions on the free troposphere (850-350hPa, ~1.5-8km) over the air-

Almond Stöcker

École Polytechnique Fédérale de Lausanne, e-mail: almond.stoecker@epfl.ch

Alessia Caponera

LUISS Guido Carli e-mail: acaponera@luiss.it

port of Frankfurt, [5] analyze regional anomalies over Western North America and Western Europe using additive mixed models. Most of the O_3 (around 90% of the atmosphere's O_3 , but still a small share of the atmospheric composition), however, is concentrated at higher altitude in the “ozone layer” where it plays an important role in protecting life on the Earth's surface by absorbing most of the ultraviolet radiation of the sun [8]. Natural dynamics here are governed by the sun's radiation via the ozone-oxygen cycle, producing about 12% of the ozone layer each day [10, Chap. 5]. The O_3 concentration takes its maximum at $\sim 20\text{km}$ above the surface in the stratosphere, but increases already from around 10km in the upper troposphere.

The main motivation of this work is to investigate in how far reactions to the COVID-19 pandemic also affected higher O_3 concentrations, focusing on the upper troposphere just at the transition to the ozone layer, and using aircraft-based ozone measurements (of the IAGOS project [1]) consulted in studies referred to above due to their long history and reliability. This comes with different statistical challenges: anticipating a potential impact to be less distinct than in lower altitudes and a generally high fluctuation in O_3 concentration, reliable modeling will have to include the spatio-temporal variability and will ideally be global rather than local as in previous studies to capture the globally varying O_3 concentration. This is ambitious for various reasons. Global modeling requires accounting for the (quasi-)spherical geometry, with comparably large data that is, on a global scale, sparse at the same time. While a conclusive analysis of the underlying research question is, thus, beyond of the scope of the present contribution, we illustrate some of the methods towards it, that might also inspire analysis in similar contexts. More specifically, we propose a spherical functional autoregressive model for this purpose and efficiently implement spline-type estimates of first- and second-order structures.

2 Global aircraft-based atmospheric data

The In-service Aircraft for a Global Observing System (IAGOS, [1]) unites different European research initiatives operating since 1994, where researchers equip commercial aircraft with instruments recording different atmospheric/climate variables along their flight trajectory in high resolution (for O_3 , in IAGOS Instrumentation Package1, temporal resolution 4s, accuracy ± 2 parts per billion [ppb], <https://www.iagos.org>, last access: 3 March 2024). While destination airports of the IAGOS fleet, in to date well over 60 000 flights, are widely spread around the globe, the data collection implies a particular spatial distribution: in lower altitudes, measurements are roughly vertical over airports, until reaching cruising phase where the altitude is widely constant. As the typical cruising altitude lies within the upper troposphere, we may thus, unlike studies focusing on low level ozone, optimally exploit the global distribution of IAGOS data when analyzing O_3 in this domain. Moreover, civic aircraft travel in distinct flight levels at every 1000 ft over sea level (measured in pressure altitude), such that at each flight level we may consider measurements to lie approximately on a sphere. In the

following, we analyze measures in FL370, a highly frequented flight level in the data (at 37000 ft, about 217hPa, 11.3km of altitude), in time span from 1 August 1994 (day $t = 1$) to 30 November 2019 ($t = 9253$) to exclude the eventuality of an anomaly due to reactions to the COVID-19 pandemic. At a later point, this may be used as reference for further analysis. For efficiency and due to autocorrelation, we restrict to $N_t \leq 100$ measurements per flight taken at (roughly) regularly distributed locations $U_{t1}, \dots, U_{tN_t}$ along its cruise trajectory. Given this subsampling, the flight identification is suppressed in the notation. Figure 1, illustrates their spatio-temporal distribution.

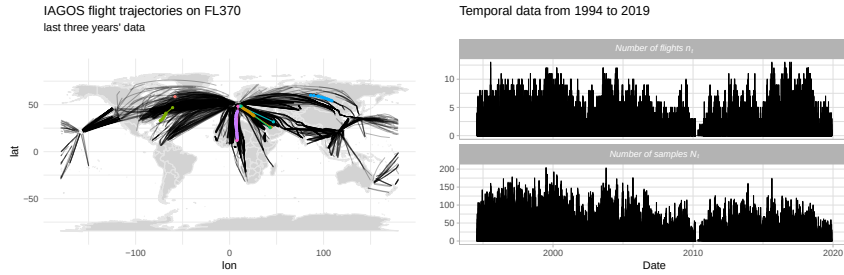


Fig. 1: *Left:* Spatial distribution of the used data (for the years 2017-19 only), corresponding to IAGOS flight trajectories (*transparent* black) at ~ 11.3 km of altitude (FL370). Flights of the last three days in color. *Right:* Temporal distribution of the used dataset over the whole time span, showing the number of flights (n_t) and selected measurements (N_t) for each day t .

3 A seasonal autoregressive model for log-ozone concentrations

Let $L^2 = L^2(\mathbb{S}^2)$ be the space of square integrable functions on the unit sphere $\mathbb{S}^2$ with inner product $\langle \cdot, \cdot \rangle_{L^2}$.

For analyzing lower stratosphere ozone, we model log ozone concentration layer $Y_t \in L^2$ on flight level FL370 (~ 11.3 km altitude), based on log ozone concentrations $Y_{tj} = Y_t(U_{tj}) + \varepsilon_{tj}$ measured by airplanes traveling at FL370 at random spherical coordinates U_{tj} on day t . The measurements are subject to noise disturbances ε_{tj} , independent from the measurement locations and the latent process.

Due to its fundamental role for the ozone concentration, different cycles of solar radiation are potentially relevant, including the seasonal cycle over the year, Quasi-Biennial Oscillation (QBO), and the 11-year solar cycle [10]. To account for both seasonality and temporal autocorrelation, ozone concentration surfaces are modeled as

$$Y_t = \mu_0 + \mu_{QBO}(t \bmod 2\text{years}) + \mu_{sol}(t \bmod 11\text{years}) + X_t,$$

with time-constant mean μ_0 , seasonal and biennial effects captured in μ_{QBO} and solar cycle effects in μ_{sol} of two and eleven (tropical) year periods, respectively. Given time t , the three mean components are assumed to belong to L^2 . We additionally model the residual process $\{X_t, t \in \mathbb{Z}\}$ as a zero-mean stationary *functional autoregressive process* on a spherical domain. A brief formal description is given in the following paragraph.

Let $\rho : L^2 \rightarrow L^2$ be a bounded linear operator with spectral norm $\|\rho\|_\infty < 1$, and $\{\xi_t, t \in \mathbb{Z}\}$ a *strong white noise* on L^2 – see [2, Definition 3.1]. We consider a collection of L^2 -valued zero-mean random variables $\{X_t, t \in \mathbb{Z}\}$, constituting the unique stationary solution of the autoregressive equation

$$X_t = \rho(X_{t-1}) + \xi_t, \quad t \in \mathbb{Z}.$$

For such model, the lag $h = 0, 1$ autocovariance operators are well defined on L^2 as $\mathcal{R}_h : f \mapsto \mathbb{E}[X_h(f, X_0)_{L^2}]$, and they are integral operators associated with the respective autocovariance functions $R_h : (u, v) \mapsto \mathbb{E}[X_{t+h}(u)X_t(v)]$ on $\mathbb{S}^2 \times \mathbb{S}^2$. Additionally, the two autocovariance operators are inherently connected to ρ through the equation $\mathcal{R}_1 = \rho \mathcal{R}_0$. This linkage points out why ρ is commonly interpreted as the *autocorrelation operator*.

If one is able to directly observe $X_{tj} = X_t(U_{tj}) + \varepsilon_{tj}$, an estimator for the lag $h = 0, 1$ autocovariance function can be obtained as the minimizer of a Tikhonov regularization problem by essentially smoothing the pairs of *raw autocovariances* $X_{t+h,j}X_{tk}$ at locations $(U_{t+h,j}, U_{tk})$. Formally, given a suitable choice of $\lambda_h > 0$,

$$\hat{R}_h = \arg \min_{R_h \in H_r^2} \sum_{t=1}^{T-h} \sum_{\substack{j,k=1 \\ j \neq k \text{ if } h=0}}^{N_t} (X_{t+h,j}X_{tk} - R_h(U_{t+h,j}, U_{tk}))^2 + \lambda_h \|R_h\|_{H_r^2}^2, \quad (1)$$

where $H_r^2 = H_r(\mathbb{S}^2) \otimes H_r(\mathbb{S}^2)$ is the tensor-product of two r -th order Sobolev spaces on $\mathbb{S}^2$ and $\|\cdot\|_{H_r^2}$ is the associated tensor-product norm, for some suitable $r > 1$. This results in a spline-type estimator with explicit expression and feasible implementation. Given $\hat{\mathcal{R}}_h$, $h = 0, 1$, it is possible to exploit the equation $\mathcal{R}_1 = \rho \mathcal{R}_0$ and define an estimator of ρ , for some suitable penalty parameter $\alpha > 0$, as

$$\hat{\rho} = \arg \min_{\rho \in \mathcal{S}_2} \|\hat{\mathcal{R}}_1 - \rho \hat{\mathcal{R}}_0\|_2^2 + \alpha \|\rho\|_2^2,$$

where $\mathcal{S}_2$ is the space of Hilbert-Schmidt operators on L^2 with norm $\|\cdot\|_2$. Consistency and rates of converge of such estimators have been studied in [3] and [4], respectively. Mean estimation is performed in the same spirit of (1), and hence, details are omitted. Additionally, note that once the mean is estimated and subtracted from Y_{tj} , we approximately obtain X_{tj} .

4 Implementation and visualization of first results

The size of the FL370 IAGOS data requires computationally efficient implementation. This holds for the mean (in particular considering smoothing parameter tuning) but even more for the estimation of the covariance functions R_0 and R_1 where the sample sizes N_t and the marginal basis dimension enter in quadratic order. $\mu_{QBO}(t, u)$ and $\mu_{sol}(t, u)$ are modeled using a tensor-product basis of cyclic cubic splines in time and second-order spherical thin-plate splines in space using component-wise penalties [12]. The latter basis is also used for $\mu_0(u)$ and as marginal basis for estimation of the covariance and cross-covariance functions $R_0(u, v)$ and $R_1(u, v)$, as resulting from a tensor-product penalty in Equation 1. For the cubic splines, instead of placing knots at all observed time points, $k_{QBO,time} = 12$ and $k_{sol,time} = 6$ knots are regularly placed over the respective period. For the spherical splines, low-rank approximations are utilized, reducing basis dimensions to $k_{space} = 50$ from $k_0 = 500$ knots originally selected at randomly among the U_{tj} .

Mean components, covariances, and the autocorrelation operator are estimated successively in that order, plugging in previous estimates. For mean estimation, we directly use the R package `mgcv` [12], also used for basis setup and low-rank approximation for $\hat{R}_0(u, v)$ and $\hat{R}_1(u, v)$. As their estimation presents the computational bottleneck, we develop a new, fast approach to covariance smoothing, based – unlike previous approaches – on the generalized linear array model [7]. This leads to a substantial efficiency gain both in terms of memory and speed. A new implementation was also required for $\hat{\rho}$. The smoothing parameter for the mean is selected via Restricted Maximum Likelihood (REML) and for the covariances based on simple split calibration. The regularization parameter for estimation of ρ is determined to ensure stationarity, i.e. $\|\rho\|_\infty \approx 0.99 < 1$, which can otherwise likely fail in high-dimensions due to randomly high correlations between nuisance directions.

To visualize the results (Fig. 2, 3), we compute the spectral decomposition of $\hat{\mathcal{R}}_0$ with leading eigenvalue-eigenfunction pair (σ_1^2, e_1) . Similarly, we decompose $\hat{\mu}_{QBO}(t, u) = \sum_j f_{QBO,j}(t) \eta_{QBO,j}(u)$ into orthogonal components explaining a decreasing share of variance, using tensor-product effect factorization [11]. Analogously for μ_{sol} and slightly adjusted for $\hat{\rho}$ yielding $\hat{X}_{t+1} = \hat{\rho}(X_t) = \sum_j \phi_j \langle \psi_j, X_t \rangle_{L^2}$. $\phi_j, \psi_j \in L^2$.

5 Discussion & outlook

In this work, we provide first methods for aircraft-based measurements, collected in approximately spherical domains naturally given by flight levels, with a solid theoretical foundation, fast practical implementation, and illustrating means of detailed visualization. While the presented first results are encouraging and seem to qualitatively align with expectations, e.g. in terms of seasonal and biennial cycle (comparing, say, with [10]) and the overall mean distribution of ozone, estimates

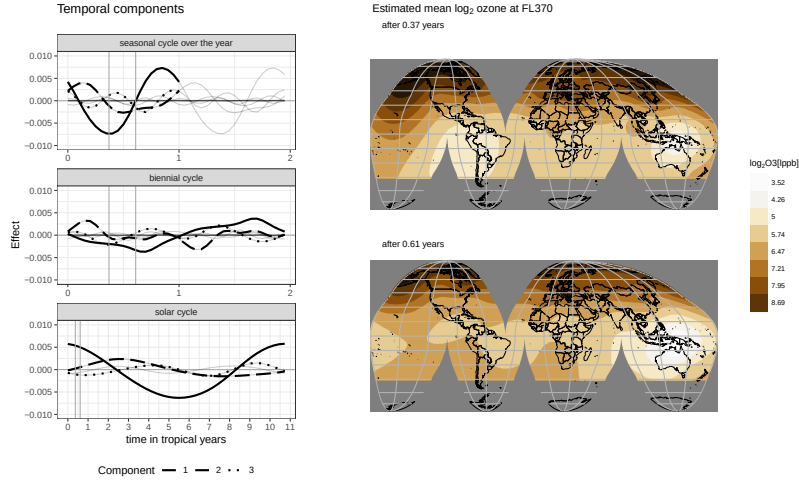


Fig. 2: *Left*: the $f_{QBO,j}$ further split into their seasonal part $f_{year,j}$ (one year period, *top*) and the remainder $f_{QBO,j} - f_{year,j}$ (*middle*), and $f_{sol,j}$ (*bottom*). *Vertical gray lines* indicate selected points in the cycle where $\hat{\mu}_0 + \hat{\mu}_{QBO} + \hat{\mu}_{sol}$ is shown (*right*).

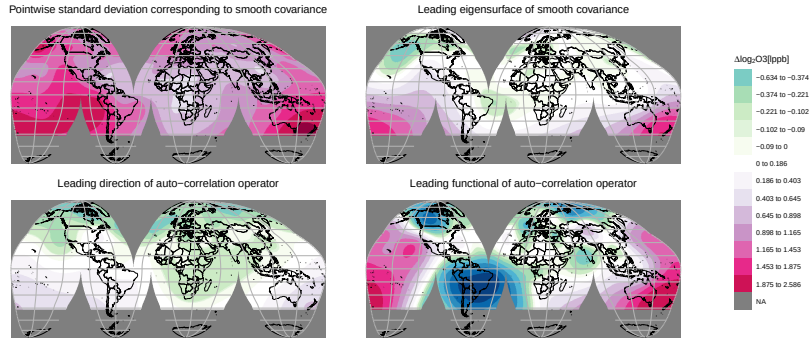


Fig. 3: *In lexical order*: Estimated square-root diagonal $\hat{R}_0^{1/2}(u, u)$ (i.e. standard deviation of $X_t(u)$), leading eigensurface $\sigma_1^2 e_1(u)$, and leading auto-correlation direction $\phi_1(u)$ and functional $\langle \psi_1, X_t \rangle$ (visualized by $\psi_1(u)/\|\psi_1\|$), for $u \in \mathbb{S}^2$.

rather outside of the flights of the IAGOS fleet and the estimate of the 11-year solar cycle should not be over-interpreted due to lack of data/repetitions. Here, potential stabilization/improvement might be achieved by combination with satellite-based data. Moreover, while accounting for well-known seasonal effects in this model, a potential trend in the ozone concentration will also need to be taken into consideration. With further refinement and comprehensive model validation, the proposed approach can then present the central basis for a careful investigation of potential anomalies due to social and political reactions to the COVID-19 pandemic up to the upper troposphere, but also for other comparable studies on the atmospheric composition.

Acknowledgments The authors wish to sincerely thank Victor M. Panaretos for his support, many insightful discussions and suggestions.

References

1. Blot, R., Nedelec, P., Boulanger, D., Wolff, P., Sauvage, B., Cousin, J.M., *et al.*: Internal consistency of the iagos ozone and carbon monoxide measurements for the last 25 years. *Atmospheric Measurement Techniques* **14**(5), 3935–3951 (2021)
2. Bosq, D.: *Linear Processes in Function Spaces: Theory and Applications*. Springer (2000)
3. Caponera, A., Fageot, J., Simeoni, M., Panaretos, V.M.: Functional estimation of anisotropic covariance and autocovariance operators on the sphere. *Electronic Journal of Statistics* **16**(2), 5080–5148 (2022)
4. Caponera, A., Panaretos, V.M.: On the rate of convergence for the autocorrelation operator in functional autoregression. *Statistics & Probability Letters* **189**, 109,575 (2022)
5. Chang, K.L., Cooper, O.R., Gaudel, A., Allaart, M., Ancellet, G., Clark, H., *et al.*: Impact of the covid-19 economic downturn on tropospheric ozone trends: An uncertainty weighted data synthesis for quantifying regional anomalies above western north america and europe. *AGU Advances* **3**(2), e2021AV000,542 (2022)
6. Clark, H., Bennouna, Y., Tsvilidou, M., Wolff, P., Sauvage, B., Barret, B., *et al.*: The effects of the covid-19 lockdowns on the composition of the troposphere as seen by in-service aircraft for a global observing system (iagos) at frankfurt. *Atmospheric Chemistry and Physics* **21**(21), 16,237–16,256 (2021)
7. Currie, I.D., Durban, M., Eilers, P.H.: Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68**(2), 259–280 (2006)
8. Dessler, A.: *Chemistry and Physics of Stratospheric Ozone*. Elsevier (2000)
9. Gkatzelis, G.I., Gilman, J.B., Brown, S.S., Eskes, H., Gomes, A.R., Lange, A.C., *et al.*: The global impacts of COVID-19 lockdowns on urban air pollution: A critical review and recommendations. *Elementa: Science of the Anthropocene* **9**(1), 00,176 (2021)
10. Hollandsworth, S.M., Binder, M.D.: Trends in stratospheric ozone. In: R.M. Todaro (ed.) *The Stratospheric Ozone Electronic Textbook*. NASA Goddard Space Flight Center Atmospheric Chemistry and Dynamics Branch (2003)
11. Stöcker, A., Steyer, L., Greven, S.: Functional additive models on manifolds of planar shapes and forms. *Journal of Computational and Graphical Statistics* pp. 1–13 (2023)
12. Wood, S.N.: *Generalized additive models: an introduction with R*. CRC press (2017)

11. Solicited - Bayesian Inference for Graphical Models

- 11.1 - Log-likelihood approximation in Stochastic EM for Multilevel Latent Class Models
- 11.2 - MCMC Sampling in Bayesian Gaussian Structure Learning

Log-likelihood approximation in Stochastic EM for Multilevel Latent Class Models

Silvia Columbu, Nicola Piras and Jeroen K. Vermunt

Abstract Multilevel cross-classified Latent Class Models are an extension of standard latent class for handling data in which each observation is simultaneously nested within two groups. The likelihood associated to the model is untractable and approximation methods such as stochastic versions of the EM can be applied. The knowledge of a final estimate of the log-likelihood can be helpful in the evaluation of parameter estimates and for selection purposes. We propose two alternative log-likelihood approximation procedures and test their performances in the Hierarchical Multilevel Latent Class Model for which a finite estimate of the likelihood is provided through a special version of the EM.

Key words: Log-likelihood, Classification Log-likelihood, Multilevel Latent Class

1 Introduction

Latent Class Models can be extended to deal with multilevel cross-classified (CC) data structures in which units are simultaneously grouped within multiple higher level units (for example, children nested within both schools and neighborhoods). A similar extension is that of Hierarchical Multilevel Latent Class (MLC) (see [3]) where it is considered the nesting of each observation within a single group. Un-

Silvia Columbu,
Department of Mathematics and Computer Science, University of Cagliari, e-mail: silvia.columbu@unica.it

Nicola Piras,
Department of Mathematics and Computer Science, University of Cagliari, e-mail: nicola.piras97@unica.it

Jeroen K. Vermunt,
Department of Methodology and Statistic, Tilburg University, e-mail: J.K.Vermunt@tilburguniversity.edu

like the MLC, in the cross-classified context the presence of a double missing data structure at the higher level causes the intractability of the derived log-likelihood that cannot be factorized as a product of the mixing probabilities. To overcome this issue, in [1] we have proposed the implementation of a Stochastic version of the EM algorithm including a Gibbs sampler between the E and the M step. The Gibbs step consists in the consecutive sampling from marginal posterior distributions of higher and lower level latent variables. This sampling procedure generates an irreducible Markov chain with a unique stationary distribution concentrated around the maximum likelihood parameter estimates. Therefore, our estimation procedure, unlike what happens in the MLC (see [3]), does not provide a finite estimate of the maximum log-likelihood.

The availability of such a finite approximation would be useful in the assessment of convergence of the estimation algorithm as well as in the definition of global model selection procedures based on the BIC. In this work we propose an approach to obtain a good approximation of the maximum of log-likelihood moving from the Markov chains obtained in the stochastic estimation algorithm. Given the finite estimate of maximum log-likelihood in the MLC setting, we propose to assess the best approximation procedure in stochastic versions of EM algorithms by performing comparative simulation studies in that frame. These results can then be extended to the Multilevel cross-classified latent class (MCCLC) formulation.

2 Log-likelihood approximation

The estimation of parameters in latent class models usually involves the introduction of the complete data log-likelihood, known in a classification context also as classification log-likelihood (CL). Its expression can be directly linked to the log-likelihood itself and therefore used to derive an approximation once that an estimation of model parameters is plugged in. In fact, following [2], the log-likelihood can be expressed as

$$\log L(\hat{\theta}) = CL(\hat{\theta}) + EN(\hat{\theta}) \quad (1)$$

where $CL(\hat{\theta})$ is the classification log-likelihood which is equivalent to the expected complete log-likelihood once the classification is given and $EN(\hat{\theta})$ is an entropy term which gives a measure of the separation of classes in the data. Their analytical expression depends on the specific model assumed. In particular, in this contribution we will consider the application of such a decomposition to estimate the maximum of the log-likelihood in two multilevel extensions (MLC and MCCLC) of latent class models. In what follows we will remind model formulations and their corresponding CL and EN expressions.

Log-likelihood approximation for MLC Let Y_{ijk} be the response on item i ($i = 1, \dots, I$) of individual or first level unit j ($j = 1, \dots, n_k$) belonging to the group level units k ($k = 1, \dots, K$). Following [3], we consider two sets of discrete latent variables X_{jk} and W_k with associated ℓ ($\ell = 1, \dots, L$) latent classes of level-1 and h

($h = 1, \dots, H$) of level-2. In this framework the model is described by two separate equations for the second level units and for the first level units, with resulting log-likelihood

$$\log L = \sum_{k=1}^K \log \sum_{h=1}^H P(W_k = h) \prod_{j=1}^{n_k} \left[\sum_{\ell=1}^L P(X_{jk} = \ell | W_k = h) P(\mathbf{Y}_{jk} | X_{jk} = \ell) \right]$$

The set of model parameters is $\theta = \{\pi_h = P(W_k = h), \pi_{\ell|h} = P(X_{jk} = \ell | W_k = h), \pi_{y_{ij\ell}} = P(\mathbf{Y}_{jk} | X_{jk} = \ell)\}$, where last ones are probability distribution parameters. Following [4] the Classification Log-likelihood for the estimated parameters, is computed as

$$\begin{aligned} CL(\hat{\theta}) &= \sum_{k=1}^K \sum_{h=1}^H P(W_k = h | \mathbf{y}_k) \log(P(W_k = h)) \\ &+ \sum_{h=1}^H \sum_{j=1}^n \sum_{\ell=1}^L P(W_k = h, X_{jk} = \ell | \mathbf{y}_k) \log(P(X_{jk} = \ell | W_k = h) P(\mathbf{Y}_{jk} | X_{jk} = \ell)) \end{aligned}$$

where $P(W_k = h, X_{jk} = \ell | \mathbf{y}_k) = P(W_k = h | \mathbf{y}_k) P(X_{jk} = \ell | \mathbf{y}_{jk}, W_k = h)$, while the Entropy (EN) term is

$$\begin{aligned} EN(\hat{\theta}) &= \sum_{k=1}^K \sum_{h=1}^H -P(W_k = h | \mathbf{y}_k) \log(P(W_k = h | \mathbf{y}_k)) \\ &+ \sum_{h=1}^H \sum_{j=1}^n \sum_{\ell=1}^L -P(W_k = h, X_{jk} = \ell | \mathbf{y}_k) \log(P(X_{jk} = \ell | W_k = h, \mathbf{y}_k)) \end{aligned}$$

Log-likelihood approximation for MCCLC The extension of MLC models to cross-classified structures requires the introduction of a third latent variable for level-2 memberships (Z_q) as well as a corresponding set of latent classes q ($q = 1, \dots, Q$) (see [1]). The expression of the log-likelihood becomes:

$$\begin{aligned} \log L(\theta) &= \log \sum_{h_1=1}^H \sum_{h_2=1}^H \cdots \sum_{h_K=1}^H \sum_{r_1=1}^R \sum_{r_2=1}^R \cdots \sum_{r_Q=1}^R \prod_{k=1}^K P(W_k = h_k) \prod_{q=1}^Q P(Z_q = r_q) \times \\ &\prod_{j=1}^{n_{kq}} \left[\sum_{\ell=1}^L P(X_{jkq} = \ell | W_k = h_k, Z_q = r_q) P(\mathbf{Y}_{jkq} | X_{jkq} = \ell) \right], \end{aligned}$$

so that the Classification Log-likelihood and the Entropy are computed as

$$\begin{aligned} CL(\hat{\theta}) &= \sum_{k=1}^K \sum_{h=1}^H P(W_k = h | \mathbf{y}_k) \log(P(W_k = h)) + \sum_{q=1}^Q \sum_{r=1}^R P(Z_q = r | \mathbf{y}_q) \log(P(Z_q = r)) \\ &+ \sum_{h=1}^H \sum_{r=1}^R \sum_{j=1}^n \sum_{\ell=1}^L P(W_k = h, Z_q = r, X_{jk} = \ell | \mathbf{y}_{jkq}) \log(P(X_{jkq} = \ell | W_k = h, Z_q = r) P(\mathbf{Y}_{jkq} | X_{jkq} = \ell)) \end{aligned}$$

with $P(W_k = h, Z_q = r, X_{jkq} = \ell | \mathbf{y}_{jkq}) = P(W_k = h, Z_q = r | \mathbf{y}_{jkq}) P(X_{jkq} = \ell | \mathbf{y}_{jkq}, W_k = h, Z_q = r)$, and

$$\begin{aligned} EN(\hat{\theta}) = & \sum_{k=1}^K \sum_{h=1}^H -P(W_k = h | \mathbf{y}_k) \log(P(W_k = h | \mathbf{y}_k)) + \sum_{q=1}^Q \sum_{r=1}^R -P(Z_q = r | \mathbf{Y}_q) \log(P(Z_q = r | \mathbf{y}_q)) \\ & + \sum_{h=1}^H \sum_{r=1}^R \sum_{j=1}^n \sum_{\ell=1}^L -P(W_k = h, Z_q = r, X_{jk} = \ell | \mathbf{y}_{jkq}) \log(P(X_{jkq} = \ell | W_k = h, Z_q = r, \mathbf{y}_{jkq})) \end{aligned}$$

3 Estimation and assessment of log-likelihood approximation

With the aim of assessing the performances of the estimation obtained through the decomposition in (1), in the MLC we use two estimation approaches, the upward-downward EM proposed in [3], and a Stochastic version of the EM (SMLC). The finite estimation obtained with the first version is compared with the approximated one in two separate simulation studies for binary and categorical observations. The SMLC algorithm scheme is the following, after initialization of $\pi_h, \pi_{\ell|h}, \pi_{y_{i|\ell}}$, iterate the following sampling steps:

SE step:

- 1) Draw $\mathbf{w}^{(t)}$ from a Multinomial distribution with probabilities

$$P(W_k = h | \mathbf{y}_k) = \frac{\pi_h P(\mathbf{Y}_k | W_k = h)}{P(\mathbf{Y}_k)}, \text{ where } P(\mathbf{Y}_k | W_k = h) = \prod_{j=1}^{n_k} P(\mathbf{Y}_{jk} | W_k = h)$$

- 2) Draw $\mathbf{x}^{(t)}$ from a Multinomial distribution with probabilities

$$P(X_{jk} = \ell | \mathbf{y}_{jk}, \mathbf{w}^{(t)}) = \frac{[\pi_{\ell|h} P(\mathbf{Y}_{jk} | X_{jk} = \ell)]^{w_{jk}^h}}{P(\mathbf{Y}_{jk})}$$

M step:

$$\pi_h = \frac{\sum_{k=1}^K w_k^{h(t)}}{K}, \quad \pi_{\ell|h} = \frac{\sum_{j=1}^n w_{jk}^{h(t)} x_{jk}^{\ell(t)}}{\sum_{j=1}^n w_{jk}^{h(t)}}, \quad \pi_{i|\ell} = \frac{\sum_{j=1}^n x_{jk}^{\ell(t)} y_{ijk}}{\sum_{j=1}^n x_{jk}^{\ell(t)}}.$$

with w_k^h binary indicator of units' membership at higher level, x_{jk}^ℓ binary indicator of units' membership at lower level and w_{jk}^h the expansion of higher level latent class indicators over the first level units j . Final estimates are then calculated as the mean over the total number of iterations after a burn-in period.

Following the Iterative scheme of the stochastic EM the final computation of CL and EN can be performed following two different approaches:

- (a) taking $\hat{\theta}$ as the mean over the iterations of the stochastic algorithm
- (b) plugging-in the current $\theta^{(t)}$ parameter values at each iteration and computing the final $CL(\hat{\theta})$ and $EN(\hat{\theta})$ as average over the iterations.

To evaluate the goodness of these different approaches we have performed two simulation studies each including 50 synthetic datasets with $K = 50$, $n_k = 30$ (number of units per group) and a fixed number of classes $L = 4$, $H = 3$. A total of 500 iterations were considered in the SMLC implementation, including 150 as burn-in period. In simulation 1 six binary indicators were generated, in simulation 2 we generated six categorical variables: two binary, two with three modalities and two with four. In Table 1 the results of both simulations scenarios are summarised through the average absolute error resulted in the comparison of the approximated values of $\log L$, as decomposition in the sum of CL and EN , obtained by applying the two approaches proposed, and the finite values of the three components in (1) obtained through the EM, computed with LatentGOLD 6.0 software (see Table 2).

Results from simulations show that the approximation of log-likelihood values computed through approach (a) are preferable, in terms of the error produced, than those given from approach (b). More specifically, we observe that the estimates though approach (b) are to be preferred in the approximation of the CL component, whereas the other approach works better for the approximation of EN .

Table 1 Average absolute errors in the comparison between Stochastic EM approximation and EM finite estimate

	Log-likelihood	CL	EN
Simulation 1			
Approach (a)	1.88	11.00	9.30
Approach (b)	11.14	7.15	14.42
Simulation 2			
Approach (a)	1.08	10.35	9.69
Approach (b)	14.79	8.76	17.67

Table 2 Average values for Log-likelihood, CL and EN computed with LatentGOLD 6.0

	Log-likelihood	CL	EN
Simulation 1	-5552.33	-6221.07	668.74
Simulation 2	-8062.26	-8655.73	593.47

4 The case of MCCLC models

In the extension to cross-classified data structures the estimation algorithm takes into consideration the simultaneous belonging of observations to W_k and Z_q level-2 latent variables, see [1]. In the stochastic algorithm the first point of SE step becomes

1.1) Draw $\mathbf{w}^{(t)}$ from a Multinomial distribution with probabilities

$$P(W_k = h | \mathbf{y}_k, \mathbf{z}^{(t-1)}) = \frac{\pi_h P(\mathbf{Y}_k | \mathbf{z}^{(t)}, W_k = h)}{P(\mathbf{Y}_k | \mathbf{z}^{(t)})},$$

$$P(\mathbf{Y}_k | \mathbf{z}, W_k = h) = \prod_{q=1}^{Q_K} \prod_{r=1}^R \left[\prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) \right]^{z_q^r};$$

1.2) Draw $\mathbf{z}^{(t)}$ from a Multinomial distribution with probabilities

$$P(Z_q = r | \mathbf{y}_q, \mathbf{w}^{(t)}) = \frac{\pi_r P(\mathbf{Y}_q | \mathbf{w}^{(t)}, Z_q = r)}{P(\mathbf{Y}_q | \mathbf{w}^{(t)})},$$

$$P(\mathbf{Y}_q | \mathbf{w}, Z_q = r) = \prod_{k_q=1}^{K_Q} \prod_{h=1}^H \left[\prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) \right]^{w_k^h}.$$

In this case, the computation of CL and EN terms requires the knowledge of the double joint posterior probability $P(W_k = h, Z_q = r | \mathbf{y}_{kq})$ which cannot be factorised as product of the conditional probability of each higher level latent class. The estimation through the approach (b) proposed in Section 3 can be performed running an additional Gibbs sampler at each iteration of the stochastic EM, with π_h and π_r set to their current values $\pi_h^{(t)}$ and $\pi_r^{(t)}$. In this case steps 1.1) and 1.2) are repeated several times and that probability is given as mean of the samplings over these iterations.

5 Conclusions

We have provided preliminary results on the study of likelihood approximations in stochastic EM for multilevel latent classes with the aim of extending these insights to situations in which no finite likelihood estimate is available. The simulations in the hierarchical model have suggested that the best likelihood approximation is obtained taking approach (a). We assume that this behaviour is also valid in the cross-classified extension where a double sampling is performed at level-2. In this case the choice of approach (a) over (b) is particularly convenient in terms of computational complexity, indeed it is sufficient a single Gibbs sampling at each E step.

References

1. Columbu, S., Piras, N., Vermunt, J. K. (2003) Multilevel Cross-Classified Latent Class Models, CLADAG Book of Abstracts and Short Papers, ISBN: 9788891935632, 390–393
2. McLachlan, G.J., Peel, D. (2000) Finite Mixture Models. New York: John Wiley & Sons, Inc.
3. Vermunt, J. K. (2008) Latent class and finite mixture models for multilevel data sets. Statistical Methods in Medical Research, 17, 33–51
4. Vermunt, J. K., Magidson, J. (2016) Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc

MCMC Sampling in Bayesian Gaussian Structure Learning

Antonino Abbruzzo, Nicola Argentino, Reza Mohammadi, Maria De Iorio, Willem van den Boom, Alexandros Beskos

Abstract This work discusses structural learning algorithms for Bayesian Gaussian Graphical models. The challenge lies in selecting and estimating the precision matrix to determine the graph structure. An algorithm based on a Birth-Death process within an MCMC (BD-MCMC) sampling process is introduced in [2]. This algorithm involves a combination of continuous and discrete Markov processes. However, combining continuous and discrete Markov processes may not guarantee convergence to the stationary distribution. To overcome this, we propose a strategy that discretises the birth-death MCMC's continuous component. A simulation study shows that this procedure helps in some extreme cases where the BD-MCMC fails and at the very least matches its performance in the other scenarios.

Key words: Bayesian Inference, Gaussian Graphical models, Markov processes

Antonino Abbruzzo

Department of Economics, Business and Statistics, UniPA, e-mail: antonino.abbruzzo@unipa.it

Nicola Argentino

Department of Economics, Business and Statistics, UniPA, e-mail: nicola.argentino@unipa.it

Reza Mohammadi

Department of Business Analytics, University of Amsterdam, e-mail: a.mohammadi@uva.nl

Maria De Iorio

Yong Loo Lin School of Medicine, National University of Singapore; Singapore Institute for Clinical Sciences, A*STAR, Singapore e-mail: mdi@nus.edu.sg

Willem van den Boom

Yong Loo Lin School of Medicine, National University of Singapore, e-mail: vandenboom@nus.edu.sg

Alexandros Beskos

Department of Statistical Science, University College London, London, UK; Alan Turing Institute, London, UK, e-mail: a.beskos@ucl.ac.uk

1 Introduction

Bayesian structure learning is a complex problem that consists of retrieving the true structure of conditional dependencies of a set of variables using probabilistic graphical models. This task is based on exploring the graph space which grows exponentially for the number of random variables involved. Exploring the model space in this particular task means searching through an almost infinite space. Thus, there is a need to speed up computations and fast convergence to a region of high posterior probability. The first attempt in trans-dimensional MCMC is the well-known Reversible Jump MCMC (RJCMCMC) [3]. The algorithm proposed by Green is a general sampler in which it is theoretically possible to jump from whatever model to another as long as we can find a good mapping function and efficient proposals. The process uses an acceptance probability that guarantees the convergence to the target distribution and makes the process efficient if reasonably high.

An alternative to the RJCMCMC is the Birth-Death MCMC (BDMCMC) [1, 5, 6]. It is based on a continuous birth and death Markov process. The algorithm jumps to higher (birth) or smaller (death) dimensions with a one-dimensional jump. In contrast with the RJCMCMC, there is no such thing as an acceptance rate because moves between models are always accepted making the process efficient.

In the context of Bayesian structure learning for Gaussian graphical models, [2] is considered the gold standard due to its speed and accuracy. However the BDMCMC proposed in [2] is based on a continuous process. Still, it also incorporates a discrete process to update a precision matrix K (inverse of the variance-covariance matrix of the multivariate Gaussian distribution) and a graph G . In some extreme cases, this algorithm can lead to an incorrect posterior distribution. We propose a discretization of the continuous process needed to obtain correct posterior distribution estimation.

2 Bayesian Gaussian Graphical model

Let $G = (V, E)$ be an undirected graph, where $V = \{1, 2, \dots, p\}$ is the set of nodes and $E \in V \times V$ is the set of existing edges. Let $W = \{(i, j) | i, j \in V, i < j\}$ and $\bar{E} = W \setminus E$ denotes the set of non-existing edges. We define a zero mean Gaussian graphical model with respect to the graph G as:

$$M_G = \{N_p(0, \Sigma) | K = \Sigma^{-1} \in P_G\} \quad (1)$$

where P_G denotes the set of $p \times p$ positive definite matrices with entries (i, j) equal to zero whenever $(i, j) \in \bar{E}$.

Let $x = (x^1, \dots, x^n)$ be an independent and identically distributed sample size n from model M_G , then the likelihood is:

$$P(x|K, G) \propto |K|^{n/2} \exp \left\{ -\frac{1}{2} \text{tr}(KS) \right\} \quad (2)$$

where $S = \mathbf{x}'\mathbf{x}$.

The target of Bayesian inference is the posterior distribution, which is given by :

$$P(G, K|x) \propto P(x|K, G)P(K|G)P(G). \quad (3)$$

Choices for priors on K are G are multiple, the most common being the *G-Wishart distribution* for K and the *Uniform distribution* for G .

3 Proposal

As discussed in Section 1, the BDMCMC fails to retrieve the posterior distribution in some extreme cases due to its mixing nature between a continuous and a discrete processes. This is not impossible as seen in [1] but needs a discretization of the continuous step. Thus, instead of running the continuous process until an edge flips, we propose it to run it for a fixed amount of time. Due to this, the waiting time $w(K)$ should be drawn from an exponential distribution with rate $\sum_{e \in E \cup \bar{E}} r_e(K)$ where $r_e(K)$ indicates birth and death rates.

About birth and death rates, an alternative derivation based on the Cholesky decomposition is provided. As in [4] Let $K^e = \Phi^T \Phi$, then:

$$\delta_e(\Phi) = \min \left\{ 1, \frac{P(\Phi^{-e}, G^{-e}|x)}{P(G, \Phi \setminus \phi_e|x)} \right\}$$

and

$$\beta_e(\Phi^{-e}) = \min \left\{ 1, \frac{P(G, \Phi \setminus \phi_e|x)}{P(\Phi^{-e}, G^{-e}|x)} \right\}$$

where $\Phi^{-e} = \Phi \setminus \phi_e$ and K^e is the precision matrix rearranged in such a way the edge e corresponds to nodes in the last two rows(columns) ϕ_e is the $(p-1, p)$ element. The pseudo-code for the algorithm is reported in Algorithm 1.

4 Results

We provide here an example of the extreme cases mentioned in the introduction. We consider a 4-nodes graph with the following precision matrix:

$$K = \begin{bmatrix} 0.716 & 0.447 & 0.334 & 0 \\ 0.447 & 0.446 & 0.488 & 0 \\ 0.334 & 0.488 & 0.574 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}$$

Algorithm 1 Resulting Algorithm**Require:** G

- 1: Sample an initial $K \sim W_G(b^*, D^*)$
- 2: Run the birth and death process for a fixed amount of time τ . Set $t = 0$ and
- 3: **repeat**
- 4: Compute the birth and death rates

$$r_e(K) = \min \left\{ \frac{I_G(b, D)}{I_G^{-e} N(\Phi^{-e}, D^*)} \right\}^{\varepsilon_e}$$

where $\varepsilon_e = 1$ if $e \in E$ and -1 if $e \in \bar{E}$

- 5: Sample a waiting time Δt from an exponential distribution with rate $r(K) = \sum_{e \in E \cup \bar{E}} r_e(K)$
- 6: Sample an edge e from the probability mass function $P(e) = r_e(K) / r(K)$
- 7: Flip edge e in graph G . That is, add e if it is absent in G and remove e if it is present.
- 8: Update K by sampling $\phi_e = \phi_{p-1, p}$ from its full conditional distribution
- 9: Set $t = t + \Delta t$. If $t < \tau$ go to step 1
- 10: **until** $t < \tau$

In this scenario the two algorithms seem to both have few problems. BDgraph is not able to estimate the precision matrix 60% of time but being accurate when estimating it with an average Frobenius norm of the difference between the real K and $\hat{K}$ of 0.27. The new algorithm instead is always able to produce an estimate of the precision matrix but being inconsistent: on 100 replication only twice is able to go lower the average BDgraph score on the Frobenius norm with an average of 0.86. Both algorithms are set to 10000 MCMC iterations.

The proposed algorithm do also retrieve the following posterior inclusion probabilities:

$$\hat{p}_{ij} = \begin{bmatrix} 0.00 & 0.71 & 0.88 & 0 \\ & 0.00 & 0.60 & 0 \\ & & 0.00 & 0 \\ & & & 0 \end{bmatrix}$$

Although not perfect, the posterior probabilities reflect the true structure of the graph and the algorithm is at least able to estimate it always showing robustness. It is worth to notice that these results are obtained using BDgraph written in C++ while the new algorithm is written fully in R language. When using the R version of BDgraph, it was never able to converge in 10000 MCMC iterations.

We also show here an example where the two algorithms perform in a very similar way: With this being the original precision Matrix from which we originate data:

1.17	0.00	0.00	0.00	0.00	-0.45	0.00	0.00
0.00	1.32	0.00	0.00	0.00	0.64	0.00	0.00
0.00	0.00	1.39	0.00	0.00	0.74	0.00	0.00
0.00	0.00	0.00	2.34	0.00	-1.77	0.00	0.00
0.00	0.00	0.00	0.00	1.84	-1.25	0.00	0.00
-0.45	0.64	0.74	-1.77	-1.25	8.19	-3.32	1.69
0.00	0.00	0.00	0.00	0.00	-3.32	3.86	0.00
0.00	0.00	0.00	0.00	0.00	1.69	0.00	2.26

We can see that that the two precision matrices retrieved by the two algorithms are almost the same:

0.7	1.24	0.01	-0.01	-0.01	-0.01	-0.44	-0.01	0.02
	0.01	1.41	0.00	0.01	0.01	0.63	0.02	0.00
	-0.01	0.00	1.45	0.00	0.01	0.75	0.01	-0.01
	-0.01	0.01	0.00	2.44	0.01	-1.85	-0.02	-0.01
	-0.01	0.01	0.01	0.01	1.93	-1.30	0.00	0.01
	-0.44	0.63	0.75	-1.85	-1.30	8.37	-3.31	1.71
	-0.01	0.02	0.01	-0.02	0.00	-3.31	3.90	0.02
	0.02	0.00	-0.01	-0.01	0.01	1.71	0.02	2.34

Table 1 BDgraph estimate of K

0.7	1.24	0.01	-0.01	-0.01	-0.01	-0.44	-0.01	0.02
	0.01	1.41	0.00	0.01	0.01	0.63	0.02	0.00
	-0.01	0.00	1.45	0.00	0.01	0.75	0.01	-0.01
	-0.01	0.01	0.00	2.43	0.01	-1.85	-0.02	-0.01
	-0.01	0.01	0.01	0.01	1.93	-1.30	0.00	0.01
	-0.44	0.63	0.75	-1.85	-1.30	8.37	-3.30	1.71
	-0.01	0.02	0.01	-0.02	0.00	-3.30	3.90	0.02
	0.02	0.00	-0.01	-0.01	0.01	1.71	0.02	2.34

Table 2 Proposed algorithm estimate of K

This confirms that at the very least the new algorithm is able to match the performance of BDgraph without the impossibility of retrieving an estimate in extreme cases. It must be said that the proposed algorithm is substantially slower than BDgraph. The computational cost, for both algorithm written in the same language, is significantly in favor of BDgraph.

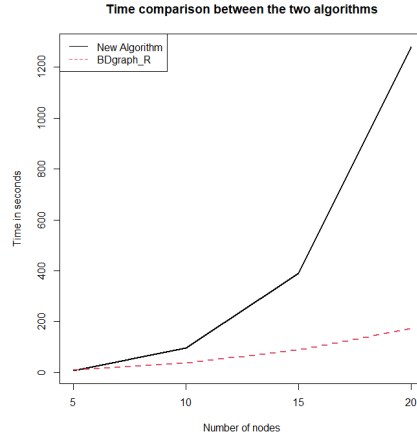


Fig. 1 Time comparison on a Intel i7-11800H CPU

5 Discussion

We presented two algorithms for Bayesian Gaussian Structure Learning which share a common underlying concept and core methodology. However, the proposed algorithm make discrete the continuous process of the BDMCMC to integrate it with every feasible discrete update. We have shown that this approach enables the consistent retrieval of the precision matrix and the associated graph, which would otherwise be unattainable without modification. However, this enhancement comes with the trade-off of significantly slower algorithmic performance.

References

1. Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components: an alternative to reversible jump methods. *The Annals of Statistics* 28 (1), 40{74}.
2. Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1), 109–138.
3. Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4), 711–732.
4. Cheng, Y. and A. Lenkoski (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electronic Journal of Statistics* 6, 2309-2331.
5. Ripley, B. D. (1977). spatial patterns.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 172-212
6. Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes.” *Scandinavian Journal of Statistics*, 359-373.
7. Lauritzen, S. (1996). *Graphical models*, volume 17. Oxford University Press, USA.

12. Contributed - Promoting Equity: Statistical Insights into Tourism, Sustainability and

- 12.1 - Lesson Learnt in the Data Science Worldview: New Dimension of Digital Divide
- 12.2 - An overview of Tourism Statistical Literacy
- 12.3 - Scalable bootstrap inference via averaged Robbins-Monro approximations
- 12.4 - The impact of sustainability on Initial Coin Offering: advantages in trading

Lesson Learnt in the Data Science Worldview: New Dimension of Digital Divide

Rita Lima¹

Abstract In this paper we discuss if it's a great time to be a data scientist entering the job market. The starting point of this discussion is the recent availability of extremely huge datasets (Big Data) in many formats and potentially in real-time, while the new data science techniques constantly being developed, for maximising the value of this data. National Statistical Offices accommodate the widespread needs to shortage qualified human resources with advanced skills in Big Data becoming “data steward” of their new data ecosystems, alongside the traditional statistical production processes. There is also evidence of the influences that this data-ecosystem has on the labor market: changes in educational and learning environments to adapt real-time training programs for the future workplace expectations and to help people stay employable. Theoretical elements on the context of the Big-data literacy, are based on interdisciplinary skills and competences to collaborate with people engaged in Big Data (not only data scientist but also statisticians, mathematicians, machine learning experts, etc.). Beyond the availability of massive raw data and the intention to understand and use them, there is potential for the implementation of policies to resolve the problems associated with the new dimension of digital divide.

Key words: Big Data, Big-Data Literacy, Data Stewardship, Data scientist, Digital Divide

1 Introduction

¹ The opinions expressed in this document are the sole responsibility of the author and do not necessarily represent the official position of the Italian National Institute of Statistics.

The growing demand for different skills and attitudes of data-science-driven work - focusing on statistics and on knowledge creation based on Big Data¹ (BD) – give new big challenges on how the relation of human capital and labour market would be widely recognized. The revolution of data-driven labour market provides opportunities for understandings of data by young people as: - Data scientist personal attributes (curious, problem focused), - Data scientist as ‘bricoleur’, - Data scientist acquisition and retention through challenging work. This is a new key qualification for employability for young people: lacking the attainment of this competence are more likely to be left in dead-end jobs (Bušelić and Zorica, 2018; Bejaković and Mrnjavac, 2020).

The concept of “data scientist (DSst)” was coined in 2008 by D.J. Patil, and Jeff Hammerbacher, who then led their teams of experts of data analysis at LinkedIn and Facebook: according to them, it referred to to best express people who used both data and science “to create something new”. Moving from the statement of the Scheveningen Memorandum on "Big Data and Official Statistics", adopted by the ESSC on 27 September 2013, the Italian National Institute of Statistics (Istat) gave its contribution both to the definition of the skill profile on data scientist and to the design of an introductory course on data science (DS). Using BD effectively requires something that Loukides calls a “holistic approach in gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others”. Anjul Bhambhri, vice president of big data products at IBM, says that “a data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization.” As the data ecosystem evolves, a mix of innovation in all education formats bring a more widely diversified world of education and training options in which various entities design and deliver different services to those who seek to learn. In this category is the idea of Data steward (DSard). While DS is defined as “the ability to handle, process and analyze data to draw insights from it”, and the required skills are: computer science, software development, statistics, data visualization, machine learning, as same as computational infrastructures, DSard is seen as “a set of skills to ensure data are properly managed, shared and preserved, both throughout the research lifecycle and for long-term preservation”. Some examples of its demanded skills are: information management, data cleaning, data management. Only recently National Statistical Offices (NSOs), key experts regarding on handling meaningful information from large amounts of data, have more attention paid to the computer science skills of the statisticians. Although statisticians have proven to be very good in developing and maintaining their quality assurance frameworks, relevant data need to be delivered frequently, in a timely manner, and with much detail as DSsts are available to do. High-quality and trusted statistics are crucial for the NSO to stay relevant. For this reason they play a central

¹ Generally, Big Data refers to large amounts of information that is generated continuously by various means from a quite wide range of social situations, which include web scraping, mobilephone use, online shopping, and banking transactions. For instance, huge datasets are generated by social media users who share information, write their opinions, and contact friends.

role in making more information for expanding the traditional role of statisticians in emerging fields, like DS and/or DSard. Meanwhile they train their staff in this fast-developing field for different aspects of BD related work, the NSOs collaboration with the private sector, academia, and civil society are a new way of working too determine the right mix of statisticians and any of the new categories of jobs. This exponential interest in the DSst suggests that the rich web ecosystem of unstructured data source and BD infrastructures is not ready to be meet by those who need to learn this growing field. Beyond the availability of massive raw data and the intention to utilize them, there needs to be capacity to understand and use data effectively to policy and planning of development programmes.

For this reason, we expect the education marketplace must change in an effort to accommodate these widespread needs.

2 The context of data stewardship in National Statistical Offices

Modernization of national statistical systems is an essential way to remain relevant in a fast-moving data landscape, making progress on methods, tools, and applications of the various BD sources for official statistics. Herewith integrating DS and official statistics, the NSOs leadership on data issues would be the good message upon build co-operatively with all participants involved the implementation of an interdisciplinary no traditional curriculum for the next generation.

It seems clear that the most important thing is building knowledge. Since the gamification activity of educational institutions, academic world and and statistical offices has been designed as an educational project in moving toward BD, an approach to the management of this data (particularly those that can identify individuals) forces in professionalising Data Stewardship (DSship) to apply to NSOs in relation to DS. But it is not so easy to think about DSst given that in many countries the educational institutions and universities do not even know what the means for career development of DS. DSard is the newly emerging role of NSOs, which implies creating order in a complex national ecosystem of data sources, which potentially could be used to the benefit of society. In NSOs the role of a DSard goes, therefore, beyond the current role of a chief statistician, as it includes giving instructions to all kinds of government and nongovernment institutes regarding the handling of their data. Really helpful to know what DSship is for NSOs, a lot of work that has been done already by them¹. It concerns with fiduciary (legal or ethical trust) level of responsibility toward the data, investigating the use of new data sources - including administrative data and BD - and building data science capability for public good (New Zealand Government, 2020; Plotkin, 2021; StatCan, 2021).

In this new data culture, the NSOs are facing several challenges such as: 1. Ensuring data quality, in term of validity and accuracy of their outputs, 2. Building trust and

¹ UNECE, 2022: Task Force on data stewardship. Link: <https://unece.org/statistics/task-force-data-stewardship>.

embedding auditable and transparent data life-cycles; 3. Providing adequate anonymization in respect of data subjects' privacy; 4. Creating data partnerships with BD originators; 5. Monitoring and planning for legislation's changes, in compliance with policies, directives and regulations¹. In addition, as BD are increasingly utilized and integrated into national data and statistical systems (i.e. the experimental statistics) DSship requires a "superset" of analysis capabilities that would be associated with four outcome 'pillars': governance, collaboration, methods and access which in turn are supported by underlying outputs, activities and inputs/needs. In this respect, there is a consensus that the adoption of a DSship approach enables NSOs to evolve from being a statistics producer to becoming a public service provider that facilitates a joined-up approach to data and statistics across different data and statistics communities. The FAIR principles could be the conditions to be taken into account to a responsible DSship in NSOs. FAIR stands for: F - Findable and easily searchable A - Accessible and easy to use I - Interoperable and more easily interpretable R - Re-usable data that is easy to share and use (European Commission, 2018; Wilkinson, 2019). In several countries NSOs already have the formal function of DS: many positions within their structure are de facto DS, fulfilling different tasks and responsibilities necessary to implement DSship. In other countries different approaches and practices that cover certain aspects of DSship are emerging (legal/data protection, statistician/data analyst, ect.)² In summary, all organisations take care of their own data management according to

¹ Regulations are formulated in the European Code of Conduct for Research Integrity (ESF/ALLEA, 2011) and revised in 2017 (ALLEA, 2017). They provide principles on data practices and management as follows (p. 6): "Researchers, research institutions and organisations: • Ensure appropriate stewardship and curation of all data and research materials, including unpublished ones, with secure preservation for a reasonable period. • Ensure access to data is as open as possible, as closed as necessary, and where appropriate in line with the FAIR Principles (Findable, Accessible, Interoperable and Re-usable) for data management. • Provide transparency about how to access or make use of their data and research materials. • Acknowledge data as legitimate and citable products of research. • Ensure that any contracts or agreements relating to research outputs include equitable and fair provision for the management of their use, ownership, and/or their protection under intellectual property rights."

In line with these principles, general data protection regulation rules (EU, 2016) were formulated and are applicable since May 2018. By these regulations, the European Parliament, the Council of the EU and the European Commission intend to strengthen and unify data protection for all individuals; universities work hard to implement the data management plan in the research design to meet the new regulations.

² UN, 2022: UN Statistical Commission, 53:rd session in March 2022, Background document from the UN Working group on Data Stewardship: Preliminary results from the Working Group on Data Stewardship.

best principles; their choices constitute a large part of the DSship, renewing their role as governmental DS.

3 Big-data literacy to address the emerging challenges from Big Data

After looking at some job descriptions of DSst and DSard, offered by research performing organizations, we can notice that new education and training ecosystem emerge in which some job preparation functions are performed by the academic world and educational institutions, dealing with the combination of traditional ways of learning with web-based communication environment.

Given these identified needs, to make FAIR data possible new educational and training programs must be prepared. This only means being Statistical literacy is not the only more important social imperative than ever. Statistical literacy (SL) is “typically more about words than number, more about evidence than about formulas” (Schield, 2004); it is the ability to understand and reason with statistics and data. The abilities to understand and reason with data, or arguments that use data, are necessary for citizens to understand material presented in publications such as newspapers, television, and the Internet. However statistics educators, statisticians and researchers around the world have not reached a consensus on a clear definition of it (English, 2013; Kaplan & Thorpe, 2010; Ridgway, Nicholson, & McCusker, 2011; Callingham & Watson, 2017). Hence numerous definitions of SL abound (Sharma, 2017). Main concern will be to take into account the greatly amplified role that data play in our lives, especially for the massive production and availability of BD.

In fact the massive use of BD and its impact on modern society required an extended definition of SL to address two concerns: that BD leads to the irrelevance of statistical reasoning and that statistics has to take action to remain relevant. The most important skills needed to succeed in this new workplace is flexibility and the ability to adapt and continuously learn.

Recently a form of Big-Data literacy (BDL) has been proposed by Francois, Monteiro and Allo (2020) as an important competence that individuals need to succeed in a data-intensive society; they have considered BDL as a new vocation for SL in line with Ben-Zvi (2017): “understanding big data and its powers and limitations is important to active citizenship and to the prosperity of democratic societies. Today’s students therefore need to learn to work and think with data from an early age, so they are prepared for the data-driven society in which they live”.

Big-data literacy could be a wider concept than both statistical and mathematical literacy based on interdisciplinary skills and competences to collaborate with people engaged in Big data (not only data scientist but also statisticians, mathematicians, machine learning experts, etc.).

An “extended definition of BDL” by D’Ignazio and Bhargava (2015), aiming to educate about “when and where data is being passively collected”, “algorithmic manipulation” and the “ethical impacts” of “data-driven decisions for individuals and for society” (2015, p. 1, p. 3).

Finally an extended critical BDL has developed in recent year that includes citizens “awareness, understanding and critical reflection of BD practices and their risks and implications, as well as the ability to implement this knowledge for a more empowered internet usage” (Sander, 2020).

Despite new findings about how BDL can be implemented have been under-researched, the emergent issues related to SL and BD might be a subject of fervent public and academic debates together with the cooperation of educational institutions, academic world and statistical offices¹ for improving it.

4 Conclusions

So is DS the next generation ability in a data-driven labour market? The answer is not.

The demand for DSsts as ambivalent people with a curriculum that is mainly build upon a omnia comprehensive educational fields (mathematics, statistics, computer sciences, but also sociology, ect.) will definitely narrow, because those people don't exist if not as a team of persons, which sometimes you already have in-house. Outside of the NSO mandate, DSship could potentially be facilitated by BDL; this connection would work very well. As discussed by Karen et al. (2020) it needs to include 1. a technical approach that studies algorithms as computer sciences; 2.a sociological approach that studies the interaction among programmers and designers; 3.a legal approach that studies mathematical algorithms as a figure and as an agent in law; and finally, 4.a philosophical approach that studies the ethics of algorithms. And all that done in an interdisciplinary approach. That's a good approach to catch the more common technical skills for a DSsts to have (data visualization, programming/software, statistics/mathematics). But many doubts hinge upon that the “soft” skills (communication, data-driven problem solving, critical thinking, learning to give constructive feedback, transparency amongst others,ect.), capabilities and attitudes of DSsts could be necessary in future workers are difficult to teach. Moreover the expanded mandate and function of NSOs to be the DSard could convince us that probably any teaching scheme could instill such sophisticated traits in a single worker, although efforts will provide to training.

¹ Istat promotes statistical literacy among individuals and society, including specific stakeholders with particular interest in statistics. Moreover Istat is actively extending it through its participation in national and local festivals or events. An overview of statistical literacy activities can be found at: <https://www.istat.it/it/informazioni-e-servizi/per-studenti-e-docenti>.

Therefore, although BD provides new methods and ideas for knowledge communities, becoming agents of economic change and of innovation, there is a divide between the people who “haves” and “have-nots” skills and competencies to gain new knowledge from handling BD (Rasheva-Yordanova et al. 2018). In fact, at the bottom of the education programs and human capital requirements, there are several shortcomings. Efforts are needed on other factors (e.g. institutions, infrastructure, political stability), for the implementation of policies which could resolve the problems associated with the digital divide in the era of accumulated, available, and accessible BD (Lima, 2018). Such an arrangement creates the opportunity for methodological learning to take place, manage model of social and economic change, allowing for a successful reorganization of positive network, more technologically oriented, particularly involving innovation and sharing information and data.

References

1. Ben-Zvi, D. : Big Data inquiry: Thinking with data. In R. Ferguson et al. (eds.), *Innovating pedagogy 2017. Exploring new forms of teaching, learning and assessment, to guide educators and policy makers* pp. 32–36. Milton Keynes, UK: The Open University (2017)
2. Barocas, S., Hood S. & Ziewitz, M.: Governing algorithms. A provocation piece. In *Proceedings of Governing Algorithms. A conference on computation, automation, and control* pp. 1–12. New York: New York University (2013)
3. Bejaković, P. and Mrnjavac, Ž. : The importance of digital literacy on the labour market, *Employee Relations*, Vol. 42 No. 4, pp. 921-932. (2020)
4. Bušelić V. and Zorica M.B.: Information Literacy Quest: In Search of Graduate Employability. *Information Literacy in the Workplace*, (eds). Serap Kurbanoglu et al. Cham, Switzerland: Springer International Publishing, pp: 98–108. (2018)
5. Callingham R., & Watson J.: The development of statistical literacy at school. *Statistics Education Research Journal* 16(1), pp. 181-201. (2017)
6. English, L.: Promoting statistical literacy through data modelling in the early school years. In E. J. Chernoff & B. S. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* pp. 441–457. New York, NY: Springer. (2013)
7. European Commission ; Turning Fair into reality, doi: 10.2777/1524, (2018)
8. Kaplan, J. J., & Thorpe, J.: Post-secondary and adult statistical literacy: Assessing beyond the classroom. *Data and context in statistics education: Towards an evidence-based society. Proceedings of the eighth international conference on teaching statistics*. Voorburg: International Statistical Institute. (2010)
9. Francois K., Monteiro C. & Allo P. : Big -Data literacy as a new vocation for statistical literacy. *Statistics Education Research*. pp: 194-205. (2020).
10. New Zealand Government: Data stewardship: managing New Zealand’s data better to change lives. A Data Stewardship Framework for NZ. Retrieved from: <https://www.data.govt.nz/assets/Data-stewardship-framework-and-toolkit-Nov2020.pdf>, (2020)
11. Plotkin, D. : *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance* (2nd Ed.). London, UK: Academic Press, (2021)
12. Rasheva-Yordanova K., Veselin Chantov V., Kostadinova I., Iliev E., Petrova P., Nikolova B. : Forming of Data Science Competence for Bridging the Digital Divide, *CONFERENCE PROCEEDINGS 9th International Conference The Future of Education*. Ed.Pixel, (2019)
13. Ridgway, J., Nicholson, J., & McCusker, S. : Developing statistical literacy in students and teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-*

- challenges for teaching and teacher education: A joint ICMI/IASE study, pp. 311–322, (2011) Dordrecht: Springer.
14. Sander, I. :What is critical big data literacy and how can it be implemented?. Internet Policy Review, 9(2). DOI: 10.14763/2020.2.1479. (2020)
 15. Schield M.; Information Literacy, Statistical Literacy and Data Literacy. IASSIST quarterly . International Association for Social Science Information Service and Technology, 28(2): pp:7-14 (2004)
 16. Sharma S. : Definitions and models of statistical literacy: a literature review, Open Review of Educational Research, 4:1, pp: 118-133, (2017)
 17. Statistics Canada. (2021). Statistics Canada's Approach to Data Stewardship [PDF]. Unpublished internal departmental document.
 18. Wilkinson, M. D. :Evaluating FAIR maturity through a scalable, automated, community-governed framework. bioRxiv, doi: 10.1101/649202, (2019)

An overview of Tourism Statistical Literacy

Yasir Jehan, Giuseppina Lo Mascolo, and Stefano De Cantis ¹

Abstract Tourism Statistics Literacy encompasses the ability to comprehend, analyse, and apply statistical data relevant to the tourism industry. This skill is crucial for stakeholders, including government bodies, businesses, and researchers, seeking to harness the vast potential of tourism. It involves interpreting key indicators such as tourist arrivals, accommodation occupancy rates, and economic contributions. Proficient tourism statistics literacy empowers decision-makers to formulate effective strategies, allocate resources efficiently, and respond to evolving market dynamics. With accurate data interpretation, tourism statistics literacy becomes a cornerstone for fostering resilience, innovation, and optimal management within the complex tourism ecosystem.

Key words: Statistical literacy, Tourism literacy, Tourism Statistical Literacy

1. Introduction:

Statistical literacy has become a crucial skill in an era defined by data-driven decision-making across various industries [13]. Defined as the ability to understand, critically evaluate, and interpret statistical information, it involves the capability to navigate through vast collections of datasets, to derive meaningful insights and to decipher patterns. [11, 15]

Scholars emphasize its essential role in informed decision-making, policy formulation, and societal understanding. They argue that a statistically literate population is better equipped to navigate complex issues, critically assess information, and participate meaningfully in a data-driven world. Critics, however, while recognizing the importance of statistics, raise concerns about the complexity of improving widespread statistical knowledge and express alarms about the potential for misinterpretation and manipulation of statistics. They emphasize the importance of teaching not just technical skills but also fostering a

¹ Yasir Jehan, University of Palermo; yasir.jehan@unipa.it

Giuseppina Lo Mascolo, University of Palermo, giuseppina.lomascolo@unipa.it

Stefano De Cantis, University of Palermo; stefano.decantis@unipa.it

broader understanding of context and methodology. The debate underscores the need for a balanced approach, where statistical literacy goes beyond mere number comprehension, incorporating critical thinking and contextual awareness to empower individuals in a data-centric society.

Some of the key contributions in the field of statistical literacy include:

- the importance of teaching statistics contextually, connecting statistical concepts to everyday life to enhance understanding, and with this the important role of the educational institutions (Universities and international organizations such as the OECD - Organization for Economic Cooperation and Development;
- the debate on the need for a broader approach to statistics, integrating critical thinking and contextual understanding into statistical education, and with this the promotion of "Statistical Reasoning" as an integral part of statistical education, highlighting the significance of developing statistical reasoning skills.

In the tourism sector, a specialized form of literacy is essential for stakeholders, ranging from policymakers to businesses, extending beyond traditional statistical interpretation [14]. The integration of tourism and statistical literacy is crucial for formulating effective strategies, responding to the dynamic nature of the tourism ecosystem, and judiciously allocating resources [14]. While Tourism Literacy encompasses an understanding of complexities within the industry [8], Statistical Tourism Literacy involves not only comprehending statistical data but also contextualizing it within the unique narrative of accommodation, travel, and economic contribution [9].

The relationship between statistical literacy and tourism literacy is intricate, with each influencing the other in the dynamic landscape of travel and tourism on one side and data interpretation on the other. Statistical literacy focuses on understanding and interpreting statistical information, while tourism literacy delves into the multifaceted realms of the tourism industry. Their collaborative impact is evident in decision-making, sustainability, and overall comprehension within the context of travel and tourism.

In the realm of data interpretation, statistical literacy becomes a powerful tool for individuals and organizations in the tourism sector. Understanding statistical data, including visitor patterns, economic contributions, and environmental impacts, is crucial for informed decision-making. Statistical literacy enables a nuanced understanding, providing insights that shape strategic planning, marketing efforts, and policy formulation. It also plays a vital role in fostering a culture of responsible and sustainable tourism. Travel decisions influenced by statistical insights contribute to minimizing negative environmental and cultural impacts associated with tourism. Individuals with statistical literacy skills are better equipped to navigate the complexities of sustainable travel, recognizing the importance of balancing economic growth and environmental preservation.

Conversely, tourism literacy contributes to the contextualization of statistical information within the unique framework of the tourism industry. It involves understanding diverse elements shaping travel experiences, including cultural nuances, ethical considerations, and community impacts. The interplay between statistical and tourism literacy emerges in the comprehensive approach needed to grasp the significance of statistical data in the broader context of travel and tourism.

2. Role of Tourism statistical literacy

The tourism industry, vital for global livelihoods and cultural exchange, significantly contributes to national economies. However, its susceptibility to rapid shifts in crises, influenced by socioeconomic development, political stability, and global economic conditions, makes decision-making crucial. Tourism statistical literacy becomes pivotal for stakeholders, ranging from individual firms to destination marketing and management organizations and local to central governments [13].

Understanding variables like tourist behaviour, market trends, and dynamics is essential for identifying emerging patterns and anticipating changes [4]. Tourism statistical literacy aids in strategic resource allocation, enabling decision-makers to maximize operational efficiency by allocating personnel, marketing budgets, or infrastructure investments where they have the most significant impact [13].

The industry involves both supply (transportation, accommodation, and ancillary services) and demand aspects, with close relationships among service providers. Efficient operation requires regulation and structure. Participants can be government entities or private sector representatives, emphasizing international, national, destination-specific, or sector-specific interests. In this intricate industry, statistical literacy is crucial for decision-making to adapt strategies to evolving trends and unforeseen situations.

3. Challenges in Tourism Statistical Literacy

The primacy challenge in tourism statistical literacy is ensuring the quality and reliability of data. The tourism sector generally contains vast volumes of data from different sources such as customer reviews, booking platforms and local tourism boards. Technological dependence introduces many hurdles such as outdated software, system glitches and compatibility issues. These challenges of technological dependence can compromise the quality of data collected and analysed, emphasizing the need of a robust infrastructure of technology in tourism statistical literacy. Koç et al [13] discuss the objective administration of Tourism Statistical Scale, highlighting the significance of technological proficiency, while Baggio [6] supports that it can be supervised scientifically. The integration of this varied and diverse data seamlessly can be complex, asking for standardized approaches to make sure an accurate and holistic approach understanding the landscape of tourism. Ensuring consistency, reliability and accuracy and managing this sheer volume pose significant challenges [13].

On the other side, a certain amount of data is subject to human input, therefore, to minimize errors and ensure data quality, rigorous checks and control must be implemented. Antolini and Grassini [2] argue of information using administrative data, the importance of accurate data collection to overcome human error is emphasized. Finally, Koç et al [13] highlighted anomalies in the statistics due to political changes and international relations, examining the inconsistencies in the data formats and underlining the potential discrepancies in information collected from different platforms.

The significant impact of unreliable and inaccurate data on decision-making is profound. Based on flawed information can lead the strategic decision to sub optimal outcomes, impacting the competitiveness of businesses and destinations. Overcoming these hurdles and challenges requires a strong and concrete effort to enhance data validity and reliability, leverage technology and standardize methodologies for a more robust tourism statistical literacy framework.

4. The Role of Education and Training in Improving Tourism Statistical Literacy

In shaping the landscape of tourism statistics and literacy, international organizations such as Eurostat and United Nations World Tourism Organization (UNWTO) plays a key role providing guidelines, framework and standards that significantly contribute to development of tourism statistical literacy [12].

UNWTO actively takes initiative to promote and enhance tourism statistical literacy capabilities. Its databases, guidelines and publications work as valuable resources for researchers and practitioners [20]. To advancing quality and availability of tourism statistics worldwide the organization is committed to aligns with the outreaching aim of enhancing tourism statistical literacy [7]. Documents such as definitions, guidelines on data collection, and reporting standards emanating from UNWTO serve as indispensable tools for researchers, educator, and practitioners. These documents not only contribute to the uniform development of tourism statistical literacy but also provide a common language in the diverse landscape of global tourism [16].

Eurostat, the statistical office of European Union, contributes to harmonization of statistical practices and standardization in Europe. Eurostat not only contributes to sets benchmarks but also provide publication and methodology for global best practices [19]. The emphasize on comparability and data accuracy further strengthens the basis of statistical literacy in the context of European tourism.

International Tourism Analytics Group: Enhanced Data and Statistical Literacy for Sustainable Tourism Development (i-TAG) project is funded under KA220 Cooperation Partnership in Erasmus+ Higher Education Action with the aim to increase statistical literacy in tourism and hospitality and provide training in the use of tourism analytics for various stakeholders. The partnership is composed by the University of Palermo (leading partner), four other universities: Modul University Vienna GMBH (Austria), Copenhagen Business School (Denmark), The Hebrew University of Jerusalem (Israel) and Universitat Rovira I Virgili (Spain), with a non-governmental organization (NGO), CESIE (Italy). Considering the growing need for tourism professionals to have adequate statistical literacy skills to make informed decisions based on data and analytics, understanding and analysing data is becoming increasingly important in a wide range of industries, including tourism, the objective of the project is to gain this development through the organization of workshops and Boot Camp for students and tourism professionals, the development of a Tourism Analytics Platform with data and tools to support decision making and the organization of Massive Open Online Course (MOOC) which cantered on tourism analytics and statistical literacy. The project addresses the priority of promoting inter-connected

higher education systems by involving universities and research centres from different countries and regions and promoting joint research, training and capacity building activities in tourism and hospitality. For details, visit: www.itag-project.eu.

5. The Role of Education and Training in Improving Tourism Statistical Literacy

To conclude, the current research aims to delve into the intricate tapestry of tourism statistical literacy, unravelling their related associations and highlighting the pivotal role in the tourism industry. Tourism statistical literacy emerges as tool across the industries for the navigation of the decision making in data-driven era, with involving a specific form that excels typical numerical comprehension. Tourism statistical literacy, a nuanced subset, connects understanding of statistical understanding with an exceptional concept of travel, economic contribution, and accommodations [19].

The overview of statistical literacy of tourism emphasizes its importance in strategic planning, decision making, sustainability, fostering innovation and resilience and real time adaptation in the competitive market dynamics. This research draws through an exploration of literature the insights from seminal works, such as those by [18],[10], [2] and others examining the diverse methodology and prospectives while [1], [19] and others related to the big data that intern contribute to the enhancement of tourism statistical literacy.

Hurdles and challenges in Tourism Statistical Literacy are acknowledged, emphasizing the importance for addressing methodological challenges and ensuring the unambiguous definition and procedures. [6],[14] leverage the wisdom, the current research explored the importance of combining big data and conventional data to map global tourism destinations.

The integration of international organizations like Eurostat and UNWTO, along with their relevant documents, further strengthens the discourse. Drawing inspiration from [8] and [9], aligns this research with standardize terminologies and global efforts fostering in-depth analysis and comparability.

As this research navigates the future of tourism, armed with tourism and statistical literacy, the current research contributes to advocating for holistic approach and ongoing dialogue that equips the stakeholders to plan strategically, make more informed decisions and foster innovation within the dynamic industry of tourism.

References

- [1] Adamiak C, Szyda B. Combining conventional statistics and big data to map global tourism destinations before Covid-19. *Journal of Travel Research*. 2022 Nov;61(8):1848-71.
- [2] Antolini, F., & Grassini, L. Issues in tourism statistics: A critical review. *Social Indicators Research*, 150(3), 1021-1042. (2020)

- [3] Aroca P, Brida JG, Volo S. Tourism statistics: Correcting data inadequacy. *Tourism Economics*. 2017 Feb;23(1):99-112.
- [4] Aroca, P., Brida, J. G., & Volo, S. Tourism statistics: Correcting data inadequacy. *Tourism Economics*, 23(1), 99-112. (2017)
- [5] Baggio, R. Improving tourism statistics: merging official records with big data. *Big Data and Business Intelligence in the Travel and Tourism Industry*, Mid Sweden University, Östersund, 89-92. (2016).
- [6] Baggio, R.. Measuring tourism: methods, indicators, and needs. *The future of tourism: Innovation and sustainability*, 255-269. (2019)
- [7] Bianchino A, Fusco D, Giordano P, Liguori MA, Palma MC, Summa D. TOURISM AS SUPPORT IN ECONOMIC DEVELOPMENT OF INNER AREAS: A MULTI-SOURCES APPROACH. (2023).
- [8] Chang KT, Huang CC, Tsaor SH. Tourist geographic literacy and its consequences. *Tourism Management Perspectives*. 2019 Jan 1; 29:131-40.
- [9] Cladera M, Rejón-Guardia F, Vich-i-Martorell GÀ, Juaneda C. Tourism students' attitudes toward statistics. *Journal of Hospitality, Leisure, Sport & Tourism Education*. 2019 Jun 1;24:202-10.
- [10] Eurostat. Methodological manual for tourism statistics. <https://ec.europa.eu/eurostat/documents/3859598/6454997/KS-GQ-14-013-EN-N.pdf#page=127> (2014)
- [11] De Cantis, S., Parroco, A. M., Ferrante, M., & Vaccina, F.. Unobserved tourism. *Annals of Tourism Research*, 50, 1-18. (2015)
- [12] Frechtling, D. C., & Hara, T.. State of the world's tourism statistics and what to do about it. *Tourism Economics*, 22(5), 995-1013. (2016)
- [13] Koç H, Karacabey FA, Demir Yurtseven E. Developing a scale for tourism literacy: validity and reliability study. *Current Issues in Tourism*. 2023 Oct 11:1-5.
- [14] Padrón Ávila, H., & Hernández Martín, R Bibliometric analysis of concepts in tourism: Destination, attraction, accommodatn, and point of interest. *Tourism Science Journal*. (2023).
- [15] Pratt, S., & Tolkach, D. Improvement in tourism statistics: Delineation of political boundaries, misrepresentation of trips, and anomalies. *UNWTO Yearbooks and Compendiums of Tourism Statistics*. (2017).
- [16] REGULATION (EU) No 692/2011 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 July 2011. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R0692&from=EN>
- [17] REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the implementation of Regulation (EU) No 692/2011 of the European Parliament and of the Council concerning European statistics on tourism (2021-2016)
- [18] Van Truong, N., Shimizu, T., Kurihara, T., & Choi, S.. Accommodation statistics: The current issues and an innovation. *Current Issues in Tourism*, 25(11), 1731-1747. (2022)
- [19] Volo, S., & Giambalvo, O. Methodological difficulties in statistical studies in tourism: A qualitative study on hidden tourism. *Tourism Methodologies Journal*. (2008)
- [20] United Nations. Statistical Division. International recommendations for tourism statistics 2008 (No. 83). *Statistical Papers (Ser. M)*. (2010)

Scalable bootstrap inference via averaged Robbins-Monro approximations

Giuseppe Alfonzetti, Ruggero Bellio

Abstract Bootstrap procedures are well-known for providing a straightforward approach to assess the uncertainty around quantities of interest in statistical models. However, with the rising prevalence of massive datasets in statistical problems, the computational cost of bootstrap methods can quickly become prohibitive in applied settings. To this end, this paper proposes the Averaged Robbins-Monro Bootstrap (ARM-B), a scalable tool for estimating parameters variability via multiple chains of Robbins-Monro updates. The simulation experiments show ARM-B to have comparable accuracy with bootstrap while significantly less computationally demanding.

Key words: Bootstrap, Stochastic approximations, Ruppert-Polyak averaging

1 Introduction

Bootstrap methods are powerful techniques to assess the quality of estimators in statistical problems and have been widely spread among practitioners since the early work of [2]. By definition, bootstrap algorithms require refitting the same model many times, which might be extremely demanding on massive datasets, even when parallelising computations across multiple resamples. For this reason, techniques like the ‘Bag of Little Bootstraps’ [3] have been proposed in the literature to enable bootstrap methods scaling in modern statistical applications. In this contribution, we tackle the scaling problem of bootstrap methods from a different perspective. Precisely, by taking advantage of the intrinsic variability of stochastic approximations [5, 4] around their numerical target (i.e. the maximum likelihood estimator), we propose to approximate the variance of the maximum likelihood estimator using multiple independent chains of averaged Robbins-Monro updates, where inputs to

Giuseppe Alfonzetti · Ruggero Bellio
Department of Economics and Statistics, e-mail: giuseppe.alfonzetti@uniud.it

each chain are resampled from the empirical distribution of the data. We denote the proposed method as the Averaged Robbins-Monro Bootstrap (ARM-B).

2 Methods

2.1 Background

Let the data (Y, X) be distributed according to the unknown distribution $\mathbf{F}$. Assume to consider a sample of dimension n of independently and identically distributed observations $(Y_1, X_1), \dots, (Y_n, X_n)$. We denote its empirical distribution with $\mathbb{F}_n$. Assume to model the conditional distribution of Y given X with $p(Y; X, \theta)$, where $\theta \in \mathbb{R}^d$ is the vector of model parameters. Then, the maximum likelihood estimator, $\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MLE}}(\mathbb{F}_n)$, is computed by solving

$$\mathbb{E}_{\mathbb{F}_n} \{ \nabla \log p(Y; X, \hat{\theta}_{\text{MLE}}) \} = 0. \quad (1)$$

Consider now the task of assessing the uncertainty around the estimate $\hat{\theta}_{\text{MLE}}$. A nonparametric bootstrap approach [1] requires drawing from $\mathbb{F}_n$ R independent samples, indexed by $r = 1, \dots, R$, and computing $\hat{\theta}_{\text{MLE}}^{(r)}$ on each of them. Then, the empirical variance of the sample $\hat{\theta}_{\text{MLE}}^{(1)}, \dots, \hat{\theta}_{\text{MLE}}^{(R)}$ converges in probability to $\text{Var}_{\mathbf{F}}(\hat{\theta}_{\text{MLE}})$.

It is well known that, under standard regularity conditions but without assuming $p(Y; X, \theta)$ to be correctly specified, the estimator $\hat{\theta}_{\text{MLE}}$ converges almost surely to θ^* , the parameter vector minimising the Kullback-Leibler distance from the true distribution, i.e. $\hat{\theta}_{\text{MLE}} \xrightarrow[n \rightarrow \infty]{a.s.} \theta^*$, and behaves asymptotically according to $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \xrightarrow[n \rightarrow \infty]{d} N\{0; H^{*-1} J^* H^{*-1}\}$ [9], where $H^* = -\mathbb{E}_{\mathbf{F}}\{\nabla^2 \log p(Y; X, \theta^*)\}$ and $J^* = \text{Var}_{\mathbf{F}}\{\nabla \log p(Y; X, \theta^*)\}$, with the symbol $\xrightarrow{d}$ referring to convergence in distribution. It follows that under the regularity conditions outlined in [9], the variance of the sample $\hat{\theta}_{\text{MLE}}^{(1)}, \dots, \hat{\theta}_{\text{MLE}}^{(R)}$ approximates $n^{-1} H^{*-1} J^* H^{*-1}$ when both n and R are large enough.

Nonetheless, the computation of $\hat{\theta}_{\text{MLE}}$ requires to solve (1) by iteratively re-evaluating $\mathbb{E}_{\mathbb{F}_n} \{ \nabla \log p(Y; X, \theta) \}$, whose computational complexity is $O(nd)$. Thus, obtaining the maximum likelihood estimator becomes more and more demanding as the dimensions of the dataset increase. Consequently, replicating the estimation on R samples often involves prohibitive computational times.

To tackle the computational bottleneck of (1), stochastic optimisation has gained much traction in the last decades since the seminal work of [5], both in the machine learning and statistics communities. Consider the simple Robbins-Monro update [5] where, given a starting point θ_0 and a suitable decreasing scheduling for the stepsize η_t , we iteratively update the estimates via

$$\theta_t = \theta_{t-1} + \eta_t \nabla \log p(Y_t; X_t, \theta_{t-1}), \quad \text{for } t = 1, \dots, \infty, \quad (2)$$

where, at each iteration, a new data point (Y_t, X_t) is sampled from some suitable distribution $\mathbf{P}$. It is apparent that such an updating rule is computationally appealing since it requires evaluating the gradient on a single data point rather than the full sample and thus only costs $O(d)$. It has been shown in [5] that, under suitable conditions for η_t , $\nabla \log p(Y; X, \theta)$ and $\log p(Y; X, \theta)$, the estimate θ_t converges almost surely to the solution of the equation $\mathbb{E}_{\mathbf{P}}\{\nabla \log p(Y; X, \theta)\} = 0$ as t diverges. It follows immediately that, when using $\mathbf{P} \equiv \mathbb{F}_n$, the Robbins-Monro estimate converges to $\hat{\theta}_{\text{MLE}}$.

While the updating rule provided by (2) is computationally very efficient, the values of θ_0 and η_0 play an essential role in practice since they affect the stability of the trajectories and the number of iterations needed to converge.

2.2 Averaged Robbins-Monro Bootstrap (ARM-B)

In the following, we describe the proposed Averaged Robbins-Monro Bootstrap (ARM-B) algorithm, an estimator of the variance of $\hat{\theta}_{\text{MLE}}$ based on multiple independent chains of Robbins-Monro updates.

Assume to compute the estimate $\hat{\theta}_{\text{MLE}}(\mathbb{F}_n)$ according to some numerical method. Starting from $\theta_0 = \hat{\theta}_{\text{MLE}}$, iteratively update R chains of Robbins-Monro updates via

$$\begin{aligned} \theta_t^{(r)} &= \theta_{t-1}^{(r)} + \eta_t \nabla \log p(Y_t^{(r)}; X_t^{(r)}, \theta_{t-1}^{(r)}); \\ (Y_t^{(r)}, X_t^{(r)}) &\stackrel{\text{iid}}{\sim} \mathbb{F}_n \text{ for } r = 1, \dots, R \text{ and } t = 1, \dots, \infty. \end{aligned} \quad (3)$$

Note that, since $(Y_t^{(r)}, X_t^{(r)})$ are independently and identically distributed according to the empirical distribution of the original data, when stopped at a specific t , the sampling component in (3) has drawn R samples of dimension t from $\mathbb{F}_n$, equivalently to a nonparametric t -out-of- n bootstrap procedure. It is generally convenient to set $t = n^\alpha$, with $\alpha > 0$, such that the number of iterations diverges as the sample size increases. The ARM-B estimator for the variance of $\hat{\theta}_{\text{MLE}}(\mathbb{F}_n)$ is given by

$$\hat{V}_{t,R} = \frac{t}{n} \frac{1}{R-1} \sum_{r=1}^R \left\{ \bar{\Delta}_t^{(r)} - \bar{\Delta}_{t,R} \right\} \left\{ \bar{\Delta}_t^{(r)} - \bar{\Delta}_{t,R} \right\}^\top, \quad (4)$$

where $\bar{\Delta}_t^{(r)} = t^{-1} \sum_{s=1}^t \{\theta_s^{(r)} - \hat{\theta}_{\text{MLE}}\}$ and $\bar{\Delta}_{t,R} = R^{-1} \sum_{r=1}^R \bar{\Delta}_t^{(r)}$. Let $\mathcal{F}_t^{(r)}$ be the filtration on the r -th chain at iteration t , and denote the collection of such filtrations with $\mathcal{F}_t^R = \{\mathcal{F}_t^{(1)}, \dots, \mathcal{F}_t^{(R)}\}$. Since the observations $\bar{\Delta}_t^{(1)}, \dots, \bar{\Delta}_t^{(R)}$ are independent and identically distributed conditioned on $\mathcal{F}_t^R$, given the sampling procedure in (3), the estimator in (4) satisfies $n\hat{V}_{t,R}/t \xrightarrow{P} \text{Var}_{\mathbb{F}_n}\{\bar{\Delta}_t | \mathcal{F}_t^R\}$, where $\bar{\Delta}_t$ is a generic observation from $\bar{\Delta}_t^{(1)}, \dots, \bar{\Delta}_t^{(R)}$. The quantity $\bar{\Delta}_t$ is constructed as the average error of the Robbins-Monro procedure, and its statistical properties

are well-known since [6] and [4]. In particular, under suitable assumptions, [4] show that $\sqrt{t}\bar{\Delta}_t = t^{-1/2} \sum_{s=1}^t \hat{H}_n^{-1} \nabla \log p(Y_s; X_s, \hat{\theta}_{\text{MLE}}) + o_p(1)$ when $(Y_s; X_s) \stackrel{\text{iid}}{\sim} \mathbb{F}_n$, and $t \text{Var}_{\mathbb{F}_n} \{\bar{\Delta}_t | \mathcal{F}_t\} \xrightarrow{p} \hat{H}_n^{-1} \hat{J}_n \hat{H}_n^{-1}$, where $\hat{H}_n = -\mathbb{E}_{\mathbb{F}_n} \{\nabla^2 \log p(Y; X, \hat{\theta}_{\text{MLE}})\}$ and $\hat{J}_n = \text{Var}_{\mathbb{F}_n} \{\nabla \log p(Y; X, \hat{\theta}_{\text{MLE}})\}$. Furthermore, by the continuous mapping theorem [8], it is straightforward to show that, since $\hat{\theta}_{\text{MLE}} \xrightarrow{a.s.} \theta^*$ [9], it holds that $\hat{H}_n^{-1} \hat{J}_n \hat{H}_n^{-1} \xrightarrow{a.s.} H^{*-1} J^* H^{*-1}$. Hence, with t and n diverging, the estimator provided by (4) is consistent for the asymptotic variance of $\sqrt{n}\hat{\theta}_{\text{MLE}}$ outlined in [9], namely $n\hat{V}_{t,R} \xrightarrow{p} H^{*-1} J^* H^{*-1}$.

2.3 Implementation remarks

From an implementation point of view, the performance of stochastic algorithms based on (2) is sensitive to the specification of the starting point θ_0 , the stopping criterion used and the scheduling for η_t . First, the choice of θ_0 is not a relevant issue for the ARM-B algorithm, since it is natural to take advantage of the available estimate of the parameter of interest by setting $\theta_0 = \hat{\theta}_{\text{MLE}}$. Second, we let the chains of Robbins-Monro updates run for a fixed number of iteration, namely $t = n^\alpha$, with $\alpha > 0$. This choice is convenient from a theoretical perspective since it allows all convergence statements from the previous section to be expressed as $n \rightarrow \infty$. Furthermore, by fixing $\alpha = 1$, both the variances of $\bar{\Delta}_t$ and $\hat{\theta}_{\text{MLE}}$ converges exactly to the same quantity, i.e. $n^{-1} H^{*-1} J^* H^{*-1}$, without the need for the scaling factor t/n in (4). Finally, the stepsize scheduling is set according to [10], namely $\eta_t = \eta_0(1 + \lambda \eta_0 t)^{-c}$. Following [10], the optimal choice for c with an objective function neither linear nor quadratic is $c = 3/4$. The hyperparameter λ instead accounts for information about the smallest eigenvalue of H^* . Since such value is unknown in practical applications, we fix it arbitrarily low at $\lambda = 10^{-5}$. Unfortunately, no prescription is provided about the hyperparameter η_0 . Thus, we tune it on the data as follows: 1. Given the original dataset, compute $\bar{\Delta}_n$ using $\eta_0 = 1$ and evaluate the likelihood at $\bar{\Delta}_n + \hat{\theta}_{\text{MLE}}$. 2. Halve the value of η_0 , re-compute $\bar{\Delta}_n$ and re-evaluate the likelihood on the original data. 3. Iteratively repeat step 2 until the likelihood stops increasing (or a maximum number of attempts is reached). 4. The hyperparameter η_0^* associated with the maximum likelihood value computed is then used in the ARM-B algorithm.

Note that the tuning procedure is performed just once on the original data. Hence, its computational complexity grows with $O(nd)$. The final cost of the ARM-B procedure becomes $O(nd) + O(Rnd) = O(Rnd)$. It is straightforward to see that ARM-B is less computationally demanding than bootstrap by design. Furthermore, if the number of iterations needed to compute $\hat{\theta}_{\text{MLE}}$ is in the same order of R , the computational time needed for the entire ARM-B procedure is comparable to the one for the original fit $\hat{\theta}_{\text{MLE}}(\mathbb{F}_n)$.

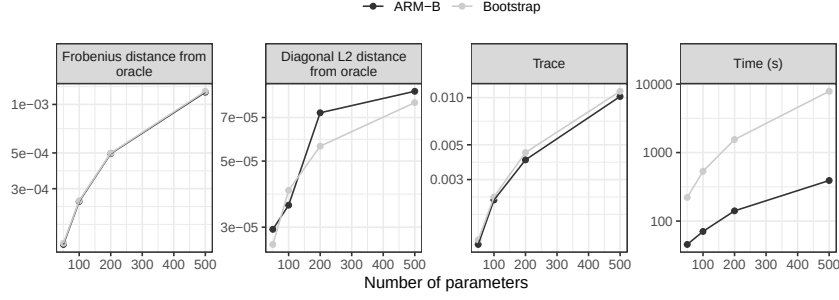


Fig. 1 Performance of ARM-B and plain bootstrap in the estimation of the covariance matrix of $\hat{\theta}_{MLE}$, for $n = 10^5$ and $d \in \{50, 100, 200, 500\}$. Oracle covariance matrix computed by generating samples from the true distribution of the data.

3 Simulation Experiments

To assess the empirical performance of the ARM-B algorithm, let us borrow one of the simulation settings presented in [3]. Namely, assume to fit a logistic regression on $(Y_1, X_1), \dots, (Y_n, X_n)$, where the data are generated as follows. Each component of the vector $X_i = (X_{i1}, \dots, X_{id})^\top$ is drawn from $X_{ij} \stackrel{\text{iid}}{\sim} \text{StudentT}(3)$ for $j = 1, \dots, d$ and $i = 1, \dots, n$, while Y_i is a binary variable with conditional distribution given by $Y_i \sim \text{Bernoulli}[\{1 + \exp(X_i^\top \mathbf{1}_d / \sqrt{d})\}^{-1}]$, where $\mathbf{1}_d$ is the d -dimensional unit vector. We fix $n = 10^5$ and let the number of parameters vary with $d \in \{50, 100, 200, 500\}$. We compare the performance of ARM-B with the conventional nonparametric bootstrap. As an oracle, we evaluate the empirical variance of $\hat{\theta}_{MLE}$ using datasets drawn from the true data generating process. The number of replications is set at $R = 250$ for all procedures. The performance is evaluated according to multiple metrics, namely: *a*) the Frobenius distance of the estimated covariance matrix from the oracle; *b*) The L_2 distance of the diagonal of the estimated covariance matrix from the oracle; *c*) The trace of the estimated covariance matrix; *d*) The computational time needed. Results are reported in Figure 1. While the estimation of the diagonal appears more accurate using the bootstrap, due in particular to ARM-B slightly underestimating the variances, the performance on the overall covariance matrix is slightly in favour of ARM-B. Nevertheless, while differences in accuracy between the two methods are minimal, ARM-B is increasingly more computationally efficient than bootstrap as the number of parameters increases.

4 Discussion and Ongoing Work

The proposed ARM-B algorithm aims to represent a practical inference tool in settings where a numerical estimate of $\hat{\theta}_{MLE}$ is possible, but the dimensions and com-

plexity of the problem prevent the computational viability of traditional procedures for the quantification of statistical uncertainty. We provide an application to logistic regression, but the algorithm described in (3)-(4) can directly be applied to more complex models satisfying the assumptions in [4]. The current software implementation is limited by heavily relying on the R package *sgd*, the companion for [7]. However, we are planning the development of an R package with a more flexible custom implementation. Note that, although currently relying on *sgd*, the use of an implicit version of (3) as proposed in [7] has not been investigated in this contribution because of mere software limitations related to the tuning procedure. Nevertheless, the implicit counterpart of (3) should inherit, in principle, the appealing properties of the implicit SGD, improving the robustness of ARM-B to hyperparameters misspecification and ease the current tuning procedure. Furthermore, it is our intention to benchmark ARM-B performances with methods other than plain bootstrap, like for example the Bag of Little Bootstraps strategy proposed in [3]. Finally, a potential future research direction is the investigation of possible extensions of the ARM-B algorithm to quantify the uncertainty around penalised estimators, allowing the proposed technique to deal with large-scale applications.

Acknowledgments

This research was funded by the European Union - NextGenerationEU (MUR DM funds 737/2021) and the Interdepartmental Project “Digital Governance and Public Administration”, Strategic Plan of the University of Udine (Italy), 2022-2025.

References

1. Davison, A. C. and Hinkley, D. V.: *Bootstrap Methods and their Application*. Cambridge University Press (1997)
2. Efron, B.: Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979)
3. Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I.: A scalable bootstrap for massive data. *J. R. Statist. Soc. B* **76**, 795–816 (2014)
4. Polyak, B. T. and Juditsky, A.B.: Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30**, 838–855 (1992)
5. Robbins, H. and Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**, 407–407 (1951)
6. Ruppert, D.: Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Dept. Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY. (1988)
7. Toulis, P., Airolidi, E.: Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Stat.* **45**, 1694–1727 (2017)
8. van der Vaart, A. W.: *Asymptotic Statistics*. Cambridge University Press (1998)
9. White, H.: Maximum likelihood estimation of misspecified models. *Econometrica*. **50**, 1–25 (1982)
10. Xu, W.: Towards optimal one pass large scale learning with averaged stochastic gradient descent. Preprint. Available at arXiv:1107.2490. (2011)

The impact of sustainability on Initial Coin Offering: advantages in trading

Alessandro Bitetto and Paola Cerchiello

Abstract Initial Coin Offerings (ICOs) have emerged as a novel way of start-up funding based on blockchain technology and this paper aims at explaining the nexus between ICOs success and underpricing, i.e. when the price of the offered token is lower than the one traded in the market. In particular, we focus on the impact of the Environmental, Social and Governance (ESG) pillars onto the ICOs' underpricing. Therefore, we built up a big and comprehensive dataset comprising ~ 8000 ICOs spanning from 2015 through 2023, containing both technical and financial information. Moreover, we assessed an ESG score by means of advanced textual analysis performed over the whitepapers. The main results show that a higher ESG orientation leads to less underpricing, especially in the early trading days.

Key words: Initial coin offering (ICO), ESG, Text analysis, Underpricing

1 Literature Review and Motivation

In recent years digital technology has progressed substantially. Besides artificial intelligence, blockchain technology can be identified as one of the pillars of these advancements. In particular, Initial Coin Offerings (ICOs) are a novel way of gathering capital for young firms and start-ups, that utilise so-called “Tokens”, a kind of crypto asset sold to investors. Afterwards, the former can either trade them on exchanges or use them for buying products and services developed by the firm conducting the ICOs Benedetti and Kostovetsky (2021). The focus of this paper deals with the well-known phenomenon of underpricing, i.e. when “the price jumps substantially on the first day of trading” compared to the price at which it was sold during the offering. This issue per se is well studied in the literature (i.e., Ljungqvist

Alessandro Bitetto, Paola Cerchiello
University of Pavia, Via San Felice al Monastero 5, 27100 Pavia,
e-mail: alessandro.bitetto@unipv.it, paola.cerchiello@unipv.it

(2007)), as it typically occurs in a similar context represented by Initial Public Offerings (IPOs). Authors have found that firms conducting an IPO can be underpriced for various reasons, e.g., country-specific regulations. The question of whether those factors also influence ICO underpricing has been studied in recent years by some authors, together with the question of whether a reduction of information asymmetry reduces ICO underpricing as it does with IPO underpricing (e.g., Benedetti and Kostovetsky (2021)).

Many papers in the emerging field of ICOs literature have focused on the drivers of Initial Coin Offering success (e.g., Belitski and Boreiko (2022), Campino et al. (2022)). They examine which variables, such as characteristics of firms, the time in which the ICO happened or the environments a firm operates in, influence a predefined measure of ICO success. Other authors focused on the contribution of information contained in the so-called 'whitepaper'. In particular, the availability of the paper, its length, the number of figures/tables, the technical jargon and the sentiment towards financial terms have been investigated (Belitski and Boreiko (2022), Campino et al. (2022)).

Literature on the influence of Environmental, Social and Governance (ESG) criteria on ICOs in general is rather scarce. Some papers have dealt with the issue in the context of ICO success. A few papers have dealt with the influence of the ESG orientation of firms on the success of ICOs (Mansouri and Momtaz (2022)). Past literature on ICO underpricing has found that many of the factors influencing IPO underpricing are also relevant in the domain of ICO underpricing (e.g., Benedetti and Kostovetsky (2021)) who found out that measures for information asymmetry influence IPO underpricing. Cao et al. (2021) proves that the average investor takes ESG criteria into account when investing on the stock market and seeks for socially responsible investment options both for the returns they yield and for social responsibility reasons per se.

Our main theoretical contribution relies on the quantitative assessment of how the ESG orientation of a start-up influences underpricing. We hypothesise that ICOs disclosing more information about their ESG orientation (also referring to the specific dimensions E, S and G) in their whitepaper should have, *ceteris paribus*, lower underpricing compared to ICOs that disclose less information - or are less ESG-oriented. We also prove that a higher ESG score also has a negative effect on the longer-term returns of tokens. "Long-term underpricing" refers to the price of a token after 14 or 60 days of trading to its price during the ICOs. Moreover, we have contributed to the debate by building up a novel, large and fine-tuned database on ICOs, containing several verified and ready-to-use information that we can make available on a public repository.

2 Data and Methodology

To achieve robust and reliable results and, given the complexity and heterogeneity of the data sources, we drafted and implemented a protocol for scraping data from the

website `ICOMarks.com`, through Python Selenium. We downloaded 8279 units between 7574 Initial Coin Offerings (ICO), 177 Security Token Offerings (STO) and 528 Initial Exchange Offerings (IEO). The available information includes *ICO details*, such as Website, Whitepaper, Whitelist and MVP, Bounty and Bonus, start/end date, country, *ICO classification*, such as Category (Tech, Finance, Energy, Infrastructure), *ICOMarks rating*, *Token details*, such as Ticker, Platform, Amount available for sale, Technology involved, *Financials*, such as ICO's Token price, (crypto)-currency accepted, Total funds raised, Hard/Soft cap for the funding round, *Team and Advisors size* and *Social Media details*, such as media on which the ICO is advertised or where the investors can discuss. We decided to focus only on ICO and we downloaded all the available whitepapers. Moreover, GDP and the Rule of Law¹ rating are also included as a proxy for a country's economic development and the soundness of its legal system.

By examining each whitepaper, we were able to learn about the company's industry, the number of words, the level of technical and financial expertise and, most importantly, whether it is compliant with the ESG principles or carrying a related business idea. The downloaded whitepapers have been analysed through advanced textual analysis techniques based on Bidirectional Encoder Representations from Transformers (BERT) architecture, to extract information about the characteristics of the proposed business idea. In particular, we use pre-trained models specifically tailored to ESG indicators and financial-related vocabularies Huang et al. (2023). The outcome of the model is a probability score for each classification class, e.g. Environmental, Social, Governance, estimating how pertinent the whitepapers text is with the topic and we created the dummy *ESGFLAG* with value 1 if at least one of the three probabilities (E, S or G) is greater than the probability of non-relevance with the topics. Additionally, we included the logarithm of the number of words in the whitepaper.

The dependent variable is the underpricing after the ICOs listing on the crypto exchange, evaluated as the ratio of the trading price and the ICOs price during the funding round. In particular, we tested different time horizons when evaluating underpricing: the price on the first-day opening and closing, after 14 days and after 60 days.

After screening for available whitepapers and relative ICOs with no missing data for the other variables of interest, we ended up with a database containing 471 ICOs, spanning from 2017 to early 2023. We then fitted a logit model with OLS estimation, taking into account year-quarter, country and category fixed effects, as well as clustering the error by country. We divide our analysis into four scenarios, one for each of the underpricing definitions described above. Table 1 reports the results. Columns (1)-(2) show that a higher ESG score decreases underpricing directly after listing. In particular, increasing the ESG score by 1% leads to a 4% decrease in underpricing when examining both underpricing after opening and after the first-day closing. This indicates that the effect of disclosing the underpricing in the S1 filing, namely a mitigation of information asymmetry and a following reduction in under-

¹ <https://worldjusticeproject.org/rule-of-law-index/>

pricing for US IPOs, found by Fenili and Raimondo (2021), translates to the domain of ICOs and whitepapers. Columns (3)-(4) report the results of the models using the underpricing after 14 days and 60 days as a dependent variable, respectively, and show that the ESG score should decrease the long-term underpricing. Columns (5)-(6) report the models where the logarithm of the underpricing on the 1st day is included as an independent variable while predicting the underpricing after 14 and 60 days. Results show that the underpricing on the listing day (relative to the ICO Price) is negatively impacting long-term underpricing. Indeed, the price increase on the first day persists over a longer timeframe, suggesting that prices do not fluctuate dramatically in the longer term and are largely explained by the happenings on the first day of trading.

Table 1: Predicting underpricing with logit model.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	1st day open	Underpricing 1st day close	14 days	60 days	Underpricing 14 days	60 days
ESGFLAG	-3.912**	-4.053***	-2.314*	-2.545**	-0.214*	-0.507*
log(ICO Price)	-0.153*	-0.152*	-0.151	-0.107*	-0.074***	0.006***
log(nwords)	0.593	0.589	0.582	0.603	0.041	0.016
log(underpricing 1st day closing)					0.915***	0.869**
TeamSize	-0.005*	-0.001	-0.006*	0.042	0.005*	0.005
AdvisorSize	-0.053	-0.048	-0.055	-0.037	-0.003*	0.006
SocialCount	0.0349	0.0392	0.0312	0.062	-0.003	-0.035*
Pre Sale Dummy	-0.225*	-0.182*	-0.216*	-0.186	0.011*	-0.018
ERC20Dummy	-0.319	-0.27	-0.372	-0.292	-0.028	-0.051**
KYCDummy	0.139	0.156	0.176	0.163	0.002	-0.119
Bonus Dummy	-0.169	-0.152	-0.147	-0.144	-0.174*	-0.217*
Bounty Dummy	-0.027*	-0.035	-0.039*	0.009**	0.034	0.088*
GDP	0.123*	0.098	0.227*	0.099	0.67*	0.149
RuleOfLaw	-0.041**	-0.102**	-0.089*	-0.131	-0.21*	-0.131
Observations	471	471	471	471	471	471
Adjusted R^2	0.58	0.67	0.66	0.67	0.59	0.61
Quarter-Year effects				yes		
Country effects				yes		
Category effects				yes		
Clustered Std. Err.				Country		

Notes: Testing different time horizons when evaluating underpricing: the price on the first-day opening and closing, after 14 days and after 60 days. The *, ** and *** symbols denote the p-values at 10th, 5th and 1st significance level, respectively.

3 Conclusion

In this paper, we explore the nexus between the attention of an Initial Coin Offering to Environmental, Social and Governance (ESG) topics and the phenomenon of its underpricing after the funding round. According to our definition, underpricing occurs when the price of the ICO's token is lower than the one traded on the market.

We contribute to the research field by investigating whether attention to ESG topics may reduce underpricing when the start-up deals with ESG topics, especially in

the long run. To this aim, we leverage Natural Language Processing tools, based on recent Artificial Intelligence techniques, to evaluate the pertinence of ICO's whitepaper to ESG subjects and to estimate the relative probability of each single dimension E, S and G. These probabilities are used, in addition to other independent and control variables, to predict the underpricing at different time horizons, i.e. at first-day open and close price and after 14 or 60 days. Underpricing is evaluated as the ratio of the trading price and the ICO price during the funding round.

Our analysis supports the main hypothesis. In line with recent Initial Public Offering (IPO) literature (Fenili and Raimondo, 2021, Ferri et al., 2023) we prove a negative relationship between higher ESG scores and underpricing. This fills the gap between the results found in the IPO underpricing literature and the ICO literature. Our results indicate that start-ups with a higher ESG orientation raising capital through ICOs tend to suffer less from underpricing on the first day of trading compared to lower or non-ESG start-ups, both after the opening and the closing of the trading day.

Further, we found that an increase in the length of the whitepaper goes along with an increase in underpricing, both in the short and longer run and that underpricing at the closing of the first day positively influences long-term returns, both after 14 and 60 days.

However, there are some important limitations of this work, that must be stressed. First of all, the size of the usable sample of ICOs is rather small. Despite the capability of putting in place a protocol for collecting a wide, comprehensive and wide database of ICOs, the main bottleneck consists of the limited availability of whitepapers and trading prices. Such bottlenecks can be mitigated in the future by increasing the quality of data sources (public platforms) and possibly by a legislation that imposes disclosure of whitepapers and other relevant information to ICO issuers.

The importance of ESG topics is undisputed and the awareness around climate change, social justice and regulatory frameworks is constantly gaining relevance to regulators and policy-makers. In this regard, we contribute to the debate by extending the attention from the IPOs underpricing phenomenon to the similar one occurring in ICOs, with a specific emphasis on those undertaking ESG-related actions. Decision-makers need to be fueled with robust and data-driven evidence to influence and shape a still unregulated market, that anyhow attracts many investors. Future research should expand on those findings and test their validity on larger samples including data on more recent and more ESG-oriented ICOs.

References

- Belitski, M. and Boreiko, D., December 2022. Success factors of initial coin offerings. *The Journal of Technology Transfer*, 47(6):1690–1706. doi: 10.1007/s10961-021-09894-.
- Benedetti, H. and Kostovetsky, L., 2021. Digital Tulips? Returns to investors in initial coin offerings. *Journal of Corporate Finance*, 66(C). doi: 10.1016/j.jcorpfin.

- 2020.1.
- Campino, J., Brochado, A., and Álvaro Rosa, December 2022. Initial coin offerings (ICOs): Why do they succeed? *Financial Innovation*, 8(1):1–35. doi: 10.1186/s40854-021-00317-.
- Cao, J., Titman, S., Zhan, X., and Zhang, W. E., 2021. Esg preference, institutional trading, and stock return patterns. *Available at SSRN*. doi: <http://dx.doi.org/10.2139/ssrn.3353623>.
- Fenili, A. and Raimondo, C., 2021. Esg and the pricing of ipos: Does sustainability matter. *Available at SSRN*. doi: <http://dx.doi.org/10.2139/ssrn.3860138>.
- Ferri, S., Tron, A., Colantoni, F., and Savio, R., March 2023. Sustainability Disclosure and IPO Performance: Exploring the Impact of ESG Reporting. *Sustainability*, 15(6):1–15.
- Huang, A. H., Wang, H., and Yang, Y., 2023. Finbert: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2):806–841. doi: <https://doi.org/10.1111/1911-3846.12832>.
- Ljungqvist, A. Chapter 7 - ipo underpricing. In Eckbo, B. E., editor, *Handbook of Empirical Corporate Finance*, Handbooks in Finance, pages 375–422. Elsevier, San Diego, 2007. doi: <https://doi.org/10.1016/B978-0-444-53265-7.50021-4>.
- Mansouri, S. and Momtaz, P. P., 2022. Financing sustainable entrepreneurship: Esg measurement, valuation, and performance. *Journal of Business Venturing*, 37(6): 106258. ISSN 0883-9026. doi: <https://doi.org/10.1016/j.jbusvent.2022.106258>.

13. Contributed - High dimensional and functional data

- 13.1 - Analysis of Brain-Body Physiological Rhythm Using Functional Graphical Models
- 13.2 - A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects
- 13.3 - Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models
- 13.4 - Investigating the association between high school outcomes and university enrolment choice: a machine learning approach

Analysis of Brain–Body Physiological Rhythm Using Functional Graphical Models

Fici R., Augugliaro L. and Wit E.C.

Abstract This paper presents an analysis of physiological data derived from a recent investigation on network physiology, adopting the conceptual framework that views the human organism as a complex network of interacting organs. The study explores coordinated interactions among organs using functional conditional Gaussian Graphical Models (fcGGM). Organ functions are modelled as networks with individual regulatory mechanisms, forming a broader system through continuous interactions. The focus of the analysis is on the interactions within and between two subnetworks: brain activity and a composite network comprising the RR interval of the electrocardiographic waveform, respiration amplitude and blood volume pulse.

Key words: EEG waves, network physiology, Functional Data Analysis, conditional functional Graphical Models

1 Introduction

The paper considers the intricate dynamics of physiological interactions, offering insights into the temporal synchronization of organ functions at the whole organism level, using functional data analysis (FDA). This analysis contributes to exploring new statistical tools to advance our comprehension of physiology interaction, shedding light on the intricate mechanisms that underlie the orchestration of physiological states within the human organism. Adopting a functional framework, the

Rita Fici

University of Palermo, Viale delle Scienze - Building 13, Palermo, e-mail: rita.fici@unipa.it

Luigi Augugliaro

University of Palermo, Viale delle Scienze - Building 13, Palermo, e-mail: luigi.augugliaro@unipa.it

Ernst-Jan Camiel Wit

Università della Svizzera italiana, Via Buffi 13, Lugano, e-mail: wite@usi.ch

vector of records for each variable and each statistical unit represents a discretized observation of a functional variable. Both the response and the explanatory sets of variables constitute two functional processes which take place in an interval of the real line $\mathcal{T} \in \mathbb{R}$ and denoted by $\mathbf{Y}(t)$ and $\mathbf{X}(t)$, where $t \in \mathcal{T}$. This analysis aims to recover the conditional independent structure of brain waves and the effect of some external covariates on the expected function of each brain wave. These relations can be represented by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of variables given by the response and the explanatory functions, and $\mathcal{E}$ represents underacted edges between functions. We apply the conditional functional Gaussian Graphical model proposed by [1], where the cross-covariance operator is considered as in [4] and the functional processes are assumed to have a partially separable covariance operator defined in [7].

2 Analysis

In this section, we first provide details on the analyzed data. Then, in Section 2.2, we offer a brief overview of the applied model and the estimation procedure used. Lastly, in Section 2.3, the estimation results are presented.

2.1 Brain-Body data

The data utilized in this study originate from a recent investigation on network physiology [3]. Within their conceptual framework, the human organism is viewed as a network where coordinated interactions among organs are essential for generating distinct physiological states and maintaining health. Each organ is considered as a network with its own regulatory mechanisms, continuously interacting with other body organs to contribute to the broader system. Physiological interactions occur at multiple levels across the temporal scale to synchronize organ functions at the whole organism level. Our analysis focuses on the interaction within and between two sub-networks; $\mathbf{Y}(t)$ is the network consisting of the α -waves of brain activity, and $\mathbf{X}(t)$ describes the activity of three organs, namely: the time interval between two successive R-waves on the electrocardiographic waveform (RR), respiration amplitude, and blood volume pulse (BVP). Brain activity is recorded using the EPOC PLUS wireless headset, consisting of 14 electrodes that record signals from different brain regions. The ECG (sampled at 250 Hz) and respiration signals (sampled at 25 Hz) are acquired through a sensorized t-shirt, while the E4 wristband captures the BVP signal (sampled at 64 Hz) with a photoplethysmographic (PPG) sensor. Each electrode records the amplitude (α) of EEG waves. As described by [8], the data were collected by monitoring eighteen young, healthy volunteers during a measurement protocol consisting of three experimental conditions corresponding to different levels of mental stress. In this study, we analyze the data corresponding to a resting

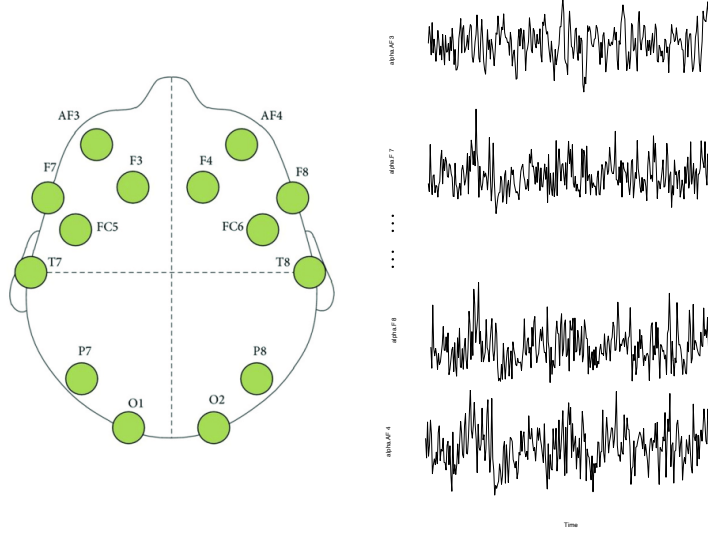


Fig. 1 Left: Graphical representation of the positioning of the 14 EEG electrodes over the scalp.

condition, lasting 12 min and consisting of watching a video showing landscapes with relaxing background music. The data consists of 300 time instants. Figure 1 shows the placement of the EEG electrodes over the scalp and some of the lines designed by the observed 300-dimensional discretized version of the brain waves.

2.2 Functional analysis methods

To recover the edge set representing the relation within and between the two processes, we analyze the relations among the coefficients resulting from the truncated joint vectorized Karhunen-Loève expansion (see [1]):

$$\mathbf{Z}(t) = \sum_{m=1}^M \boldsymbol{\zeta}_m \varphi_m(t)$$

where $\mathbf{Z}(t) = [\mathbf{Y}(t)^\top \mathbf{X}(t)^\top]^\top$, and each element in $\boldsymbol{\zeta}_m$ is equal to $\zeta_{m,i} = \int_{\mathcal{T}} Z_i(t) \varphi_m(t) dt$ where $\{\varphi_m\}_{m=1}^M$ is a system of orthonormal basis. To estimate the basis, we perform the eigen-decomposition on the observed matrices of data (see [2]). We analyze $\{\boldsymbol{\zeta}_m\}_{m=1}^{18}$, which constitute 90% of the total variability. Figure 2 shows the kernel estimations of the densities of the scores for $M = 1$ for some of the variables under examination. The assumption of Gaussianity underlying the fcMMs appears quite appropriate for the scores of the explanatory variables. However, an observed asymmetry is noted in the scores of the response variables. To recover the edge-set, we

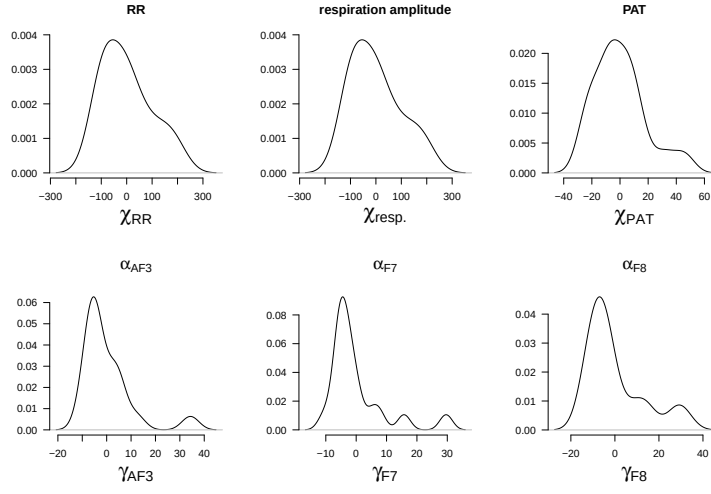


Fig. 2 Estimated densities for the scores of the explanatory variables and some of the response variables.

use the double-penalized joint group lasso estimator proposed by [5]. The optimization problem is performed iteratively, fixing one of the two tuning parameters and varying the other one with 20 values. The tuning parameter value for the next iteration is selected using the joint version of Kullback-Leibler cross-validation (KLCV, [6]) proposed in [1]. The total number of iterations is 3.

2.3 Results

Figure 3 shows the selected links representing the response conditional independence structure. The right side of the scalp exhibits a higher number of links compared to the left side. The most connected alpha wave is $P8$ with 12 links. The waves $T8$ and $F8$ have 11 connection each. Table 1 provides details on the joint KLCV (jKLCV) regressors selection, indicating which explanatory variables have impacts on each of the response variables. The RR result affects all the α waves. The respi-

	AF3	F7	F3	FC5	T7	P7	O1	O2	P8	T8	FC6	F4	F8	AF4
RR	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Resp	1	1	1	1	1	1	1	1	0	1	1	1	1	1
PAT	0	0	0	1	1	1	0	0	0	0	0	0	0	0

Table 1 jKLCV: regressors selection as results of jKLCV selection method.

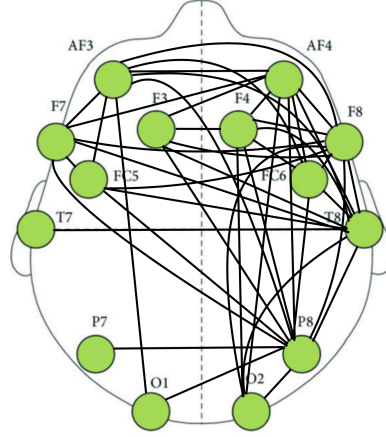


Fig. 3 jKLCV: conditional independent response network.

ration amplitude affects all the response waves except $P8$. Lastly, the blood volume pulse affects the α waves of $FC5$, $T7$ and $P7$ sensors.

3 Conclusion

The present work aims to contribute to the refinement and development of methods that allow addressing questions related to physiology by adopting a functional point of view.

This study delves into the intricate dynamics of physiological interactions, employing functional data analysis (FDA) to explore the temporal synchronization of organ functions at the whole organism level. By applying the conditional functional Gaussian Graphical model, we aimed to uncover the conditional independence structure of brain waves and the influence of external covariates. The adopted methodology includes the truncated joint vectorized Karhunen-Loève expansion and the double-penalized joint group lasso estimator. Those tools allow us to recover the conditional independence structure and identify influential factors on brain wave activities. The results highlighted the significance of the RR interval and respiration amplitude in shaping the expected value of almost all the α brain waves. The blood volume pulse has a significant effect on some of the sensors.

References

1. Fici, R., Augugliaro L., Wit E.C.: Functional Conditional Gaussian Graphical Models. (2024) doi: 10.48550/arXiv.2401.10196
2. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons. (2015) doi: 10.1002/9781118762547
3. Pernice R, Antonacci Y, Zanetti M, Busacca A, Marinazzo D, Faes L, Nollo G.: Multivariate Correlation Measures Reveal Structure and Strength of Brain-Body Physiological Networks at Rest and During Mental Stress. *Front Neurosci. Frontiers in Neuroscience*. **14** (2021) doi: 10.3389/fnins.2020.602584
4. Qiao, X., Guo, S., James, G.M.: Functional Graphical Models. *Journal of the American Statistical Association*. **114** (525) pp. 211-222 (2019) doi: 10.1080/01621459.2017.1390466
5. Sottile G., Augugliaro L., Vinciotti V, Arancio W., Coronnello C.: Sparse inference of the human hematopoietic system from heterogeneous and partially observed genomic data. (2022) doi: 10.48550/arXiv.2206.09863
6. Vujačić, I., Abbruzzo, A., Wit E.C.: A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *Journal of Statistical Computation and Simulation*. **85** (18) pp. 3628-3640 (2015) doi: 10.1080/00949655.2014.992020
7. Zapata J., Oh S.Y., Petersen A.: Partial separability and functional graphical models for multivariate Gaussian processes. Oxford University Press (OUP). **109**(3) pp. 665–681 (2021) doi: 10.1093/biomet/asab046
8. Zanetti, M., Mizumoto, T., Faes L., Fornaser, A., Cecco, M., Maule, L. Valente, M., Nollo, G.: Multilevel assessment of mental stress via network physiology paradigm using consumer wearable devices. *Journal of Ambient Intelligence and Humanized Computing*. **12** pp. 4409–4418 (2021) doi: 10.3389/fnins.2020.602584

A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects

Ruggero Bellio and Cristiano Varin

Abstract Parameter estimation of generalized linear models with crossed random effects for large-scale settings is hampered by challenging numerical hindrances. This contribution focuses on logistic regression with crossed-random intercepts and it investigates the properties of two estimation methods for which a scalable software implementation exists, namely the all-row-column and penalized quasi-likelihood methods. The results of a simulation study for sparse settings inspired by e-commerce data, with sample sizes up to 10^6 , suggest that the all-row-column method is preferable over penalized quasi-likelihood.

Key words: all-row-column estimation, e-commerce, penalized quasi-likelihood.

1 Introduction

Generalized linear models with crossed random effects arise in a variety of settings such as Agriculture, Biology, Epidemiology, Education, Genetics, and Linguistics. Here we are mostly concerned with the analysis of user-generated content arising in e-commerce or recommender systems, where the typical sample sizes largely exceed what is normally encountered in other applications of crossed random effects.

An illustrative example of the setting of interest is reported in Bellio et al. [1] where it is analysed a sample of data from an online personal styling service. The data include a binary response variable y_{ij} that denotes whether customer i rates item j as top fit or not. In the sample data there are around 5,000,000 user ratings,

Ruggero Bellio

University of Udine (Italy), Department of Economics and Statistics, e-mail: ruggero.bellio@uniud.it

Cristiano Varin

Ca' Foscari University of Venice, Department of Environmental Sciences, Informatics and Statistics. e-mail: cristiano.varin@unive.it

produced by more than 700,000 customers for about more than 3,000 clothes. A reasonable statistical model for this data assumes a generalized linear model for binary responses with linear predictor

$$\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + a_i + b_j \quad (1)$$

where $i = 1, \dots, R$ is the index for customers, $j = 1, \dots, C$ is the index for clothes, $\mathbf{x}$ is a vector of covariates with coefficients $\boldsymbol{\beta}$ and, finally, a_i and b_j are mutually independent normally-distributed random intercepts for customers and clothes with variances σ_A^2 and σ_B^2 , respectively.

The estimation of a generalized linear model for binary responses with crossed random effects is notoriously challenging. The corresponding likelihood function requires to integrate out the random effects, entailing an integral of dimension $R+C$, which is very hard to approximate in broad generality, and even more so when R or C is very large as in e-commerce applications. In such settings, methods requiring Monte Carlo approximations are readily ruled out, since they usually imply a computational cost exceeding $O(N)$, where N is the sample size. Some scalable alternatives to Monte Carlo methods have been developed recently:

1. the pseudo quasi-likelihood method [2, 3] using the scalable algorithm proposed in Ghosh et al. [4];
2. the all-row-col method proposed by Bellio et al. [1];
3. the combination of Gaussian variational approximation and composite likelihood proposed in Xu et al. [5].

While there is software available for the first two methods, there is not yet any software provided for the variational approximation of Xu et al. Moreover, Xu et al. do not give any example of application of their variational approximation for large-scale problems.

We have not included in the list of methods for inference in crossed-random effects the popular first-order Laplace's approximation implemented in very efficient R packages such as `lme4` [6]. Indeed, the results in [1] show that such method may have a computational cost exceeding $O(N)$ in sparse settings, coupled with some possible (minor) estimation bias arising even for very large sample sizes. The latter is far less critical than the computational issue, which rules out this method for the large datasets met in e-commerce applications. Therefore, in this short paper we compare penalized quasi-likelihood and all-row-call for logistic regression with crossed random effects, since they are both scalable methods with public software available.

2 Two scalable estimation methods

Penalized quasi-likelihood is a well-established estimation method, which consists in maximising an objective function corresponding to a simplified version of Laplace's approximation [2]. The method entails the iterated resolution of two kinds

of estimating equations, one which jointly updates fixed effects and random effects, while the other one updates the variance components. The method is consistent only when both the number of random effects and the number of observations per random effect increase, and for binary data substantial finite-sample bias may arise. The method is rather general, and can be applied to virtually any mixed-effect model.

The application of the method to large-dimensional settings illustrated in [4] requires some ingenuity to handle the large number of random effects. Indeed, the authors proposed a modified version of Schall's estimation approach [3], making use of an ad-hoc version of backfitting to keep the computational complexity within the $O(N)$ bound. The code supporting [4] is for logistic regression only.

The all-row-column method was recently proposed by Bellio et al. [1] for probit regression with crossed-random effects. The method exploits the fact that the likelihood for the fixed effects in the marginalized model has a simple form in the case of probit link. The all-row-column method proceeds in three steps:

1. the *all step*. The starting point is the marginal regression model

$$Pr(Y_{ij} = 1) = \Phi(\mathbf{x}_{ij}^\top \boldsymbol{\gamma}), \quad (2)$$

where $\boldsymbol{\gamma} = \boldsymbol{\beta} / (1 + \sigma_A^2 + \sigma_B^2)^{1/2}$. The $\boldsymbol{\gamma}$ coefficients are estimated by fitting a probit regression model for independent data;

2. the *row step*. A probit regression model with only the a_i (the 'rows') effects is considered

$$Pr(Y_{ij} = 1 | a_i) = \Phi(\mathbf{x}_{ij}^\top \boldsymbol{\gamma}_A + u_i), \quad (3)$$

where $u_i = a_i / (1 + \sigma_B^2)^{1/2} \sim \mathcal{N}(0, \tau_A^2)$, with $\tau_A^2 = \sigma_A^2 / (1 + \sigma_B^2)$, $\boldsymbol{\gamma}_A = \boldsymbol{\gamma} (1 + \tau_A^2)^{1/2}$. The parameter τ_A^2 is estimated by fitting a random intercept model including only the u_i effects with $\boldsymbol{\gamma}$ held fixed at the value $\hat{\boldsymbol{\gamma}}$ estimated at the all step;

3. the *col step*. A probit regression model with only the b_j (the 'columns') effects is considered

$$Pr(Y_{ij} = 1 | b_j) = \Phi(\mathbf{x}_{ij}^\top \boldsymbol{\gamma}_B + v_j), \quad (4)$$

where $v_j = b_j / (1 + \sigma_A^2)^{1/2} \sim \mathcal{N}(0, \tau_B^2)$, with $\tau_B^2 = \sigma_B^2 / (1 + \sigma_A^2)$, $\boldsymbol{\gamma}_B = \boldsymbol{\gamma} (1 + \tau_B^2)^{1/2}$. The parameter τ_B^2 is estimated by fitting a random intercept model including only the v_j effects with again $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$. The estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ are obtained from those of τ_A^2 and τ_B^2 . Finally, the estimate of $\boldsymbol{\beta}$ is computed as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\gamma}} (1 + \hat{\sigma}_A^2 + \hat{\sigma}_B^2)^{1/2}$.

Bellio et al. [1] proved that method is consistent and with a computational complexity of order $O(N)$. To proceed with a direct comparison of all-row-column with penalized quasi-likelihood, however, one needs either to implement the probit version of penalized quasi-likelihood, or to adapt the all-row-column method to logistic regression. In this short paper, we consider the latter route because we think the logit link may be of interest to practitioners.

In principle, the extension of all-row-column to logit link may proceed by following the results about marginalized models (see for example [7]), but the resulting

numerical burden might be demanding, since N one-dimensional equations involving a convolution integral would have to be solved numerically for every evaluation of the marginalized likelihoods obtained by omitting one of the two random effects. Even though Bellio et al. [1] conjectures that the resulting computation complexity may be still within the $O(N)$ bound, such a route appears impractical at best, so here we suggest an approximate method.

Like before, let γ be the coefficients of the marginalized model, namely the logit counterpart of the probit marginalized model (2). Even though there is no exact explicit relation linking γ to the parameters of the linear predictor (1), some approximate relations exist. For example, one may follow Wang and Luis [8], obtaining

$$\gamma \doteq \frac{\beta}{\sqrt{1 + (3/\pi^2)(\sigma_A^2 + \sigma_B^2)}}. \quad (5)$$

Similar relationships can be defined for the models omitting one of the two random effects that are used in the row and column steps of the all-row-column method, in this way allowing to extend the all-row-column method to logistic regression.

3 Some simulation results

Some simulations following closely the studies in Ghosh et al. [4] and Bellio et al. [1] have been carried out. We considered sample sizes N ranging from 10^3 to 10^6 and generated 1,000 datasets from model (1) with logit link at each sample size. We simulated seven standard normal correlated predictors and set the intercept to $\beta_0 = -2$ and the other regression coefficients to $\beta_\ell = -2 + 0.5\ell$ for $\ell = 1, \dots, 7$. For the random effects, we considered a sparse design with the number of random effects given by $R = N^{0.88}$ and $C = N^{0.53}$. This is a much more demanding setting than that considered in Ghosh [4] to investigate the properties of their penalized quasi-likelihood estimator. Finally, the random effects standard deviations were set to $\sigma_A = 0.8$ and $\sigma_B = 0.4$, two realistic values for applications in e-commerce. Besides penalized quasi-likelihood and all-row-column, also an *approximate oracle* method was considered, where the $\hat{\gamma}$ coefficients estimated by a standard logistic regression were adjusted by inverting the relationship (5) using the true values of the variance components.

The analysis of the average estimation times (not reported here) confirmed that the two methods scale to $O(N)$. The statistical performances instead show some marked differences. Figure 1 displays the scatterplots of the $\log_{10}$ of the mean squared errors against $\log_{10} N$ for the three methods under comparison. The scatterplots are reported for the intercept, one of the β coefficients and the two standard deviations σ_A and σ_B . Results for the other six β coefficients are similar to those of the β coefficient show in Figure 1. Below we summarize the interesting findings we obtained from this simulation study:

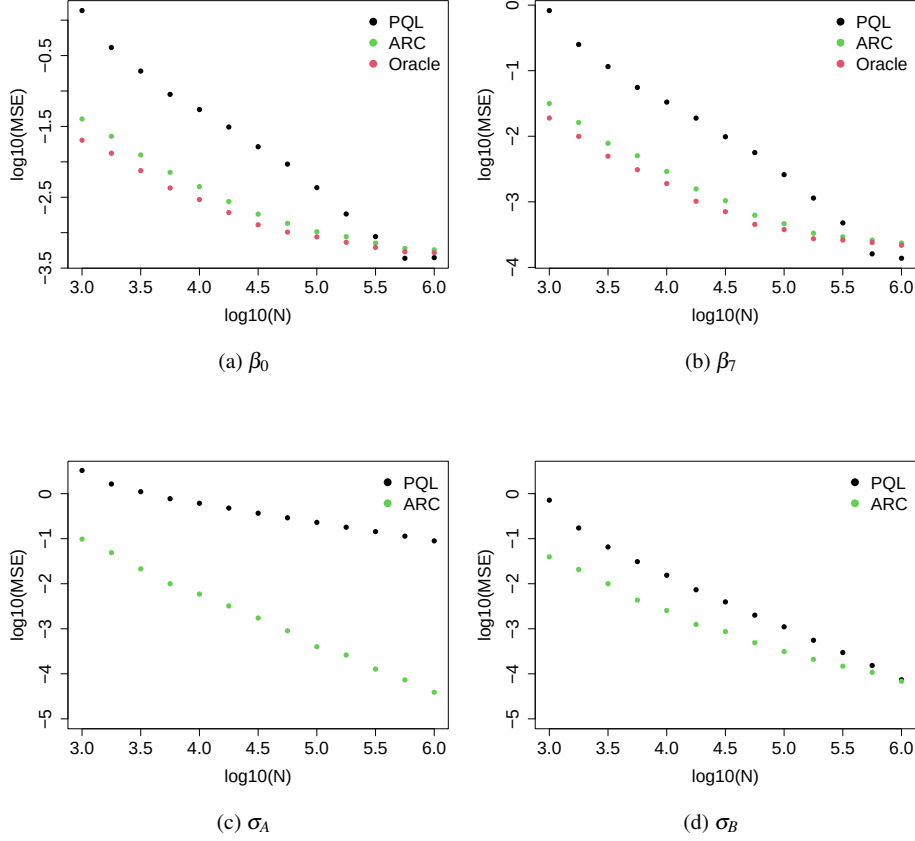


Fig. 1: Simulation study. Scatterplots of $\log_{10}$ of the mean squared errors against $\log_{10}N$ for the penalized quasi-likelihood (PQL), all-row-column (ARC) and oracle estimators of four selected parameters.

- the all-row-column and oracle methods provide rather similar results for the β coefficients. Indeed, all-row-column estimates well the variance components that are needed to rescale the marginal estimates and thus it can match the oracle results to a quite good extent;
- penalized quasi-likelihood gives estimates of β coefficients with much higher mean square error than all-row-column, except for very large sample sizes, where the approximate nature of relation (5) kicks in, resulting in some asymptotic bias for the all-row-column and oracle estimators. From other plots (not shown here), it is also apparent that the all-row-column bias is however always small in size;

- penalized quasi-likelihood exhibits a very large bias and variance in the estimation of σ_A . This bias follows from the fact that with $R = N^{0.88}$, each a_i is estimated using very few observations; indeed, on the average, $N/R = N^{0.12} = 3.98$ for $N = 10^5$ and $N^{0.12} = 5.25$ for $N = 10^6$. Things are better for estimation of σ_B , since each b_j is estimated using larger groups.

Some further simulations were carried out for another setting with much larger variances. As expected the results were worse for all the estimators, but with a pattern quite similar to that discussed above. The mean square error of the all-row-column estimator for the β coefficients was higher than that of the penalized quasi-likelihood for $N > 10^5$. This is not surprising since (5) is valid under *small-sigma asymptotics*. Yet all-row-column was still preferable to penalized quasi-likelihood for the estimation of variance components.

4 Conclusions and ongoing research

The results of this note support the use of the all-row-column estimator also for the logit case. Despite its approximate nature, all-row-column seems preferable to penalized quasi-likelihood for practical use. Ongoing work will focus on the extension of the all-row-column methodology to crossed random effects with random slopes, which may be valuable for large-scale analyses of user-generated content.

Acknowledgements Ruggero Bellio was supported by the research project *Latent Variable Models for Complex Data* funded by the European Union - NextGenerationEU (MUR DM funds 737/2021).

References

1. Bellio, R., Ghosh, S., Owen, A.B., Varin, V.: Scalable estimation of probit models with crossed random effects. arXiv:2308.15681 (2023). <https://arxiv.org/abs/2308.15681>
2. Breslow, N.E., Clayton, D.G: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25 (1993).
3. Schall, R.: Estimation in generalized linear models with random effects. Biometrika 78, 719–727 (1991).
4. Ghosh, S., Hastie, T., Owen, A.B.: Scalable logistic regression with crossed random effects. Electron. J. Stat. 16, 4604–4635 (2022).
5. Xu, L., Reid, N., Kong, D.: Gaussian variational approximation with composite likelihood for crossed random effect models. arXiv:2310.12485 (2023). <https://arxiv.org/abs/2310.12485>
6. Bates, D., Maechler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using `lme4`. J. Stat. Softw. 67, 1–48 (2015).
7. Heagerty, P.J.: Marginally specified logistic-normal models for longitudinal binary data. Biometrics 55, 688–698 (1999).
8. Wang, Z., Louis, T.A.: Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. Biometrika 90, 765–775 (2003).

Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models

A. Gagliano, G. Sottile, N. Sciaraffa, C. Coronello, L. Augugliaro

Abstract In the past decade, advances in single-cell RNA sequencing technologies have radically improved the comprehension of cell biology. Pseudo-time computation has revealed greater heterogeneity in cell differentiation, which is important for the study of various diseases. In this paper we evaluate pseudo-time calculation methodologies and propose statistical models for analyzing the relationship between pseudo-time and gene expression.

Key words: Single-Cell, Pseudo-time, B-spline, Quantile Regression

1 Introduction

Thanks to the introduction of the new technology *scRNA-seq* (Single Cell RNA Sequencing), it is possible to study cells individually, reaching a level of detail and heterogeneity never seen before. One of the problems in dealing with these data is to understand the conditional dependence structure between gene expression and cell time because of the huge number of zeros. To solve this problem, two key issues need to be addressed: the first is to derive a time measure from the data under consideration, and the second is to define statistical models suitable for analyzing the trajectory of gene expression along the obtained time path. The first step is achieved exploiting that cells are not static but follow a developmental process along trajectories, and this process is strongly correlated with the maturity state of the cell, which allows the gene expression data used to obtain pseudo-time.

Once pseudo-time trajectories are reconstructed, in the second step it will be possible to analyze the relationship between gene expression and pseudo-time to

A. Gagliano, G. Sottile, L. Augugliaro, University of Palermo, Department of Economics, Business and Statistics, Viale delle Scienze 8, Palermo, e-mail: antonino.gagliano@unipa.it
N. Sciaraffa, C. Coronello, Fondazione Ri.MED, Via Bandiera 11, Palermo

understand the trajectory of differentially expressed genes. In order to do this a penalized nonparametric statistical models are applied, highlighting the difficulties in implementing traditional methods because of the peculiarity of scRNA-seq data to lean toward very low expression values. We will see an application in healthy and myelofibrosis-affected cells dataset.

2 Dataset description

The analyzed dataset was obtained by using the 10x genomic platform to profile the single-cell transcriptome of hematopoietic stem/progenitor cells (HSPCs) CD34+Lineage derived from 16 human myelofibrosis tissue samples and 5 healthy donors. Specifically, the data are composed by a count matrix of 33538 rows, i.e., genes, and 122154 columns, i.e., HSPCs cells with the ability to differentiate into a variety of mature blood cells. TARGET-seq single cell RNA-sequencing data is available at Accession Number GEO: GSE122198 [1].

3 Methods

The methodology will follow the two phases structure at Results. The first phase concerns a review of the three main techniques for computing pseudo-time (TSCAN, Slingshot and Monocle 3), the second one analyzes the conditional dependence structure of gene expression (response variable) respect to the obtained pseudo-time (explanatory variable) using penalized B-spline and quantile regression.

3.1 Pseudo-time calculation

TSCAN. TSCAN (*Tools for Single Cell Analysis*) is structured in three key steps. The first step is to group cells into clusters, the second step is to construct the Minimum Spanning Tree (MST) to connect the centers of each cluster and to establish their order in the differentiation process, and finally, the third step is to project cells along the tree structure to establish their pseudo-temporal value [2].

Slingshot. Slingshot in the first step is similar to TSCAN: it identifies lineages by MST, but after that the principal curve approach is used. Specifically, if the MST identifies only one lineage, then that lineage is defined by the standard principal curve, but if the MST identifies multiple lineages then the simultaneous principal curve method is introduced. Once the set of principal curves is defined, the cells are projected onto each curve, and the pseudo-time will be assigned based on the distance from the root ($t = 0$) [3].

Monocle 3. Monocle 3 defines cellular trajectories based on the relationship between individual cells. The first step is to define partitions between cells through the PAGA algorithm. This process begins with the identification of clusters through k-nearest neighbors (KNN) and then communities through Louvain algorithm, such

that the clusters obtained represent nodes in the PAGA. These nodes are connected only if the clusters are close in the KNN graph more frequently than would be predicted with a simple binomial model: these connections determine the partitions of cells in the data. The second step is to construct a principal graph via MST on reference cells found by K-Means. The third step is to extend the principal graph to the relationship between individual cells: cells are projected onto the nearest arc of the principal graph and, based on the order they have in that arc, they are connected, allowing cyclic trends to be determined as well. Finally, the pseudo-time is calculated based on the geodetic distance of the cell from the initial point of the trajectory [4].

3.2 Nonparametric models

B-spline. B-spline is a very flexible nonparametric approach to fit a piecewise polynomial regression. This is possible by setting different nodes, i.e different bases, that define a local fit. These bases, can be denoted as $B_i(x)$ and constitute the columns of the model matrix, which will take non-zero value only at the nodes where they are defined. The final curve will be $\mu(x) = \sum_j B_j(x_i, p) \beta_j$ and the estimation of the β_j coefficients is done by minimizing the following objective function:

$$\mathcal{L}_\lambda(\beta) = \sum_i (y_i - \sum_j B_j(x_i, p) \beta_j)^2 + \lambda \sum_j (\Delta^d \beta_j)^2 \quad (1)$$

where $\lambda_j (\Delta^d \hat{\beta}_j)^2$ represents the *Penalty Function*, i.e the sum of the differences of order d between the β coefficients. The hyperparameter λ is estimated through the *Harville, Fellner and Schall* (HFS) algorithm.

Quantile Regression. Quantile Regression, estimated through penalized B-spline, allows the focus on the conditional quantiles of the distribution of Y (gene expression) given X (pseudo-time). The final curve will be given by $Q_Y(\tau | x) = \sum_j B_{\tau j}(x_i; p) \hat{\beta}_{\tau j}$. Now, the penalty function $\sum_j \lambda_j \|\Delta^d \beta_j\|_1$ will be estimated for each quantile and through a L_1 penalty on the sum of the absolute value differences of the β coefficients:

$$\mathcal{L}_\lambda(\beta) = \sum_i \rho_\tau(y_i - \sum_j B_{\tau j}(x_i; p) \beta_{\tau j}) + \sum_j \lambda_j \|\Delta^d \beta_j\|_1 \quad (2)$$

where ρ_τ represents the *check function*. In addition, the optimal value of the hyperparameter vector λ_j will be estimated by *K-Fold Cross Validation*.

4 Results

Following a meticulous data preparation phase involving quality check, normalization, batch effect correction, and selection of the most variable genes, the dimensionality reduction was performed (via PCA and UMAP) with the purpose of clustering and annotating cells in the data. The cell types detected are *Hematopoi-*

etic Stem Cells (HSCs), *Common Myeloid Progenitors* (CMPs), *Megakaryocyte-Erythroid Progenitors* (MEPs), *Megakaryocyte* and *Erythroid*.

The aim is to explore the genetic trajectory of genes that manifest different expression levels between healthy and myelofibrosis subjects. The analysis is developed in two key phases, in the first phase, pseudo-time per single cell will be determined, in the second one we will focus on using this pseudo-time to define the genetic evolution through non parametric models appropriate for single cell expression data.

Phase 1. The three pseudo-time algorithms in Figure 1 reveal the importance of choosing a trajectory based on the type of data being observed and the expectation of cell differentiation pathways. TSCAN, is the simplest algorithm of the three, because it identifies a linear trajectory designated by two different pathways, one terminating on Megakaryocytes (1a) and one on Erythroids (1b), with the ability to separate MEPs cell progeny for the two pathways. The trajectories obtained with Slingshot show a path similar to TSCAN, but with a nonlinear fit that can more consistently reflect the cellular behaviour of the data at the two identified Lineages (1c and 1d). Monocle 3 (Figure 1e), on the other hand, shows different results from previous algorithms because, based on the relationship between single cells and not on clusters, it identifies a highly branched pathway with two bifurcations terminating on Erythroid and Megakaryocyte cells.

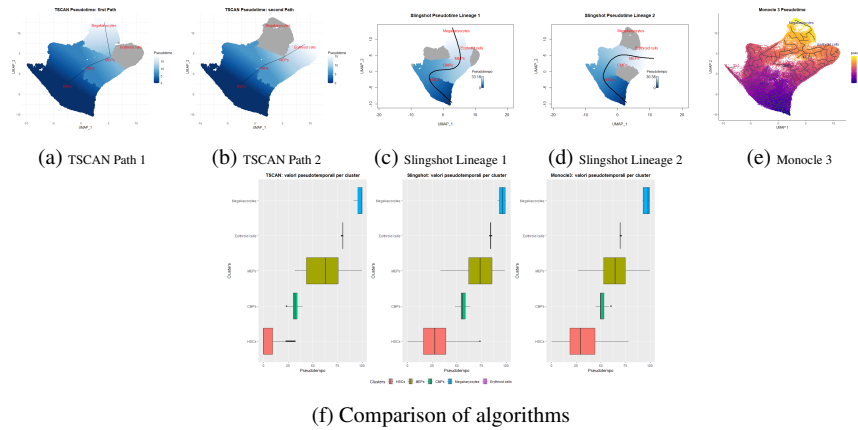


Fig. 1: Comparison of the three pseudotemporal calculation algorithms: TSCAN, Slingshot and Monocle 3.

The pseudo-temporal values, obtained from the three algorithms, conditional on each cluster (Figure 1f), are valid since they are consistent with the hierarchy of cellular differentiation known in biology. Indeed, they are able to define the initial state at HSCs cells and identify bifurcations at the terminal differentiation states such as Erythroids and Megakaryocytes. However, while Slingshot and TSCAN pre-

dict a proportional and increasing temporal progression between different stages of differentiation, the tree structure of Monocle 3 shows the presence of similar pseudo-temporal values between CMPs and MEPs cells, but also between MEPs and Erythroids, reflecting the oscillatory trajectories generated. This leads to consider whether the complexity shown by Monocle 3 is due to the biological pathway of the cells or whether it is due to variability in cell-cell structure.

Phase 2. Once the pseudo-temporal value of cells has been extracted from the three algorithms, it will be possible to analyze the statistical nonparametric models most suited to follow the genetic trajectory of the differentially expressed genes along the pseudo-temporal pathway defined by TSCAN, Slingshot and Monocle 3. For detecting differentially expressed genes we have used *Wilcoxon Rank Sum test* and *Fold Change* revealing that the *PPBP* gene is the one with the most differences between myelofibrosis and healthy subjects. The genetic trajectory was analyzed by considering the conditional dependence of *PPBP* gene expression respect to the pseudo-time, first through penalized B-splines, fitted on the average of the data (at Figures 2a, 2b, 2c, 2d and 2e), then through a quantile regression based again on B-splines (at Figures 2f, 2g, 2h, 2i and 2j). The first element that needs to be distinguished is that TSCAN and Slingshot identify a trajectory only for pathways (or lineages) ending in Megakaryocytes, this means that the *PPBP* gene is not expressed in Erythroid cells, a key element that does not emerge in Monocle 3 because it predicts a unique trajectory.

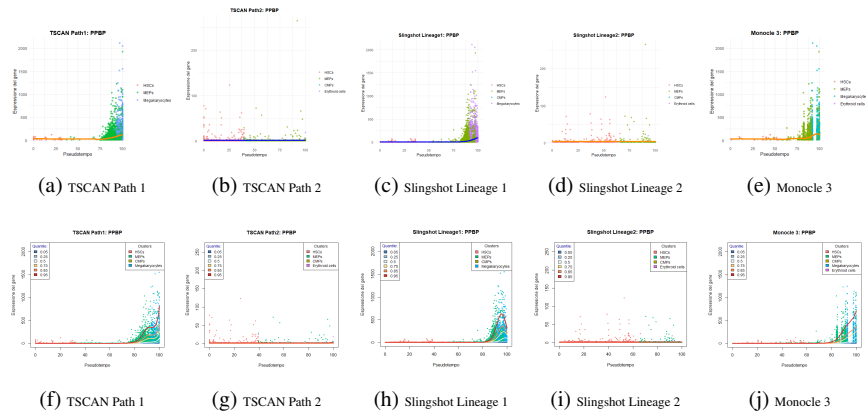


Fig. 2: Relationship between gene expression of the *PPBP* gene in different cells and pseudo-time, fitted by B-Spline and Quantile Regression.

Looking at Figures 2f, 2g, 2h, 2i, and 2j it can be seen that the use of a B-spline on the mean of the data is not appropriate for plotting the trajectory of the *PPBP* gene: the curve is flat on values close to zero throughout the pseudo-temporal pathway (except for a slight increase in expression on Megakaryocytes). This is due to a

large number of cells in which this gene have low expression masking possible variations at small clusters where they are highly expressed. Another critical issue is due to the need, for the pseudo-time algorithms, to plot pseudo-time trajectories in a reduced-dimensional space, which poses the problem of maintaining the initial geometries of the high-dimensional space: clusters may not respect the right distances from each other leading to a strong overlap in pseudo-time allocation between different clusters. Specifically, the low expression of MEPs cells for the PPBP gene overlaps with the high expression in Megakaryocytes that have a small number of cells, this results in the fact that the B-spline on the average expression does not allow highlighting the variations that are revealed in small clusters of cells.

A possible solution is to place the focus on the expected quantile of gene expression conditional on pseudo-time, looking at the highest quantiles to avoid underestimation of the expression data. The Figures 2f, 2g, 2h, 2i, and 2j show changes from the lowest quantiles, but significantly relevant only after the median expression quantile, which allows us to understand that the PPBP gene trajectory, along the pseudo-temporal pathway, tends to increase at Megakaryocyte cells. However, Slingshot, unlike the other algorithms, following initial growth shows a decline in gene expression.

5 Conclusions

The analysis performed allowed the evaluation of different pseudo-time calculation methods and leads to the conclusion that TSCAN and Slingshot prove to be the most suited and easy to interpret for this type of data. In fact, by defining two different pathways (or lineages) they allow us to highlight how the gene *PPBP*, differentially expressed between myelofibrosis patients and healthy subjects, has growth tendencies at Megakaryocytic cells, on which more attention should be paid for the treatment of the myelofibrosis disease. This emerged from the application of quantile regression estimated through B-spline, which, in contrast to B-spline on the mean, was useful for this purpose, highlighting how, for small clusters of cells, there is a need for statistical techniques that can go beyond cellular abundance and identify crucial details in genetic trajectories that might be hidden by the complexity of single-cell data.

References

1. Rodriguez-Meira, A. et al.: Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular cell*, **73**(6), pp. 1292 – – 1305 (2019)
2. Ji, Zhicheng and Ji, Hongkai: TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, **44**, pp. e117. Oxford University Press (2016)
3. Street, Kelly et al.: Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.*, **19**, pp. 1 – – 16. Springer (2018)
4. Cao, Junyue et al.: The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, pp. 496 – – 502. Nature Publishing Group UK London (2019)

Investigating the association between high school outcomes and university enrolment choice: a machine learning approach

Andrea Priulla¹, Alessandro Albano^{1,2}, Nicoletta D’Angelo¹, Massimo Attanasio¹

Abstract This paper investigates how proficiency in mathematics and Italian tests in high school affect university enrolment choices in Italy. We distinguish between students from scientific and humanistic backgrounds, providing valuable insights into their enrolment choices. We employ gradient boosting methodology, adjusting for students’ sociodemographic characteristics and previous educational attainment. Results shed light on the interplay between student performance, sex, and the type of high school attended in shaping enrolment choices.

Key words: gender gap, university enrolment, STEM, gradient boosting

1 Introduction

Despite recent advancements in female educational attainment at the secondary and tertiary levels, there has been limited success in reducing gender vertical and horizontal segregation in education (Barone et al., 2019), with females accounting for only 31% of new STEM entrants at the bachelor’s level among OECD countries (OECD, 2022). The international literature has extensively highlighted the importance of adopting statistical models capable of considering the intricate structure of educational inequalities (Giambona and Porcu, 2018; Priulla et al., 2023). In this framework, machine learning methods have also extended to the realm of education, where they play a pivotal role in analyzing educational data (Hilbert et al., 2021).

In this paper, we employ a machine learning approach to discern the influences of students’ previous educational attainment on the choices of enrolling at university and, subsequently, selecting a specific field of study. To this aim, we use administrative longitudinal micro-data from the *Anagrafe Nazionale Studenti* and INVALSI.

2 Data

The data utilized for the analysis comes from two distinct national administrative sources: INV-S and ANS-U. INV-S comprises micro-data from INVALSI, a research institution which administers standardized tests across five educational levels (grades 2, 5, 8, 10, and 13) to assess students’ proficiency in mathematics, Italian

¹Department of Business, Economics and Statistics, University of Palermo

²Sustainable Mobility Center (Centro Nazionale per la Mobilità Sostenibile—CNMS)
e-mail: andrea.priulla@unipa.it

language, and English reading and listening skills. For this study, we focus on the grade 13 tests in mathematics and Italian. Alongside test scores, INVALSI collects sociodemographic data, family background information, and other indicators of previous educational attainment. The ANS-U database from the Anagrafe Nazionale Studenti provides detailed longitudinal information about the university trajectories of all students enrolled in Italian universities from 2010 to 2020. The linkage of these databases allows for a thorough investigation of the school-to-university transition at the individual level.

We focus on the subset of students attending humanistic and scientific curricula in high school in 2018/19, and we study their school-to-university transition after graduation. We further distinguish the traditional scientific and the applied science tracks within the scientific curriculum (Priulla et al., 2023).

We consider the following set of covariates: HS macroregion; HS socioeconomic status (SES)¹; a variable indicating whether the student attended a public or private high school; sex; HS curriculum; score in INVALSI maths and Italian tests².

3 Methods

We use Gradient Boosting (GBM) (Friedman et al., 2000) to analyze students’ enrollment choices based on a set of predictors. GBM is an ensemble learning technique that sequentially combines predictions from weak learners, typically decision trees, to minimize a loss function $L(y, \hat{F}(\mathbf{x}))$ in expectation, where $\hat{F}(\mathbf{x})$ approximates the output variable y using input variables. The method approximates $\hat{F}(x)$ as a weighted sum of B functions $h_m(x)$ from weak learners: $\hat{F}(x) = \sum_{b=1}^B \alpha_b h_b(x)$, where b denotes the iteration. The process begins with a simple model, $F_0(x)$, typically a shallow decision tree. At each iteration, a weak learner is fitted to the negative gradient of the loss function with respect to the current model’s predictions:

$$F_b(x) = F_{b-1}(x) + \left(\arg \min_{h_b} \left[\sum_{i=1}^n L(y_i, F_{b-1}(x_i) + h_b(x_i)) \right] \right) (x)$$

Given computational challenges in selecting the optimal h_b for an arbitrary loss function L , the steepest descent via functional gradient descent is employed. Popular implementations include GBM, XGBoost, LightGBM, and CatBoost. Understanding the inner workings of a machine learning model and interpreting its predictions are critical aspects of model deployment and decision-making. In this section, we delve into two essential model interpretability tools: Accumulated Local Effects (ALE), and Multi-dimensional Partial Dependence Plots (PDP). These techniques offer valuable insights into the importance and impact of individual features on a model’s predictions, aiding in the interpretation and selection of relevant variables

¹ The OECD PISA surveys and INVALSI tests use the ESCS (Index of Economic, Social, and Cultural Status) that synthetically defines the socioeconomic status of students’ families. A negative (positive) value of the index indicates a lower (higher) SES than the Italian average.

² INVALSI, similarly to PISA tests, reports all the scores using the Rasch metric, setting the average for Italy at 200 and the standard deviation at 40.

for further analysis. The PDP is a visualization tool to investigate how features collectively shape a model's predictions. It operates by marginalizing the model's predictions over the distribution of features not of interest, revealing the relationship between variables of interest and the predicted outcome. Multi-dimensional PDP plots visualize the joint effects of multiple features on the predicted outcome, providing a comprehensive understanding of feature interactions. ALE plots offer a fast and unbiased alternative to one-dimensional PDP plots, allowing interpretation of the effects of individual features on a model's predictions. ALE plots provide the main effect of a feature at a certain value compared to the average prediction of data.

4 Results

We built two GBM models utilizing the same set of predictors listed in Section 2. In Model 1, we predict the probability of enrollment at an Italian university. Therefore, the reference population for this model comprises all fifth-year high school students in 2018/19. Then, In Model 2, we shift our focus to predicting enrollment in a STEM program. This model is built on the subset of students who enrolled at the university. For both models, the dataset is partitioned into training (25%) and test (75%) sets. The model undergoes hyperparameter tuning on the training set through a 10-fold cross-validation procedure. After model selection, evaluation on the test set reveals an AUC of 0.71 for Model 1 and 0.69 for Model 2. These values indicate good discriminatory power for both models.

In Figure 1, the ALE of the three continuous covariates are shown. The most important variables tend to display more varied patterns. As for school SES (Figures 1a and 1d), the observed variations are not drastic, suggesting that the influence of a very high SES on enrolment probabilities becomes more nuanced beyond a certain point. Focusing on STEM enrolment, the result indicates that an overall higher school SES negatively influences students' decisions to pursue a STEM career at university, which is reasonable since students from upper classes are more inclined to attend the humanistic curriculum (Panichella and Triventi, 2014). As for the Italian scores (Figures 1b and 1e), an increase is associated with a higher effect on university enrolment. Conversely, beyond a critical threshold (175), proficiency in Italian reduces the probability of enrolling in STEM before a further increase is observed among top performers. Finally, Figures 1c and 1f unveil an overall positive effect of maths scores on both outcomes. As for university enrolment, the effect of maths scores steadily increases, while it sharply increases after a specific threshold for STEM enrolment. Nevertheless, particular caution is warranted in the range of 70-100, as the data indicates a scarcity of students in this segment.

In Figures 2 and 3, we compute multi-dimensional Partial Dependence Plots (PDP), focusing on the interplay of high school curriculum, sex, and performance in the INVALSI Italian and math tests. The three top panels show the estimated probabilities for female students, while the bottom ones show those for males, both conditional on HS curricula. As expected, the overall probability of university enrolment is notably high (Figure 2) and increases as scores in both tests increase.

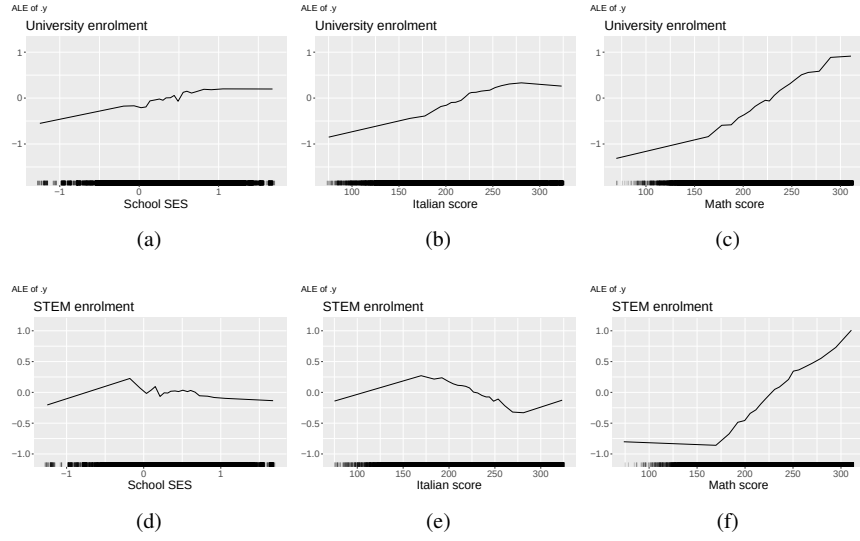


Fig. 1: One-dimensional ALE plots of continuous predictors.

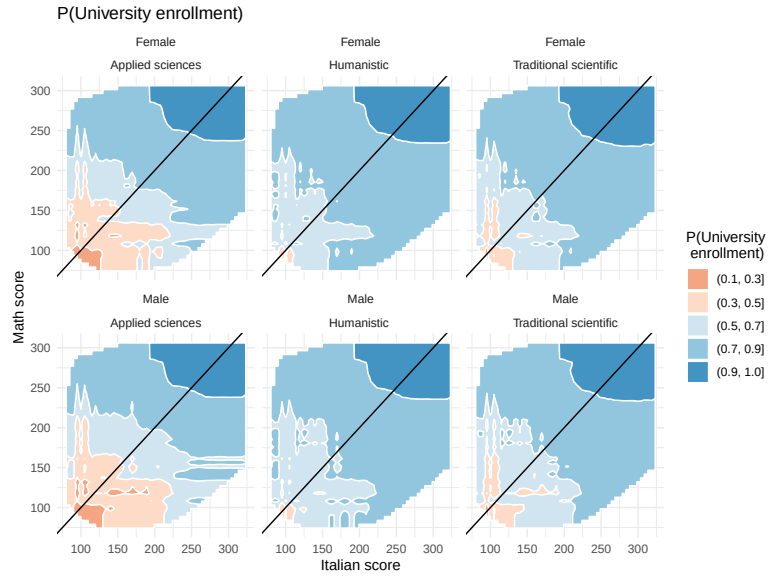


Fig. 2: Probability of university enrolment based on Italian and math scores, sex, and HS track.

Notably, low-achieving students attending applied sciences exhibit a lower probability of university enrolment compared to their low-achieving peers from the other

curricula. Moreover, males show a slightly lower probability of enrolling in university than females, especially when they perform badly in Italian tests.

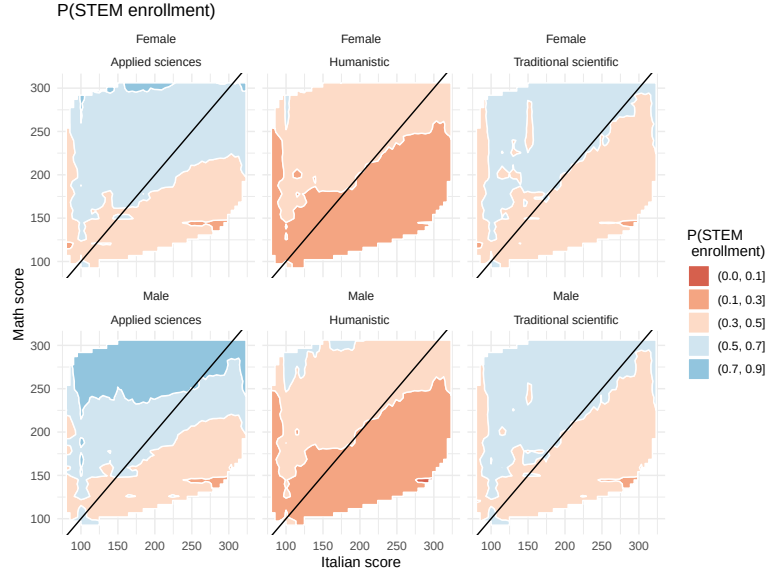


Fig. 3: Probability of STEM enrolment based on Italian and math scores, sex, and HS track.

The scenario radically changes for the choice to enrol in STEM (Figure 3). Students from applied sciences display a higher probability of enrolling in STEM, especially top-performing males in maths tests. On the other hand, a small subset of female students have a high probability of enrolling in STEM when attaining maximum math scores and performing average in Italian tests. Conversely, students with a humanities background consistently present lower probabilities of STEM enrolment. It is noticeable that a subset of male students with strong quantitative skills demonstrates a noteworthy interest in STEM fields. Finally, students from the traditional scientific track fall in between the other two groups of students.

5 Conclusions

In this work, we use GBM models to investigate the interplay between academic performance, sex, and high school background in shaping students' choices. The statistical analysis underlines the significant influence of proficiency in both Italian and mathematics on university enrolment decisions. Importantly, the study reveals

a different impact of previous educational attainment on the enrollment patterns of males and females.

In summary, while no significant gender differences are observed in the choice to enroll at university, it becomes evident that students with high math scores are more associated with higher probabilities of enrolling in STEM. This trend is particularly pronounced, especially when coupled with lower or moderate Italian language scores, and it is more prominent among male students with a scientific background.

This study provides a further understanding of the factors driving academic success and contributes to the ongoing discourse on gender differences in education. Moreover, the insights into high school backgrounds shed light on the importance of tailoring educational pathways to individual strengths and interests.

Fundings

The research work of Alessandro Albano has been supported by the European Union - NextGenerationEU - National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033— 17/06/2022, Spoke 2, CUP B73C2200076000. The research work of Nicoletta D'Angelo has been supported by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001).

References

- Barone, C., Schizzerotto, A., Assirelli, G., and Abbiati, G. (2019). Nudging gender desegregation: A field experiment on the causal effect of information barriers on gender inequalities in higher education. *European Societies*, 21(3):356–377.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.
- Giambona, F. and Porcu, M. (2018). School size and students' achievement. empirical evidences from pisa survey data. *Socio-Econ. Plan. Sci.*, 64:66–77.
- Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., and Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, 9(3):e3310.
- OECD (2022). *Education at a Glance 2022*.
- Panichella, N. and Triventi, M. (2014). Social inequalities in the choice of secondary school: Long-term trends during educational expansion and reforms in Italy. *European Societies*, 16(5):666–693.
- Priulla, A., Vittorietti, M., and Attanasio, M. (2023). Does taking additional maths classes in high school affect academic outcomes? *Socio-Economic Planning Sciences*, page 101674.

14. Contributed - Statistical Analysis in economic and market dynamics

- 14.1 - A comparison of multi-factor stochastic models for commodity prices C3
- 14.2 - Nonparametric ranking estimation with application to the propensity for Circular Economy of Italian economic sectors
- 14.3 - Impact of the Russian invasion of Ukraine on coal markets: Evidence from an event-study approach
- 14.3 - Labour market and time series: a forecast approach for European countries from 1995 to 2022

A comparison of multi-factor stochastic models for commodity prices

Luca Vincenzo Ballestra^a, Christian Tezza^a, Paolo Foschi^a

^aDepartment of Statistical Sciences, Alma Mater Studiorum University of Bologna, Italy

Abstract

We focus on stochastic multi-factor models for commodity pricing. Specifically, we conduct an empirical analysis where we compare the pioneering approach initially developed by Schwartz (1997), the three-factor model proposed by Hughen (2010) and the four-factor models presented in Yan (2002) and Schone and Spinler (2017). An in-sample and an out-of-sample analyses focusing on futures prices for copper traded on the London Metal Exchange indicate that while the model proposed by Schone and Spinler (2017) demonstrates superior in-sample performance, the three-factor model by Schwartz (1997) and the four-factor model by Yan (2002) exhibit the lowest out-of-sample root mean square errors.

1 Introduction

In this paper, we conduct a comparative analysis of several stochastic multi-factor models for commodity pricing. Specifically, we examine the pioneering work initially developed by [Schwartz \(1997\)](#), which encompasses up to three stochastic factors governing the dynamics of the commodity price, convenience yield, and the instantaneous interest rate. Additionally, we consider the three-factor model proposed by [Hughen \(2010\)](#), where the commodity price, its volatility, and the convenience yield are assumed to be stochastic. Furthermore, we evaluate two popular four-factor models: the model introduced by [Yan \(2002\)](#), which incorporates stochastic dynamics for the commodity price, volatility, convenience yield, and interest rate, including a stochastic term to account for jumps in price and volatility trajectories; and the more recent approach by [Schöne and Spinler \(2017\)](#), which models the stochastic dynamics of the commodity price, its long-term mean, volatility, and cost of carry.

All of these models provide closed-form solutions (up to Runge-Kutta integration) for pricing futures, enabling us to estimate their parameters through calibrations to real future prices. However, since not all the stochastic factors are observable, we employ a Kalman filter procedure to estimate the models.

To assess the performance of these models, we conduct both in-sample and out-of-sample analyses, focusing on futures prices for copper traded on the London Metal Exchange (LME). Our findings indicate that while the model proposed by [Schöne and Spinler \(2017\)](#) demonstrates superior in-sample performance, the three-factor model by [Schwartz \(1997\)](#) and the four-factor model by [Yan \(2002\)](#) exhibit the lowest out-of-sample root mean square errors (RMSE). This suggests the significance of incorporating a stochastic interest rate and a stochastic volatility in commodity pricing models.

2 Stochastic models for commodity prices

2.1 Schwartz's (1997) one-factor model

The one-factor model of [Schwartz \(1997\)](#), hereafter *SCH-1factor*, assume that the commodity log spot price, $X \equiv \ln S$, can be characterized by an Ornstein-Uhlenbeck stochastic process, under the physical measure, as:

$$dX = \kappa_x \left(\mu_x - \frac{\sigma_x^2}{2} - X \right) dt + \sigma_x dZ_x \quad (1)$$

The magnitude of the speed of adjustment $\kappa_x > 0$ measures the degree of mean reversion to the long run mean log price, $\mu_x - \frac{\sigma_x^2}{2}$. The second term in equation characterizes the volatility of the process, with dZ_x being an increment to a standard Brownian motion.

In this model, the commodity is not an asset in the usual sense and the log of the spot price, plays the role of an underlying state variable upon which contingent claims can be written.

According to equation (8) of [Schwartz \(1997\)](#), we can compute the log-future price with time to maturity $\tau \equiv T - t$, as:

$$\ln F(\tau) = \beta_0(\tau) + \beta_x(\tau)X \quad (2)$$

where $\beta_x(\tau) = e^{-k\tau}$ and $\beta_0(\tau) = (1 - e^{-k\tau})(r - \frac{\sigma_x^2}{2}) + \frac{\sigma_x^2}{4k}(1 - e^{-2k\tau})$, r denotes the constant risk-free rate.

2.2 Schwartz's (1997) two-factor model

The two-factor model of [Schwartz \(1997\)](#), denoted *SCH-2factor*, models the log-spot price X and convenience yield δ via the following equations:

$$\begin{aligned} dX &= \left(\mu_x - \delta - \frac{\sigma_x^2}{2} \right) dt + \sigma_x dZ_x \\ d\delta &= \kappa_\delta (\mu_\delta - \delta) dt + \sigma_\delta dZ_\delta \end{aligned} \quad (3)$$

where the increments to standard Brownian motion are correlated with: $\mathbb{E}[dZ_x dZ_\delta] = \rho_{x\delta} dt$.

The log-future price with time to maturity τ , is given as in equation (19) of [Schwartz \(1997\)](#):

$$\ln F(\tau) = \beta_0(\tau) + X + \beta_\delta(\tau)\delta \quad (4)$$

where $\beta_0(\tau) = \left(r - \mu_\delta + \frac{\sigma_\delta^2}{2\kappa_\delta^2} - \frac{\sigma_x \sigma_\delta \rho_{x\delta}}{\kappa_\delta} \right) \tau + \frac{\sigma_\delta^2}{4} \frac{(1 - e^{-2\kappa_\delta \tau})}{\kappa_\delta^3} + \left(\mu_\delta \kappa_\delta + \sigma_x \sigma_\delta \rho_{x\delta} - \frac{\sigma_\delta^2}{\kappa_\delta} \right) \frac{(1 - e^{-\kappa_\delta \tau})}{\kappa_\delta^2}$ and $\beta_\delta(\tau) = -\frac{1 - e^{-\kappa_\delta \tau}}{\kappa_\delta}$.

2.3 Schwartz's (1997) stochastic interest rate model

The three-factor model of [Schwartz \(1997\)](#), *SCH-3factor*, for commodity contingent claims is composed of following stochastic differential equations, where r denotes the (instantaneous) interest rate,

$$\begin{aligned}
dX &= \left(r - \delta - \frac{\sigma_x^2}{2} \right) dt + \sigma_x dZ_x \\
d\delta &= \kappa_\delta (\mu_\delta - \delta) dt + \sigma_\delta dZ_\delta \\
dr &= \kappa_r (\mu_r - r) dt + \sigma_r dZ_r
\end{aligned} \tag{5}$$

where the Brownian motion increments are correlated via $\mathbb{E}[dZ_x dZ_\delta] = \rho_{x\delta} dt$, $\mathbb{E}[dZ_\delta dZ_r] = \rho_{\delta r} dt$ and $\mathbb{E}[dZ_x dZ_r] = \rho_{xr} dt$.

The log-future price with time to maturity τ is given by equation (27)-(28) of [Schwartz \(1997\)](#), that is:

$$\ln F(\tau) = \beta_0(\tau) + X + \beta_\delta(\tau)\delta + \beta_r(\tau)r \tag{6}$$

where $\beta_\delta(\tau) = -\frac{1-e^{-\kappa_\delta\tau}}{\kappa_\delta}$ and $\beta_r(\tau) = \frac{1-e^{-\kappa_r\tau}}{\kappa_r}$ and

$$\begin{aligned}
\beta_0(\tau) &= \frac{(\kappa_\delta\mu_\delta + \sigma_x\sigma_\delta\rho_{x\delta})((1-e^{-\kappa_\delta\tau}) - \kappa_\delta\tau)}{\kappa_\delta^2} - \frac{\sigma_\delta^2(4(1-e^{-\kappa_\delta\tau}) - (1-e^{-2\kappa_\delta\tau}) - 2\kappa_\delta\tau)}{4\kappa_\delta^3} \\
&\quad - \frac{(\kappa_r\mu_r + \sigma_x\sigma_r\rho_{xr})((1-e^{-\kappa_r\tau}) - \kappa_r\tau)}{\kappa_r^2} - \frac{\sigma_r^2(4(1-e^{-\kappa_r\tau}) - (1-e^{-2\kappa_r\tau}) - 2\kappa_r\tau)}{4\kappa_r^3} \\
&\quad + \sigma_\delta\sigma_r\rho_{\delta r} \left(\frac{(1-e^{-\kappa_\delta\tau}) + (1-e^{-\kappa_r\tau}) + (1-e^{-(\kappa_\delta+\kappa_r)\tau})}{\kappa_\delta\kappa_r(\kappa_\delta + \kappa_r)} \right) \\
&\quad + \frac{\kappa_\delta^2(1-e^{-\kappa_r\tau}) + \kappa_r^2(1-e^{-\kappa_\delta\tau}) - \kappa_\delta\kappa_r^2\tau - \kappa_r\kappa_\delta^2\tau}{\kappa_\delta^2\kappa_r^2(\kappa_\delta + \kappa_r)}
\end{aligned} \tag{7}$$

2.4 Hugen (2010) stochastic volatility model

The model of [Hugen \(2010\)](#), *HU-3factor*, introduces the stochastic volatility among the state variables vector, that is $\mathbb{X} \equiv [X \ \delta \ V]^T$ with corresponding Brownian motion vector $\mathbb{Z} \equiv [Z_X \ Z_\delta \ Z_V]^T$, so that the model comprises the following equations:

$$d\mathbb{X} = (A + B\mathbb{X}) dt + \Sigma^{1/2} d\mathbb{Z} \tag{8}$$

$$\Sigma = \Omega_0 + \Omega_1 V \tag{9}$$

where $\Sigma^{1/2}$ denote the Cholesky decomposition of Σ and drift parameters

$$A = \begin{bmatrix} 0 \\ \mu_\delta \\ \mu_V \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -1 & -1/2 \\ \kappa_{\delta X} & \kappa_\delta & \kappa_{\delta V} \\ 0 & 0 & \kappa_V \end{bmatrix} \tag{10}$$

and diffusion coefficients

$$\Omega_0 = \begin{bmatrix} 0 & \omega_{12} & 0 \\ \omega_{21} & \omega_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Omega_1 = \begin{bmatrix} 1 & \sigma_{X\delta} & \sigma_{XV} \\ \sigma_{\delta X} & \sigma_\delta^2 & \sigma_{\delta V} \\ \sigma_{VX} & \sigma_{V\delta} & \sigma_V^2 \end{bmatrix} \tag{11}$$

The log-future price with time to maturity τ is given by equation (27) of [Hugen \(2010\)](#), that is:

$$\ln F(\tau) = \alpha(\tau) + \beta(\tau)\mathbb{X} \quad (12)$$

where $\alpha(\tau)$ is scalar and $\beta(\tau) = [\beta_1(\tau), \beta_2(\tau), \beta_3(\tau)]$. The coefficient values can be obtained as done in (28)-(31) via Runge-Kutta numerical integration of the following system of differential equations:

$$\begin{aligned} \frac{d\alpha(\tau)}{d\tau} &= \beta(\tau)A + \frac{1}{2}\beta(\tau)\Omega_0\beta(\tau)^T \\ \frac{d\beta_1(\tau)}{d\tau} &= \beta(\tau)B_{(1)} \\ \frac{d\beta_2(\tau)}{d\tau} &= \beta(\tau)B_{(2)} \\ \frac{d\beta_3(\tau)}{d\tau} &= \beta(\tau)B_{(3)} + \frac{1}{2}\beta(\tau)\Omega_1\beta(\tau)^T \end{aligned} \quad (13)$$

where $B_{(i)}$ denotes the i -th column of B and the terminal solutions are $\alpha(0) = 0, \beta_1(0) = 1, \beta_2(0) = 0, \beta_3(0) = 0$.

2.5 Yan (2002) stochastic volatility with jumps model

The four-factor model of Yan (2002), denoted *YAN-4factor*, comprises the following set of equations:

$$\begin{aligned} dX &= \left(r - \delta - \lambda\mu_J - \frac{\sigma_x^2}{2} - \frac{V}{2} \right) dt + \sigma_x dZ_x + \sqrt{V} dZ_V + J dq \\ dr &= (\mu_r - \kappa_r r) dt + \sigma_r \sqrt{r} dZ_r \\ d\delta &= (\mu_\delta - \kappa_\delta \delta) dt + \sigma_\delta dZ_\delta \\ dV &= (\mu_V - \kappa_V V) dt + \sigma_V \sqrt{V} dZ_V + J_V dq \end{aligned} \quad (14)$$

where q denotes the number of jumps such that $P(dq = 1) = \lambda dt$ and J denotes the jump intensity, with $\ln(1 + J) \sim N\left(\ln(1 + \mu_J) - \frac{\sigma_J^2}{2}, \sigma_J^2\right)$. The increments are correlation as $\mathbb{E}[dZ_x dZ_\delta] = \rho_{x\delta} dt$ and $\mathbb{E}[dZ_x dZ_V] = \rho_{xV} dt$. Lastly, the jump size of volatility follows an exponential distribution, that is $J_V \sim \text{Exp}(\theta), \theta > 0$.

The futures pricing formula is derived in Equation (13) of Yan (2002), that is:

$$\ln F(\tau) = X + \beta_0(\tau) + \beta_r(\tau)r + \beta_\delta(\tau)\delta \quad (15)$$

where $\beta_\delta(\tau) = \frac{-(1-e^{-\kappa_\delta\tau})}{\kappa_\delta}$, $\beta_r(\tau) = \frac{2(1-e^{-\xi_r\tau})}{2\xi_r - (\xi_r - \kappa_r)(1-e^{-\xi_r\tau})}$ where $\xi_r = \sqrt{\kappa_r^2 - 2\sigma_r^2}$ and

$$\begin{aligned} \beta_0(\tau) &= -\frac{\theta_r}{\sigma_r^2} \left(2 \ln \left(1 - \frac{(\xi_r - \kappa_r)(1 - e^{-\xi_r\tau})}{2\xi_r} \right) + (\xi_r - \kappa_r)\tau \right) + \frac{\sigma_\delta^2\tau}{2\kappa_\delta^2} - \frac{\sigma_x\sigma_\delta\rho_{x\delta} + \theta_\delta}{\kappa_\delta}\tau \\ &\quad - \frac{(\sigma_x\sigma_\delta\rho_{x\delta} + \theta_\delta)e^{-\kappa_\delta\tau}}{\kappa_\delta^2} + \frac{4\sigma_\delta^2e^{-\kappa_\delta\tau} - \sigma_\delta^2e^{-2\kappa_\delta\tau}}{4\kappa_\delta^3} + \frac{\sigma_x\sigma_\delta\rho_{x\delta} + \theta_\delta}{\kappa_\delta^2} - \frac{3\sigma_\delta^2}{4\kappa_\delta^3} \end{aligned} \quad (16)$$

We note that the futures price is not a function of spot volatility, jumps or their associated parameters.

2.6 Schöne and Spinler (2017) four factor model

The four-factor model of Schöne and Spinler (2017), denoted *SN-4factor*, is composed of four state variables, so let the state vector be $\mathbb{X} \equiv [X \ \theta \ q \ V]^T$ where θ represent the long-term level of X and q is a factor related to the cost of carry¹. The Brownian motions vector is $\mathbb{Z} \equiv [Z_X \ Z_\theta \ Z_q \ Z_V]^T$, so that the model is represent by the following equations

$$d\mathbb{X} = (A + B\mathbb{X}) dt + \Sigma^{1/2} d\mathbb{Z} \quad (17)$$

$$\Sigma = \Omega_0 + \Omega_1 V \quad (18)$$

where

$$A = \begin{bmatrix} 0 \\ \mu_\theta \\ \mu_q \kappa_q \\ \mu_v \kappa_v \end{bmatrix}, \quad B = \begin{bmatrix} -\kappa_x & \kappa_x & -1 & \kappa_{xv} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\kappa_q & 0 \\ 0 & 0 & 0 & -\kappa_v \end{bmatrix} \quad (19)$$

and

$$\Omega_0 = \begin{bmatrix} 0 & \omega_{12} & \omega_{13} & 0 \\ \omega_{21} & \omega_{22} & \omega_{23} & 0 \\ \omega_{31} & \omega_{32} & \omega_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Omega_1 = \begin{bmatrix} 1 & \sigma_{X\theta} & \sigma_{Xq} & \sigma_{XV} \\ \sigma_{\theta X} & \sigma_\theta^2 & \sigma_{\theta q} & \sigma_{\theta V} \\ \sigma_{qX} & \sigma_{q\theta} & \sigma_q^2 & \sigma_{qV} \\ \sigma_{VX} & \sigma_{V\theta} & \sigma_{Vq} & \sigma_V^2 \end{bmatrix} \quad (20)$$

In this way a time-varying correlation structure is allowed.

The log-future price with time to maturity τ is given by Equation (11) of Schöne and Spinler (2017), that is

$$\ln F(\tau) = \alpha(\tau) + \beta(\tau)\mathbb{X} \quad (21)$$

where $\alpha(\tau)$ is scalar and $\beta(\tau) = [\beta_1(\tau), \beta_2(\tau), \beta_3(\tau), \beta_4(\tau)]$. The coefficient values can be obtained as done in Equation (13)-(17) of Schöne and Spinler (2017) via Runge-Kutta numerical integration of the following system of differential equations:

$$\begin{aligned} \frac{d\alpha(\tau)}{d\tau} &= \beta(\tau)A + \frac{1}{2}\beta(\tau)\Omega_0\beta(\tau)^T \\ \frac{d\beta_1(\tau)}{d\tau} &= \beta(\tau)B_{(1)} \\ \frac{d\beta_2(\tau)}{d\tau} &= \beta(\tau)B_{(2)} \\ \frac{d\beta_3(\tau)}{d\tau} &= \beta(\tau)B_{(3)} \\ \frac{d\beta_4(\tau)}{d\tau} &= \beta(\tau)B_{(4)} + \frac{1}{2}\beta(\tau)\Omega_1\beta(\tau)^T \end{aligned} \quad (22)$$

where $B_{(i)}$ denotes the i -th column of B and the terminal solutions are $\alpha(0) = 0, \beta_1(0) = 1, \beta_2(0) = 0, \beta_3(0) = 0$ and $\beta_4(0) = 0$.

¹The cost of carry in Huguen (2010) is defined as the net convenience yield, that is $q_t = r_t - \delta_t$, where as usual r_t denotes the instantaneous interest rate and δ_t the convenience yield

<i>SCH-1factor</i>		<i>SCH-2factor</i>		<i>SCH-3factor</i>	
μ_x	8.0951(0.7797)	μ_x	0.8547(0.7067)	μ_δ	-2.6285(0.7791)
r	0.2232(0.0872)	r	0.0140(0.0036)	μ_r	-2.5451(0.8106)
κ_x	0.0025(0.0000)	μ_δ	0.0027(0.0007)	κ_δ	0.2025(0.0337)
σ_x	0.2769(0.0028)	κ_δ	0.0053(0.0032)	κ_r	3.1865(0.1200)
σ_{ϵ_1}	0.0176(0.0003)	σ_x	0.4245(0.1210)	σ_x	0.8928(0.0073)
σ_{ϵ_2}	0.0106(0.0002)	$\rho_{x\delta}$	0.9311(0.0098)	$\rho_{x\delta}$	0.3866(0.0175)
σ_{ϵ_3}	0.0041(0.0000)	σ_δ	0.0157(0.0009)	ρ_{xr}	-0.2320(0.0831)
σ_{ϵ_4}	0.0118(0.0001)	σ_{ϵ_1}	0.0022(0.0003)	σ_δ	0.1433(0.0019)
		σ_{ϵ_2}	0.0021(0.0004)	ρ	0.0259(0.0092)
		σ_{ϵ_3}	0.0079(0.0000)	σ_r	0.1257(0.0062)
		σ_{ϵ_4}	0.0000(0.0000)	σ_{ϵ_1}	0.0001(0.0000)
				σ_{ϵ_2}	0.0000(0.0000)
				σ_{ϵ_3}	0.0032(0.0000)
				σ_{ϵ_4}	0.0002(0.0001)
LL:	1.0403×10^4		1.5245×10^4		1.6506×10^4
BIC:	-2.0740×10^4		-3.0399×10^4		-3.2896×10^4

Table 1: In-sample estimated coefficients. The standard errors, reported in parenthesis, are computed by inverting the negative Hessian matrix evaluated at the optimum parameter values.

3 Empirical analysis

In this section, we examine the performance of the models in both the in-sample and out-of-sample contexts. The empirical study utilizes daily high-grade copper prices for the settlement of future contracts on the London Metal Exchange (LME). Specifically, we consider daily quotations within the time interval from June 6, 2019, to April 5, 2023, encompassing a total of 1000 days.

The in-sample analysis involves estimating model parameters based on futures with various maturities, namely 2, 7, 13, and 17 months, denoted as F_2, F_7, F_{13}, F_{17} , respectively. The estimation procedure employed is the Kalman filter, which is a widely used technique for parameter estimation in commodity price models (see [Schwartz \(1997\)](#) or [Schöne and Spinler \(2017\)](#)). It is important to note that in the observation equation, we introduce a measurement error term $\epsilon_{i,t}$, where $\mathbb{E}[\epsilon_{i,t}] = 0$ and $\mathbb{V}[\epsilon_{i,t}] = \sigma_{\epsilon_i}^2$. Consequently, we estimate the variance parameters for each future, as presented in the tables [1-2](#).

The "out-of-sample" exercise involves evaluating the futures prices for maturities $F_1, F_3, F_5, F_9, F_{11}, F_{19}$ using the parameters obtained from the in-sample analysis. The results are illustrated in table [3](#).

References

- W.K. Hughen. A maximal affine stochastic volatility model of oil prices. *Journal of Future Markets*, 30(2):101–133, 2010.
- E.S. Schwartz. The stochastic behavior of commodity prices: Implications for valuation and hedging. *Journal of Finance*, 52(3):923–973, 1997.

<i>HU-3factor</i>		<i>YAN-4factor</i>		<i>SN-4factor</i>	
μ_δ	0.0134(0.0034)	λ	0.4853(0.0132)	μ_θ	0.4079(0.2622)
μ_v	0.2572(0.0651)	μ_J	0.7252(0.0090)	μ_q	0.1139(0.3078)
$\kappa_{\delta x}$	0.0041(0.0002)	σ_J	0.3374(0.0060)	μ_v	0.9541(0.1787)
κ_δ	0.1083(0.0137)	θ	0.5793(0.0038)	κ_x	0.4141(0.0265)
$\kappa_{\delta v}$	0.0243(0.0032)	σ_x	0.6366(0.0054)	κ_{xv}	1.0991(0.6399)
κ	0.6764(0.1446)	μ_r	0.7899(0.0385)	κ_θ	0.3426(0.0093)
ω_{12}	0.0530(0.0004)	κ_r	0.8536(0.0370)	κ_v	0.4365(0.0927)
ω_{22}	0.0491(0.0006)	σ_r	0.3610(0.0278)	ω_{12}	0.0540(0.0215)
$\rho_{x\delta}$	0.2012(0.0104)	μ_δ	0.5741(0.0349)	ω_{13}	0.6585(0.0258)
ρ_{xv}	0.2262(0.0224)	κ_δ	0.3907(0.0329)	ω_{22}	0.7904(0.0636)
σ_δ	0.6779(0.0012)	σ_δ	0.0469(0.0030)	ω_{23}	0.0410(0.0673)
$\rho_{\delta v}$	-0.5963(0.0170)	σ_V	0.2671(0.0057)	ω_{33}	0.6661(0.1157)
σ_v	0.5740(0.0444)	$\rho_{x\theta}$	0.4545(0.5484)	ρ_{xv}	0.5159(0.3907)
ρ_{xv}	0.2262(0.0224)	ρ_{xq}	0.1716(0.5532)	σ_θ	0.3275(1.7508)
σ_{ϵ_1}	0.0365(0.0001)	$\rho_{x\delta}$	0.1294(0.0024)	$\rho_{\theta q}$	0.2537(0.0001)
σ_{ϵ_2}	0.0336(0.0001)	ρ_{xV}	0.0062(0.0048)	$\rho_{\theta v}$	0.4521(0.2120)
σ_{ϵ_3}	0.0239(0.0004)	σ_{ϵ_1}	0.0001(0.0000)	σ_q	0.2909(0.1398)
σ_{ϵ_4}	0.0167(0.0001)	σ_{ϵ_2}	0.0052(0.0002)	ρ_{qv}	0.6618(1.1108)
		σ_{ϵ_3}	0.0002(0.0000)	σ_q	0.0514(0.0029)
		σ_{ϵ_4}	0.0011(0.0000)	σ_{ϵ_1}	0.0056(0.0001)
				σ_{ϵ_2}	0.0182(0.0004)
				σ_{ϵ_3}	0.0604(0.0017)
				σ_{ϵ_4}	0.0711(0.0016)
LL:	1.6670×10^4		1.5509×10^4		1.6903×10^4
BIC:	-3.3191×10^4		-3.0852×10^4		-3.3615×10^4

Table 2: In-sample estimated coefficient. The standard errors, reported in parenthesis, are computed by inverting the negative Hessian matrix evaluated at the optimum parameter values.

Future	<i>SCH-1factor</i>	<i>SCH-2factor</i>	<i>SCH-3factor</i>	<i>HU-3factor</i>	<i>YAN-4factor</i>	<i>SN-4factor</i>
F_1	0.009 11	0.0053	0.0041	0.0063	0.0041	0.0041
F_3	0.009 07	0.0022	0.0016	0.0042	0.0016	0.0016
F_5	0.009 07	0.0032	0.0035	0.0047	0.0036	0.0035
F_9	0.009 20	0.0024	0.0015	0.0042	0.0039	0.0027
F_{11}	0.009 54	0.0021	0.0018	0.0047	0.0017	0.0091
F_{19}	0.009 90	0.0017	0.0017	0.0024	0.0019	0.0010
Mean*100:	0.9310	0.2855	0.2408	0.4473	0.2800	0.3711

Table 3: Out of sample results: RMSE pricing errors in valuing futures contracts for the considered models.

- M.F. Schöne and S. Spinler. A four-factor stochastic volatility model of commodity prices. *Review of Derivatives Research*, 20:135–165, 2017.
- X.S. Yan. Valuation of commodity derivatives in a new multi-factor model. *Review of Derivatives Research*, 5:251–271, 2002.

Nonparametric ranking estimation with application to the propensity for Circular Economy of Italian economic sectors

Stefano Bonnini, Michela Borghesi and Massimiliano Giacalone

Abstract This work concerns an innovative nonparametric method to estimate the ranking of multivariate populations applied to an economic problem. Circular Economy is spreading very rapidly because the idea that a linear production system is no longer sustainable, from an environmental point of view, is taking hold. The main goal of this work is to determine the ranking of Italian economic sectors according to the tendency to implement various Circular Economy activities, in other words on the propensity towards Circular Economy, based on sample data. The dataset is original and it refers to a survey carried out in Italy in 2020.

Abstract *Questo lavoro riguarda un metodo nonparametrico innovativo di stima di ranking per popolazioni multivariate applicato ad un problema economico. L'Economia Circolare si sta diffondendo molto rapidamente perché sta prendendo piede l'idea che un sistema di produzione lineare non sia più sostenibile dal punto di vista ambientale. L'obiettivo principale di questo lavoro è quello di determinare, sulla base di dati campionari, la graduatoria dei settori economici italiani basata sulla tendenza ad implementare varie attività di Economia Circolare, in altre parole sulla propensione verso l'Economia Circolare stessa. Il dataset è originale e si riferisce ad un'indagine condotta in Italia nel 2020.*

Key words: Ranking, Multivariate Statistics, Nonparametric Test.

Stefano Bonnini

University of Ferrara, Via Voltapaletto 11, Ferrara, e-mail: stefano.bonnini@unife.it

Michela Borghesi

University of Ferrara, Via Voltapaletto 11, Ferrara, e-mail: michela.borghesi@unife.it

Massimiliano Giacalone

University of Campania "Luigi Vanvitelli", Corso Gran Priorato di Malta I, Capua (NA), e-mail: massimiliano.giacalone@unicampania.it

1 Introduction

This work concerns the application of an innovative nonparametric statistical method to a Circular Economy (CE) problem. The goal is to estimate the ranking of Italian economic sectors according to their propensity towards CE, represented by multivariate sample data. The term Circular Economy is a paradigm that operates on the principles of the 3Rs: Reduce, Reuse and Recycle. One of this decade's goals is to foster sustainable development to reverse the inevitable course of man-made climate change. One of the effective strategies, which makes the CE essential for this goal, concerns the revolution of the economic system to replace the "end-of-life" concept with the reduction and reuse of materials [3]. The foundation of this concept is that a linear economic model (of the take-make-dispose type) is no longer sustainable. On the other hand, we have a circular economic system, which is considered one of the most effective solutions to environmental sustainability problems [7]. Furthermore, it is well known that the manufacturing sector is the most involved in the transition to the CE [1].

This work aims to represent a contribution to the empirical literature on the comparison of economic sectors in terms of their propensity towards CE. In particular, we focus on Italian Small and Medium Enterprises and, as said, we propose the use of an innovative inferential approach, never used in this framework before.

Section 2 focuses on the presentation of the statistical problem and of the methodological solution. The results of the application of the proposed method to the Italian case study are shown and discussed in Section 3. Finally, Section 4 is dedicated to concluding remarks.

2 Statistical problem and methodological solution

As said in Section 1, the goal of this study is to compare economic sectors on the basis of the adoption of CE practices. In particular, we want to create a ranking of $C > 2$ sectors based on the proportions of companies that have implemented some specific CE activities. Each sector represents a population and the response variable is multidimensional. Each component of the multivariate response corresponds to a specific CE activity.

Let us consider a random sample from each of the C multivariate populations $\Pi_1, \dots, \Pi_C$ with $C > 2$. According to the proposed method, we have to test the null hypothesis of equality in distribution of the C multivariate populations, versus the alternative hypothesis that the equality is not true. The inferential results of this test, are used to obtain a relative ranking of the C populations, based on pre-specified criteria [5]. The ranking is in decreasing order of propensity for CE, i.e. the first sector has the maximum rank and, consequently, the highest propensity. Similarly, the last sector has the minimum rank and the lowest propensity.

Let $\mathbf{Y}$ be the p -dimensional vector representing the response variable from population Π . Each component of $\mathbf{Y}$ is a dichotomous variable. In the case study, such

dichotomous variable takes 1 if a company undertakes the corresponding CE activity and 0 otherwise. The data observed on the C samples, are used to rank the populations $\Pi_1, \dots, \Pi_C$, with respect to the p dichotomous components of the multivariate response. The rank of the j -th population, $r(\Pi_j)$, represents the propensity of such sector towards CE and can be computed as follows:

$$\begin{aligned} r_j = r(\Pi_j) &= 1 + \sum_{h \neq j} I(\mathbf{Y}_j \stackrel{d}{>} \mathbf{Y}_h) = \\ &= 1 + \left\{ \#(\mathbf{Y}_j \stackrel{d}{>} \mathbf{Y}_h), h = 1, \dots, C, h \neq j \right\}, j = 1, \dots, C, \end{aligned} \quad (1)$$

where $I(\cdot)$ is the indicator function and $\#$ means the number of times [2, 5]. In Equation 1, the rank is derived by using the concept of multivariate stochastic ordering defined according to stochastic dominance pairwise comparisons. The first step of the method consists of obtaining the matrix $A = [a_{ij}]$ of the p -values resulting from the directional pairwise comparisons. Since the pairwise comparisons consist of one-tailed tests, the matrix is not symmetric. Furthermore, unlike in the univariate case, $a_{ij} + a_{ji}$ could be less than 1 for some (i, j) . The pairwise comparisons represent partial tests in a C -sample multiple test. Hence, to control the family-wise error rate and avoid the inflation of the type-I error probability, the elements of A (p -values of such partial tests) must be adjusted [12]. Matrix $\tilde{A} = [\tilde{a}_{ij}]$ of adjusted p -values is transformed into the significance matrix $B = [b_{ij}]$, such that $b_{ij} = 1$ if $\tilde{a}_{ij} \leq \alpha$ and $b_{ij} = 0$ if $\tilde{a}_{ij} > \alpha \forall i, j$, where α is the significance level. Finally, using Equation 1, the ranks of sectors are obtained by summing up the elements of each row of B [6].

For each multivariate pairwise comparison, the methodology of Combined Permutation Tests (CPTs) is applied [4]. CPTs consist of a family of permutation tests, suitable for complex testing problems that can be broken down into partial tests (sub-problems). This approach is distribution-free and robust with respect to the underlying distribution. The assumption of exchangeability of the individual vectors of observed values is the unique condition for the test, making the stronger and sometimes unrealistic iid condition unnecessary. Exchangeability, under the null hypothesis of equality in distribution, is satisfied [10, 11].

The procedure consists in carrying out the univariate two-sample permutation tests on proportions and then combining the p -values of the partial tests corresponding to the different variables, for each directional pairwise comparison. The p -values of such combined tests are used to determine the elements of matrix $\tilde{A}$. A further combination of these tests provides an overall p -value for the C -sample multivariate test. Formally, the test statistics for the pairwise comparisons are the difference of the sample proportions, and the combined test statistics are based on the combination of the partial p -values obtained through the application of the significance level function according to the permutation distribution. Following the Fisher's rule, the combined test statistic for the comparison between the j th and the h th sample and for the overall test are $T'_{jh} = -2 \sum_v \ln(\lambda_{jh,v})$ and $T'' = -2 \sum_{h \neq j} \ln(\lambda'_{jh})$ respectively,

where $\lambda_{jh,v}$ and λ'_{jh} are the p -values of the comparison $j - h$ for the v -th variable and in general, respectively.

An overall p -value less than α , provides empirical evidence in favour of the hypothesis of difference of the C populations and implies the need of analyzing matrix $\tilde{A}$ of the adjusted partial p -values to determine the ranking. If the overall p -value is greater than α then no ranking is admissible because there is no significant difference between the sample proportions.

3 Case study

The ranking method presented in the previous section was applied to the original dataset collected in a sample survey carried out in Italy in 2020. As mentioned above, the case study involves an application concerning Circular Economy. The data refer to a sample of Italian SMEs operating in different sectors. The dataset is original and refers to a sample survey on innovations for CE adopted by Italian SMEs in the two-year period 2017-2018. The data were collected through an interview carried out, with the CATI method, in January 2020. The sample of 3029 interviewed firms was selected through a stratified random sampling. The response consists of ten dummy variables related to different CE practices adopted or not in the period 2017-18. The CE practices are the following: reduction of water use in the production process, reduction in the use of materials, use of energy generated from renewable sources, reduction in the use of electricity, reduction of emitted waste, reuse of waste in its production cycle, transfer of waste to other companies, which use them in their production cycle, change in product design to minimize the use of raw materials, change in product design to maximize their recyclability, change in the production process to reduce greenhouse gas emissions.

We compared the seven sectors (listed in detail in the Appendix) and, by applying the methodology presented in Section 2, we obtained an overall p -value equal to 0.001. Since such p -value is significant at the level $\alpha = 0.10$, there is empirical evidence in favor of the hypothesis that the adoption of Circular Economy practices is affected by the company's sector of activity.

The significance of the overall test can be attributed to one or more specific pairwise comparisons, after adjustment of the partial p -values to control the family-wise error (FWE) for the multiplicity [12]. Hence, the p -values of the partial tests were adjusted with the Bonferroni-Holm method, which controls the FWE in a strong sense [8, 9] (see Table 1).

Then we replaced the significant p -values with 1 and the others with 0, to obtain the matrix B and estimate the ranking of the seven sectors as described above.

According to the rank of the seven groups (see Table 2), we can conclude that, when a firm operates in the sector concerning chemical-pharmaceutical, plastic and refined petroleum products, there is the highest propensity to adopt the considered CE practices. All the other sectors considered are tied for second place in the ranking.

Table 1 Table of adjusted p -values with Bonferroni-Holm method (significant in bold).

>	food-bev	text-wear	wood-print	chem-plast	metals	computer	motors
food-bev	-	1.000	0.590	1.000	0.261	1.000	1.000
text-wear	1.000	-	1.000	1.000	1.000	1.000	1.000
wood-print	1.000	1.000	-	1.000	1.000	1.000	1.000
chem-plast	0.427	0.146	0.094	-	0.036	0.123	0.068
metals	1.000	1.000	1.000	1.000	-	1.000	1.000
computer	1.000	1.000	1.000	1.000	1.000	-	1.000
motors	1.000	1.000	1.000	1.000	1.000	1.000	-

Table 2 Table for the computation of the rank.

>	food-bev	text-wear	wood-print	chem-plast	metals	computer	motors	Rank
food-bev	-	0	0	0	0	0	0	1
text-wear	0	-	0	0	0	0	0	1
wood-print	0	0	-	0	0	0	0	1
chem-plast	0	0	1	-	1	0	1	4
metals	0	0	0	0	-	0	0	1
computer	0	0	0	0	0	-	0	1
motors	0	0	0	0	0	0	-	1

4 Conclusions

The presented nonparametric methodology for ranking multivariate populations, based on the application of combined permutation tests and pairwise comparisons, is a distribution-free, robust and flexible statistical solution. Its application to an original dataset concerning a survey about Italian small and medium enterprises provides empirical evidence in favour of the hypothesis that the adoption of CE practices depends on the company's activity sector. In particular, we can conclude that, when a firm operates in the sector concerning chemical-pharmaceutical, plastic and refined petroleum products, there is the highest propensity to adopt the considered CE practices.

Acknowledgements The authors thank the Italian Ministry of Education, University and Research that funded the departmental development program (DEM - University of Ferrara) for the period 2018-2022, to promote excellence in education and research ("Dipartimenti di Eccellenza"). This paper represents the output of such a project.

Appendix

Stratification of sectors defined according to the ATECO 2022 classification:

- *food-bev*: food products, beverages and tobacco products (sectors 10-11-12 of the Ateco classification)
- *text-wear*: textiles and wearing apparel (sectors 13-14 of the Ateco classification)
- *wood-print*: wood, paper, printing and reproduction (sectors 16-17-18 of the Ateco classification)
- *chem-plast*: chemical-pharmaceutical, plastic and refined petroleum products, coke (sectors 19-20-21-22 of the Ateco classification)
- *metals*: basic metals and fabricated metal products (sectors 24-25 of the Ateco classification)
- *computer*: computer, electronic and optical products, machinery and equipment (sectors 26-28 of the Ateco classification)
- *motors*: motor vehicles, trailers, semitrailers and other transport equipment (sectors 29-30 of the Ateco classification)

References

1. Acerbi, F., and Taisch, M.: A literature review on circular economy adoption in the manufacturing sector. *J Clean Prod* 273, 123086 (2020) <https://doi.org/10.1016/j.jclepro.2020.123086>
2. Arboretti Giancristofaro, R., Bonnini, S., Corain, L., Salmaso, L.: A permutation approach for ranking of multivariate populations. *J. Multivar. Anal.*, **132**, 39-57 (2014)
3. Austin, A., and Ur Rahman, I.: A triple helix of market failures: Financing the 3Rs of the circular economy in European SMEs. *J. Clean. Prod.* **361**, 132284 (2022)
4. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric hypothesis testing. Rank and permutation methods with applications in R. Wiley (2014)
5. Corain, L., Arboretti, R., Bonnini, S.: Ranking of multivariate populations: A permutation approach with applications. CRC press (2017)
6. Corain, L., Salmaso, L.: A nonparametric method for defining a global preference ranking of industrial products. *J. Appl. Stat.* **34**(2), 203-216 (2007)
7. Cristoni, N., Tonelli, M.: Perception of firms participating in a circular economy. *Eur. J. Sustain. Dev.*, **7** (4), 105-118 (2018)
8. Giacalone, M., Zirilli, A., Cozzucoli, P., Alibrandi, A.: Bonferroni-holm and permutation tests to compare health data: methodological and applicative issues, *BMC Med. Res. Methodol.* 18:81 (2018)
9. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 65-70 (1979)
10. Pesarin, F.: Nonparametric Combination Methodology. In *Multivariate Permutation Tests with Applications in Biostatistics*, 2nd ed.; Wiley: Chichester, UK (2001)
11. Pesarin, F., Salmaso, L.: Permutation tests for complex data: applications and software. Wiley series in probability and statistics (2010)
12. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment; Wiley-Interscience: New York, NY, USA (1992)

Impact of the Russian invasion of Ukraine on coal markets: Evidence from an event-study approach

Yana Kostiuk, Paola Cerchiello and Arianna Agosto

Abstract We study the immediate market response of the coal-related industry to the onset of the Russian full-scale invasion of Ukraine. Our findings, based on a rigorous event study analysis, indicate a significant and negative reaction in the European region on the 21st and 24th of February, followed by an increase in average abnormal returns (AARs) on the first day following the invasion. As for Emerging countries, no results were observed for China, while India registered positive AARs on the second and third day of the war. Over the designated 14-day event period, positive Cumulative Average Abnormal Returns (CAARs) are evident in all examined sub-samples, though statistical significance is observed only within the Titans sub-sample.

Key words: Russo-Ukrainian war, event study, coal, market reaction

1 Introduction

While the world was still recovering from the consequences of the COVID-19 pandemic, Russian troops initiated the full-scale invasion of Ukraine on the 24th of February 2022.

Yana Kostiuk
IUSS, Palazzo del Broletto, Piazza della Vittoria, 15, 27100, Pavia (PV), e-mail: yana.kostiuk@iusspavia.it

Paola Cerchiello
Department of Economics and Management, UNIPV, Via S. Felice Al Monastero, 5, 27100, Pavia
e-mail: paola.cerchiello@unipv.it

Arianna Agosto
Department of Economics and Management, UNIPV, Via S. Felice Al Monastero, 5, 27100, Pavia
e-mail: arianna.agosto@unipv.it

The outbreak of the war exacerbated the energy crisis that had been evolving since the second half of 2021 due to the rapid post-COVID economic rebound and reduced supplies of Russian gas to Europe¹.

With the beginning of the full-scale invasion, the EU committed to reducing its energy dependency on Russia, with a major focus on natural gas, and imposed sanctions on the Russian energy sector. The supply-side constraints led to skyrocketing fossil fuel prices, forcing governments to look for readily available substitutes, such as coal, and to reshape supply chains. In 2022, global coal consumption reached a new all-time high, rising by 4% year-on-year to 8.42 billion tonnes, resulting in a 1.6% increase in coal-related emissions for the same year [?].

The ongoing global energy crisis has affected the political economy of the coal sector not only by reshuffling the resources available to pro-coal actors due to the windfall profits, but also by bolstering the pro-coal narrative [?]. Notably, several countries intensified their coal mining operations due to increased demand from countries looking for an alternative to Russian coal.

The analysis of the financial impact of the black-swan events has been gaining more interest recently with the outbreak of the COVID-19 pandemic, while the literature with a focus on military conflicts is still relatively limited. [?] analyzed the value of the government bonds of countries affected by WWII. In a similar vein, [?] and [?] analyzed the impact of WWII on the British and USA stock markets, respectively.

When it comes to the full-scale Russo-Ukrainian war, several studies have been performed so far. [?] observed a negative and statistically significant European banks stocks' market reaction to the invasion of Ukraine. [?], [?], and [?] studied the implications for food security.

Finally, several studies place their primary emphasis on energy markets. In particular, [?] studied the impact of the shock to energy prices induced by the Russian invasion on the financial performance of the Eurostoxx 600 index. [?], in their analysis of the impact on the energy and metal markets, demonstrated a significant increase in abnormal returns related to the renewable energy sector. Many studies focus on energy markets as a whole or oil and gas markets with limited attention to coal. However, as suggested by [?], even a short-term relapse to coal could lead to a vicious cycle and undermine the global just transition process.

In light of the above-mentioned considerations, the present study aims at expanding the existing literature by investigating the impact of the invasion on the global coal market. By applying the event study approach to a sample of 213 firms drawn from the Global Coal Exit List (GCEL) 2022, we analyze abnormal returns for the pre and post event days. We expect that Russia's full-scale invasion of Ukraine is associated with positive abnormal returns for the coal mining sector and that the magnitude of the impact varies among defined groups of interest.

The rest of the paper is organized as follows: Section ?? explains the data and methodology, and Section ?? presents and discusses our findings.

¹ Pipeline deliveries from Russia to Europe declined by 25% year-on-year in Q4 2021[?].

2 Data and methodology

2.1 Data

We define Russia's full-scale invasion of Ukraine as *event* and thus the 24th of February 2022 as *event day* ($t=0$). The event window consists of 14 days, from $t-3$ to $t+10$, while the estimation window spans 120 days, from $t-125$ to $t-5$. Only trading days are taken into consideration. To handle the discrepancies in trading days among the considered markets, we calculate both trade-to-trade and lumped returns. For brevity, we only report results based on the trade-to-trade technique of adjustment.

To generate the sample of firms, we use the Global Coal Exit List updated in 2022². The list is provided by the German NGO Urgewald and includes almost 3000 companies and their subsidiaries covering more than 90% of global thermal coal production. To eliminate any possible source of bias and confounding, we obtain the final sample through the following steps. First, we remove all not-public firms reducing the sample to 375 firms. As the next step, observations characterized by thin trading or with incomplete stock mark data are eliminated. Finally, the firms with the announcement date of dividend payments as of 24 February 2022 are also removed. The final sample includes 213 firms representing 23 countries.

We use daily closing prices of the stocks subsequently adjusted for dividend payments. Both share prices and dividend payments are retrieved from Refinitiv. Furthermore, we download the daily closing prices of three regional market indexes to be used as benchmarks in the calculation of abnormal returns (ARs): EU-ROSTOXX600, Dow Jones Global Titans 50, and S&P EM 100. Based on the country of the firms' headquarters and according to the reference benchmark index, we obtain three sub-samples for the calculation of the AARs at the aggregated level: *Europe* (Czech Republic, Finland, Germany, Italy, Turkey, Warsaw), *Titans* (US, UK, Canada, Australia), and *Emerging* (China, Taiwan, India, Pakistan, Indonesia, Malaysia, Thailand, Vietnam, Morocco, Chile, Brazil, South Africa, South Korea).

2.2 Methodology

Central in the event study is the calculation of the abnormal returns (ARs) which can be defined as follows:

$$AR_{it} = R_{it} - E(R_{it}), \quad (1)$$

where AR_{it} is the abnormal return at time t for the firm i , R_{it} represents the actual return at time t , and $E(R_{it})$ is the expected return at time t for the same firm.

The actual returns are defined as follows:

² Available for download at <https://www.coalexist.org/>

$$R_{it} = \ln \left(\frac{P_{it} - P_{it-1}}{P_{it}} \right), \quad (2)$$

where P_{it} is the dividend and split-adjusted firm's share price at time t , and P_{it-1} is the price at time $t - 1$. To estimate the expected returns we employ the Market Model which considers the firm's past performance and its sensitivity to the general market by using stock market indexes as benchmarks. The model has been estimated using the following equation:

$$\mathbb{E}(R_{it}) = \alpha_i + \beta_i R_{mt}, \quad (3)$$

where α_i and β_i are the Ordinary Least Square (OLS) estimators of the model's parameters, and R_{mt} is the market return of the benchmark index m at time t . As previously mentioned, we use three benchmark indexes, namely EUROSTOXX600, DJ Global Titans 50 and S&P EM100.

The next step is the calculation of the average abnormal returns (AARs) which measure the daily average of ARs at aggregated level using the equation:

$$\text{AAR}(t) = \frac{1}{N} \sum_{i=1}^N AR_i \quad (4)$$

where N is the total number of firms in the cluster.

Finally, we calculate the Cumulative average abnormal returns (CAARs) defined as follows:

$$\text{CAAR}_{(t_1, t_2)} = \sum_{t_1}^{t_2} \text{AAR}_{it} \quad (5)$$

To assess the statistical significance of AARs we use both parametric and non-parametric tests. For parametric testing we use the adjusted standardized cross-sectional test, which is based on the Patell Z test, but additionally accounts for event-induced volatility and cross-serial correlation. As for the non-parametric tests, we report the generalized rank test. The results are discussed in the next section.

3 Results

Table ?? displays the AARs through the 14-day event window at the sub-sample level. Considering the pre-event window, we observe a negative and statistically significant Europe's stock price reaction on day -3, that is the 21st of February when Russia's president announced the independence of Donetsk and Luhansk regions. The precise impact of this announcement cannot be estimated for the *Titans* sub-sample given the closure of the US and Canadian stock exchanges due to the national holidays, while the *Emerging* group has not registered any significant change. On the event day, Europe is the only region to react showing at least 10% significant AARs of -4,1%.

Table 1 Average abnormal returns at sub-sample level

Day	<i>Titans</i>			<i>Europe</i>			<i>Emerging</i>		
	AAR	Adj.Std CSect Z	Gen. Rank T	AAR	Adj.Std CSect Z	Gen. Rank T	AAR	Adj.Std CSect Z	Gen. Rank T
-3	0.006	0.678	0.358	-0.031	-3.619**	-2.747***	0.003	0.214	-0.004
-2	-0.008	-0.699	-0.772	0.014	1.452	0.994	-0.002	-0.454	-0.683
-1	0.009	-0.114	-0.259	-0.005	-0.665	-0.589	0.008	0.989	0.460
0	-0.021	-1.391	-1.231	-0.041	-2.347**	-1.918*	-0.008	-0.829	-0.779
1	0.018	2.592**	1.702**	0.022	2.226**	1.738*	0.013	1.111	0.448
2	0.001	0.026	-0.290	0.035	1.229	0.606	0.021	1.842*	1.043
3	0.010	-0.401	-0.571	-0.000	-0.588	-0.148	0.010	0.946	0.504
4	0.023	3.206***	2.178*	0.038	1.167	0.818	0.011	0.923	0.352
5	0.035	2.330**	1.755**	0.001	-0.606	-0.310	0.016	1.325	0.614
6	0.014	2.229**	1.552	0.01	-0.180	-0.344	-0.004	-0.462	-0.484
7	-0.000	0.706	0.552	0.008	0.161	0.115	0.002	0.108	-0.108
8	-0.005	-1.503	-1.400	0.019	2.143**	1.734*	-0.021	-1.513	-1.327
9	-0.011	-2.322**	-2.051**	-0.006	0.577	0.399	0.003	-0.065	-0.173
10	0.020	1.947*	1.339	0.011	0.572	0.712	0.013	1.420	0.771

Notes: *, **, *** indicate 10, 5 and 1 percent level of significance.

The first day after the invasion brought a significant and positive increase in the share price of 1,8% and 2,2% in *Titans* and *Europe* sub-samples, respectively. Additionally, we observe a significant and positive increase in AARs in Europe on the 8th post-event day, which coincides with the EU Commission's call for a rapid phase-out of Russia's fossil fuels. Three consecutive days of positive AARs can be observed for *Titans* starting from the 4th day post-even. Turning our attention to the *Emerging* sub-group, we can note that firms in the sample were almost unaffected when considered at the aggregated level. However, results change when we pull out China and India, which represent 60% and 16% of the sub-sample respectively, and which are the two largest coal-producing countries. In particular, we can observe at least a 10% significant change in AARs for India confirmed by both tests: a 5,6% decrease on the event day and a 5% and 3,2% increase in the share prices on days 1 and 2. China shows no significant changes in returns, while the remaining countries in the sub-sample exhibit a 1% significant increase in AARs on days 2 (2.7%) and 4 (2.2%).

The analysis of cumulative average abnormal returns for the 14-day event window shows positive CAARs across all examined sub-sampled, albeit achieving statistical significance exclusively within the category of *Titans*. Thus, we can confirm that coal-related firms headquartered in the USA, the UK, Canada and Australia outperformed the stock market on a cumulative basis following the outbreak of the war. To enhance the robustness of our findings, we conducted a comprehensive analysis,

exploring various sub-sample configurations and employing diverse benchmark indexes for abnormal return calculations. No contradictory outcomes were observed.

4 Conclusions

This paper analyses the impact of the Russo-Ukrainian war on the global coal markets by using the event-study approach. The results show significant and positive ARs for the Titans sub-sample on a cumulative basis within the window $[-3;10]$. The European region has reacted negatively at $t=-3$ and $t=0$. In the longer run, the CAARs are positive, albeit not statistically significant. As for the *Emerging* sub-sample, no significant reaction has been detected for the defined event window.

The post-event window analysed in the paper is relatively brief, thus is focused on the immediate market response and does not include the impact of supply chain reshuffling and other processes that followed the EU's decision to ban coal imports from Russia. The results suggest a further investigation with an extended event window for a more comprehensive understanding of the multifaceted impact of the Russo-Ukrainian war on global coal markets.

References

1. Bentley, A. R., Donovan, J., Sonder, K., Baudron, F., Lewis, J. M., Voss, R., and Govaerts, B. (2022). Near-to long-term measures to stabilize global wheat supplies and food security.
2. Campbell C.J., Cowan A.R., Salotti V. Multi-country event-study methods, *Journal of Banking & Finance*, Volume 34, Issue 12, 2010, <https://doi.org/10.1016/j.jbankfin.2010.07.016>.
3. Choudhry, T. (2010). World War II events and the Dow Jones industrial index. *Journal of Banking & Finance*, 34(5), 1022–1031.
4. European Commission (2022). Communication from the Commission: Joint European Action for more affordable, secure and sustainable energy, 8 March 2022, COM/2022/108 final
5. Ferriani, F. and Gazzani, A.G.: The Impact of the War in Ukraine on Energy Prices: Consequences for Firms' Financial Performance (September 12, 2022).
6. Frey, B., and Kucher, M. (2001). Wars and markets: How bond values reflect the Second World War. *Economica*, 68(271), 317-333.
7. Hudson, R., and Urquhart, A. (2015). War and stock markets: The effect of World War Two on the British stock market. *International Review of Financial Analysis*, 40, 166-177.
8. IEA (2022), Coal 2022, IEA, Paris <https://www.iea.org/reports/coal-2022>
9. Martins, A. M., Correia, P., and Gouveia, R. (2023). Russia-Ukraine conflict: the effect on European banks' stock market returns.
10. Martin, W., and Minot, N. (2022). The impacts of price insulation on world wheat markets during the 2022 food price crisis. *Australian Journal of Agricultural and Resource Economics*.
11. Poursina, D., Schaefer, K. A., Hilburn, S., and Johnson, T. (2023). Economic impacts of the Black Sea Grain Initiative. *Journal of Agricultural Economics*.
12. Umar, M., Riaz, Y., Yousaf, I.: Impact of Russian-Ukraine war on clean energy, conventional energy, and metal markets: Evidence from event study approach, *Resources Policy*, Vol.79
13. Yanguas-Parra, P., Malz, N., Corral M., Hauenstein, C., Furnaro, A. (2023). Perspective: How a short-term relapse to coal could put exporting countries and just transition processes at risk. *Energy Research & Social Science*. 97. 10.1016/j.erss.2023.102989.

Labour market and time series: a forecast approach for European countries from 1995 to 2022

Paolo Mariani, Andrea Marletta and Piero Quatto

Abstract In statistical literature, many contributions investigated the relation between macro-economic variables and the employment rates obtained from the Labour Force Survey. The main aim of the work is to detect the presence of an economic growth, measured by GDP, followed by a positive dynamic of the considered employment rates in some EU countries. A three-way data analysis approach has been proposed to track the evolution of the relationship during time in each country. An original approach based on the superior influence of the most recent observations has been used to obtain predictions for the future coordinates of the trajectories. The proposed method also provided prediction intervals in order to display a measure of the prediction error.

Key words: labour market, trajectory analysis, three-way data, inclusiveness

1 Introduction

This study provides a preliminary analysis to detect potential relationships between economic growth and a number of economic outcomes related to the job market and inequality. The idea is that if a relationship exists, this could represent a signal of an economic inclusiveness in some European countries.

The aim is to establish whether the behaviour of selected variables support the hypothesis of inclusive growth dynamics, as already proposed by Europe 2020, and confirmed by Agenda 2030 in the Sustainable Development Goal 8 providing a strategy for a smart, sustainable and inclusive growth.

This work investigates the relationship between economic growth and factors promoting inclusiveness in the EU countries. According to the European

Commission, inclusiveness is defined as "How citizens and groups can interact and participate in open policy and decision making". In this context, inclusiveness is intended as the capability to make young and women participant to the production of national wealth and the decrease of inequality. For this reason, three main dimensions of inclusiveness has been considered: employment, gender equal opportunities and income distribution.

2 Data and methods

Gross Domestic Product is here intended as indicator of economic growth, it is achieved in chain linked volumes and divided for the total population. On the other hand, the following economic indicators here are used as inclusiveness measures: Gini Index, Total Employment Rate, Female Employment Rate, Young (15-24) years Employment Rate.

Data source for Gross Domestic Product and Employment Rates (total, female and young) is Eurostat. Data about Gini Index are from EU-SILC EUropean Survey on Income and Living Conditions. Data are selected for 7 European Countries (Italy, France, Germany, Greece, Ireland, Portugal, Spain) from 1995 to 2022.

From a methodological point of view, this is a customary situation in which, the data structure is on 3 levels: units, variables and time. In presence of three-way data, a typical approach is given by D'Urso (2000), that represents data using a "slice" scatterplot (See Figure 1). For each occasion (year) k , we have a matrix given by I units $\times$ X variables, named "slice" (D'Urso and Vichi, 1998).

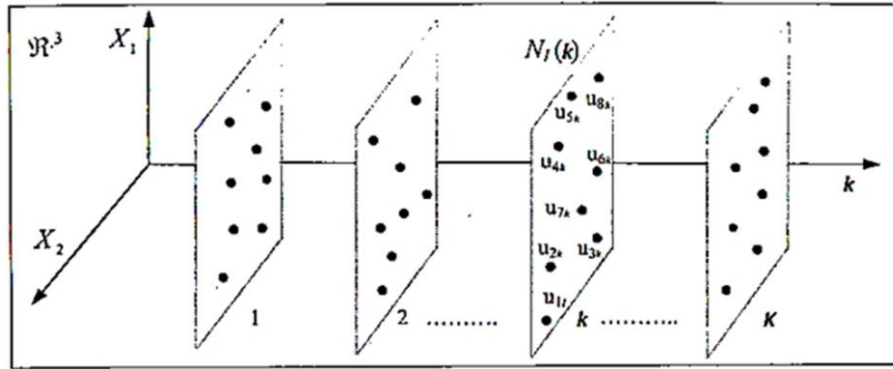


Figure 1: Example of scatter plot of three-way data. Source: D'Urso (2000)

Before applying some statistical models, data need to be re-arranged in a bi-dimensional pattern. The K slices are stacked to obtain a single matrix M , whose dimension is $IK \times X$. Once obtained, data in two dimensions, a principal component

analysis (PCA) is performed to reduce the number of variables. The distribution of units for each occasion can be visualized in the space spanned by the first two principal components, i.e. the compromise plane (Carlier, 1986; D'Urso and Vichi, 1998).

The dynamic path of a unit can be displayed by connecting its points for different occasions (Coppi and D'Urso, 2001; Liberati and Mariani, 2012).

The scores obtained from PCA could be considered a short time series of vectors $x_0, x_1, \dots, x_t$ and the average velocity can be defined as:

$$V_t = \frac{1}{t} \sum_{i=0}^{t-1} v_i^t$$

with $v_0^0 = 0$ and

$$v_i^t = \frac{x_t - x_i}{t - i}$$

representing the rate of change.

In this framework, a possible forecast is proposed:

$$\hat{x}_{t+1} = x_t + V_t = x_t + \frac{H_t}{t} (x_t - G_t)$$

in which V_t expresses the tendency of the time series, H_t is the t -th harmonic number and G_t provides a suitable synthesis assuming that the farther the observations are from the time t , the less they should influence it.

In addition, a prediction interval is given by:

$$x_t + V_t \pm z_{1-\alpha/2} \hat{\sigma}_t$$

for the case of a one-dimensional Gaussian process such that

$$x_{t+1} - x_t - V_t \sim N(0, \sigma^2),$$

where

$$\sigma^2 = \text{Var}(x_{t+1} - x_t - V_t)$$

can be unbiasedly estimated by

$$\hat{\sigma}_t^2 = \frac{1}{t-1} \sum_{i=1}^{t-1} (x_{i+1} - x_i - V_i)^2 \quad (t = 2, 3, \dots),$$

because of

$$E(\hat{\sigma}_t^2) = \frac{1}{t-1} \sum_{i=1}^{t-1} E[(x_{i+1} - x_i - V_i)^2] = \sigma^2.$$

In this case, the level of the prediction interval is given by

$$P(x_t + V_t - z_{1-\alpha/2} \hat{\sigma}_t < x_{t+1} < x_t + V_t + z_{1-\alpha/2} \hat{\sigma}_t) = P\left(\frac{|x_{t+1} - x_t - V_t|}{\hat{\sigma}_t} < z_{1-\alpha/2}\right) \cong 1 - \alpha$$

since $z_{1-\alpha/2}$ represents the standard Normal quantile of order $1 - \alpha/2$.

3 An application on some European countries

Using the proposed dataset, let $I = 7$ the number of European countries, let $X = 5$ the number of macroeconomic variables and $K = 28$ years from 1995 to 2022. Instead of the simple PCA, a Weighted Factor Analysis (WFA) has been applied using the country population as a weight. This technique combined with the re-arrangement of the three-way data allows to achieve three possible evidences: similarities between countries, similarities between macroeconomic variables and evolution of the countries over the considered period using the time trajectory.

The WFA conducted on the entire dataset, detected 2 latent components grouping the macroeconomic variables. The first component involves the three employment rates, and the second one is related to GDP and Gini index. The total variance explained by the first 2 components is 74.6%.

For space reasons, here in Figure 2, results are represented only for Ireland.

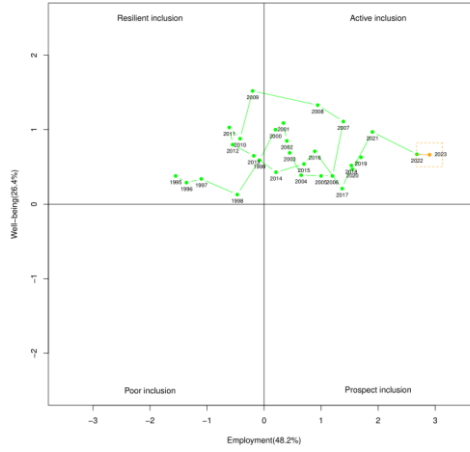


Figure 2: Time trajectory for Ireland from 1995 to 2020 of the two components given by WFA. Source: Own elaborations on Eurostat data.

From the trajectory analysis, it is possible to note that positive movements are present until 2007, after there are ten years of cyclic movements, while in last five years, the trajectory enters in the active inclusion quarter. The obtained forecast for 2023 returns very positive values and the 95% confidence interval forecast for 2023 implies the coordinates of 2022. The motivations of these movements could be due to expansive policies thanks to favourable taxation for foreign enterprises and to new treaties for a durable peace between the Eire and Northern Ireland.

4 Conclusions

The dynamic factor analysis has been used as an exploratory approach to verify the presence of an inclusive growth for some European countries from 1995 to 2022. The trajectory analysis showed interesting results synchronized to macro-economic events e.g. the impact of economic crisis is evident on the trajectories which tend to shrink at that time. Only some countries show a clear path toward inclusive growth but forecast for 2023 underline a tendency of improvement. The application of this approach is dependent on the share of variance explained by the first two principal components. Future research could concern the study of the co-movements of the economic time series using a cointegration approach and introduction of concept of acceleration to obtain long-run forecasts.

References

1. Carlier, A. (1986). Factor analysis of evolution and cluster methods on trajectories. In *COMPSTAT: Proceedings in Computational Statistics, 7th Symposium held in Rome 1986* (pp. 140-145). Physica-Verlag HD.
2. Coppi, R., & D'Urso, P. (2001). The geometric approach to the comparison of multivariate time trajectories. In *Advances in Classification and Data Analysis* (pp. 93-100). Springer Berlin Heidelberg.
3. D'Urso, P.: Dissimilarity measures for time trajectories. *J. Ital. Stat. Soc.* 6, 53–83 (2000)
4. D'Urso, P., & Vichi, M. (1998). Dissimilarities between trajectories of a three-way longitudinal data set. In *Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98) Università "La Sapienza", Rome, 21–24 July, 1998* (pp. 585-592). Springer Berlin Heidelberg.
5. Liberati, C., & Mariani, P. (2012). Banking customer satisfaction evaluation: a three-way factor perspective. *Advances in Data Analysis and Classification*, 6, 323-336.

15. Contributed - Innovations in cluster and latent class models

- 15.1 - Biclustering of discrete data by extended finite mixtures of latent trait models
- 15.2 - Seismic events classification through latent class regression models for point processes
- 15.3 - Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization

Biclustering of discrete data by extended finite mixtures of latent trait models

Dalila Failli, Maria Francesca Marino, and Francesca Martella

Abstract We aim at performing a joint clustering of units and variables in a binary data matrix in a biclustering perspective. In this framework, units are partitioned into clusters (components) via a finite mixture approach; in each component, variables are partitioned into clusters (segments) by adopting a flexible specification of the linear predictor. Dependence between variables is modeled via a multidimensional, continuous, latent trait. The proposed model is applied to the Regensburg Pediatric Appendicitis data set, with the aim of identifying homogeneous groups of pediatric patients with respect to subsets of clinical features.

Key words: Model-based clustering, Finite mixtures, Concomitant variables, Gaussian quadrature, EM algorithm

1 Introduction

Biclustering [6, 7, 8, 2] entails the joint clustering of units and variables in a data matrix. The literature on this topic is now rather extensive and offers the possibility to choose between exploratory and model-based approaches. When looking at these latter, biclustering can be considered as an extension of standard finite mixture-based clustering methods. In this paper, we focus the attention on the Mixtures of

Dalila Failli

Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze e-mail: dalila.failli@unifi.it

Maria Francesca Marino

Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze e-mail: mariafrancesca.marino@unifi.it

Francesca Martella

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Roma e-mail: francesca.martella@uniroma1.it

Latent Trait Analyzers (MLTA) model, introduced in the literature by [5]. Here, the aim is that of clustering units via a finite mixture specification and model the residual dependence between variables via a continuous, multidimensional, latent trait. We propose to modify the MLTA to allow for a joint clustering of units and variables: the former are partitioned into subsets called components and, in each of them, the latter are partitioned into subsets called segments. This goal is achieved by considering a finite mixture of latent trait analyzers providing a partitioning of units, and a flexible specification of the linear predictor for partitioning variables, following an approach similar to that detailed by [10, 1]. Furthermore, according to the proposal by [4], we allow for the inclusion of covariates on the latent layer of the model to detect how these influence component formation.

The proposal is applied to the Regensburg Pediatric Appendicitis data set [9], entailing information on clinical features from a cohort of pediatric patients with suspected appendicitis admitted to Children’s Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021. The aim is twofold: 1) identifying groups of patients with similar clinical conditions for specific subsets of features; 2) taking into account how patients’ characteristics influence the probability of belonging to a particular component.

2 Data

The Regensburg Pediatric Appendicitis data set involves 782 pediatric patients and 54 clinical features. Information on the results of ultrasound (US) images, laboratory tests, and clinical scores are recorded for each patient. By excluding missing data, we obtain a data matrix of 542 patients and 24 clinical features. Table 1 shows a detailed description of the features considered in the analysis. As regards patients’ characteristics, the following covariates are considered: age, BMI, gender, height, weight, length of stay in the hospital, management (conservative, primary surgical, or secondary surgical), severity (uncomplicated, complicated).

3 The proposal

Let Y_{ik} denote the k -th binary random variable (item) and y_{ik} the corresponding observed value for the i -th unit, with $i = 1, \dots, N$, and $k = 1, \dots, R$. Without loss of generality, we assume that y_{ik} is equal to 1 if unit i provides a positive answer to the binary item k and $y_{ik} = 0$ otherwise. The original MLTA model assumes that every unit belongs to one out of G unobserved components identified by a discrete latent variable $\mathbf{z}_i = (z_{i1}, \dots, z_{iG})' \stackrel{iid}{\sim} \text{Multinomial}(1, (\eta_1, \dots, \eta_G))$, whose generic element is $z_{ig} = 1$ if unit i belongs to component g , and $z_{ig} = 0$ otherwise. Parameter η_g denotes the probability that a randomly selected unit belongs to component g , under the constraints that $\sum_{g=1}^G \eta_g = 1$ and $\eta_g \geq 0$, $g = 1, \dots, G$.

Table 1 Variables description.

Alvarado Score	Score based on symptoms, signs, and laboratory test. Value 1 is assigned if the score is greater than 3.
Pediatric Appendicitis Score	Score based on patient's history and examination. Value 1 is assigned if the score is greater than or equal to 6.
Appendix on US	Detectability of the vermiform appendix during sonographic examination.
Migratory Pain	Abdominal pain from epigastrium to the right lower quadrant.
Lower Right Abdominal Pain	Right iliac fossa pain detected on palpation.
Contralateral Rebound Tenderness	Pain of the contralateral side during the release of pressure over the abdomen.
Coughing Pain	Abdominal pain by forced cough.
Nausea	Feeling of sickness/ejection of contents from stomach through the mouth.
Loss of Appetite	Lack or reduction of appetite.
Body Temperature	Measured by a thermometer placed in the rectum or in the auditory canal. Value 1 is assigned if the body temperature is greater than 37.36.
WBC Count	Number of leucocytes in a unit volume of blood. The normal number is 4,500 to 11,000 WBCs per microliter. Value 1 is assigned if the count is greater than 11 or lower than 4.5.
Neutrophil Percentage	Mature WBC in the granulocytic series. A normal result is between 40% and 60%. Value 1 is assigned if the percentage is greater than 60 or lower than 40.
Neutrophilia	Relative neutrophilic leucocytosis, often a result of a bacterial infection.
RBC Count	Number of erythrocytes in a unit volume of blood. The standard value for children is between 4.0 million to 5.5 million red blood cells per microliter of blood. Value 1 is assigned if the count is greater than 5.5 or lower than 4.
Hemoglobin	Hemoglobin level. Standard value for children is between 10 g/dL and 15 g/dL. Value 1 is assigned if the value is greater than 15 or lower than 10.
RDW	A blood test that measures the differences in the volume and size of the erythrocytes. The normal RDW in children is between 12.3 and 14.1. Value 1 is assigned if it is greater than 14.1 or lower than 12.3.
Thrombocyte Count	Number of platelets in a unit volume of blood. In healthy pediatric subjects normal count ranges between 250,000 μ L and 450,000 μ L. Value 1 is assigned if the count is greater than 450 or lower than 250.
CRP	Protein produced by the liver. Standard values are no higher than 10 mg/L. Value 1 is assigned if the value is greater than 10.
Dysuria	Pain or other difficulty during urination.
Stool	Characteristics of bowel movements. Value 1 is assigned if the patient has diarrhea or constipation.
Peritonitis	Spasm of abdominal wall muscles detected on palpation.
Psoas Sign	Abdominal pain produced by extension of the hip.
Free Fluids	Free fluids inside the abdomen.

Furthermore, the MLTA assumes the existence of a D -dimensional, unit-specific, continuous, latent trait ($\mathbf{u}_i$) distributed according to a standard D -variate Gaussian density, i.e. $\mathbf{u}_i \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I})$. This is assumed to capture the heterogeneity of units in the way they respond (positively or not) to the different items. Conditional on $\mathbf{z}_i$ and $\mathbf{u}_i$, responses in the vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iR})'$ are assumed to be independent Bernoulli random variables with probabilities $\pi_{gk}(\mathbf{u}_i)$, $k = 1, \dots, R$, modeled through the following logistic function:

$$\text{logit}(\pi_{gk}(\mathbf{u}_i)) = b_{gk} + \mathbf{w}'_{gk} \mathbf{u}_i. \quad (1)$$

To perform a joint clustering of units and variables, we propose to modify the logistic function in Equation (1) as

$$\text{logit}(\pi_{gk}(\mathbf{u}_i)) = b_g + \mathbf{a}'_{gk}(\boldsymbol{\mu} + \mathbf{u}_i).$$

Here, the parameter b_g is a component-specific effect providing a baseline propensity for units in the g -th component to positively respond to the R items, while $\boldsymbol{\mu}$ is a D -dimensional vector of fixed parameters which are assumed to be constant across components. Last, $\mathbf{u}_i \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I})$ is a D -dimensional, continuous latent trait associated to unit i and $\mathbf{a}_{gk}$ is a D -dimensional membership indicator for the k -th variable in the g -th component. In detail, $\mathbf{a}_{gk}$ is a row stochastic vector ($D \leq R$), whose generic element, a_{gkd} , is equal to 1 if within the g -th component, the k -th variable belongs to segment d , and is equal to 0 otherwise. That is, it allows to select one out of the D segment-specific effects in the vectors $\boldsymbol{\mu}$ and $\mathbf{u}_i$. Overall, for a variable k belonging to the d -th segment, success probabilities are modeled as

$$\pi_{gk}(u_{id}) = \frac{1}{1 + \exp[-(b_g + (\boldsymbol{\mu}_d + \mathbf{u}_{id}))]}.$$

To conclude, we follow the proposal by [4] and allow the component probabilities η_g to vary across units via the following logit specification:

$$\eta(\mathbf{x}_i; \boldsymbol{\beta}_g) = \Pr(z_{ig} = 1 \mid \mathbf{x}_i; \boldsymbol{\beta}_g) = \frac{\exp\{\mathbf{x}_i' \boldsymbol{\beta}_g\}}{1 + \sum_{g'=2}^G \exp\{\mathbf{x}_i' \boldsymbol{\beta}_{g'}\}}, \quad g = 2, \dots, G.$$

Here, $\boldsymbol{\beta}_g$ denotes a J -dimensional vector of coefficients measuring the impact of covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{iJ})'$ on $\eta(\mathbf{x}_i; \boldsymbol{\beta}_g)$.

Let $\boldsymbol{\theta}$ represent the vector of all free model parameters. The likelihood function of the model can be written as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N L_i(\boldsymbol{\theta}) = \prod_{i=1}^N \left(\sum_{g=1}^G \eta(\mathbf{x}_i; \boldsymbol{\beta}_g) \int_{R^D} f(\mathbf{y}_i \mid \mathbf{u}_i, z_{ig} = 1) f(\mathbf{u}_i) d\mathbf{u}_i \right), \quad (2)$$

where $f(\mathbf{y}_i \mid \mathbf{u}_i, z_{ig} = 1) = \prod_{k=1}^R f(y_{ik} \mid \mathbf{u}_i, z_{ig} = 1) = \prod_{k=1}^R (\pi_{gk}(\mathbf{u}_i))^{y_{ik}} (1 - \pi_{gk}(\mathbf{u}_i))^{1-y_{ik}}$ denotes the product of Bernoulli probability mass functions. Since the D -dimensional integral in Equation (2) cannot be computed analytically, we approximate it via Gauss-Hermite quadrature. Maximum likelihood estimates of model parameters can be obtained by maximizing the log-likelihood function via an extended EM algorithm [3].

The central issue of partitioning units and variables is addressed by assigning the former to components according to a Maximum a Posteriori rule based on the posterior component probabilities $\hat{z}_{ig}$; variables are instead partitioned by directly relying on the maximum likelihood estimate of $\mathbf{a}_{gk}$.

Simulation results have shown the effectiveness of the proposed approach in recovering the true value of model parameters, as well as in partitioning the rows and the columns of the observed data matrix.

4 Results

The proposed model is estimated for a number of components, G , and segments, D , ranging from 1 to 4. The optimal specification is the one providing the smallest BIC value, and corresponds to $G = 2$ components and $D = 3$ segments. In Figure 1 and 2, we report the original data matrix and the predicted probabilities $\hat{\pi}_{gk}(\mathbf{u}_i)$, respectively. By looking at this latter figure, we may recognize two homogeneous components and three homogeneous segments.

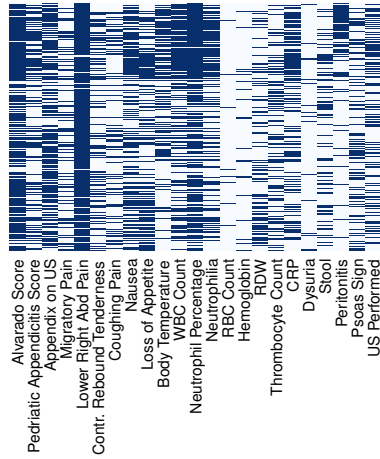


Fig. 1 Data matrix.

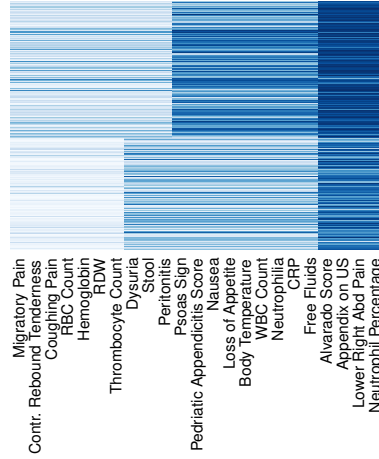


Fig. 2 Predicted probabilities $\hat{\pi}_{gk}(\mathbf{u}_i)$.

For a better interpretation of such partitions, we report in Figure 3 the distribution of the estimated probabilities $\hat{\pi}_{gk}(\mathbf{u}_i)$. Looking at these distributions, it is evident that, given a segment, the probability of clinical manifestations is generally lower for patients belonging to the first component and higher for those belonging to the second one. On the other hand, by focusing on clinical features, we identify a first segment consisting of the least manifested conditions in each component, a second segment consisting of slightly more manifested conditions, and a third segment consisting of the most manifested conditions.

Regarding the effect of concomitant variables on component formation, results of the analysis suggest that the chance of belonging to the second component (more severe clinical conditions) is higher for males, primary and secondary surgical patients, and increases with weight. In contrast, this probability decreases with BMI, height, and length of stay and it is lower for patients with uncomplicated severity.

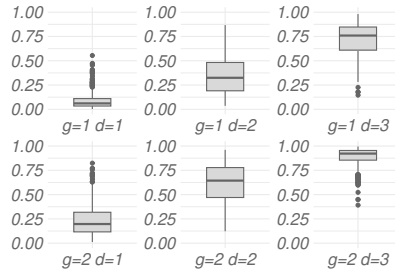


Fig. 3 Distribution of the estimated probabilities $\hat{\pi}_{gk}$ for $g = 1, \dots, 2$ and $d = 1, \dots, 3$.

5 Conclusions and further developments

We extend the MLTA to perform a joint clustering of units and variables of a binary data matrix, in a model-based biclustering perspective. A finite mixture specification allows us a clustering of units and, within each of them, a flexible specification of the linear predictor including a unit-specific, continuous, latent trait, allows for a clustering of variables too. Simulation results and the application of to data on pediatric patients potentially affected by appendicitis show the effectiveness of the proposed approach. An extension of this allowing for the biclustering of both qualitative (binary) and quantitative data is a current research topic.

References

1. Alfö, M., Marino, M. F., Martella, F.: Biclustering multivariate discrete longitudinal data, *Stat. Comput.* **34**, 1–21 (2023)
2. Bock, H.: Simultaneous clustering of objects and variables. In: Tomassone, R. (eds.) *Anal. des donnees es et Inform.*, 187–204. INRIA, Le Chesnay, France (1979)
3. Dempster, A.P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc., Ser. B*, **39**, 1–38 (1977)
4. Failli, D., Marino, M.F., Martella, F.: Finite mixtures of latent trait analyzers with concomitant variables for bipartite networks: an analysis of COVID-19 data. *Multivariate Behavioral Research*, *in press* (2024)
5. Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* **24**, 569–588 (2014)
6. Good, I.: *Categorization of Classification. Mathematics and Computer Science in Biology and Medicine*. Her Majesty's Stationary Office, London (1965)
7. Hartigan, J.A.: Direct clustering of a data matrix. *J. Am. Stat. Assoc.* **67**, 123–129 (1972)
8. Hartigan, J.A.: *Clustering Algorithms*. Wiley, London (1975)
9. Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Ozkan, E., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Knorr, C., Vogt, J. E.: *Regensburg Pediatric Appendicitis Dataset (1.01)*. Zenodo. (2023) <https://doi.org/10.5281/zenodo.7669442>
10. Martella, F., Alfö, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. *J. Stat. Comput. Simul.* **87**, 2186–2206 (2017)

Seismic events classification through latent class regression models for point processes

Giada Lo Galbo, Giada Adelfio, Marcello Chiodi

Abstract We are trying to identify sub-processes of seismic events from the point processes' point of view and according to the latent class regression approach. Each seismic event is classified as membership of one of the 4 identified sub-classes of seismic sequences, each defined by particular and well-defined characteristics. So far, seismic sub-sequences have been identified and described according to several declustering methods. In this application, we show how sub-processes can be identified starting from the definition of a spatio-temporal intensity function for point processes, assuming independence of the past.

Key words: Latent class, mixture model, spatio-temporal point process, clustering, earthquake, seismic sequence

1 Introduction

The study of seismic sequences has been approached to identify sub-processes characterizing the background and the induced components [1]. What has been studied so far concerns the characteristics of the magnitude distribution corresponding to the cited above components [11], or the covariates affecting the induced component [3]. What has not been taken into account is the importance of covariates related to the occurrence of earthquakes. This study aims to identify the components of seismic processes, through the application of a latent class regression model, by analyzing the spatio-temporal intensity dependence on covariates.

Giada Lo Galbo · Giada Adelfio · Marcello Chiodi
Department of Economics, Business and Statistics, University of Palermo e-mail: giada.logalbo@unipa.it, e-mail: giada.adelfio@unipa.it, e-mail: marcello.chiodi@unipa.it

2 Methodology

2.1 Spatio-temporal point process

A spatio-temporal inhomogeneous Poisson point process has parametric probability density function, $f_X(X; \boldsymbol{\theta})$, defined within a spatio-temporal bounded window, $|W \times T|$, with volume $W > 0$ and length $T > 0$ as in Eq. (1)[4]:

$$f_X(X; \boldsymbol{\theta}) = e^{\int_W \int_T [1 - \lambda_{\boldsymbol{\theta}}(\mathbf{u}, t)] dt d\mathbf{u}} \prod_{(\mathbf{u}, t) \in X} \lambda_{\boldsymbol{\theta}}(\mathbf{u}, t) \quad \{(\mathbf{u}, t)\} \in X \quad \boldsymbol{\theta} \in \Theta \quad (1)$$

where: $\boldsymbol{\theta}$ is the set of regression parameters; Θ is the parameter space and $\lambda_{\boldsymbol{\theta}}(\mathbf{u}, t)$ is the parametric first-order intensity function describing the point pattern X .

Assuming log-linear dependence from a set of covariates, $\mathbf{Z}(\mathbf{u}, t)$, the first-order intensity function of the $(\mathbf{u}, t)^{th} \in X$ point is specified as in Eq. (2) [10]:

$$\lambda_{\boldsymbol{\theta}}(\mathbf{u}, t) = \exp \{ \boldsymbol{\theta}' \mathbf{Z}(\mathbf{u}, t) + \varrho(\mathbf{u}, t) \} \quad (2)$$

with $\varrho(\mathbf{u}, t)$ a scalar offset. The parameters, $\boldsymbol{\theta}$, are estimated by maximizing the log-likelihood function, $\ell_X(\boldsymbol{\theta}; X)$. Berman and Turner[5] propose a device, which adds a set of dummy points, X_d , within the spatio-temporal window, $|W \times T|$, where there are no observed points; i.e., the latter can be approximated as in Eq. (3):

$$\ell_X(\boldsymbol{\theta}; X) = \sum_{(\mathbf{u}, t) \in X_D} \ell_X(\boldsymbol{\theta}; (\mathbf{u}, t)) \approx \sum_{(\mathbf{u}, t) \in X_D} w_{(\mathbf{u}, t)} \{ y_{(\mathbf{u}, t)} \log [\lambda_{\boldsymbol{\theta}}(\mathbf{u}, t)] - \lambda_{\boldsymbol{\theta}}(\mathbf{u}, t) \} \quad (3)$$

where: $\sum_{(\mathbf{u}, t) \in X_D} w_{(\mathbf{u}, t)} = |W \times T|$; $X_D = X \cup X_d$; $w_{(\mathbf{u}, t)}$ is a quadrature weight of the $(\mathbf{u}, t)^{th}$ point; $y_{(\mathbf{u}, t)} = \frac{z_{(\mathbf{u}, t)}}{w_{(\mathbf{u}, t)}}$, where $z_{(\mathbf{u}, t)} = 1$ if $(\mathbf{u}, t) \in X$, $z_{(\mathbf{u}, t)} = 0$ if $(\mathbf{u}, t) \in X_d$.

2.2 Finite mixtures of log-linear regression models

A finite mixture model of R components has conditional density function, $g_X(X | \boldsymbol{\psi})$, defined by a whole set of parameters $\boldsymbol{\psi} = \{\boldsymbol{\theta}_r, \pi_r\}_{r=1}^R$, as expressed in Eq. (4)[7, 8]:

$$g_X(X | \boldsymbol{\psi}) = \sum_{r=1}^R \pi_r f_X(X; \boldsymbol{\theta}_r) \quad \text{with: } \sum_{r=1}^R \pi_r = 1; \quad \pi_r > 0 \quad (4)$$

where: $\{\boldsymbol{\theta}_r, \pi_r\}$ are the regression and weight parameters describing the r^{th} cluster.

The posterior probability of j^{th} cluster membership, is defined as in Eq. (5):

$$\rho_{(\mathbf{u}, t)_j} = P(j | (\mathbf{u}, t), \boldsymbol{\psi}) = \frac{\pi_j f_X((\mathbf{u}, t); \boldsymbol{\theta}_j)}{\sum_{r=1}^R \pi_r f_X((\mathbf{u}, t); \boldsymbol{\theta}_r)} \quad \forall (\mathbf{u}, t) \in X \quad (5)$$

where: $\{f_X(X; \boldsymbol{\theta}_r)\}_{r=1}^R$ is the set of components belonging to the finite mixture.

The parameters estimate, $\boldsymbol{\Psi}$, is obtained by maximizing the ‘complete data log-likelihood’, $c\ell_X(\boldsymbol{\Psi}; X)$, which is expressed as in Eq. (6):

$$c\ell_X(\boldsymbol{\Psi}; X) = \sum_{(\mathbf{u}, t) \in X} \log \left[\sum_{r=1}^R \pi_r f_X((\mathbf{u}, t); \boldsymbol{\theta}_r) \right] \quad (6)$$

The maximization of the complete data log-likelihood function is carried out through the iterative Expectation-Maximization (EM) algorithm [6]. For the r^{th} cluster, at the $(i+1)^{th}$ iteration, given: $\hat{\boldsymbol{\Psi}}^{(i)} = \{\hat{\boldsymbol{\pi}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}\}$, the Expectation and the Maximization steps consist of, respectively [9]:

- **E-step** $\mathbb{E}[\hat{\pi}_r^{(i+1)}] = \frac{1}{n(X)} \sum_{(\mathbf{u}, t) \in X} \hat{\rho}_{(\mathbf{u}, t)_r}^{(i)} = \frac{1}{n(X)} \sum_{(\mathbf{u}, t) \in X} \hat{P}(r | (\mathbf{u}, t), \hat{\boldsymbol{\Psi}}^{(i)})$
- **M-step** $\hat{\boldsymbol{\theta}}_r^{(i+1)} = \arg \max_{\boldsymbol{\theta}_r} \left\{ \sum_{(\mathbf{u}, t) \in X} \hat{\rho}_{(\mathbf{u}, t)_r}^{(i)} \ell_X(\boldsymbol{\theta}_r^{(i)}; (\mathbf{u}, t)) \right\}$

3 Application to seismic data

We are interested in classifying events characterizing a spatio-temporal clustered point pattern by a probabilistic clustering approach [8]; i.e., according to an estimated probability of latent class membership [2]. The data are provided by the Istituto Nazionale di Geofisica e Vulcanologia (INGV), and belong to the Catalogo delle Localizzazioni ASSolute (CLASS). They refer to Italian earthquakes that occurred between 1980 and 2018, with information on: depth, $D_p(\mathbf{u}, t)$; root mean square error from P/S arrival time, $R(\mathbf{u}, t)$; hypocentral error on vertical, $E_v(\mathbf{u}, t)$, and horizontal, $E_h(\mathbf{u}, t)$, coordinates; gap azimuth, $G(\mathbf{u}, t)$; distance from the nearest station, $D(\mathbf{u}, t)$; distance between probabilistic and expected hypocenter, $D(\mathbf{u}, t)$; quality location, $Q(\mathbf{u}, t)$; radius of a sphere with PDF volume, $P(\mathbf{u}, t)$.

By using the intensity function in Eq. (7), we fit models with a number of latent classes from $R = 2$ to $R = 8$ and choose the model with $R = 4$ latent classes, according to the AIC and BIC criteria:

$$\lambda_{\boldsymbol{\theta}}(\mathbf{u}, t) = e^{\{\theta_0 + \theta_1 D_p(\mathbf{u}, t) + \theta_2 R(\mathbf{u}, t) + \theta_3 E_v(\mathbf{u}, t) + \theta_4 E_h(\mathbf{u}, t) + \dots \\ \dots + \theta_5 G(\mathbf{u}, t) + \theta_6 D(\mathbf{u}, t) + \theta_7 L(\mathbf{u}, t) + \theta_8 Q(\mathbf{u}, t) + \theta_9 P(\mathbf{u}, t)\}} \quad (7)$$

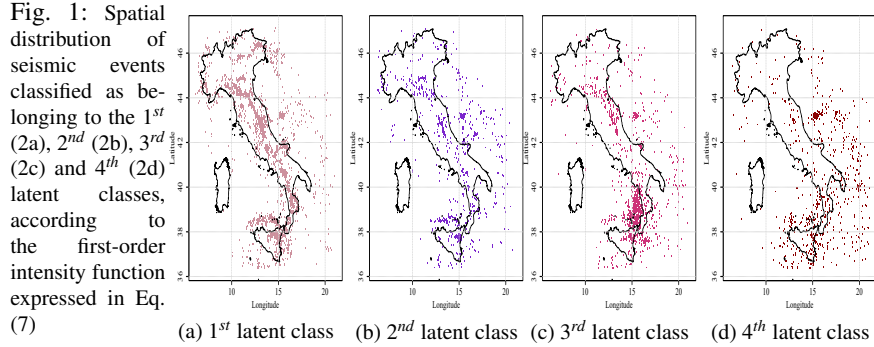
Tab. 1 shows the results of the latent class regression model fitting. As shown in Tab. 1, the covariates have the same positive (R, E_v, G, D) or negative (D_p, E_h, L, Q, P) effects, respectively, between the 1st and the 2nd latent classes, on the intensity of seismic events, although the different estimates’ values; some covariates have similar positive (D_p, G, D) or negative (Q, P) effects, on the intensity of seismic

events classified as belonging to the 3rd and the 4th latent classes. According to the regression parameters' estimates, the remaining covariates have null or opposite effects (and different magnitudes), among 3rd and 4th latent classes' intensities.

Table 1: Regression parameters estimates, $\hat{\theta}$, with standard errors, $\hat{\sigma}_{\hat{\theta}}$, and p -values, p , for the model with $R = 4$ latent classes and first-order intensity function defined as in Eq. (7)

r	θ_0	D_p θ_1	R θ_2	E_v θ_3	E_h θ_4	G θ_5	D θ_6	L θ_7	Q θ_8	P θ_9
$\hat{\theta}$	-14.677	-0.057	0.541	0.118	-0.003	0.008	0.011	-0.330	-3.086	-0.134
1 $\hat{\sigma}_{\hat{\theta}}$	0.094	0.004	0.038	0.014	0.008	0.001	0.002	0.021	0.240	0.023
p	< 0.001	< 0.001	< 0.001	< 0.001	0.718	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
$\hat{\theta}$	-14.624	-0.044	0.153	0.204	-0.301	0.006	0.012	-0.199	-2.770	-0.170
2 $\hat{\sigma}_{\hat{\theta}}$	0.094	0.003	0.067	0.010	0.020	0.001	0.002	0.013	0.279	0.030
p	< 0.001	< 0.001	0.022	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
$\hat{\theta}$	-14.788	0.004	0.089	0.062	0.056	0.011	0.005	-0.275	-4.657	-0.169
3 $\hat{\sigma}_{\hat{\theta}}$	0.096	0.001	0.034	0.006	0.007	0.001	0.001	0.018	0.255	0.021
p	< 0.001	< 0.001	0.009	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
$\hat{\theta}$	-14.946	0.006	-2.255	0.000	0.000	0.014	0.005	0.004	-4.205	-0.059
4 $\hat{\sigma}_{\hat{\theta}}$	0.115	0.000	0.281	0.001	0.001	0.001	0.001	0.001	0.256	0.007
p	< 0.001	< 0.001	< 0.001	0.847	0.977	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Figs. 1 and 2 show the spatial and the temporal distributions, respectively, of the 1st (2a, 3a), 2nd (2b, 3b), 3rd (2c, 3c) and 4th (2d, 3d) latent classes. As shown in Figs. 1 and 2, the 3rd and 4th latent classes are defined by seismic events loosely aggregated and nearer the tectonic zones around the Messina area. The 1st and 2nd latent classes identify seismic sequences which occurred in Central Italy and Northern Apennine chain.



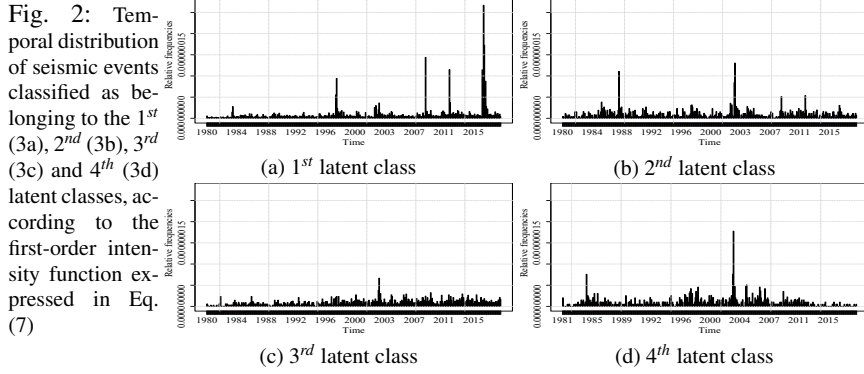


Fig. 3 and Tab. 2 show, respectively, the density functions of depth (D_p , 4d), magnitude (m , 4c), inter-time (D^T , 4a), inter-distance (D^S , 4b) and the weight parameters estimates, $\{\pi_r\}_{r=1}^R$, with the summary statistics of magnitude, m , depth, D_p , inter-time, D^T , and inter-distance, D^S , for each of the $R = 4$ latent classes.

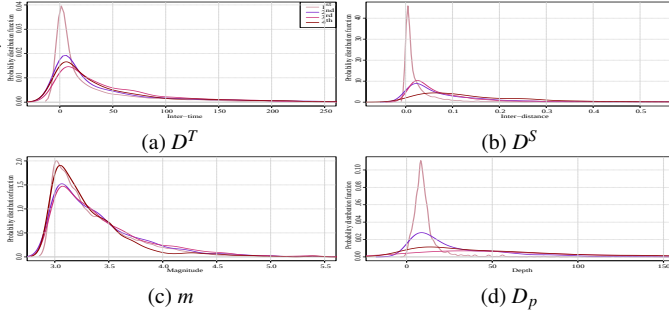
Table 2:

Weight parameters, $\{\pi_r\}_{r=1}^R$, and summary statistics of magnitude, m , depth, D_p , inter-time (hours), D^T , and inter-distance (km), D^S , for the model with $R = 4$ latent classes and first-order intensity function defined as in Eq. (7)

r	π_r	Quantile					Moment				
		Q_0	Q_1	Q_2	Q_3	Q_4	μ	σ	μ_3	β_2	
1	0.65	m	3.00	3.10	3.20	3.50	6.10	3.36	0.40	1.96	8.54
		D_p	-1.40	6.36	8.96	12.08	216.56	11.66	12.60	5.83	54.90
		D^T	0.00	0.62	8.18	38.15	1241.34	32.80	3749.80	4.94	52.49
		D^S	0.00	0.00	0.01	0.03	1.78	0.03	0.00	8.15	132.47
2	0.11	m	3.00	3.10	3.30	3.60	5.70	3.40	0.43	1.71	6.67
		D_p	-0.93	8.00	17.18	41.10	205.69	28.41	29.20	1.87	7.48
		D^T	0.00	2.76	17.15	58.85	1549.91	49.40	8070.56	6.53	87.12
		D^S	0.00	0.02	0.05	0.13	1.24	0.10	0.02	3.33	19.64
3	0.15	m	3.00	3.10	3.30	3.60	5.80	3.44	0.45	1.63	6.17
		D_p	-1.05	28.82	61.96	198.53	615.93	111.35	107.81	1.15	3.74
		D^T	0.00	8.08	29.81	69.92	690.53	52.50	4589.54	2.87	15.87
		D^S	0.00	0.02	0.05	0.11	1.89	0.10	0.02	4.77	38.36
4	0.09	m	3.00	3.10	3.20	3.50	5.40	3.32	0.37	2.11	8.72
		D_p	-0.69	13.98	40.94	82.42	558.78	64.81	75.59	2.48	11.47
		D^T	0.00	3.54	22.63	61.81	574.52	48.78	5057.82	2.71	12.80
		D^S	0.00	0.06	0.11	0.23	2.09	0.17	0.04	4.00	28.7

As shown in Fig. 3 and Tab. 2, the 1st and 2nd latent classes are in that order the most spatio-temporally aggregated and characterized by highest (lowest) maximum values of magnitude (depth). The 1st latent class is characterised by the majority of events belonging to the catalog. The 3rd and 4th latent classes of events are the least clustered and are characterized by events occurring at the highest depths. The latter groups of events correspond to deep earthquakes occurring around the Calabro-Ionian slab.

Fig. 3: Density function of inter-time, D^T (4a, hours), inter-distance, D^S (4b, km), magnitude, m (4c), and depth, D_p (4d), conditional to latent class membership



4 Conclusions

Despite the experimental purpose of the analysis, the application of a latent class regression model for point process, assuming independence of the past, allowed us to identify seismic sub-patterns. The results highlighted by the subsequent exploratory analysis on the events characterizing each latent class, are interesting. Further investigation of the identified sub-sequences, or accounting for different subsets of covariates, could lead to in-depth results concerning the seismic phenomenon.

References

- Adelfio, G., et al.: Earthquakes clustering based on maximum likelihood estimation of point process conditional intensity function. *Proceedings of the "The 4th International Workshop on Statistical Seismology"* (StatSei4), **13**, (2006)
- Adelfio, G., Chiodi, M.: Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stoch. Environ. Res. Risk Assess.*, **29**, pp. 443–450. Springer (2014)
- Adelfio, G., Chiodi, M.: Including covariates in a space-time point process with application to seismicity. *Stat. Methods. Appl.*, **30**, pp 947 – –971 (2021)
- Baddeley, A., Rubak, E., Turner, R.: *Spatial Point Patterns Methodology and Applications* with R. Chapman and Hall/CRC, New York (2015)
- Berman, M., Turner, T.R.: Approximating point process likelihoods with GLIM. *Appl. Stat.*, **41**:31 – –38, Wiley (1992)
- Dempster, A. P., Nan M. L., and Donald B. R.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**:1, pp. 1 – –22 (1977)
- DeSarbo, W. S., and William L. C.: A maximum likelihood methodology for clusterwise linear regression. *J. Classif.*, **5**, pp. 249 – –282 (1988)
- Fraley, C., Raftery, A. E.: Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.*, **97**:458, pp. 611–631. Taylor & Francis (2002)
- Leisch, F.: FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J. Stat. Softw.* **11** (2004)
- Lewis, P., A. W.: Recent results in the statistical analysis of univariate point processes. *N. P. S.* (1971)
- van Stiphout, T., Zhuang, J., Marsan, D.: Seismic declustering, Community Online Resource for Statistical Seismicity Analysis. (2012)

Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization

Stavolo Agostino¹, Grassia Maria Gabriella, Marino Marina, Mazza Rocco, Paesano Simone, Sacco Dario

Abstract Cluster analysis, as a form of unsupervised learning, has been developed to group observations by leveraging application-specific similarity measures. This study investigates matrix factorization techniques, with a specific focus on analyzing lexical tables within the framework of term-document matrices. Symmetric Non-Negative Matrix Factorization (SNMF) takes center stage as an effective tool for clustering operations. The primary challenge addressed is the automated determination of the optimal number of clusters.

Key words: clustering, symmetric non-negative matrix factorization

1 Introduction

The transition from analog to digital data, driven by advancements in information technology and the Internet's growth, has generated vast data volumes across domains. Textual datasets pose challenges due to high dimensionality and sparsity, demanding significant computational resources and impacting result generalizability. Analyzing large corpora faces high dimensionality issues, as lexical tables form large, sparse matrices. This data noise must be minimized to reduce computational complexity and ensure reliable results in associative text relationship

¹ Stavolo Agostino, University of Naples Federico II, agostino.stavolo@unina.it
Grassia Maria Gabriella, University of Naples Federico II, mariagabriella.grassia@unina.it
Marino Marina, University of Naples Federico II, marina.marino@unina.it
Mazza Rocco, Univeristy of Bari Aldo Moro, rocco.mazza@uniba.it

analyses. Various approaches exist for matrix dimensionality reduction and automatic information extraction from document collections.

Factorial-based Approach: Utilizes factorization techniques like LCA [9] and LSA [3] based on GSVD to minimize the number of terms needed for describing documents in a low-rank vector space.

Network-based Approach: Represents matrices as graphs to visualize contextual relationships between terms [12], overcoming Bag-of-Words encoding limitations. Community detection methods are commonly used [5].

Probabilistic Approach: Addresses dimensionality reduction through probabilistic models such as topic modeling (pLSA [6] and LDA [1]), enabling the discovery of underlying topics within a corpus.

Non-Negative Factorization (NMF): Utilizes NMF to achieve reduced dimensionality through additive combinations of original terms, with applications in various fields [10]. However, its effectiveness in clustering is limited for nonlinear cluster structures, leading to the introduction of symmetric Non-Negative Matrix Factorization (SNMF).

The contribution suggests utilizing consensus clustering to determine the ideal number of semantic clusters for delineating the primary themes within Ursula von der Leyen's communication, employing Symmetric Non-Negative Matrix Factorization.

2 Symmetric Non-Negative Matrix Factorization

The objective of Symmetric Non-Negative Matrix Factorization (SNMF) is to approximate a non-negative symmetric matrix A by expressing it as the product of a non-negative matrix H and its transpose H^T , resulting in $A \approx HH^T$.

During clustering operations, outcomes are obtained directly by marking the number of columns corresponding to the maximum value in each row of H [15], making H also known as the clustering assignment matrix. SNMF, falling under the category of soft graph clustering, relaxes the orthogonality constraint on the H matrix, resulting in approximately orthogonal matrices and fuzzy clustering results, distinguishing it from traditional hard graph clustering methods. Due to its advantages in data clustering, SNMF has gained significant attention, leading to the development of various algorithms. Researchers have recognized SNMF's superior performance in clustering tasks [14] [2], and it has been noted for its effectiveness in extracting topics from lexical matrices [18]. Recent studies [16], have shown that SNMF outperforms both k-means and spectral clustering methods, indicating its potential in diverse applications. SNMF has found success not only in clustering tasks but also in community detection within graphs, as demonstrated [11]. The formalization of SNNMF in the following equation:

$$\min_{H \geq 0} \|A - HH^T\|_F^2 \quad (1)$$

It is considered a generalized form for solving clustering objectives [7]. Involving the minimization of the Frobenius norm, the non-negative matrix H plays a crucial role, with dimensions $n \times k$, where k is the number of required clusters. It is essential for capturing the cluster structure. SNMF is essentially a form of clustering on graphs, demonstrating similarities to kernel k -means and spectral clustering, given its relaxation of non-negativity orthogonality on H . The performance of SNMF clustering relies on the construction of the similarity matrix A . Various methods, proposed by [8], include considering the affinity matrix $\tilde{A}$, where elements represent weights of edges between data points, and constructing A based on ratio correlation or normalized cut. The SNMF algorithm is designed for decomposing completely positive similarity matrices, presenting challenges when negative eigenvalues are involved. To address this, the single-loss weight SNMF model was developed by [4].

3 Automatic number of clusters in SNMF

A significant challenge for clustering methods is determining the number of clusters in the data. For the SNMF we use the consensus clustering to define the automatic number of k clusters [13]. Consensus clustering formalizes the concept of amalgamating diverse clusters into a unified representative, or consensus, to highlight shared structures across various datasets and unveil substantial differences among them. The objective of consensus clustering is to identify a representative consensus that reflects the provided cluster of a particular dataset.

Formally, the vector ID_q comprises labels, $ID_{qj} \in [1, \dots, k]$, indicating the cluster memberships for each row in X . The binary connectivity matrix, B_q , is derived from the values in the ID_q vector, where B_{ij} equals 1 if the i th and j th elements belong to the same cluster and 0 otherwise. Formally, $B^{(q)}_{(i,j)} = \{1 \text{ if } x_i \text{ and } x_j \text{ belong to the same cluster, } 0 \text{ otherwise}\}$. The consensus matrix $M^{(k)}$ is computed for a fixed number of clusters k by averaging all binary connectivity matrices:

$$M^{(k)} = \frac{\sum_{q=1}^r B^q}{r} \quad (2)$$

Each element in $M^{(k)}$ (known as the consensus index), indicates how frequently the i th and j th elements are clustered together and r is the number of resamples. A perfect clustering results in a consensus matrix with only 0 or 1 values. Deviations from these values quantify the stability of clusters for a given number of clusters, and by comparing consensus matrices $M^{(k)}$ for various $k \in [k_l, k_u]$, the number of clusters corresponding to the minimum deviation from a perfect consensus is identified as the potential number of clusters in the actual data.

4 Empirical study

To empirically showcase the practical value of consensus clustering in automatically identifying the optimal number of clusters in the SNMF, we conducted an analysis on Instagram posts made by Ursula von der Leyen, the President of the European Commission. Our objective was to delineate the primary topics of discussion spanning from her inauguration year in 2018 to January 1, 2024. We use Crowd Tangle to extract 708 posts by the official profile of the European President. The preprocessing phases of cleaning textual data involve several essential steps to prepare the text for analysis. Firstly, normalization ensures consistency by converting text to a standard format. This may involve converting all letters to lowercase and removing punctuation marks. Next, tokenization breaks the text into individual words or tokens, separating them into distinct units. Removing stopwords is another crucial step, where common words that don't carry significant meaning, such as "the" or "and" are eliminated to reduce noise. Lemmatization helps further refine the text by reducing words to their root form. Stemming removes affixes from words, while lemmatization maps words to their base form [12]. After that we create the vocabulary, and we apply the cosine similarity to create a term-term matrix. In Figure 1 we use the consensus clustering to cosine similarity matrix and identify five topics (Table 1), according to the lower value of the consensus measure.

Figure 1: Optimal number of clusters

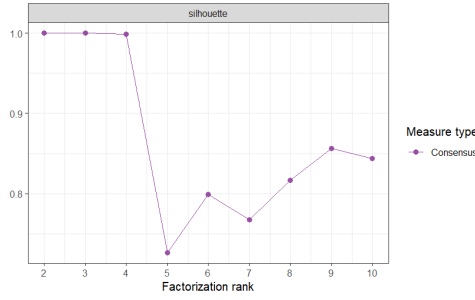


Table 1: Semantic clusters of Ursula von der Leyen communication

<i>Clusters</i>	<i>Words</i>
Balkan policies	Albania, Machedonia, Heads state, implement, government, earthquake, solidarity, balkan, call, trip
Ukrainian war	Respond, dependency, consumer, russian, belarus, propose, attack, sanction, diplomatic, Ukraine
Green economy	Green, plan, invest, digital, resilience, hydrogen, recovery, project, next_generation, sustainable
Refugees	Foster, family, housing, child, lose, house, damage, home, responsibility, refugee,
Covid-19	Vaccination, vaccine, pandemic, Europe, order, dose, announce, test, safe vaccine

Considering the top ten terms appearing in the cluster, the first topic delves into Balkan support policies, specifically focusing on the comparison between Macedonia and Albania, alongside the management of earthquake funds. The discussion likely centers around the initiatives and strategies implemented by international organizations and governments to provide support and assistance to the Balkan region, with a specific focus on Macedonia and Albania. The second topic involves the ongoing conflict between Ukraine and Russia and the role of Europe in maintaining stability and balance in international relations. The conflict has been a significant geopolitical issue, with implications for regional security and global politics. Europe, as a key player in international affairs, has a vested interest in resolving the conflict peacefully and preventing further escalation. Europe's role may include diplomatic efforts, economic sanctions, peacekeeping initiatives, and support for Ukraine's sovereignty and territorial integrity. The third theme pertains to government policies and initiatives aimed at promoting a green economy and transitioning to renewable energy sources. A green economy focuses on sustainable development, environmental protection, and reducing carbon emissions to combat climate change. Policies related to this may include incentives for renewable energy production, investments in clean technologies, regulations to reduce greenhouse gas emissions, and support for sustainable practices in various sectors such as transportation, agriculture, and industry. The use of renewable energies, such as solar, wind, hydroelectric, and geothermal power, is a key component of transitioning to a green economy, as it reduces reliance on fossil fuels and mitigates environmental impacts. The last topic defines the vaccination strategy of Covid-19 pandemic. The focus is on vaccination efforts in Europe, encompassing discussions on vaccine procurement, distribution, and the scientific campaign to support vaccination efforts.

5 Conclusions and limitations

The paper delves into various matrix factorization techniques, with a specific focus on utilizing lexical tables to unveil latent themes within textual data. Its primary objective is to effectively decompose these matrices, with a particular emphasis on symmetric non-negative matrix factorization (SNMF), renowned for its effectiveness in clustering tasks. The central aim of the study is to develop an automated method for determining the optimal number of clusters within a lexical matrix. Among the methodologies investigated, consensus clustering emerges as a promising approach, backed by the analysis results, supported by the results of the analysis, which showed the main themes of Ursula von der Leyen's communication on Instagram. However, the paper recognizes the computational challenges

associated with consensus clustering, especially when handling large datasets or a high number of resamples, as it can lead to a substantial computational burden.

References

1. Blei D.M., Ng A.Y., Jordan M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022 (2003)
2. Chen, B. W.: Symmetric nonnegative matrix factorization based on box-constrained half-quadratic optimization. *IEEE Access*, 8, 170976–170990. (2020)
3. Deerwester, S. C., Dumais S. T., Landauer T. K., Furnas G. W., Harshman. R. A.: Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407. (1990).
4. Ding, C., He, X., Simon, H. D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining* (pp. 606–610). Society for Industrial and Applied Mathematics. (2005).
5. Fortunato, S.: Community detection in graphs. *Phys. Rep.* 486, 75–174 (2010)
6. Hofmann T.: Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1–2):177–196. (2001)
7. Kuang D, Ding C, Park H.: Symmetric nonnegative matrix factorization for graph clustering. In: *Proceedings of SIAM international conference on data mining (SDM)*, pp 106–117. (2012)
8. Kuang D., Choo J., Park H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitioned Clustering Algorithms* (pp. 215–243). Springer, Cham. (2015)
9. Lebart L., Salem A., Berry L.: *Correspondence Analysis of Lexical Tables., Exploring Textual Data. Text, Speech, and Language Technology*, vol 4. Springer, Dordrecht. (1998)
10. Lee D. D., Seung H. S.: Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. (1999).
11. Luo, X., Liu, Z., Jin, L., Zhou, Y., Zhou, M.: Symmetric nonnegative matrix factorization-based community detection models and their convergence analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3), 1203–1215. (2021).
12. Misuraca, M., Scepi, G., Spano, M.: Network-Based Dimensionality Reduction for Textual Datasets. In *Models for Data Analysis: SIS 2018, Palermo, Italy, June 20–22* (pp. 175–190). Cham: Springer International Publishing. (2018).
13. Vangara, R., Rasmussen, K. Ø., Chennupati, G., Alexandrov, B.: Determination of the number of clusters by symmetric non-negative matrix factorization. In *Big Data III: Learning, Analytics, and Applications* (Vol. 11730, pp. 104–113). SPIE. (2021).
14. Wu, W., Jia, Y., Kwong, S., Hou, J.: Pairwise constraint propagation-induced symmetric nonnegative matrix factorization. *IEEE transactions on neural networks and learning systems*, 29(12), 6348–6361. (2018)
15. Yan, W., Zhang, B., Yang, Z., Xie, S.: Similarity learning-induced symmetric nonnegative matrix factorization for image clustering. *IEEE Access*, 7, 166380–166389. (2019).
16. Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y.: Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 749–757). Society for Industrial and Applied Mathematics. (2013).

16. Contributed - Modelling on spatial phenomena

- 16.1 - Integrating computational and statistical algorithms in RT-GSCS for spatial survey administration
- 16.2 - Sensitivity mapping as a tool to support siting of offshore wind farms and increase citizens' acceptability
- 16.3 - Investigating hotel consumers' purchase intention on web analytics data through PLS-SEM
- 16.4 - Spatio-temporal analysis of lightning point process data in severe storms

Integrating computational and statistical algorithms in RT-GSCS for spatial survey administration

Yuri Calleo, Simone Di Zio, and Francesco Pilla

Abstract RT-GSCS is a web-based open platform facilitating dynamic spatial surveys and consensus-building. It supports multiple surveys, where users engage through a user-friendly map interface, contributing opinions in real-time. RT-GSCS employs a spatial consensus algorithm inspired by the Delphi method, visually representing experts' opinions on the map. The system's real-time analyses include heat maps, dynamic clustering, and textual analysis, offering immediate insights. It proves valuable for decision-making, allowing users to collaboratively identify spatial information for effective policy implementation. This paper details RT-GSCS's administration procedure, consensus algorithm, and spatial support analysis, highlighting its unique contributions to real-time spatial collaboration systems.

Key words: RT-GSCS, spatial survey, survey administration, participation

1 Introduction

Real-Time Geo-Spatial Consensus System (RT-GSCS, <http://www.rtgscs.com/>) is a web-based open platform designed for the administration of spatial surveys aimed at collecting and visualizing spatial data through the aggregation of subjective opinions [1]. The system establishes a dynamic, real-time process that facilitates group collaboration in addressing spatial queries, with the final aim of achieving a convergence of opinions [2]. RT-GSCS is provided with a virtual environment capable of supporting multiple surveys simultaneously. It includes a dedicated

¹ Yuri Calleo, University College Dublin, Ireland; email: yuri.calleo@ucdconnect.ie
Simone Di Zio, University "G. d'Annunzio", Chieti-Pescara, Italy; email: s.dizio@unich.it
Francesco Pilla, University College Dublin, Ireland; email: francesco.pilla@ucd.ie

backend panel and supports numerous responses through an interactive frontend map interface. Its utility lies in pinpointing precise locations or areas within a territory for the effective implementation of policies. The platform’s primary interface features an interactive map designed for “user-friendliness”, catering to individuals unfamiliar with online spatial survey tools. The interface incorporates a left-hand sidebar as a central hub for crucial information, including attachments, questions, types of opinion points, comments, consensus information, and advanced analysis. Users contribute their responses in the form of N_e points (with $e = 1, 2, \dots, E$ the label of the respondents), and these judgments are acquired in the contributors’ panel and synthesized into the main user interface in real-time [2]. The distinctive feature of the new version of the system (Fig. 1) lies in its incorporation of real-time spatial and textual support analyses, coupled with an innovative statistical algorithm. RT-GSCS not only enables immediate collaboration for data acquisition and visualization but also fosters the exploration of spatial consensus among users [3]. In this context, the concept of “consensus”, derives from the Delphi method [4] [5], particularly useful for achieving a convergence or stability among a group of privileged experts.

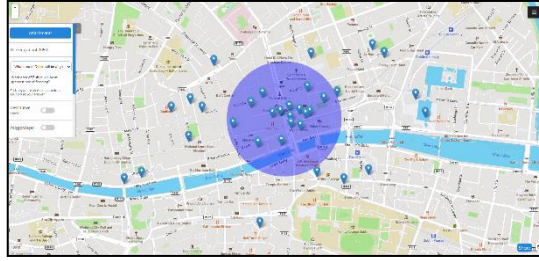


Figure 1: System’s interface (v3.0)

From the initial version of the system, various updates have evolved, incorporating new computational and statistical algorithms to enhance its operational procedures. For this reason, this paper seeks to showcase the capabilities of RT-GSCS (v3.0), specifically by elucidating the algorithms and advanced analysis employed in data collection, analysis, data visualization, and spatial consensus.

2 Method foundation

The system is currently interfacing with a web server (v3.0), the most recent release (<http://www.rtgscs.com/>). A spatial survey can be created on the main website. Contributors start by making an account and filling out a form with necessary details like personal information, institutions, and study purpose. At this point, the administrator reviews the request and then grants the contributor authorization to start a new survey. The crafting of a new survey can be executed by accessing the dedicated panel within the system and inputting the credentials provided during the registration

process. Upon logging in, the contributor's control panel encompasses a comprehensive set of tools for configuring a new survey, which includes: 1) *Survey description*; 2) *Attachments uploading*; 3) *Spatial questions input*; 4) *Approve/Reject users*; 5) *Real-time results*. Once the survey is ready to be administered, selected users receive a private link. Upon registering with their details (e.g., name, surname, affiliation, and email) and obtaining approval from contributors, they gain access to initiate their work within the virtual environment. Respondents can answer the uploaded questions anonymously, aligning with the principles of the Delphi method. This maintains confidentiality, reducing potential cognitive biases in the decision-making process. Within the system, users encounter an interactive map as the primary interface along with a left-hand hub containing all essential information pre-uploaded by contributors, including attachments, introductions, and questions. Starting from that, users can interact with the map, choosing their preferred question and providing responses with a point (which we indicate here with the generic notation n_i) by clicking on any area of the map. In this case, we “transform” a subjective opinion into spatial coordinates, directly acquired within the system and readily available in the contributor panel and the main interface. Users possess the capability to respond to various questions by providing three types of spatial information: 1) *Opinion-points* (n_i): inserted on the map with x, y coordinates part of the dynamic process of convergence. 2) *Descriptive points*: inserted on the map with x, y coordinates to acquire further information but not directly involved in the dynamic convergence process. 3) *Polygons*: inserted in the form of polygons on the territory, users can identify a geographical area to respond to the question. For all three types of spatial information, users can anonymously comment on the opinions of others, providing valuable feedback to the decision-making process at any given time. All the information provided by the users is acquired in the backend but also displayed in the main interface, including the possibility to consult the experts' textual comments during the process.

2.1 Computational algorithm

The algorithm implemented in this system aligns with the principles of the Spatial Delphi [2] and the Real-Time Spatial Delphi [6] [1], aiming to acquire a convergence of opinions on geographic areas. As users respond to a question on the platform, a set of N judgments are automatically gathered (where $N = \sum_{e=1}^E N_e$, N_e is the number of answers given by user e , and E is the number of contributors). Notably, from the initial placement of each point for a given question, the algorithm highlights the convergence of opinions – on that question – using a geometric element identified as a circle (C). While the choice of the geometric element is subjective, the circle is deemed appropriate for accurately representing a portion of the territory. Analogously to one of the widely used indicators in the statistical summary of the Delphi method [4], namely the interquartile range (IQR), the algorithm determines a circle C , the smallest among the possible C_i circles ($i = 1, 2, \dots, N$), each centred on one point and containing half of the N opinion points. Following this logic, a real-time

representation of the *IQR* is presented on the interactive map with C , encapsulating 50% of the N judgments, along with their corresponding x, y coordinates. Throughout the survey, as users respond to different questions, for each question a spatial distribution of $N = n_1, n_2, \dots, n_N$ points emerge in real-time on the map. For each question, the algorithm identifies a minimum area A_i for each circle C_i that covers 50% of the points, denoted as $A_i \supseteq T_{N/2}$, where the set $T_{N/2}$ represents the 50% of the N opinion points. However, in this case, the key challenge lies in the choice of one of the N circles calculated. Since the aim is to seek consensus, which means a small area where an intervention or decision can be concentrated, the algorithm computes, for each question, the minimum of the vector $\mathbf{A} = [A_1, A_2, \dots, A_N]$ of the N areas $\min(\mathbf{A})$. The circles C with area $\min(\mathbf{A})$ represents the spatial consensus achieved on the territory. This methodology, employed in RT-GSCS, fosters active collaboration among experts or users for immediate decision-making. This approach is crucial because a mere visualization of the N points without a convergence circle C could result in disparate clusters of points on the territory, hindering the formulation of decisions and policy implementations. Similar to the traditional Delphi method (see among many [7]), RT-GSCS allows users to revise their opinions at any time, incorporating feedback from other users, thereby promoting convergence, and acknowledging the possibility of points being placed outside territorial boundaries or in illogical areas. To facilitate this aspect, if a user places a point outside of the consensus circle C_i an automatic window appears asking for the motivation and the comment is labelled with a red flag, thus denoting to other users this information.

The distinct feature of version 3.0 lies in the ability to execute support real-time analyses, facilitating immediate visualization of user-provided judgments, thereby fostering an interactive and dynamic process (Fig. 2). The conducted analyses encompass: 1) *Heat map analysis*; 2) *Dynamic cluster analysis*; 3) *Textual analysis*. The heat map analysis serves the purpose of revealing the spatial distribution of N points, aiming to identify hotspots, areas of unusually high concentration or density of opinion-points. Assuming a distribution of N points denoted by coordinate pairs (x, y) , the algorithm autonomously divides the space into a grid of cells, where each cell $G_{\{i,j\}}$ is linked to a random variable representing the count of data points falling within it ($N_{\{i,j\}}$). As new data points are gathered, real-time updates continuously adjust these counts. The resultant heat map visually communicates this distribution, utilising colour intensity to signify varying densities of data points across the grid. This analysis can be enabled from the main panel and provides users with an overview of opinions on the map, pinpointing potential hotspots and promoting an understanding of concentrated decision zones. Moreover, the real-time cluster analysis, which dynamically adjusts in response to user actions such as zooming, is executed by considering the distance between points. Let N denote the set of data points represented by geographical coordinates on the map, where each n_i point is defined as $n_i = (x_i, y_i)$, for $i = 1, 2, \dots, N$. With $d_{(n_i, n_j)}$ we denote the Euclidean distance between points n_i, n_j . The algorithm conducts clustering of data points based on their spatial proximity, denoting S as the set of clusters, where S_k represents the k -th cluster with $k = 1, 2, \dots, K$. Given the interactive nature of the map within the system, the clustering is not static; it dynamically adjusts with user interactions,

particularly in response to specific zoom levels. Let $K_{(z)}$ denote the number of clusters when the user is at a particular zoom level z . In this context, $K_{(z)}$ is dynamic and adapts as the user zooms in or out on the interactive map. The assignment of an opinion point n_i to a specific cluster S_k (say $n_i \in S_k$) undergoes changes based on the zoom level and the inclusion of new points on the map (per question). In summary, Z serves as the primary parameter representing the zoom level on the map and can vary continuously. The dynamic cluster assignment is articulated as $S_{k(z)}$, where $n_i \in S_{k(z)}$ indicates that at zoom level z , data point n_i belongs to cluster S_k , reflecting the visual representation of clusters at a specific zoom level, encompassing adjustments in cluster boundaries, colours, and labels. This analysis proves especially valuable in guiding users towards existing clusters of opinions, enabling them to comprehend viewpoints prior to sharing their own.



Figure 2: Real-time analysis

Finally, the textual analysis focuses on the users' comments associated with each question and the N points. Within the system, users have the flexibility to provide comments on other points, fostering collaboration and allowing for the reconsideration of decisions, akin to the traditional version of the Delphi method. However, in a decision-making process, reviewing these comments can be time-consuming, particularly for survey respondents with limited time, such as experts, local authorities, or governmental officials. To mitigate this challenge, the system employs a textual analysis methodology for each survey question, following a specific approach. For a single question, considering a collection of m comments/remarks, each denoted as R_i (with $i = 1, 2, \dots, m$) the algorithm tokenizes each remark into words, getting a vector of terms V_{ij} in the i -th remark. To refine the analysis and eliminate less meaningful or irrelevant words, a common stop-words list (English language) is applied, resulting in a set matrix of V'_{ij} terms. Subsequently, the algorithm calculates the term frequency (TF) for each term in V'_{ij} , determining the frequency of term t in comment i . Finally, the total term frequency across all comments related to a question is calculated using the provided equation: $TF_{(t,i)} = \sum_{j=1}^{n_i} I(t_j = t)$, where n_i denotes the number of terms in remark i . Following the ranking process, we narrow our focus to the initial 50 words to generate a real-time word-cloud. In this visual representation, the size of each word is proportionate to its term frequency (TF) within the context of the question. The same procedure applies to all questions, so the system returns as many word-clouds as questions. This strategic

approach is designed to enhance the efficiency of textual comment analysis, offering a more concise and meaningful interpretation of user feedback. By condensing the information into a word cloud format, the system facilitates a quick and visually intuitive understanding of the most frequently used terms, contributing to a more streamlined and effective assessment of user input.

3 Conclusion and future works

In summary, RT-GSCS (v3.0) is a versatile and powerful real-time system for spatial surveys, fostering innovative group collaboration. The system's hierarchical structure, intuitive survey administration, and advanced spatial analyses make it a valuable resource for decision-makers, experts, and citizens. RT-GSCS's ability to foster spatial consensus through dynamic convergence circles sets it apart from other tools, proving instrumental in decision-making and policy implementation. Indeed, in decision-making processes involving spatial decisions, various variables must be assessed to arrive at an appropriate choice. Often, workshops are convened, but they may reveal different cognitive biases arising from expressing opinions face-to-face. Furthermore, the platform serves as an indispensable tool in Futures Studies context, facilitating the development of spatial scenarios. In instances where spatial data is either partially or entirely inaccessible, thereby hindering the utilization of spatial reference models, the engagement of domain experts becomes paramount.

In future works, the platform could incorporate new innovations from both computational and participatory perspectives, including a new algorithm that calculates different types of circles based on research objectives, addressing potential limitations from territorial dislocation influence. Finally, it could also involve the integration of new internal tools, such as artificial intelligence models [8], to assist experts in visualizing real images of their decisions.

References

1. Calleo, Y., Di Zio, S., Pilla, F.: Facilitating spatial consensus in complex future scenarios through Real-Time Spatial Delphi: A novel web-based open platform. *Futures & Fore-sight Science*, 5, e155 (2023)
2. Di Zio, S., & Pacinelli, A.: Opinion convergence in location: A spatial version of the Delphi method. *Technological Forecasting and Social Change*, **78**(9), 1565--1578 (2011).
3. Mekonnen, A. D., Gorsevski, P. V.: A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio. *Renewable and Sustainable Energy Reviews*, **41**, 162-177 (2015)
4. Linstone, H. A., Turoff, M.: *The delphi method*. Reading, MA: Addison-Wesley. 3-12 (1975)
5. Calleo, Y., Pilla, F.: Delphi-based future scenarios: a bibliometric analysis of climate change case studies. *Futures*, 103143 (2023)
6. Di Zio, S., Rosas, J. D. C., Lamelza, L.: Real Time Spatial Delphi: Fast convergence of experts' opinions on the territory. *Technological Forecasting and Social Change*, **115**, 143--154 (2017)
7. Grime, M. M., & Wright, G.: Delphi method. *Wiley statsref: Statistics reference online*, **1**, 16 (2016)
8. Di Zio, S., Calleo, Y., & Bolzan, M.: Delphi-based visual scenarios: An innovative use of generative adversarial networks. *Futures*, **154**, 103280 (2023).

Sensitivity mapping as a tool to support siting of offshore wind farms and increase citizens' acceptability

Cilluffo G., Sottile G., Ciriminna L., Signa G., Tomasello A. and Vizzini S.

Abstract

Wind power is one of the fastest growing renewable energy technologies. Environmental impact assessment is crucial for the sustainable development of offshore wind farms (OWF). One of the objectives of the Wind In My Backyard project is to understand and find ways to raise public understanding and engagement with wind power. This preliminary study focuses on the island of Pantelleria. Ecological niche models were estimated for three target marine species (*Caretta caretta*, *Stenella coeruleoalba* and *Tursiops truncatus*). An additive value model was used to produce a sensitivity map. ENMs showed good predictive performance. The sensitivity map was able to identify potential sensitive areas. Our study may be useful for the identification of potential sensitive areas, drive selection of suitable sites for wind farm installation and improving citizens' understanding of the complex processes involved in marine spatial planning

Key words: Ecological Niche Models, Sensitivity map, Pantelleria, marine fauna.

¹ Giovanna Cilluffo, Department of Earth and Marine Sciences, University of Palermo; National Inter-University Consortium for Marine Sciences CoNISMa, Rome, Italy; giovanna.cilluffo@unipa.it;
Gianluca Sottile, Department of Economics, Business and Statistics, University of Palermo, Italy; gianluca.sottile@unipa.it
Geraldina Signa, Department of Earth and Marine Sciences, University of Palermo; National Inter-University Consortium for Marine Sciences CoNISMa, Rome, Italy; geraldina.signa@unipa.it;
Laura Ciriminna, Department of Earth and Marine Sciences, University of Palermo; National Inter-University Consortium for Marine Sciences CoNISMa, Rome, Italy; laura.ciriminna@unipa.it;
Agostino Tomasello, Department of Earth and Marine Sciences, University of Palermo; National Inter-University Consortium for Marine Sciences CoNISMa, Rome, Italy; agostino.tomasello@unipa.it;
Salvatrice Vizzini, Department of Earth and Marine Sciences, University of Palermo; National Inter-University Consortium for Marine Sciences CoNISMa, Rome, Italy; salvatrice.vizzini@unipa.it;

1 Introduction

According to the European Climate Change Act, the total installed capacity of wind energy is expected to increase in the coming years. Increasing the number of wind farms is important for reducing greenhouse gas emissions. However, wind farms can also have an impact on biodiversity. There is also a lack of understanding of how offshore wind farms (OWF) and related infrastructure can be built, operated, and decommissioned, and how this could affect marine ecosystems and their functions. [1, 2]. In fact, while wind energy is a key sector for renewable energy production, OWF can have both positive and negative impacts on marine fauna. Impacts may vary depending on several factors, including siting and specific project design.

There are many potential pathways through which OWF can have positive impacts for marine fauna. First, the underwater structures of offshore wind turbines provide new habitats for benthic marine species. They may attract fish and other marine organisms, creating new ecosystems [3]; Second, the restricted access around wind turbine foundations can serve as marine protected areas, helping to preserve certain marine habitats and reducing human disturbance in those areas [4, 5]. Third, the turbine foundations can act as artificial reefs fish aggregate, as they provide structure in the open ocean. This can attract fish and enhance local biodiversity [2].

On the other hand, OWF may negatively influence marine fauna as the construction and operation of offshore wind farms can generate underwater noise and vibrations. This may affect marine species that are sensitive to sound, such as marine mammals and fish, potentially leading to changes in behaviour, migration patterns, and communication [6]. Marine species, particularly birds, may be at risk of collision with the turbine blades [7]. However, advancements in technology, such as radar systems and shut down-on-demand features, are being implemented to mitigate these risks. In addition, the installation of turbines and the associated infrastructure may alter the seabed and sediment transport, affecting benthic organisms and the composition of the ocean floor [6].

Therefore, environmental impact assessments and careful planning are crucial to understanding and mitigating the potential effects on marine species in specific locations. Regulations and guidelines are continually evolving to address these concerns and promote sustainable offshore wind energy development.

Ecological Niche Models (ENMs) help to identify the key environmental factors that influence the distribution of a species, and they are widely used in conservation biology for various purposes, such as biodiversity conservation.

The current study is a part of the Wind In My Backyard project (WIMBY; HORIZON-RIA_CL5) aimed to understand and find ways to to advance social awareness and engagement in Europe. In particular, the aim of the current preliminary results is to identify sensitivity areas for siting and citizens' acceptability of OWF in the pilot site of the Island of Pantelleria.

2 Methods

2.1 Study Area

The study was carried out in Pantelleria, a small volcanic island 110 km off the coast of Sicily with high potential for renewable energy production. The island's own electricity grid is mainly powered by diesel generators, but due to its location in one of the windiest areas in Italy, offshore wind energy would be a sustainable alternative.

2.2 Target species selection

In the current preliminary study, we report the results on three species of high conservation interest *Caretta caretta*, *Stenella coeruleoalba* and *Tursiops truncatus*. Target species were selected based on Descriptor 1 (Marine Biodiversity) of the Marine Strategy Framework Directive (Council Directive 92/43/EEC); for the assessment of good environmental status, while species occurrence data were downloaded from the Global Biodiversity Information Facility (GBIF) database.

2.3 Marine environmental data

We considered nine environmental variables: Nitrates (micromol/L), Chlorophyll-*a* (mean at mean depth) (mg/m³), Phosphates (micromol/L), Silicates (micromol/L), Dissolved oxygen (ml/l), and pH obtained from Bio-ORACLE v.2.1[8], bathymetry (radians), sea surface salinity (Practical Salinity Scale-PSS) and sea surface temperature (°C) obtained from MARSPEC [9]. The raster files for each variable were converted to a WGS84 coordinate system with a spatial resolution of ~9.2 km.

2.4 Ecological niche modeling

ENMs were estimated using four different algorithms: generalized linear model, random forest, max entropy and supporting vector machine. The dataset was presence-only, and 1000 background points were randomly sampled throughout the study area. The presence data were randomly split into 30% and 70% for model testing and calibration, respectively, and a 5-fold cross-validation procedure was used to estimate predictive performance of the models. We evaluated the predicted habitat suitability using the area under the Receiver Operating Curve (ROC), known as the Area Under the Curve (AUC). Habitat suitability was mapped using the ensemble method combining the predictions of multiple individual base models to achieve a more accurate and reliable prediction. Data were analysed using the sdm R package. Once obtained the habitat suitability maps for each species, the raster files were imported in QGIS. From each map the probabilities of occurrence were extracted using a point reticules and a composite indicator was created.

2.5 Sensitivity map

A composite indicator, using the Additive Value Model (AVM), was built to combine multiple measures deriving from ENMs ensemble prediction. Once obtained the composite indicator, the Inverse Distance Weighted (IDW) interpolation was used to estimate values at unmeasured locations with the aim of creating the sensitivity map.

3 Results

Occurrence of each considered species was reported in Figure 1. *Caretta caretta* was more abundant in south-west of Pantelleria, *Tursiops truncatus* was found sporadically in the north-east and *Stenella coeruleoalba* in north and north-west. Marine environmental data used in ENMs were illustrated in Figure 2. All ENMs performed better than the random models. Based on the AUC, the predictive performance ranged from good to excellent in terms of discrimination power; AUC scores ranged from 0.70 to 0.90 (Table 1). For each species an ensemble prediction map was created (Figure 3.). The composite indicator produced by the three ensemble predictions was illustrated in Figure 4. The sensitivity map shows a potentially sensitive area extending from the north-west to the north of Pantelleria.

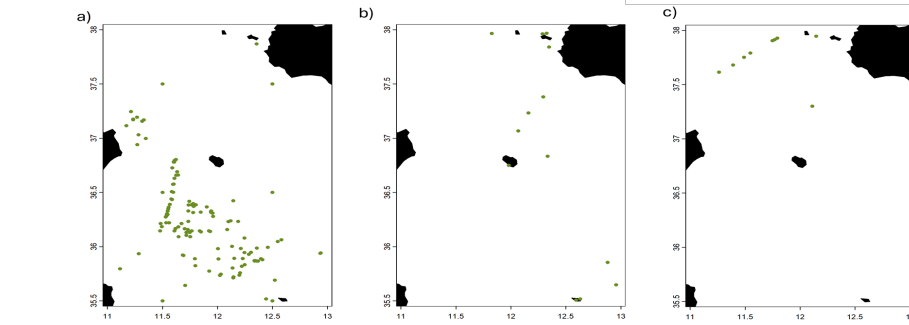


Figure 1: Occurrence data of *Caretta caretta*, (a); *Tursiops truncatus* (b) and *Stenella coeruleoalba* (c).

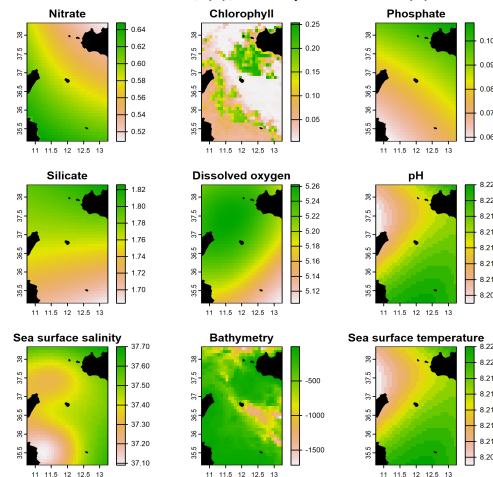


Figure 2: Marine environmental data

Table 1: Predictive performance of the models: Area Under the Curve (AUC); GLM: generalized linear model; RF: random forest; MaxEnt: Max entropy; SVM: Support Vector Machine.

Species	GLM	RF	MaxEnt	SVM
<i>Caretta caretta</i>	0.85	0.89	0.9	0.87

Contribution Title	5			
<i>Tursiops truncatus</i>	0.76	0.81	0.81	0.70
<i>Stenella coeruleoalba</i>	0.90	0.87	0.92	0.81

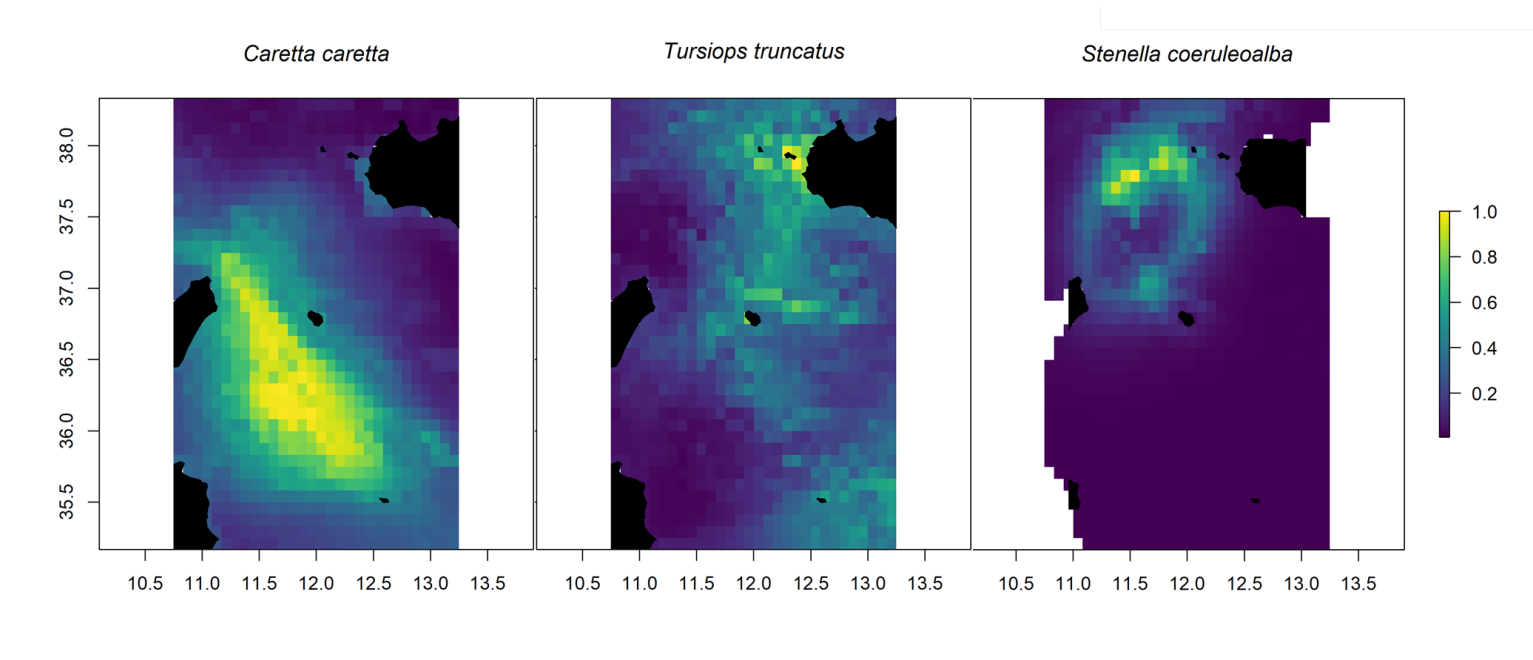


Figure 3: Habitat suitability maps of *Caretta caretta*, (a); *Tursiops truncatus* (b) and *Stenella coeruleoalba* (c).

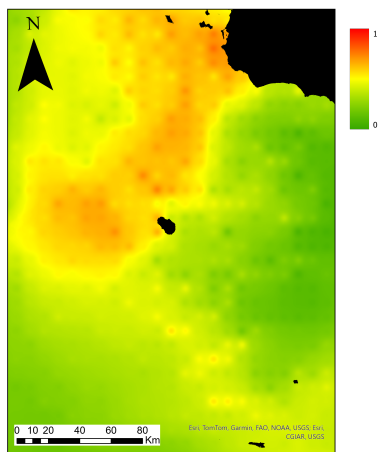


Figure 4: Sensitivity map.

4 Conclusion

This study underlines the value of using ENM for estimating habitat suitability maps. To combine results deriving by ENM for each target species can help in the process of biodiversity conservation, supporting strategic environmental assessments and spatial planning. Furthermore, the sensitivity map created can be used to identify key marine areas for different species, with the aim of supporting OWF siting within the framework of sustainable development, and for improving citizens' understanding of the complex processes involved in marine spatial planning.

References

1. Papathanasopoulou, E., Beaumont, N., Hooper, T., Nunes, J., Queirós, A.M.: Energy systems and their impacts on marine ecosystem services. *Renew. Sustain. Energy Rev.* 52, 917–926 (2015)
2. Raoux, A., Tecchio, S., Pezy, J.-P., Lassalle, G., Degraer, S., Wilhelmsson, D., Cachera, M., Ernande, B., Le Guen, C., Haraldsson, M., others: Benthic and fish aggregation inside an offshore wind farm: which effects on the trophic web functioning? *Ecol. Indic.* 72, 33–46 (2017)
3. Werner, K.M., Haslob, H., Reichel, A.F., Gimpel, A., Stelzenmüller, V.: Offshore wind farm foundations as artificial reefs: The devil is in the detail. *Fish. Res.* 272, 106937 (2024)
4. Hammar, L., Perry, D., Gullström, M.: Offshore wind power for marine conservation. *Open J. Mar. Sci.* 6, 66–78 (2015)
5. Ashley, M., Mangi, S., Rodwell, L.: The potential of offshore windfarms to act as marine protected areas—A systematic review of current evidence. *Mar. Policy.* 45, 301–309 (2014)
6. Watson, S.C., Somerfield, P.J., Lemasson, A.J., Knights, A.M., Edwards-Jones, A., Nunes, J., Pascoe, C., McNeill, C.L., Schratzberger, M., Thompson, M.S., others: The global impact of offshore wind farms on ecosystem services. *Ocean Coast. Manag.* 249, 107023 (2024)
7. Duriez, O., Pilard, P., Saulnier, N., Boudarel, P., Besnard, A.: Windfarm collisions in medium-sized raptors: even increasing populations can suffer strong demographic impacts. *Anim. Conserv.* 26, 264–275 (2023)
8. Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., De Clerck, O.: Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Glob. Ecol. Biogeogr.* 21, 272–281 (2012)
9. Sbrocco, E.J., Barber, P.H.: MARSPEC: ocean climate layers for marine spatial ecology: Ecological Archives E094-086. *Ecology.* 94, 979–979 (2013)

Investigating hotel consumers' purchase intention on web analytics data through PLS-SEM.

Giuseppina Lo Mascolo - Chiara di Maria - Marcello Chiodi – Arabella Mocciano Li Destri ¹

Abstract In an era of rapid technological advancement, the hospitality industry has witnessed a significant shift towards online booking platforms. This shift necessitates an exploration of psychological and behavioural factors influencing online purchase intention. This research investigates the relationships between trust, brand orientation, and engagement attitude in hotel consumers' online purchase intention. Utilizing Partial Least Square Structural Equation Modelling (PLS-SEM) on web analytics data, the study offers a comprehensive analysis of factors influencing consumers' decision-making in online hotel bookings. The findings enhance understanding of consumer behaviour in the digital hospitality industry, with implications for operations, strategy, and management.

Key words: *Google Analytics, PLS-SEM, Hospitality Management, Consumer Behaviour*

¹

Giuseppina Lo Mascolo – University of Palermo – giuseppina.lomascolo@unipa.it
Chiara Di Maria – University of Palermo – chiara.dimaria@unipa.it
Marcello Chiodi – University of Palermo – marcello.chiodi@unipa.it
Arabella Mocciano Li Destri – University of Palermo – arabella.mocciano@unipa.it

1 Introduction

The consumer purchasing process encompasses personal, psychological, social, and cultural factors (Kotler et al., 2010), but our focus shifts to the increasing number of online hotel bookers, with identified influential factors like information quality, website quality, perceived price, brand, and trust.

Brand orientation is a strategic direction influencing customer satisfaction and value through brand awareness. It shapes the customer-company relationship by addressing needs. Despite being introduced over two decades ago, it is still an evolving concept impacting online purchase intentions. Trust, a critical factor in online shopping, reduces perceived risk, fostering positive word-of-mouth. Online trust, especially in the hospitality industry, instills confidence in the provided information and reservation process. Trust is crucial in mitigating uncertainties and psychological concerns, making consumers more likely to engage in online purchases. Scenario attitude, reflecting anticipated experiences, is influenced by trust, brand orientation, and fear of uncertainty. Trust and brand orientation are interconnected, with trust enhancing the willingness to book directly through e-vendors' websites (Polites et al., 2012)

Consumer engagement research has traditionally used surveys for emotional and cognitive insights, but they may be biased (Roseman et al., 2011). A trend now involves analyzing behavioral records on the web and social media. Various studies focus on website evaluation and the link between well-performing websites and consumer behaviours. Web analytics tools, particularly Google Analytics, are crucial for understanding customer behavior and improving e-commerce performance. Kaur (2021) notes Google Analytics' role in capturing accurate data for monitoring website performance and user behavior. These tools offer insights into users' cognitive processing, aiding decision-making through indicators like clicks and time spent (Tam and Ho, 2006).

Therefore, our aim is to use websites analytics as items, to analyse the following hypothesis suggested by extant literature:

H1: Brand orientation (BO) influences customer purchase intention (PI)

H2: Trust (TRUST) is a critical determinant in the PI

H3: Consumers tend to engage (Engagement) if they perceive the relationship as trustworthy (TRUST)

H4: Engagement influences PI

H5: Scenario Attitude (SA) is influenced by TRUST and BO

H6: SA influences PI

2 Data collection

Data were collected on Google Analytics G4 of one hotel located in Palermo city centre and focused on six months of web traffic from different sources (direct, organic, referral, social, paid search) that describe the behaviour of about 145000 visitors of which 57.7% are female coming mostly from Europe, United States and other countries worldwide. In addition to the web analytics data, to proceed with the estimation of scenario attitude, we added the number of nights booked directly and the average booking window for reservations made within the website and during the same period of observation. Our dataset counts 888 observations and 10 items, so the indicators of the formatively measured constructs for Brand Orientation (BO) were, Visitors (Bo_1), Sessions (Bo_2), Session with engagement (Bo_3), Numbers of events (Bo_4); for TRUST, engagement rate (Tr_1); for Engagement, Average duration of the engagement (En_1), Session with engagement per visitor (En_2), Event per session (En_3); for Scenario Attitude (SA), Total number of reserved nights (Tr_3), Average booking window (Tr_4). We excluded the numbers of reservations per visits (Tr_2) because with a preliminary analysis they were not significant and their absence didn't cause any change in the final model.

3 Model and estimation method

Structural Equation Models (SEM) allow for the simultaneous modelling and estimation of complex relationships among multiple variables, including unobservable ones measured indirectly by multiple indicators. This overcomes the limitations of first-generation multivariate statistical techniques such as multiple regression, logistic regression, and analysis of variance (Haenlein and Kaplan, 2004). These limitations include the assumption of a simple model structure, where all variables are considered observable, and the assumption that all variables are measured without error.

In our research, we chose to apply Partial Least Squares SEM (PLS-SEM), that uses a composite-based approach (Jöreskog and Wold, 1982). This is in contrast to common factor-based SEM, which uses weighted composites accounting for measurement error.

PLS-SEM is considered efficient for small sample sizes and complex models, making no distributional assumptions. It handles both reflective and formative models, including single-item constructs, providing higher statistical power compared to other methods.

A PLS path model consists of a structural model linking latent constructs and displaying relationships and measurement models illustrating relationships between constructs and indicators. Error terms are associated with constructs and reflectively

measured indicators to represent unexplained variance. The model development involves testing measurement theory for reliability and validity and then testing structural theory.

In the hospitality and tourism research context, PLS-SEM is commonly used due to its versatility across various approaches (confirmatory, predictive, and descriptive) and its ability to handle theoretical concepts like perceptions and attitudes. The two-step testing process of measurement and structural theory makes it a valuable tool for testing complex chains and nomological networks in fields dealing with theoretical concepts.

To apply the model to the data, we transformed data into a Likert scale from 1 (low visits or time) to 9 (high visits and time spent), and we used the *SeminR* package in R Studio.

The model that we have used was as follow:

$$PI = \beta_E Engagement + \beta_{SA} SA + w_{BO} BO + w_{Tr} TRUST + \varepsilon_{PI},$$

Where:

$$Engagement = \sum_{i=1}^3 \lambda_i^{En} En_i + \beta_{Tr} TRUST + \varepsilon_{En}$$

$$SA = \sum_{i=1}^2 \lambda_i^{SA} Tr_i + \beta_{BO} BO + \beta_{Tr} TRUST + \varepsilon_{SA}$$

$$BO = \sum_{i=1}^4 \lambda_i^{BO} BO_i + \beta_{Tr} TRUST + \varepsilon_{BO}$$

$$TRUST = \lambda^{Tr} Tr_1 + \varepsilon_{TRUST}$$

and $\lambda_i^{En}, \lambda_i^{SA}, \lambda_i^{BO}, \lambda^{Tr}$ are the loading factors, $\beta_E, \beta_{SA}, \beta_{Tr}, \beta_{BO}$ are the correlation weights, and w_{BO}, w_{Tr} are the regression weights.

The estimated model (figure 1) shows clearly that TRUST and BO are related positively with customers' PI (H1 and H2) with respectively $w_{BO} = 0.225$ and $w_{Tr} = 0.867$, TRUST influences positively customers' Engagement (H3) with $\beta_{Tr} = 0.883$ and the latter has effect on PI (H4) with $\beta_E = 0.881$. In the model, the path coefficients shows that SA has a poor influence over PI (H6) with $\beta_{SA} = 0.028$ and that TRUST is negatively related to SA with $\beta_{Tr} = -0.126$ and positively related to BO with $\beta_{BO} = 0.505$ (H5). This could be intended as the presence of a certain fear about the future in the market that in any case has a poor effect, for this hotel on the intention of making a reservation.

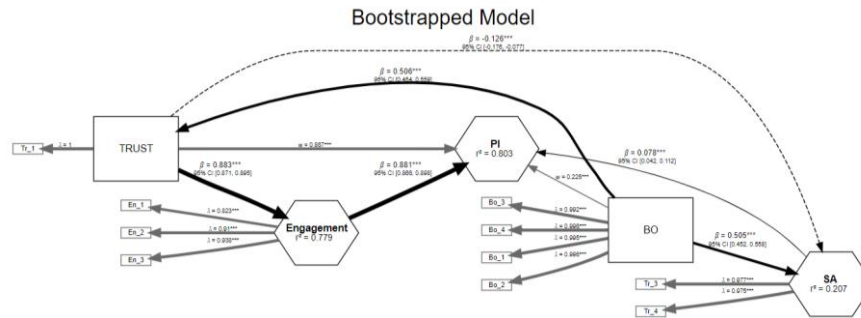


Figure 1 – Estimate model

As far as concerns the model's explanatory power (Shmueli & Koppius, 2011), R^2 represents the variance explained in each of the endogenous constructs. Our model can explain in a substantial way (Hair, Ringle, & Sarstedt, 2011) PI ($R^2 = 0.803$) and Engagement ($R^2 = 0.779$), while it explicates weakly SA ($R^2 = 0.207$) and TRUST ($R^2 = 0.256$).

The results show that the model estimation converged after twelve iterations. Bootstrapping the model, and selecting 10,000 boot samples, we assessed the formative indicator weights' significance.

All the formative weights are significant as can be also confirmed by t-values that are all above 1.96 (at $\alpha = 5\%$).

The bootstrapped loadings' t-values are clearly above 2.576, suggesting that all indicator loadings are significant at a level of $\alpha = 1\%$ and showing that we acted in the right way considering all of them in the formatively constructed.

Furthermore, the model can be considered reliable as a matter of fact that all the measures are higher than their recommended level and even if the Cronbach's alpha for PI is almost equal to 0.7, we can consider it reliable as stated also by Lee et al (2011).

4 Conclusion

This study reveals the usefulness of web analytics and PLS-SEM to explain, and estimate the Purchase Intention of hotel consumers. In particular, the application of PLS-SEM to web analytics data has shown that the method is reliable and that the model is a comprehensive analysis of the factors influencing consumers' decision-making process when booking hotels online. The importance of this study lies on the fact that it is built starting from web analytics and consumer navigation data and use. It is an extension of the study of Xun (2015) about key performance metrics offering a new point of view and a new way of applying the structural equation model using data not based on self-reported data.

From a managerial point of view, the study underlines the importance of measuring consumer purchase intention through online behavioural data for proactive sales and marketing strategies, particularly in the hospitality industry. It reveals the need to use big data to enable comprehensive predictions and adaptive solutions, enhancing operational efficiencies, but underlines also the challenge for hospitality firms lacking technology access. The peculiar scenario that managers are facing enhances the importance of strategic decision execution, therefore personalized marketing driven by big data insights aids in understanding customer preferences and crafting targeted campaigns, facilitating efficient resource allocation and responsiveness to market dynamics. Nonetheless, the study's findings are relevant across industries, emphasizing the significance of data-driven strategies for competitive advantage. Understanding and measuring consumer purchase intention are vital for meeting market needs in complex environments.

5 Acknowledgement

The authors thanks Quintocanto Hotel & SPA (Palermo), Visioni srl, and Be My Guest, for the availability of the data on which we conduct the analysis.

Reference

1. Haenlein, Michael, and Andreas M. Kaplan. "A beginner's guide to partial least squares analysis." *Understanding statistics* 3.4 (2004): 283-297.
2. Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook* (p. 197). Springer Nature.
3. Joreskog, K.G. and Wold, H. (1982) The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. In: Joreskog, K.G. and Wold, H., Eds., *Systems under Indirect Observation: Causality, Structure, Prediction, Part I*, Elsevier, Amsterdam, 263-270.
4. Kaur, J. (2021) Significance of Google Analytics to E-commerce Companies to design customer acquisition and customer strategies - *Vidyabharati International Interdisciplinary Research Journal*, 12(2), 778-790
5. Kotler, Philip. "The prosumer movement: A new challenge for marketers." *Prosumer Revisited: Zur Aktualität einer Debatte* (2010): 51-60.
6. Lee W-Y, Ahn J, Kim J-H, et al. Reliability and validity of a self-reported measure of medication adherence in patients with type 2 diabetes mellitus in Korea. *Journal of International Medical Research*. 2013;41(4):1098-1110. doi:10.1177/0300060513484433
7. Polites, Greta L., et al. "A theoretical framework for consumer e-satisfaction and site stickiness: An evaluation in the context of online hotel reservations." *Journal of Organizational Computing and Electronic Commerce* 22.1 (2012): 1-37.
8. Roseman, Ira J. "Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses." *Emotion Review* 3.4 (2011): 434-443.
9. Shmueli, Galit, and Otto R. Koppius. "Predictive analytics in information systems research." *MIS quarterly* (2011): 553-572.
10. Tam, Kar Yan, and Shuk Ying Ho. "Understanding the impact of web personalization on user information processing and decision outcomes." *MIS quarterly* (2006): 865-890.
11. Xun, J. (2015). Return on web site visit duration: Applying web analytics data. *Journal of Direct, Data and Digital Marketing Practice*, 17, 54-70.

Spatio-temporal analysis of lightning point process data in severe storms

Nicoletta D'Angelo, Milind Sharma, Marco Tarantino and Giada Adelfio

Abstract This paper deals with the analysis of the 19 May 2013 tornadic supercell in central Oklahoma, exploiting point process theory to estimate the intensity of lightning flashes. We develop a computational strategy to fit spatio-temporal Poisson models, including external covariates, to understand the role of environmental spatio-temporal covariates on the occurrence of such lightning events.

Key words: Lightning data, Spatial Statistics, Point processes, Spatio-temporal analysis

1 Introduction

This paper analyses data from the 19 May 2013 tornadic supercell in central Oklahoma, USA. Per the U.S. National Weather Service (NWS) storm survey reports, the supercell produced multiple tornadoes throughout its lifetime. The first tornado had a damage rating of 0 on the enhanced Fujita (EF) scale (McDonald and Mehta, 2006) around 2122 UTC. Subsequently, two more EF1 tornadoes were reported at 2133 and 2153 UTC near Arcadia and Fallis, Oklahoma. The most intense tornado occurred around 2213 UTC, inflicting an EF3 damage as it passed over the city of Carney, Oklahoma. A detailed study of the electrical and polarimetric characteristics of this storm can be found in Sharma et al. (2021). A comparison of the polarimetric indicators of the updraft intensity with vertical velocity retrieved from dual-Doppler wind synthesis for the same storm can be found in Wienhoff et al. (2018).

Nicoletta D'Angelo, Marco Tarantino, Giada Adelfio
Department of Business, Economics and Statistics, University of Palermo
e-mail: nicoletta.dangelo@unipa.it; marco.tarantino01@unipa.it; giada.adelfio@unipa.it

Milind Sharma
Department of Atmospheric Sciences, Texas A&M University, USA
e-mail: milindsharma@tamu.edu

A lightning mapping array (LMA) can map the initiation and subsequent propagation of a lightning discharge in four dimensions (space-time; x , y , z , and t). A lightning flash propagates as a bidirectional leader, emitting very high frequency (VHF) signals from both negative and positive leaders. These signals, falling in the 60-66 MHz frequency range, can be detected by the LMA, which then triangulates the spatio-temporal location of each VHF source (Rison et al., 1999). Lightning data can be thought as spatio-temporal point patterns in different ways: (1) raw VHF source points (can be binned in fixed size time window; too many); (2) flash initiation centroids (one point per flash); (3) gridded flash products (e.g., flash extent or flash initiation density). Scenarios (2) and (3) can further be classified as marked point patterns since we have additional metadata for each flash or each grid point.

Point process methodology, rooted in statistical theory, enables the modelling of events occurring in space and time, making it an ideal tool for studying flash initiation patterns. In this paper, we exploit point process theory to estimate the intensity of lightning flash initiation. In particular, the aim of the paper is to understand the role of environmental spatio-temporal covariates on the occurrence of flash initiation events. To this end, we develop a computational strategy to fit spatio-temporal Poisson models, including external spatio-temporal covariates.

The structure of the paper is as follows. Section 2 introduces the data. In Section 3 we illustrate the proposed methodology to fit spatio-temporal Poisson processes depending on external covariates. Then, Section 4 contains the results of the application to the lightning data, and the paper ends with conclusions in Section 5.

2 Lightning data and spatio-temporal covariates

Since each lightning flash comprises multiple points, we usually use spatio-temporal clustering techniques like DB-SCAN (density-based clustering algorithm; Ester et al., 1996) to cluster points close to each other (in space-time) into lightning flashes (Bruning, 2015). We retain all lightning flashes comprising ten or more VHF sources when all sources occur within 3 km distance and 150 ms from the first identified source in that flash (MacGorman et al., 2008).

Figure 1 (a) shows the plot of the total lightning flash rate (sum of intra-cloud and cloud-to-ground flashes per minute) with three distinct phases in the lightning flash rate: (First) Monotonic increase in flash rate between 20:02:00 and 20:30:00 UTC; (Second) Minor fluctuations but stable flash rate of ~ 150 flashes per minute between 20:30:01 and 21:05:00 UTC; (Third) Gradual decline in flash rate from 21:05:01 UTC onwards until the demise of the storm around 22:30:00 UTC. Figure 1 (b) depicts the multitype spatial point pattern, with the number of flashes in each time frame being 2278, 5687 and 3457, respectively.

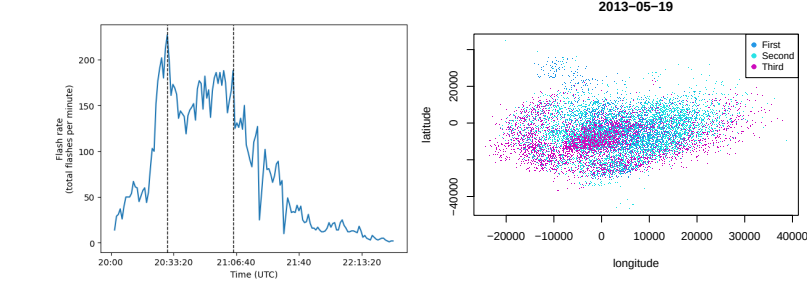


Fig. 1: (a) total count of flashes per minute; (b) multitype spatial point pattern.

We use the surface meteorological variables data from the Oklahoma mesonet as environmental covariates. For the purposes of this study, we use the 5-minute interval data for dry bulb temperature (TAIR), relative humidity (RELH), and equivalent potential temperature (THETA_E), displayed in Figure 2. All of them are available at 5-minute intervals within a $60 \text{ km} \times 60 \text{ km}$ box around the storm centre throughout the analysis period and were measured at 1.5 m above the ground level.

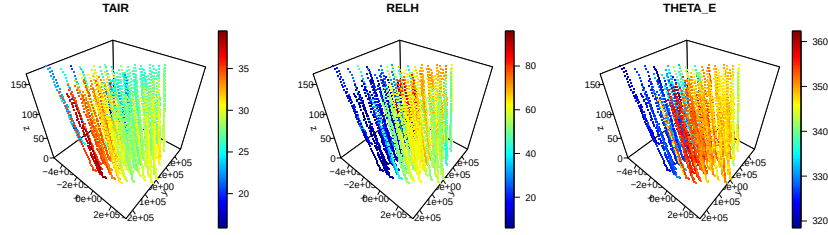


Fig. 2: Environmental spatio-temporal covariates

3 Methodology

We assume that the template model is a Poisson process, with a parametric intensity or rate function $\lambda(u, t; \theta) = \exp(\theta^T Z(u, t))$ with spatial and temporal coordinates $u \in W, t \in T$, unknown parameters $\theta \in \Theta$, and some spatio-temporal covariates $Z(u, t)$ (D'Angelo and Adelfio, 2023). The log-likelihood is

$$\log L(\theta) = \sum_i \lambda(u_i, t_i; \theta) - \int_W \int_T \lambda(u, t; \theta) dt du \quad (1)$$

up to an additive constant, where the sum is over all points u_i in the point pattern $\mathbf{x}$.

For estimation purposes, we use a finite quadrature approximation of the log-likelihood. Renaming the data points as $x_1, \dots, x_n$ with $(u_i, t_i) = x_i$ for $i = 1, \dots, n$, then generate m additional “dummy points” $(u_{n+1}, t_{n+1}) \dots, (u_{m+n}, t_{m+n})$ to form a

set of $n+m$ quadrature points (where $m > n$). Then we determine quadrature weights $a_1, \dots, a_m$ so that integrals in (1) can be approximated by a Riemann sum

$$\int_W \int_T \lambda(u, t; \theta) dt du \approx \sum_{k=1}^{n+m} a_k \lambda(u_k, t_k; \theta)$$

where a_k are the quadrature weights such that $\sum_{k=1}^{n+m} a_k = l(W \times T)$ where l is the Lebesgue measure. Then, writing $y_k = e_k/a_k$, the log-likelihood (1) of the template model can be approximated by

$$\log L(\theta) \approx \sum_j a_k (y_k \log \lambda(u_k, t_k; \theta) - \lambda(u_k, t_k; \theta)) + \sum_k a_k.$$

Apart from the constant $\sum_k a_k$, this expression is formally equivalent to the weighted log-likelihood of a Poisson regression model with responses y_k and means $\lambda(u_k, t_k; \theta) = \exp(\theta Z(u_k, t_k))$. This means that the model can be maximised using the standard `glm` function, but also that covariate values must be known in every data and dummy point location. As this is typically unfeasible in practice, we first interpolate covariate values at a very fine regular grid, and then attribute to each data or dummy points the value of the closest point in three dimension. As the covariate location within the analysed region is just 205 sites, the interpolation is performed on a $12^3 = 1728$ point grid. Then, the interpolation at a (data or dummy) point location x_k is performed through the inverse-distance weighting smoothing procedure of the covariate values $Z(x_j)$ at their sampling locations $j = 1, \dots, J$. In such a case, the smoothed value at location x_k is

$$Z(x_k) = \frac{\sum_j w_j Z(x_j)}{\sum_j w_j},$$

where the weight w_j is the j -th element of the inverse p th powers of distance, $\mathbf{w} = \{w_j\}_{j=1}^J = \left\{ \frac{1}{d(x_k - x_j)^p} \right\}_{j=1}^J$, with $d(x_k - x_j) = \|x_k - x_j\|$ the Euclidean distance from x_k to x_j . To fit such a model, the spatio-temporal quadrature scheme is obtained by defining a spatio-temporal partition of $W \times T$ into cubes C_k of equal volume v , assigning the weight $a_k = v/n_k$ to each quadrature point (dummy or data), where n_k is the number of points that lie in the same cube as the point u_k . The number of dummy points should be sufficient for an accurate estimate of the likelihood. We start with a number of dummy points $m \approx 4n$, increasing it until $\sum_k a_k = l(W \times T)$.

4 Results

We propose to model the flash lighting data by a Poisson model with a linear predictor including a non-parametric term for spatio-temporal coordinates and parametric expression for the spatio-temporal covariates, excluding the time phases, as

$$\lambda(u, t) = \exp(f(u, t) + \theta_1 Z_{TAIR}(u, t) + \theta_2 Z_{RELH}(u, t) + \theta_3 Z_{THETA.E}(u, t))$$

where $f(\cdot)$ is a nonparametric function for $(u, t) \in W \times T$, estimated here through thin plate regression splines (Wood, 2003).

Coefficients	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-189.342	31.497	-6.011	0.000***
TAIR	-5.033	0.689	-7.302	0.000***
RELH	-0.893	0.109	-8.224	0.000***
THETA_E	1.039	0.161	6.469	0.000***

Table 1: Parametric Coefficients of the fitted Poisson model

The results in Table 1 indicate that even though the individual effects of dry air temperature and relative humidity tend to have a negative correlation with the total lightning flash rate, equivalent potential energy (which represents a combined effect of dry air temperature and humidity) has a mild positive effect. This makes physical sense since air parcels with higher potential energy being lifted upward by the storm updraft are more likely to reach cold temperatures and since lightning is a manifestation of interaction between graupel and ice crystals, both of which are active between 0 and -40 °C, more surface air parcels (with higher energy), reaching to higher altitudes can increase the lightning flash rate.

5 Conclusions

The analysed dataset represents just one supercell storm (single sample point), with limited spatiotemporal covariate data. Thereby, the preliminary results presented here should not be generalized. We plan to expand this study by applying the methodology developed for this case to a larger dataset representing multiple convective modes in different ambient environments. We would also like to compare the effect of environmental covariates for thunderstorms that predominantly lower either positive or negative charges to the ground. Other future paths include exploring more complex models, like the log-Gaussian Cox processes, multitype Poisson models, and local ones (D’Angelo et al., 2023). Indeed, multitype Poisson models could be useful to characterize the different phases of the storm by assessing the significance of the covariates’ random effects and running a test of segregation.

6 Acknowledgments

The lightning data used in this study were obtained from the Oklahoma lightning mapping array data (OKLMA), provided by the U.S. NOAA National Severe Storms Laboratory. Environmental covariate data were obtained from the Oklahoma Mesonet (Brock et al., 1995; McPherson et al., 2007) stations can be found at <https://doi.org/10.15763/dbs.mesonet.standard>. This work was supported by the Mobilità e Formazione Internazionali - Miur INT project “Sviluppo di metodologie

per processi di punto spazio-temporali marcati funzionali per la previsione probabilistica dei terremoti” and the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001).

References

- Brock, F. V., Crawford, K. C., Elliott, R. L., Cuperus, G. W., Stadler, S. J., Johnson, H. L., and Eilts, M. D. (1995). The oklahoma mesonet: a technical overview. *Journal of Atmospheric and Oceanic Technology*, 12(1):5–19.
- Bruning, E. (2015). Imatools: Imatools-v0.5z-stable.
- D'Angelo, N. and Adelfio, G. (2023). *stopp: Spatio-Temporal Point Pattern Methods, Model Fitting, Diagnostics, Simulation, Local Tests*. R package version 0.1.0.
- D'Angelo, N., Adelfio, G., and Mateu, J. (2023). Locally weighted minimum contrast estimation for spatio-temporal log-gaussian cox processes. *Computational Statistics & Data Analysis*, 180:107679.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.
- MacGorman, D. R., Rust, W. D., Schuur, T. J., Biggerstaff, M. I., Straka, J. M., Ziegler, C. L., Mansell, E. R., Bruning, E. C., Kuhlman, K. M., Lund, N. R., et al. (2008). Telex the thunderstorm electrification and lightning experiment. *Bulletin of the American Meteorological Society*, 89(7):997–1014.
- McDonald, J. R. and Mehta, K. C. (2006). *A recommendation for an Enhanced Fujita scale (EF-Scale)*. Wind Science and Engineering Center, Texas Tech University.
- McPherson, R. A., Fiebrich, C. A., Crawford, K. C., Kilby, J. R., Grimsley, D. L., Martinez, J. E., Basara, J. B., Illston, B. G., Morris, D. A., Kloesel, K. A., et al. (2007). Statewide monitoring of the mesoscale environment: A technical update on the oklahoma mesonet. *Journal of Atmospheric and Oceanic Technology*, 24(3):301–321.
- Rison, W., Thomas, R. J., Krehbiel, P. R., Hamlin, T., and Harlin, J. (1999). A gps-based three-dimensional lightning mapping system: Initial observations in central new mexico. *Geophysical research letters*, 26(23):3573–3576.
- Sharma, M., Tanamachi, R. L., Bruning, E. C., and Calhoun, K. M. (2021). Polarimetric and Electrical Structure of the 19 May 2013 Edmond–Carney, Oklahoma, Tornado Supercell. *Mon. Wea. Rev.*, 149(7):2049–2078.
- Wienhoff, Z. B., Bluestein, H. B., Wicker, L. J., Snyder, J. C., Shapiro, A., Potvin, C. K., Houser, J. B., and Reif, D. W. (2018). Applications of a Spatially Variable Advection Correction Technique for Temporal Correction of Dual-Doppler Analyses of Tornado Supercells. *Mon. Wea. Rev.*, 146(9):2949–2971.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114.

17. Contributed - Statistical machine learning for predictive modelling

- 17.1 - Application of statistical techniques to predict the effective temperature of young stars
- 17.2 - Topological Attention for Denoising Astronomical Images
- 17.3 - LSTM-based Battery Life Prediction in IoT Systems: a proof of concept
- 17.4 - Predictive modeling of drivers' brake reaction time through machine learning methods

Application of statistical techniques to predict the effective temperature of young stars

Marco Tarantino, Loredana Prisinzano, Giada Adelfio

Abstract One of the current topics of discussion in astrophysical research revolves around the duration of the stellar cluster formation process: some theoretical models postulate formation through a single event (rapid process), others suggest multiple formation events (slow process), implying different ages for stars within the same cluster. A crucial variable for deriving age of young stars is the effective temperature. This study focuses on applying various statistical techniques, such as GLM, SVM, PLSR, Boosting, and Random Forest, to better predict the effective temperature of young stars. The results obtained from the analysis highlight that the Random Forest model outperforms other models in terms of predictive performance.

Key words: stellar cluster formation, young stars, effective temperature, Random Forest, Cross-Validation, reddening

1 Introduction

The rate of the star formation process is a key question in astrophysics, particularly regarding whether stars within the same molecular cloud form from a singular event or through multiple events over time [4–6]. Distinguishing between these possibilities is crucial for refining the physical models that describe the process of star formation. Formally, let Δt_i be the observed age spread within a stellar cluster and Δt_e the spread due to measurement errors. It follows that $\Delta t_i = \Delta t_i + \Delta t_e$, where Δt_i represents the true intrinsic age spread. The goal is to determine whether the observed spread is solely due to measurement errors or if there is indeed an intrinsic age spread among young stars within a cluster. This investigation aims to provide

M. Tarantino, G. Adelfio, University of Palermo, Department of Economics, Business and Statistics, Viale delle Scienze 8 e-mail: marco.tarantino01@unipa.it

L. Prisinzano, Istituto Nazionale di Astrofisica - Osservatorio Astronomico di Palermo, Piazza del Parlamento 1

support for the hypothesis of either a single event process ($\Delta t_i = 0$) or a gradual formation process characterized by successive events ($\Delta t_i > 0$). One of the fundamental astrophysical variables for deriving the age of stars through appropriate stellar evolution models is the effective temperature of the stars. Accurate measurements of the effective temperatures can be obtained through spectroscopic observations. These, however, can be acquired only on targeted and generally very bright objects and, therefore, on a very small fraction of stars. In contrast, through photometric observations, positions and magnitudes of several thousands of stars are extracted in a single pointing and used to derive the stellar photometric effective temperatures. However, from photometric observations it is very difficult, at least with the techniques used so far, to measure accurate effective temperatures, due to the absorption effects on the magnitudes. It consists in a reduction in the intensity of radiation from a source, attributed to the absorption by interstellar dust situated between the source and the observation point, causing a variation in the magnitudes. The objective of the present analysis is to identify a statistical method that can predict the proper effective temperatures by using photometric variables as explanatory factors. Then, the same model can be used to predict spectroscopic measurements. The stars of the training set have been selected from the catalogue of stars observed with the Gaia-ESO survey and published in [1], considering only young stars, i.e. with an age of less than about ten million years and a probability of membership higher than 0.9 (a probability value of belonging to a young cluster assigned to each star). Photometric data on the selected stars were extracted from the Gaia DR3, Gaia-ESO, Pan-STARRS and 2MASS surveys. In particular, magnitudes in different filters and colours, calculated as differences in magnitudes in different filters, were considered. Magnitudes and colours are corrected for absorption and reddening respectively using the extinction map by [2]. The final testing dataset includes 1169 young stars, for which 16 explanatory are considered. The primary challenge inherent in this type of data lies in the substantial correlation among the explanatory variables. Magnitudes, despite being measured in different filters, inherently capture luminosity, while colours are derived from the differences between magnitudes in distinct filters.

2 Methods

Regarding the conducted analysis, five techniques were employed for predicting effective temperature: *Generalised Linear Model*, *Support Vector Machine*, *Partial Least Square Regression*, *Boosting* and *Random Forest*. Respect to the training of all five methods, the technique of *Cross Validation*, also known as *k-fold*, is used. This statistical method involves dividing the training set into k subsets of equal size. In each iteration, the k -th subset is used as the validation set, while the remaining portions constitute the training set for that particular iteration. In this application, the dataset is divided into five folds. Each statistical method is then applied five times and the final result, in terms of predictive ability, corresponds to the average of the prediction errors across various iterations. This approach enhances the robustness and validity of the model's performance assessment. The error measures

considered in the analysis are the *Mean Square Error*, *Root Mean Square Error* and *Mean Absolute Error*. In the following, only the Random Forest is presented, as it is the model with the greatest predictive capability, according to the methodology described above. The Random Forest is an *Ensemble method* that uses decision trees to solve classification or regression problems. Trees are obtained through a procedure of segmentation of the sample, dividing it into homogeneous groups with respect to the response variable, through subdivisions based on a single explanatory or linear combination of several explanatory variables. This overcomes the problem of strong correlations in the set of predictors, as in the case of this application. The key approach of the Random Forest lies in bootstrap aggregation: several sub-samples of the training set are created, on which independent trees are constructed. Moreover, in the construction of each tree, a random subset of explanatory variables is selected, breaking the correlation between the trees and helping to overcome the correlation between the predictor variables themselves. This strategy makes the Random Forest robust and suitable for modelling complex and non-linear relationships between the response variable and the predictors. It particularly excels in situations where there is a strong correlation between the explanatory variables, proving to be the most effective approach for addressing the complexities inherent in the current problem.

3 Results

In Table 3, we report results in terms of prediction errors of the five approaches used in forecasting effective temperature. The Root Mean Squared Error (RMSE) is in the same scale as the temperature values and indicates that on average, for the Random Forest, predictions may deviate by approximately 109 degrees Kelvin. This level of error is comparable to the typical inaccuracies observed in spectroscopic sample temperatures. Regarding the Random Forest, we can derive the importance of each

Table 1 Comparison of the five approaches used in terms of MSE, MAE and RMSE to assess their predictive performance.

model	MSE	RMSE	MAE
<i>GLM (Gamma)</i>	33669.19	180.96	114.65
<i>PLSR (Gamma)</i>	55037.16	234.6	182.92
<i>SVR</i>	37214.86	192.01	131.92
<i>Boosting</i>	82543.19	264.27	124.22
<i>Random Forest</i>	13268.14	109.25	55.21

explanatory variable in obtaining the predictions. The importance can be attributed to the number of times the variable is used as one of the first splits of a tree, as it is identified by the algorithm as a discriminative variable for predicting the response. Fig.1 shows the minimum depth and average depth of each variable. By depth is meant at which tree split that variable is used. Specifically, in the graph the average depth is represented by the value in the white box. This average is calculated from the minimum depths observed on the five hundred trees. From the results, it can be

seen that the most important variables are the colors $G - G_{rp}$ and $r - i$, as well as T_{eff}^{phot} , which is an ‘initial’ measure of the actual temperature, obtained on the basis of astrophysical notions. In order to verify the validity of the temperature predictions, we consider two tests: the estimation of reddening and the verification of the predictive ability of the chosen model on an independent test set.

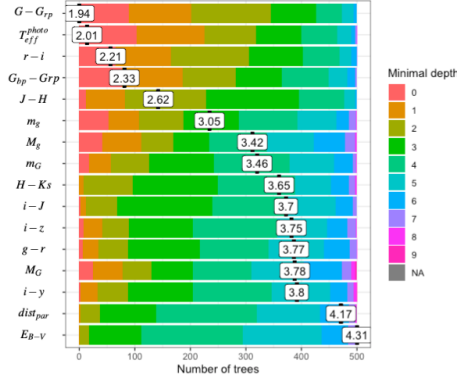


Fig. 1 Average depth values of each variable versus the number of trees used in the Random Forest. Variables are sorted in ascending order of minimum average depth, i.e., by importance.

Reddening estimation. The spectral energy distribution of a star with a given effective temperature can be convolved in different filters. For this reason, theoretical or empirical relationship can be found between the effective temperatures and the intrinsic colors of the stars. To obtain the reddening estimates we refer to the photometric color ($G_{BP} - G_{RP}$). This is obtained as the difference between the magnitudes G observed in the filters BP and RP . Starting from the effective temperatures predicted by the Random Forest, we derived the expected intrinsic color ($G_{BP} - G_{RP}$)₀, i.e. the colour not affected by reddening, by using the relationship given by [3]. In particular, we have interpolated the values of the intrinsic colour by inverting the fifth degree polynomial, using the effective temperatures predicted by the Random Forest. The reddening estimates $E(G_{BP} - G_{RP})$ are obtained as $E(G_{BP} - G_{RP}) = (G_{BP} - G_{RP}) - (G_{BP} - G_{RP})_0$, i.e. the difference between the values of the observed colour ($G_{BP} - G_{RP}$), i.e. affected by reddening and obtained from the Gaia Archive, and the intrinsic values of the colour of the stars ($G_{BP} - G_{RP}$)₀, obtained as described above. As an external test, to check the goodness of the predictions obtained by Random Forest, the estimated $E(G_{BP} - G_{RP})$ values were compared with the values predicted by the map of [2], as shown in Fig. 1. It is important to underline that the values of reddening obtained by Vergely et al. are statistical estimates, because they are based on the deviations of the density of stars obtained from various catalogues as a function of distance. The Vergely et al. values are used in our analysis as initial corrections for reddening, but they are less accurate than the ones obtained from the process described above, because our estimated are based on the individual effective temperature values calculated star by

star with the Random Forest algorithm. As evident from the figure, there is greater variability in the estimates based on the temperature values predicted by Random Forest, because several of the young stars included in our sample are affected by local extinction due to the circumstellar material.

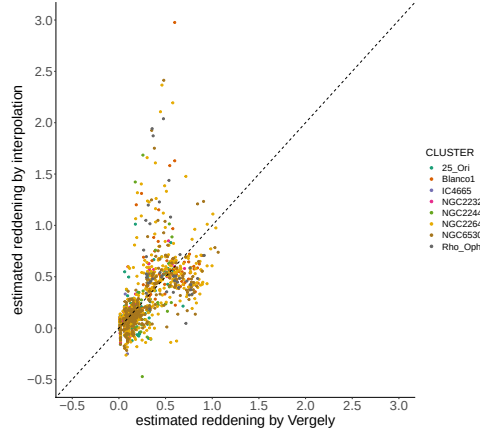


Fig. 2 Comparison of reddening values obtained by Vergely et al. and those obtained through the relationship given by [3] and our results. Different colors correspond to the various stellar clusters included in our sample.

Test set prediction. For diagnostics purpose, we evaluate the predictive capability of the model on stars that are not part of the training dataset. Then, in the subsequent analysis, temperatures of stars selected from the study by [1] with a membership probability exceeding 0.9 and associated with GES clusters situated in the Southern Hemisphere are taken into account. The test set is therefore composed of 1408 stars, for which all the magnitudes and colours necessary for the application of the model are available. As shown in Fig. 2, the model succeeds in obtaining good temperature estimates, especially for temperature values below 5 000 degrees Kelvin, whereas the estimates for higher temperatures are less accurate.

4 Conclusion

The aim of the work was to identify the most effective statistical model for predicting photometric effective temperatures. This involved magnitude and color values in various photometric bands, aiming to closely align with the spectroscopic temperature values provided for specific clusters of young stars. The choice of the best model was based on the error obtained in terms of the predictive capability of the model (MSE, RMSE and MAE), assessing the discrepancies between the temperature values predicted by the model and the reference spectroscopic temperature values. From the results obtained through Cross-Validation, the Random Forest is the best model to predict the temperature values of young stars. In order to assess

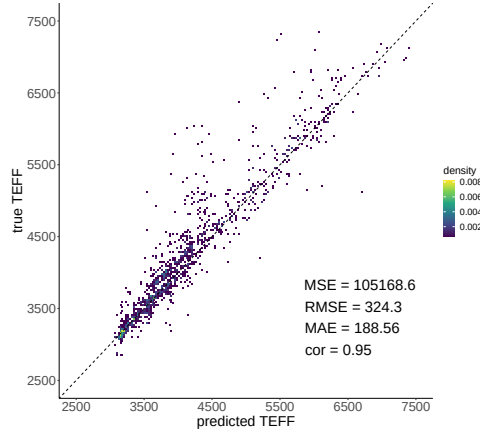


Fig. 3 Comparison of the photometric temperature values of the test set obtained from GES and the predictions obtained through Random Forest.

the goodness of the obtained temperature predictions, two tests were carried out: the reddening estimates and the generalisation capability of the model. Both tests further confirm the Random Forest’s ability to predict the temperature of such stars using photometric variables as explanatory variables. It is crucial to highlight certain limitations of this approach, including challenges in predicting higher values of effective temperature (beyond 5000 Kelvin degrees) and the inability to interpret the impact of explanatory variables on the response. In summary, the Random Forest emerges as the most effective approach for analyzing these astrophysical, characterized by substantial correlations among the involved photometric variables.

References

1. Jackson, R.J. et al.: The Gaia-ESO Survey: Membership probabilities for stars in 63 open and 7 globular clusters from 3D kinematics. Volume **509**, number **2**, pp. 1664–1680, [adsurl = https://ui.adsabs.harvard.edu/abs/2022MNRAS.509.1664J\(2022\)](https://ui.adsabs.harvard.edu/abs/2022MNRAS.509.1664J(2022))
2. Vergely, J.L. et al.: Three-dimensional extinction maps: Inverting inter-calibrated extinction catalogues. Volume **664**, eid **A174**, pp. A174.(2022)
3. Montalto, M. et al.: The all-sky PLATO input catalogue. Volume **653**, eid **A98**, pp. A98, [adsurl = https://ui.adsabs.harvard.edu/abs/2021AA...653A..98M\(2021\)](https://ui.adsabs.harvard.edu/abs/2021AA...653A..98M(2021))
4. Palla, Francesco and Stahler, Steven W.: Accelerating Star Formation in Clusters and Associations, journal = *Astrophysical Journal*, volume **540**, number **1**, pp. 255–270, [adsurl = https://ui.adsabs.harvard.edu/abs/2000ApJ...540..255P\(2000\)](https://ui.adsabs.harvard.edu/abs/2000ApJ...540..255P(2000))
5. Kroupa, Pavel et al.: Evidence for feedback and stellar-dynamically regulated bursty star cluster formation: the case of the Orion Nebula Cluster, journal = *Astronomy & Astrophysics*, volume **612**, eid **A74**, pp. A74(2018)
6. Karam, Jeremy and Sills, Alison: Modelling star cluster formation: mergers, journal = *Monthly Notices of the Royal Astronomical Society*, volume **513**, number **4**, pp. 6095–6104, [adsurl = https://ui.adsabs.harvard.edu/abs/2022MNRAS.513.6095K\(2022\)](https://ui.adsabs.harvard.edu/abs/2022MNRAS.513.6095K(2022))

Topological Attention for Denoising Astronomical Images

Riccardo Ceccaroni and Pierpaolo Brutti

Abstract Astronomical observations often involve capturing faint signals from celestial objects, such as distant galaxies or dim stars. These signals can be easily overwhelmed by noise, which includes electronic noise, atmospheric interference, and other artifacts. It is then clear that using an effective denoising algorithm is crucial to improve the accuracy of data interpretation by extracting the actual astronomical signals from the noisy background. In this work, we introduce a novel denoising approach based on deep neural networks with pixel attention derived from the topological structure extracted by 0-dimensional Persistent Homology (PH). Our network architecture combines an AutoEncoder (AE) with a Topological Module (TM) that assigns weights to pixels based on PH, providing nuanced attention to the relevant image regions. Our methodology is evaluated against widely used denoising methods, showcasing its potential for enhancing the quality of astronomical images.

Key words: Denoising, Astronomical Image, Persistent Homology, Topological Attention

1 Introduction

In the ever-expanding field of astronomical data analysis, obtaining high-quality images is key. However, atmospheric fluctuations, optical distortions, instrument calibration, and inherent noise during acquisition, may strongly degrade the quality

Riccardo Ceccaroni
Università di Roma “La Sapienza”, P.le Aldo Moro 5, Roma, 00185, Italy
e-mail: riccardo.ceccaroni@uniroma1.it

Pierpaolo Brutti
Università di Roma “La Sapienza”, P.le Aldo Moro 5, Roma, 00185, Italy
e-mail: pierpaolo.brutti@uniroma1.it

of observations and, in turn, adversely impact the overall accuracy of the scientific analyses. Consequently, a post-acquisition image processing phase becomes essential. In fact, astronomical images undergo a meticulous cleaning procedure that typically goes from bias and dark frame subtraction to flat field calibration, all steps aimed at rectifying errors usually arising from instrument-related conditions. On the other hand, removing the intrinsic noise that affects the image acquisition process remains a challenging problem and the development of effective denoising algorithms is an active area of research.

2 Related works

Numerous classical denoising methods, such as Perona-Malik (PM) filtering [1], bilateral filtering [2], total variation denoising [3], non-local means [4] and wavelets [5], have been proposed to effectively reduce noise in astronomical images. PM method is considered the gold standard among them [6]. Recent years have witnessed experiments utilizing neural networks for denoising, demonstrating promising results [7]. In this study, we introduce a novel denoising algorithm based on deep neural networks with pixel attention derived from the topological structure extracted from the image using a 0-dimensional Persistent Homology (PH).

3 Theoretical Background

Topological Data Analysis (TDA) is a relatively new area of research that combines principles from algebraic and computational topology with data analysis techniques. The primary goal of TDA is to extract and analyze the topological features inherent in complex data sets to provide a way to understand their qualitative shape and connectivity structure. One key ingredient of TDA is PH, a concept that provides a robust and multiscale representation of the topological features present in the data. PH captures the evolution of topological structures across various scales by identifying and tracking the birth and death of topological features, such as connected components, loops, and voids, throughout a filtration process. The filtration process involves systematically varying a parameter, typically representing a distance or scale. The 0-dimension PH closely correlates with connected components, meaning that a connected object in the image originates at a point p_{birth} with a *birth* value and concludes at point p_{death} with a *death* value (merging with another connected object). The resulting *birth* and *death* values are succinctly presented in a diagram known as a Persistence Diagram (PD). The difference between *birth* and *death* is also called *lifetime*. For more theoretical aspects see [8, 9, 10].

4 Methods

In this work, we introduce a deep neural network incorporating an AutoEncoder (AE) augmented with a Topological Module (TM), designed to emphasize pixels within the image. In the upcoming subsections, we detail the architectures of the AE and AE+TM networks, along with the process of constructing a dataset used for training and testing these models.

4.1 AutoEncoder

The AE architecture we consider is rather simple and it is composed of convolutional layers only. Since we will be testing its performance on a dataset of relatively small (140×140) images, as suggested in [11], we decided not to include Maxpool layers to reduce the unavoidable edge effects at a minimum.

More in details (see Figure 1), the network consists of an initial convolutional layer with 1 input channel and 64 output channels, followed by three convolutional layers with both input and output at 64 channels. The final layer transitions from 64 input channels to 1 output channels. All convolutional operations employ a 3×3 kernel with padding set to 1.

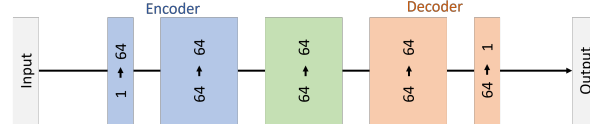


Fig. 1 AutoEncoder architecture.

4.2 Topological Module

We introduce now the TM designed to assign a weight to each pixel (x, y) of the image. The first step is to compute the PD of our input image. After that, for each object $\mathbf{p} = (p_{\text{birth}}, p_{\text{death}})$ in the PD, we build a “local” Gaussian weight map

$$w_{\mathbf{p}}(x, y) = \text{birth} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \left((x_{\text{birth}} - x)^2 + (y_{\text{birth}} - y)^2 \right) \right\},$$

where $(x_{\text{birth}}, y_{\text{birth}})$ denotes the spatial coordinates (“in-image”) of the birth point p_{birth} associated to that object, and σ is equal to the Euclidean distance between $(x_{\text{birth}}, y_{\text{birth}})$ and $(x_{\text{death}}, y_{\text{death}})$, the “in-image” counterparts of p_{birth} and p_{death} . These individual images $\{w_{\mathbf{p}}\}_{\mathbf{p}}$ are subsequently summed together to produce the

topological weight map W . An additional enhancement to the TM involves the incorporation of a filter applied to the PD. In its default configuration, this filter is formed by applying a threshold to the lifetime values, but alternative techniques like Persistent Entropy [12] can also be employed. The purpose of this filter is to remove topological noise from the PD by eliminating close to the diagonal objects and, consequently, to accelerate the construction of the W map.

4.3 Network proposal

The network we propose combines the AE and the TM described in the preceding paragraphs (see Figure 2). For this purpose, we employ three AE networks with the same structure, denoted as M_1 , M_2 , and M_3 . Hence, given an image I as input, we first calculate $W = TM(I)$. Subsequently, the image I is passed through M_1 , while the weight map W traverses the network M_2 . This yields $O_1 = M_1(I)$ and $O_2 = M_2(W)$. These outputs, having the same dimensions as I , are multiplied together. Finally, applying M_3 to the result, we get the output $O = M_3(O_1 \times O_2)$.

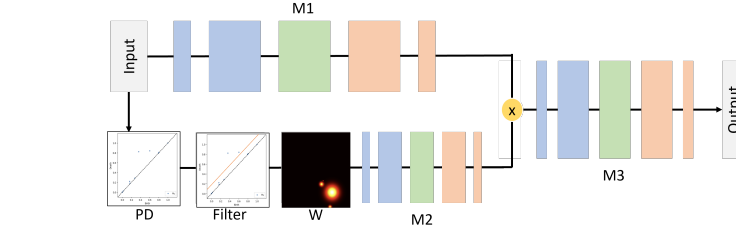


Fig. 2 AutoEncoder with Topological Attention Network.

4.4 Astronomical Datasets

The dataset was generated using the image “*hlsp_illustris_jwst_nircam_mag30-field-11-10_f200w_v1_lightcone.fits*” sourced from the *ILLUSTRIS* dataset [13], accessible through the *MAST* repository. The *ILLUSTRIS* simulation is a leading cosmological tool, aiding astronomers in studying dark matter, gas dynamics, star formation, and feedback processes shaping our universe. We used a simulation from JWST’s NIRC*am* instrument, observing field MAG30 with the F200W filter. This master image was partitioned into patches of (140×140) pixels that, after normalization to bring pixel values in $[0, 1]$, were subsequently corrupted with Gaussian noise (mean 0, variance 0.2). The resulting dataset is composed of 1000 images used for training, with an additional 222 images allocated for both, validation and testing.

5 Results

In this section we compare our proposal with well-established denoising techniques documented in the literature, both classical, like PM, and network-based, like UNet [7] and the AE described in Section 4.1. PM was tuned via a grid-search over its two parameters: κ , which controls the sensitivity to edges, and γ , which regulates the smoothing at each iteration. In particular, we tried $\kappa \in \{10, 15, 20, 25, 30\}$ and $\gamma \in \{0.1, 0.14, 0.17, 0.2\}$, obtaining $(\kappa, \gamma) = (20, 0.2)$ as optimal pair. The neural networks underwent training with 100 epochs and early stopping to mitigate overfitting. The optimizer adopted was Adam with a learning rate of $5e^{-5}$. Notably, UNet exhibited overfitting in the 9th epoch, and AE showed signs of overfitting in the 10th epoch. On the other hand, AE+TM completed training but exhibited a slight underfitting tendency¹. The loss function employed in all models is MSE.

Table 1 presents a comparative analysis of the methods using traditional metrics for image quality assessment such as MSE, SSIM, and PSNR. Additionally, Figure 3 visually illustrates the outputs of the different methods considered in this study.

Table 1 Comparison between methods

	MSE	SSIM	PSNR
PM	0.0105	0.3202	20.1539
UNet	0.0015	0.5991	29.1328
AE	0.0023	0.4885	27.6992
AE+TM (NoFilter)	0.0015	0.7062	30.3320
AE+TM (Filter)	0.0015	0.6951	30.2448

6 Conclusions

This study demonstrates the effectiveness of combining AE and TM to design a powerful tool capable of enhancing the overall image quality by suitably pushing pixel focus into the analysis.

When compared to well-established methods, our approach exhibits competitive denoising performance, highlighting the potential of incorporating topological features into image processing tasks.

To further improve the model’s capabilities, adjustments to the architecture, refinement of the “local” weight function $w_p(x, y)$, and experimentation with more extensive datasets are definitively in order.

This research lays the foundation for future studies, where the integration of topological attention mechanisms in the field of image denoising will progressively advance our understanding of the universe.

¹ For code and experiments details: <https://github.com/riccardoc95/astroTopoAttention>

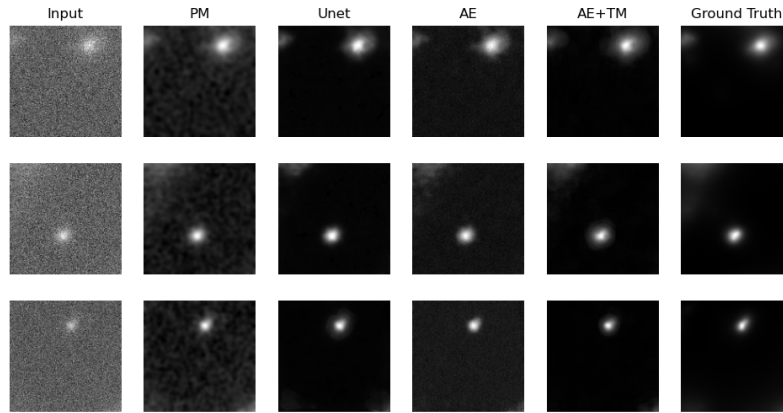


Fig. 3 Results obtained on three test images. Reconstructions from AE+TM and UNet show less background noise than those from PM and AE. In addition, in line with its higher SSIM value, AE+TM tends to better preserve shapes and structures when compared to UNet.

References

1. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.
2. C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998.
3. Leonid Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 11 1992.
4. A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 60–65 vol. 2, 2005.
5. O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(4):14–38, 1991.
6. Valerio Roscani, Silvia Tozza, Marco Castellano, Emiliano Merlin, Daniele Ottaviani, Maurizio Falcone, and A. Fontana. A comparative analysis of denoising algorithms for extragalactic imaging surveys. *Astronomy & Astrophysics*, 643, 09 2020.
7. Antonia Vojtekova, Maggie Lieu, Ivan Valtchanov, Bruno Altieri, Lyndsay Old, Qifeng Chen, and Filip Hroch. Learning to denoise astronomical images with u-nets. *Monthly Notices of the Royal Astronomical Society*, 503, 11 2020.
8. Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference*. 10 2017.
9. Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. 01 2010.
10. James Munkres. *Elements of Algebraic Topology*. 01 1984.
11. Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? 05 2018.
12. Nieves Atienza, Rocio Gonzalez-Díaz, and Matteo Rucco. Persistent entropy for separating topological features from noise in vietoris-rips complexes. *Journal of Intelligent Information Systems*, 52, 06 2019.
13. D. Nelson, A. Pillepich, S. Genel, M. Vogelsberger, V. Springel, P. Torrey, V. Rodriguez-Gomez, D. Sijacki, G.F. Snyder, B. Griffen, F. Marinacci, L. Blecha, L. Sales, D. Xu, and L. Hernquist. The illustris simulation: Public data release. *Astronomy and Computing*, 13:12–37, 2015.

LSTM-based Battery Life Prediction in IoT Systems: a proof of concept

Vanessa Verrina, Andrea Vennera and Annarita Renda

Abstract Predicting accurately battery life is crucial for applications ranging from electric vehicles and energy storage to emerging technologies in the context of Industry 4.0. The Internet of Things (IoT) has witnessed significant growth, permeating diverse domains. The burgeoning electronics industry has fuelled the demand for portable devices, with IoT devices facing a considerable challenge in ensuring reliable and prolonged battery life. Rechargeable batteries, such as Li-ion systems, are preferred for their high energy density and extended cycle life, aligning with the increasing demand for miniature wireless devices. This work presents a proof-of-concept, employing an optimised Long Short-Term Memory (LSTM) recurrent neural network model to forecast battery life in IoT systems. Utilising publicly available lab datasets and controlled conditions, the results are bench-marked against state-of-the-art models, showcasing the superior performance of the proposed approach. This methodology addresses the critical concern of IoT device battery life, thereby enhancing the reliability and longevity of IoT applications.

Key words: IoT Networks, Battery Life Prediction, Neural Networks, LSTM

1 Introduction

The rise of Internet-of-Things (IoT) has brought attention to challenges like security and energy optimisation. Energy efficiency is a major concern due to IoT de-

Vanessa Verrina

Department of Economics, Statistics and Finance “Giovanni Anania”, University of Calabria, Via P. Bucci, Cubo 0 C, Arcavacata di Rende, CS, Italy, e-mail: vanessa.verrina@unical.it

Andrea Vennera

R&D Department - SCAI Lab, 87036, Rende, CS, Italy, e-mail: andrea.vennera@grupposcai.it

Annarita Renda

R&D Department - SCAI Lab, 87036, Rende, CS, Italy, e-mail: annarita.renda@grupposcai.it

vices constantly performing computations with limited power. Various techniques aim to minimise energy consumption, but accurate prediction of device battery life is crucial for network stability. This is especially important for IoT nodes relying on effective battery conservation. Inaccurate predictions can lead to loss of critical information and undermine the potential for inferences [1]. To address these challenges, accurately predicting the battery life of IoT devices becomes imperative to enhance their lifespan. The lithium-ion battery holds a crucial role in storing and supplying energy, owing to its higher energy density, lighter weight, and extended charge and discharge cycles. In the past years, a big effort is being put in developing high capacity miniature-micro Li-ion batteries for IoT applications [2].

The challenge lies in the variation of expected battery life based on environmental factors (e.g., temperature and humidity) and usage intensity. Even batteries of the same model exhibit different service lives in practical applications. Prognostic and Health Management (PHM) of Li-ion batteries aims to detect faults early and predict the degradation progression by: i) monitoring the State of Health (SOH) of Li-ion batteries, which is determined by comparing the remaining capacity of the battery to its initial capacity; ii) predicting their Remaining Useful Life (RUL), which estimates the remaining predicted lifetime of the battery by considering its usage and degradation relative to the failure threshold [3, 4, 5, 6].

This study focuses on predicting the remaining battery life of Li-ion batteries by proposing a method that uses Long Short Term Memory (LSTM) network. The effectiveness of LSTM in time series predictions offers improved prediction accuracy compared to other data-driven methods. The aim of the work is to use this model to predict the battery life well in advance so that the batteries of any IoT node can be recharged or changed early enough. The experiments conducted in this study use the publicly available batteries dataset from the data repository of the NASA Ames Research Center. This work is conducted within the context of the IBIS ECO project [7].

2 Proposed Methodology

Dataset

The method proposed in this paper is validated by using the Li-ion battery datasets from the data repository of the NASA Ames Research Center. Each dataset records information about charging, discharging and impedance of Li-ion batteries. More details can be found in [8].

Problem Statement

As indicated on description file of the datasets, when the capacity of the battery drops to 70% of the nominal capacitance (2 Ah), the End-of-Life (EOL) of the

battery is determined. Since the capacity of batteries can only be measured in the discharge cycle data, we collect the data related to the discharge of the battery only. The SOH of the battery is defined as

$$SOH = \frac{C_n}{C_i}, \quad (1)$$

where, C_n is the capacity of n-th cycle, C_i represents the nominal capacity of the battery. At the same time, the EOL of the battery is set to 70% of the nominal capacity of the battery. In our case, $EOL = 0.7 * 2 \text{ Ah} = 1.4 \text{ Ah}$.

We indicate the RUL as the number of cycles remaining before the battery reaches its EOL.

Proposed Model

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) particularly effective in the domain of time series prediction. Unlike the original RNN, LSTM can both store and transfer cell states, endowing the network with prolonged memory capabilities. This design choice empowers LSTM-based neural networks to excel in long-term memory tasks [9].

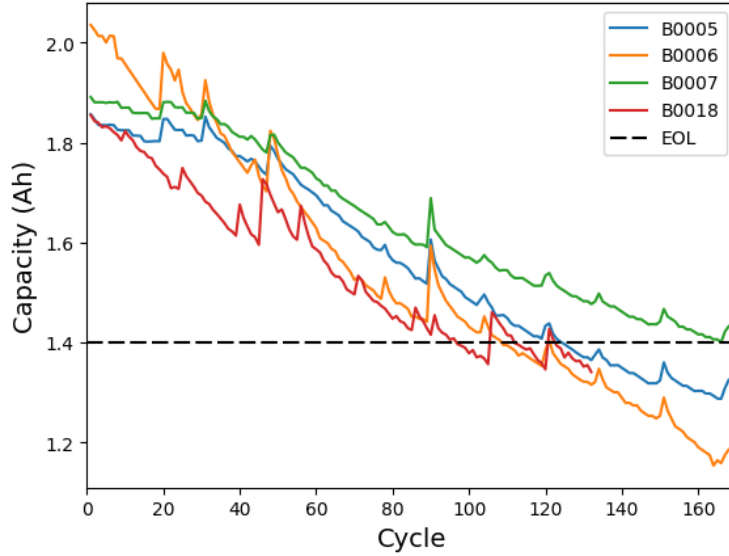


Fig. 1 Capacity as a function of the charging/discharging cycles and End-Of-Life (EOL) of lithium-ion batteries.

Pre-processing

During this phase, the dataset undergoes pre-processing to make it compatible with the prediction algorithm. This involves handling missing values and standardisation. Missing values are addressed using the padding interpolation technique to ensure data integrity. Additionally, the standard scalar technique is employed to normalise the data, allowing all values in the dataset to conform to a common scale.

Training the LSTM

Several parameters of the LSTM such as number of layers, optimisation function, activation function, number of epochs, window size etc. are chosen with the help of random search algorithm. For each battery, the data corresponding to the first fifty cycles are used for training the LSTM and the rest is used to test the performance of the LSTM.

Validating the model

In order to evaluate the performance of the model, RMSE is used as the evaluation criterion of the model. RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad (2)$$

where N is the number of data points, y_i is the observed value at data point i , and $\hat{y}_i$ is the predicted value at data point i . The outcomes produced by the LSTM model are subsequently assessed against traditional state-of-the-art machine learning models to analyse and justify its superiority.

3 Results and Discussion

We use LSTM to train the model. Our model consists of an LSTM layer composed of 200 neurons and a dropout layer with rate 0.3 to avoid overfitting. The optimizer is stochastic gradient descent (SGD), and the learning rate is 0.02. The window size is chosen to be 10. We show the results of one of the battery life cycle datasets, B0005. These results are shown in Fig. 2.

As shown in the graph, the predicted capacity (red curve) and how it behaves over time is very close to the real values (blue curve). While the EOL is reached at the cycle 128 for the real data, the predicted data reach the EOL at the cycle 127. Thus, the error in the estimation of the RUL was -1. The goodness of the model is

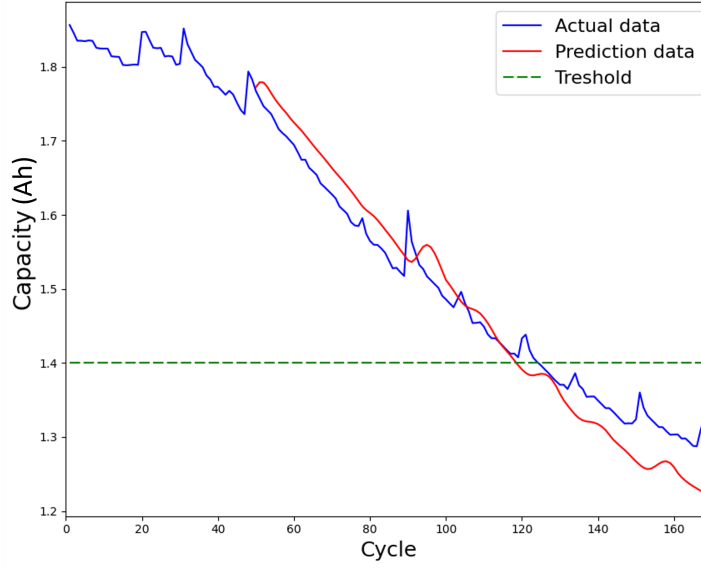


Fig. 2 Comparison between the discharge profile of the battery B0005 (blue curve) and the predicted one (red curve) obtained by training the model LSTM on the data related to the first 50 cycles.

corroborated by the value of the RMSE which is 0.038, below the values reported in literature of about 0.05 [5].

Moreover, other ML algorithms were tested against the B0005 battery dataset in order to select the one that better fit the data. Among them, we have selected Support Vector Regression (SVR) and Random Forest (RF). The results, in terms of RMSE and R^2 , are compared in Table 1.

	SVR	RF	LSTM
RMSE	0.067	0.049	0.038
R^2	0.543	0.767	0.973

Table 1 Performance analysis of proposed model vs ML models.

The effectiveness of the proposed LSTM model is evident as it consistently outperformed the other models examined in this study. Table 1 illustrates the superior performance of the proposed model across key metrics such as mean squared error and test variance score when compared to state-of-the-art models. The outcomes affirm that the LSTM algorithm not only exhibits a rapid convergence rate and substantial error reduction in estimating the global optima but also achieves superior prediction results with significantly reduced training time.

4 Conclusions

The Internet of Things (IoT) has undergone significant evolution, exerting a profound impact across personal, societal, and professional domains. Irrespective of the application area, the optimisation of battery life stands out as a paramount requirement.

In this proof-of-concept study, we advocate the use of lithium-ion batteries, where the monitoring of State of Health (SOH) and the prediction of Remaining Useful Life (RUL) assume critical importance. To address these challenges, we propose the use of a Long Short-Term Memory (LSTM) recurrent neural network. Our proposed methodology is rigorously tested using datasets from NASA's Li-ion batteries. The experimental results underscore the efficacy of our proposed approach, demonstrating superior accuracy compared to state-of-the-art machine learning models in both SOH monitoring and RUL prediction for Li-ion batteries.

As part of future studies and applications, we propose to test the chosen model on authentic data sourced from the IoT sensors currently operational within the context of the IBIS ECO project.

Acknowledgements This research was conducted as part of the IBIS ECO "IoT-based Building Information System for Energy Efficiency & Comfort" project funded under the ERDF Operational Program 2014-2020 - Action 1B.1.2.2. Public Notice: "Complex Research and Development Projects 'CORES'" - Thematic Areas: "Energy and Bioeconomy" - CUP G49J19001400004.

References

1. S. R. K. Somayaji, M. Alazab, M. MK, A. Bucchiarone, C. L. Chowdhary and T. R. Gadekallu, A Framework for Prediction and Storage of Battery Life in IoT Devices using DNN and Blockchain, IEEE Globecom Workshops, (2020).
2. Anju Toor, Albert Wen, Filip Maksimovic, Abhinav M. Gaikwad, Kristofer S.J. Pister, Ana C. Arias, Stencil-printed Lithium-ion micro batteries for IoT applications, Nano Energy, Volume 82 (2021).
3. Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, Iyapparaja Meenakshisundaram, Thippa Reddy Gadekallu, Sparsh Sharma, Mohammed Alkahtani, Mustufa Haider Abidi, Deep Neural Networks Based Approach for Battery Life Prediction, Computers, Materials & Continua, 69(2), 2599-2615, (2021).
4. Severson, K.A., Attia, P.M., Jin, N. et al. Data-driven prediction of battery cycle life before capacity degradation. Nature Energy 4, 383–391 (2019).
5. Laisuo Su, Mengchen Wu, Zhe Li, Jianbo Zhang, Cycle life prediction of lithium-ion batteries based on data-driven methods, eTransportation, Volume 10 (2021).
6. J. Qu, F. Liu, Y. Ma and J. Fan, "A Neural-Network-Based Method for RUL Prediction and SOH Monitoring of Lithium-Ion Battery," in IEEE Access, vol. 7, pp. 87178-87191, (2019).
7. <https://www.ibiseco.it/>.
8. B. Saha and K. Goebel, "Battery Data Set", NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA (2007).
9. S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

Predictive modeling of drivers' brake reaction time through machine learning methods

Alessandro Albano^{1,2}, Giuseppe Salvo^{2,3} and Salvatore Russotto^{2,3}

Abstract This paper investigates the task of predicting drivers' Brake Reaction Time (BRT) using machine learning methods in the context of driving safety. Data is collected using a driving simulator, and interpretability tools such as variable importance and multidimensional partial dependence plots are utilised to interpret the results. The study provides insights into the factors influencing driving safety, with implications for driver training and road safety interventions.

Key words: reaction time, driving simulation, road safety, machine learning gradient boosting

1 Introduction

Driving safety hinges on driver alertness, a cornerstone of this multifaceted activity. Statistics underscore “human error” as the primary culprit in accidents, constituting around 30% of incidents, with half linked to distracted driving [5]. Research by Koopmans et al. [6] highlights how fatigue and stress compromise a driver's reaction time, as evidenced by the increased trajectory variance observed in fatigued drivers during road and simulator tests. However, the intricacies of driving demand further investigation. While human error is often pinpointed as the root cause of accidents, Dekker [3] considered human error within a broader systemic framework that includes technology and organisational factors.

Driving simulators, offering controlled environments, are indispensable for studying drivers' reaction times and conducting tests safely. These simulators must ensure reliability, faithful virtual reality, scene control, motion simulation, data recording,

¹Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

²Sustainable Mobility Center (Centro Nazionale per la Mobilità Sostenibile—CNMS)

³Department of Engineering, University of Palermo, Palermo, Italy

e-mail: alessandro.albano@unipa.it; giuseppe.salvo@unipa.it; salvatore.russotto01@unipa.it

and behaviour assessment tools. Simulators hold advantages over field tests, providing controlled settings and standardised procedures for data collection. Validation, often supplemented with real accident data, is essential for simulator studies [8].

The study aims to predict drivers' brake reaction times and examine the factors influencing these reaction times within a controlled environment. Recognising the critical importance of driver alertness in ensuring safety on the roads, our research builds upon the understanding that human error, including delayed reaction times, significantly contributes to accidents. By employing machine learning techniques, we aim to uncover patterns and factors that influence brake reaction times, thereby improving our ability to address potential safety risks.

2 Data

The Department of Engineering at the University of Palermo possesses a state-of-the-art driving simulator that meets the rigorous standards outlined by the Ministry of Infrastructure and Transport [1]. Equipped with an array of sensors, cameras, and microphones, the simulator's advanced data acquisition platform captures comprehensive information regarding driver behaviour, vehicle dynamics, and environmental conditions, enabling meticulous analysis of driving performance. This driving simulator offers remarkable flexibility, allowing for seamless adjustments to the virtual environment, vehicle characteristics, and experimental parameters. Designed to emulate a realistic yet safe driving environment, the simulator ensures data reliability while minimising potential risks.

For consistency with real-world driving conditions, a specific vehicle configuration has been meticulously selected: i) Truck with a laden mass exceeding 3.5 tons, ii) 8-speed automatic transmission, iii) Unladen vehicle, and iv) Iveco Eurocargo cab equipped with comprehensive controls. Participants in the study firstly engage in 10 minutes of free driving on a motorway route characterised by moderate traffic and sunny weather conditions. This phase is designed to familiarise participants with the vehicle controls and the driving environment. Then, participants navigate an urban/suburban road under sunny weather conditions with an automatic transmission. The route involves encountering queued or isolated vehicles, requiring participants to stop promptly in response to braking by the preceding vehicle or an audible signal.

To ensure an immersive driving experience devoid of distractions, tests are conducted individually. The study recruits drivers aged between 35 and 61, possessing valid driving licenses and Driver Qualification Cards (CQC), with annual driving experience ranging from 20 000 to 65 000 kilometers.

Upon completion of each test, a detailed report is generated, documenting key metrics such as vehicle speed at braking, reaction time, and total braking distance.

3 Methods

The machine learning techniques employed to predict Brake Reaction Time (BRT) based on explanatory variables were:

Gradient Boosting Machine (GBM) [4]: GBM sequentially fits a series of weak learners, often decision trees, to the residuals of the preceding model, iteratively refining predictions by emphasizing previously mispredicted instances.

Support Vector Machine(SVM) [9]: SVM, constructs a hyperplane in a high-dimensional space to separate data points into different classes, maximizing the margin between classes while minimizing classification errors, making it effective for both linear and non-linear classification tasks through the use of various kernel functions.

Random Forest (RF) [2]: RF works by constructing multiple decision trees using random subsets of the data and features, then combining their predictions to improve accuracy and reduce overfitting, making it robust for various predictive tasks.

k-Nearest Neighbors (KNN) [10]: KNN predicts the target value of a data point by averaging the target values of its k nearest neighbors, where proximity is determined based on a chosen distance metric such as Euclidean distance.

These methods offer diverse approaches to analyzing BRT relationships, providing insights into both linear and non-linear dependencies within the data. Furthermore, interpretability tools such as variable importance and multidimensional partial dependence plots [7] are employed to interpret the results of the best-performing method.

4 Results

In order to assess the predictive accuracy of machine learning techniques for predict BRT, we conducted a 10-fold cross-validation. The results are presented in Table 1, showing the mean squared error (MSE) derived from the cross-validation process.

GBM stands out with the lowest cross-validation MSE of 0.0513, signifying robust

Table 1: Comparison of cross-validation MSE.

Method	Cross-validation MSE
GBM	0.0513
SVM	0.0515
RF	0.0533
KNN	0.0583

predictive accuracy. This low MSE suggests that GBM excels in estimating

Brake Reaction Time (BRT) based on the chosen explanatory variables, with relatively minor prediction errors. The SVM approach demonstrates competitive performance, yielding a cross-validation MSE of 0.0515, similar to GBM. RF exhibits a cross-validation MSE of 0.0533, slightly higher than GBM and SVM. However, this outcome still reflects a reasonably accurate predictive performance, suggesting that RF can effectively estimate BRT with only a slightly higher error. Finally, KNN produces the highest cross-validation MSE among the four methods, with a value of 0.0583.

We now turn to assessing the influence of predictors in the two models. Firstly, we examine the relative importance of predictors in the GBM model in Table 2.

Table 2: Relative Influence of Predictor Variables in GBM model.

Variable	Relative Influence (%)
Distance to Obstacle	26.7
Vehicle Speed at Brake Event	19.5
Average Speed	13.3
Time to Obstacle	12.2
Fuel Consumption per 100km	10.4
Maximum Speed	8.74
Distance Traveled	4.23
Fuel Consumption on the Route	3.30
CO2 Emissions	1.67

Among these variables, “Distance to Obstacle” stands out as the most crucial determinant of BRT, with a relative importance of 26.7%. In addition, “Vehicle Speed at Brake Event” also holds considerable importance in predicting BRT, with a relative importance of 19.5%. This variable represents the speed of the vehicle at the moment when the driver initiates braking. The relative importance of the remaining variables in predicting Brake Reaction Time (BRT) using the GBM model is distributed across the following factors: “Average Speed,” “Time to Obstacle,” “Fuel Consumption per 100km,” “Maximum Speed,” “Distance Traveled,” “Fuel Consumption on the Route,” and “CO2 Emissions.” While these variables exhibit lower relative importance values, their contributions should not be dismissed. They collectively represent a broad set of factors that influence BRT to varying degrees.

To provide a deeper understanding of how pairs of variables jointly influence the model’s predictions, we utilise Bidimensional Partial Dependence Plots. These plots allow for visualising the relationship between two variables and the target outcome while controlling for the effects of other variables, shedding light on potential interactions and critical thresholds in the data. Figure 1 shows bidimensional PDPs, focusing on the joint effect of Vehicle Speed at Brake Event’ and ‘Distance to Obstacle’ on Brake reaction time.

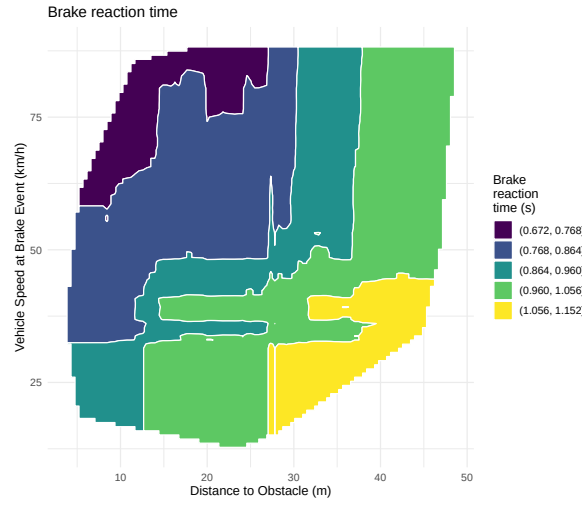


Fig. 1: Brake reaction time predictions as a function of Vehicle Speed and Distance to obstacle (2-dimensional PDP Plot).

The figure offers insights into regions where drivers exhibit shorter reaction times. Notably, it reveals that for high speeds, starting from 55 km/h onwards, and short distances to obstacles (up to 20 meters), drivers tend to have estimated reaction times spanning from 0.672 to 0.768 seconds. Conversely, for lower speeds, within the range of 20 to 40 km/h, and at a greater distance from the obstacle, typically between 35 and 45 meters, drivers tend to exhibit estimated reaction times ranging from 1.056 to 1.152 seconds.

5 Conclusions

In conclusion, our study compared four regression methods for estimating Brake Reaction Time (BRT) using selected explanatory variables. Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) demonstrated the lowest cross-validation Mean Squared Error (MSE), indicating robust predictive accuracy. Random Forest (RF) also performed well, while K-Nearest Neighbors (KNN) showed relatively less accurate predictive performance.

Key variables such as “Distance to Obstacle” and “Vehicle Speed at Brake Event” significantly influenced BRT prediction, underscoring the crucial role of these factors in modelling driver behaviour and enhancing road safety. Notably, our experiments suggest that drivers tend to brake earlier at higher speeds and shorter distances, likely indicating heightened vigilance.

Overall, our findings contribute to understanding BRT prediction and can inform the

development of road safety strategies and driver assistance systems. Indeed, in the last year, some cities have proposed the introduction of a 30 km/h speed limit in all urban areas. Our study shows that lower speeds can lead to longer reaction times, potentially influencing accident rates.

Future research may explore developing real-time prediction systems that could enhance road safety further.

Acknowledgements This work has been supported by the European Union - NextGenerationEU - National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033— 17/06/2022, Spoke 2, CUP B73C2200076000.

References

- [1] Ministero delle infrastrutture e della mobilità sostenibili, decreto 30 luglio 2021 disposizioni in materia di qualificazione iniziale e formazione periodica per conducenti professionali ai sensi della direttiva 2003/59/ce, come modificata dalla direttiva (ue) 2018/645. (21a05381) (gu serie generale n.221 del 15-09-2021), 2021.
- [2] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.
- [3] Sidney Dekker. *The Field Guide to Human Error Investigations*. 2001.
- [4] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [5] ISTAT. Rapporto BES 2022, “Il Benessere Equo e Sostenibile in Italia”, 2023.
- [6] Inge Koopmans, Renske J. Doll, Hans van der Wall, Marieke de Kam, Geert J. Groeneveld, et al. Fit for purpose of on-the-road driving and simulated driving: A randomised crossover study using the effect of sleep deprivation. *PLOS ONE*, 18(2):e0278300, 2023.
- [7] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
- [8] Orit Shechtman, Sherrilene Classen, Kezia Awadzi, and William Mann. Comparison of driving errors between on-the-road and simulated driving assessment: A validation study. *Traffic Injury Prevention*, 10(4):379–385, 2009.
- [9] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- [10] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

18. Contributed - Ordinal and preference data analysis

- 18.1 - OSILA (Order Statistics In Large Arrays): an original algorithm for an efficient attainment of the order statistics
- 18.2 - The Mallows model with respondents' covariates for the analysis of preference rankings
- 18.3 - Value-Based Predictors of Voting Intentions: An Empirical and Explainable approach
- 18.4 - A dynamic version of the Massey's rating system with an application in basketball

OSILA (Order Statistics In Large Arrays): an original algorithm for an efficient attainment of the order statistics

Andrea Cerasa

Abstract We describe the main features and empirical results of OSILA, an original approach for efficiently finding the k^{th} smallest element in an array. The algorithm's iterative framework and simple idea provide a fast solution for the selection problem, especially suitable for large datasets.

Key words: Selection problem, statistical computing, Matlab, R

1 Introduction

In statistics, the k^{th} order statistics of a sample is defined as its k^{th} -smallest value (2). The problem of computing the k^{th} smallest element of an array is called the *selection problem* and can be formalized as follows. Given an N -elements array of unsorted elements $X \equiv \{x_i, i = 1, \dots, N\}$ and a value $k \in [1; N]$, the *selection problem* consists in finding the k^{th} smallest element of X , so the element $\tilde{x} \in X$ for which there are at most $k - 1$ elements $x_i < \tilde{x}$ and at least k elements $x_i \geq \tilde{x}$. The naïve solution to this problem involves sorting the array X and then finding the element in the k^{th} position. The execution time of this operation is entirely subject to the complexity of the sorting algorithm, which is generally $O(N \log N)$.

An alternative algorithm for the solution of the selection problem is described in (1). The main idea behind this new procedure called OSILA is to overcome the redundancy of the sorting operation of the naïve approach through focusing on a (relatively) small subset of X which includes the target value $\tilde{x}$. The reduction of the number of elements to sort is expected to importantly reduce the waiting time for obtaining the result.

Andrea Cerasa

European Commission, Joint Research Centre, Via E. Fermi 2479, Ispra, 21027, Varese, Italy, e-mail: andrea.cerasa@ec.europa.eu

This article summarizes the main details and findings described in (1). In particular, Section 2 will provide the technical details of the proposed algorithm, whereas Section 3 summarizes the results obtained in several simulation exercises. Finally, Section 4 concludes.

2 Description of OSILA algorithm

To simplify the explanation of the algorithm, we introduce the *rank* function ρ . This function, given a vector X and an element $x \in X$, returns the rank of the element in the vector:

$$\rho_X(x) = \sum_{x_i \in X} I(x_i \leq x) \Rightarrow \rho_X(x_{(i)}) = i \text{ and } \rho_X^{-1}(i) = x_{(i)} \quad (1)$$

In this definition, we assume implicitly that, even if the vector X contains elements with the same value, their rank is always distinguishable. Since this assumption does not affect the accuracy and efficiency of the algorithm, we maintain it throughout the procedure. According to (1), the objective of the selection problem is to find the value $\tilde{x} = \rho_X^{-1}(k)$.

Instead of sorting the whole vector X , OSILA algorithm aims to find two values x_- and x_+ such that:

$$\rho_X(x_-) = k_- \leq k \text{ and } \rho_X(x_+) = k_+ \geq k \Rightarrow x_- \leq \tilde{x} \leq x_+ \quad (2)$$

If we define the set of $m = k_+ - k_- + 1$ elements of X that are between x_- and x_+ as $Z = \{x_i \in X : x_- \leq x_i \leq x_+\}$, we have $\tilde{x} = \rho_Z^{-1}(k - k_- + 1)$. Therefore, our focus can be narrowed down to these m elements of X that are between x_- and x_+ , sort them and find our target value $\tilde{x}$ in this reduced set of data.

Given as input: (i) a vector X of N values; (ii) an integer $k \in [1, N]$; (iii) a confidence level α ; (iv) the size n of the sample S ; the procedure for finding $\tilde{x}$ (i.e. the k^{th} order statistics) is based on the following two steps.

STEP 1: The objective of the first step is to find x_- and x_+ . We start by randomly extracting (without repetition) a sample S of $n \ll N$ observations from X . Then, we select the element in S which is expected to be the closest to the target value $\tilde{x}$, that is:

$$x_0 = \rho_S^{-1}(j_0) \text{ with } j_0 = \begin{cases} 1 & \text{if } \lceil \frac{n+1}{N+1} k \rceil = 0 \\ \lceil \frac{n+1}{N+1} k \rceil & \text{if } 1 \leq \lceil \frac{n+1}{N+1} k \rceil \leq n \\ n & \text{if } \lceil \frac{n+1}{N+1} k \rceil > n \end{cases} \quad (3)$$

where $\lceil b \rceil$ defines the arithmetic operation that rounds b to the nearest integer.

Defining $k_0 = \rho_X(x_0)$ (i.e. the rank of x_0 inside the original array X), it is possible to distinguish three cases:

- (a) $k_0 = k$: very lucky (but very unlikely at the same time) event. In this case we can stop searching, since $\tilde{x} = x_0$;
- (b) $k_0 > k$: in this case, we fix $x_+ = x_0$, $k_+ = k_0$ and we look for a lower bound x_- of $\tilde{x}$ in the subset $S_- \equiv \{x_s \in S : \rho_S(x_s) < j_0\}$. Considering the confidence level α , a suitable candidate is the biggest element of S_- that satisfies:

$$E[\rho_X(x_s) | x_s \in S_-] + z_\alpha \times \sigma[\rho_X(x_s) | x_s \in S_-] \leq k \quad x_s \in S \quad (4)$$

with z_α representing the $1 - \alpha$ percentile of the standard normal. After some mathematical derivations, the element in S_- that satisfies (4) is:

$$x_\alpha = \begin{cases} -\infty & \text{if } j_\alpha \leq 0 \\ \rho_{S_-}^{-1}(j_\alpha) & \text{if } j_\alpha > 0 \end{cases} \quad j_\alpha = \left\lfloor \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right\rfloor \quad (5)$$

where $\lfloor c \rfloor$ defines the greatest integer smaller than c , and

$$\begin{aligned} a &= -A^2 - B^2 & A &= k_0/j_0 \\ b &= B^2 j_0 + 2Ak & B &= z_\alpha \left[\frac{k_0(k_0 - j_0)}{j_0^2(j_0 + 1)} \right]^{1/2} \\ c &= -k^2 \end{aligned}$$

Defining $k_\alpha = \rho_X(x_\alpha)$ as the rank of x_α in X , there are two possibilities:

- $k_\alpha \leq k$: in this case, we fix $x_- = x_\alpha$, $k_- = k_\alpha$, and then proceed with the second step;
 - $k_\alpha > k$: in this case, the new values obtained for x_α and k_α provide a better upper bound. Then, we fix $x_+ = x_\alpha$, $k_+ = k_\alpha$ and we repeat all the operation in (b) unless the value of j_α obtained through expression (5) is 1. In this case, we fix $x_- = -\infty$, $k_- = 1$, and we proceed with the second step.
- (c) $k_0 < k$: in this case, we fix $x_- = x_s$, $k_- = k_0$. The algorithm proceeds symmetrically to case (b). We look for an upper bound x_+ of $\tilde{x}$ in the subset $S_+ \equiv \{x_s \in S : \rho_S(x_s) > j_0\}$. Considering the confidence level α , a suitable candidate is the smallest element of S_+ that satisfies:

$$E[\rho_X(x_s) | x_s \in S_+] - z_\alpha \times \sigma[\rho_X(x_s) | x_s \in S_+] \geq k \quad x_s \in S \quad (6)$$

where again z_α represents the $1 - \alpha$ percentile of the standard normal. After some mathematical derivations, the element in S_+ that satisfies (6) is:

$$x_\alpha = \begin{cases} \infty & \text{if } j_\alpha > n \\ \rho_S^{-1}(j_\alpha) & \text{if } j_\alpha \leq n \end{cases} \quad j_\alpha = \left\lceil \frac{-e + \sqrt{e^2 - 4df}}{2d} \right\rceil \quad (7)$$

where $\lceil \cdot \rceil$ defines the ceiling function as in Section 1, and

$$\begin{aligned} d &= C^2 + D^2 & C &= (N - k_0 + 1)/(n - j_0 + 1) \\ e &= 2C(k_0 - k) - nD^2 + j_0D^2 - D^2 & D &= z_\alpha \left[\frac{(N - k_0 + 1)(N - k_0 - n + j_0)}{(n - j_0 + 1)^2(n - j_0 + 2)} \right]^{1/2} \\ f &= (k_0 - k)^2 \end{aligned}$$

Defining $k_\alpha = \rho_X(x_\alpha)$ as the rank of x_α in X , there are again two possibilities:

- $k_\alpha \geq k$: in this case, we fix $x_+ = x_\alpha$, $k_+ = k_\alpha$, and then proceed with the second step;
- $k_\alpha < k$: in this case, the new values for x_α and k_α provide a better lower bound. Then, we fix $x_- = x_\alpha$, $k_- = r$ and we repeat all the operation in (c) unless the value of j_α obtained through expression (7) is n . In this case, we fix $x_+ = +\infty$, $k_+ = N$, and we proceed with the second step.

STEP 2: Once obtained x_- and x_+ from the previous step, the target value $\tilde{x}$ is calculated as follows.

- (a) Build the sample Z by filtering the values of X between x_- and x_+ , that is $Z \equiv \{x_i \in X : x_- \leq x_i \leq x_+\}$.
- (b) Sort the $m = k_+ - k_- + 1$ elements of Z .
- (c) $\tilde{x} = \rho_Z^{-1}(k - k_- + 1)$.

Comments: in the description of the algorithm, n was treated as an input. (1) provides a mathematical expression for calculating an optimal value of n for N , k and α . This value $\tilde{n}$ approximates the values of n which is expected to minimize the execution time of OSILA. (1) also shows how α provides an upper value of the probability to repeat the calculations in (b) or (c) of STEP 1 more than once. Finally, it is important to emphasise that *the random nature of OSILA does not affect its final result (which always corresponds to the target value $\tilde{n}$), but rather its execution time.*

3 Execution times

The execution times of OSILA are compared to the ones of the naïve approach in a simulation exercise that considers different values for N and k , while maintaining $\alpha = 0.01$. Specifically, the comparison focuses on:

- OSILA($\tilde{n}$): OSILA algorithm with $n = \tilde{n}$, according to the procedure described in (1);
- naïve approach: order X and take its k^{th} element.

The main advantage of OSILA is the reduction of the total number of elements to sort. This is expected to reduce the waiting time for obtaining the result, independently of the sorting algorithm adopted. For this reason, the simulation exercise involved different software environments and sorting algorithms:

- MATLAB: where its `sort` function which is a built-in code based on a stable version of the Quicksort algorithm originally proposed by (3).
- R: where the `sort` function offers the possibility to choose the sorting algorithm between:
 1. `shell`: based on the Shellsort variant proposed by (5);
 2. `quick`: uses (6) implementation of Hoare's Quicksort method;

3. `radix`: based on a variant of the radix sorting algorithm implemented by Matt Dowle and Arun Srinivasan in the `data.table` package (see 4, for a detailed description of the radix algorithm).

Table 1 shows the mean, the median and the 95th percentile of the execution times across 1,000 simulations for random arrays of respectively 10^6 and 10^7 elements extracted from the standard normal. What becomes evident from the tables is that, regardless of the value of n and the sorting algorithm employed, the execution times of OSILA are significantly smaller than those associated with the naive approach. In the case of MATLAB, this advantage reaches an order of magnitude, while for the three algorithms in R, it is less prominent but still noteworthy. Additionally, it is interesting to observe that, although the three naive versions of R routines exhibit significantly different execution times, their OSILA counterparts are remarkably similar. One possible explanation for this phenomenon could be that OSILA, by concentrating on a small sample of the original array, operates within the linear portion of the convex execution functions. This renders the choice of the underlying sorting method almost inconsequential.

The panels in Figure 1 illustrate the comparison of run times between the naive and OSILA($\tilde{n}$) procedures for smaller values of N with $k = \lceil 0.50N \rceil$. These panels clearly show that, irrespective of the sorting algorithm and software environment used, the superiority of OSILA over the naive procedure becomes evident when N is in the range of approximately 10^3 to 10^4 . Hence, while OSILA is designed to handle very large arrays, it also proves to be efficient for relatively small arrays.

Fig. 1: Run time patterns obtained with OSILA($\tilde{n}$) (grey - light) and with the Naïve procedure (blue - dark) for different values of N and for $k = \lceil 0.50N \rceil$ ($\alpha = 0.01$)

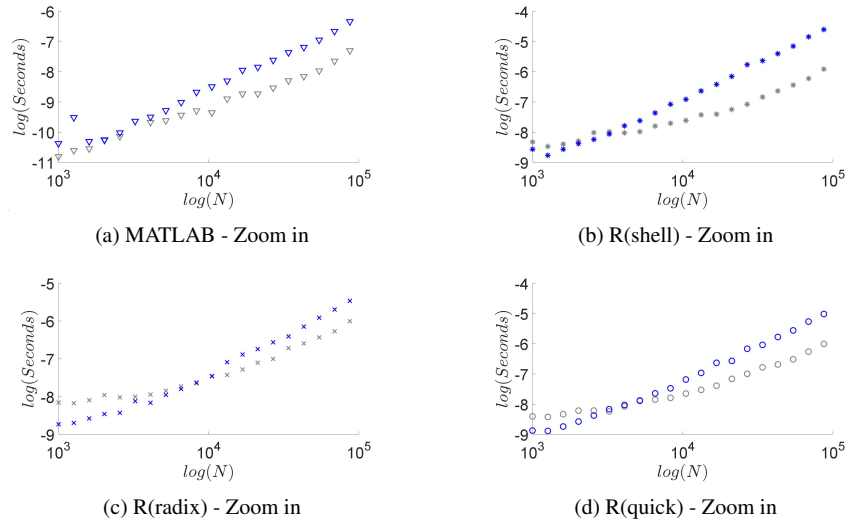


Table 1: Mean ($\bar{\mu}$), median (P_{50}) and 95th percentile (P_{95}) of the execution times for the selection algorithms across 1,000 repetitions with $N = 10^6$ and $\alpha = 0.01$.

			$k = \lceil 0.01N \rceil$			$k = \lceil 0.25N \rceil$			$k = \lceil 0.50N \rceil$		
			$\bar{\mu}$	P_{50}	P_{95}	$\bar{\mu}$	P_{50}	P_{95}	$\bar{\mu}$	P_{50}	P_{95}
$N = 10^6$	MATLAB	OSILA($\tilde{n}$)	0.034	0.025	0.152	0.035	0.026	0.114	0.036	0.026	0.141
		Naïve	0.286	0.273	0.404	0.283	0.272	0.391	0.283	0.273	0.381
	R(shell)	OSILA($\tilde{n}$)	0.211	0.190	0.365	0.243	0.220	0.406	0.260	0.235	0.415
		Naïve	1.375	1.343	1.784	1.385	1.345	1.811	1.378	1.340	1.823
	R(radix)	OSILA($\tilde{n}$)	0.207	0.205	0.214	0.235	0.233	0.244	0.251	0.250	0.261
		Naïve	0.579	0.576	0.597	0.580	0.577	0.598	0.579	0.576	0.595
	R(quick)	OSILA($\tilde{n}$)	0.207	0.205	0.214	0.235	0.234	0.246	0.251	0.250	0.263
		Naïve	0.954	0.950	0.984	0.953	0.950	0.981	0.953	0.950	0.982
$N = 10^7$	MATLAB	OSILA($\tilde{n}$)	0.032	0.031	0.037	0.036	0.035	0.044	0.037	0.037	0.043
		Naïve	0.411	0.408	0.471	0.453	0.447	0.527	0.452	0.449	0.516
	R(shell)	OSILA($\tilde{n}$)	0.211	0.210	0.230	0.239	0.238	0.260	0.254	0.253	0.264
		Naïve	1.729	1.725	1.765	1.727	1.725	1.762	1.726	1.723	1.760
	R(radix)	OSILA($\tilde{n}$)	0.205	0.204	0.211	0.221	0.219	0.235	0.235	0.233	0.252
		Naïve	0.912	0.910	0.927	0.902	0.901	0.919	0.902	0.900	0.921
	R(quick)	OSILA($\tilde{n}$)	0.209	0.208	0.217	0.236	0.236	0.243	0.250	0.250	0.258
		Naïve	1.106	1.104	1.126	1.103	1.101	1.124	1.101	1.100	1.123

In each simulation, the selection algorithms are applied on the same array of 10^6 or 10^7 elements randomly extracted from a $N(0, 1)$. The execution times for $N = 10^6$ have been multiplied by 10.

4 Conclusions

OSILA algorithm presents a novel and efficient solution for finding the k^{th} smallest element in an array. Unlike the naive approach, which involves sorting the entire array and has a time complexity of $O(N \log N)$, OSILA focuses on a smaller subset of the array. The algorithm's key idea is to identify two values, x_- and x_+ , within this subset, such that the k^{th} smallest element lies between them. By reducing the number of elements to sort, OSILA achieves a significant reduction in execution time.

Empirical results demonstrate the superior performance of OSILA compared to the naive approach, especially for large datasets. The algorithm's efficiency is highlighted through comparisons in various software environments and sorting algorithms. Even for relatively small arrays, OSILA consistently outperforms the naive method, showcasing its versatility and effectiveness across different scenarios.

Overall, OSILA stands out as a promising algorithm for efficiently obtaining order statistics in large arrays, offering practical advantages in terms of execution time and applicability to diverse datasets.

References

- [1] Cerasa, A.: Order statistics in large arrays (osila): a simple randomised algorithm for a fast and efficient attainment of the order statistics in very large arrays. *Computational Statistics* pp. 1–26 (2023)
- [2] David, H.A., Nagaraja, H.N.: *Order statistics*. John Wiley & Sons, New Jersey (2004)
- [3] Hoare, C.A.R.: Algorithm 64: quicksort. *Communications of the ACM* **4**(7), 321 (1961)
- [4] Knuth, D.E.: *The art of computer programming; Volume 3: Sorting and Searching*. Addison-Wesley Professional, Boston (1998)
- [5] Sedgewick, R.: A new upper bound for Shellsort. *Journal of Algorithms* **7**(2), 159–173 (1986)
- [6] Singleton, R.C.: Algorithm 347: An efficient algorithm for sorting with minimal storage [m1]. *Communications of the ACM* **12**(3), 185–186 (1969)

The Mallows model with respondents' covariates for the analysis of preference rankings

Marta Crispino, Lucia Modugno and Cristina Mollica

Abstract We propose an extension of the Mallows model with Spearman distance, able to incorporate respondents' covariates into the analysis of preference rankings. This is achieved by using the framework of mixtures of experts, that is, mixture models in which the parameters are functions of covariates. In particular, we postulate that the probability of each unit to belong to a given cluster is linked to its concomitant covariate values via a generalized linear model. Maximum likelihood estimation is achieved by means of a hybrid iterative procedure combining the Expectation-Maximization and Minorization-Maximization algorithms.

Key words: Mallows model, mixture of experts, rankings, clustering, covariates

1 Introduction

The Mallows model (Mallows, 1957), a popular probabilistic model for analyzing ranking data, assumes that the probability distribution of a ranking decays as the distance between the ranking and the modal (*consensus*) ranking increases. In this paper, we focus on the Mallows model with Spearman distance (MMS), that went through a recent revival thanks to Crispino et al. (2023), who noted some model simplifications enabling its application in scenarios where its use was previously impractical. Crispino et al. (2023) extended the MMS into the finite mixture framework to make inference on a sample of heterogeneous rankings by means of an

Marta Crispino and Lucia Modugno

Bank of Italy, Directorate General for Economics, Statistics and Research. Via Nazionale, 187 - 00184 Rome, e-mail: marta.crispino@bancaditalia.it, lucia.modugno@bancaditalia.it

Cristina Mollica

Department of Statistical Sciences Sapienza University of Rome, e-mail: cristina.mollica@uniroma1.it

Expectation-Maximization (EM) algorithm which allows to accurately investigate the group structure of such data.

When the interest lies in exploring possible relations between the presence of different preference patterns and the available covariates, it is common practice to first implement a clustering method, like a simple mixture model and second to investigate the association between the cluster membership and the available covariates. However, this relationship and its associated uncertainty is not incorporated in the statistical model used to fit the data (Gormley and Frühwirth-Schnatter, 2019).

In this paper, we explore an extension of the finite mixture model to jointly analyze the effect of covariates on the group structure. We do so by enriching the mixture of MMS of Crispino et al. (2023) with a mixture of experts (MoE) model (Jacobs et al., 1991; Jordan and Jacobs, 1994). In particular, we propose to incorporate covariates into the mixing weights of the mixture model, thus modelling the idea that the individual membership to specific clusters depends not only on the preferences (rankings) expressed, but also on the unique characteristics of the units themselves. Evidence on the relevance of this approach comes from works dealing with other parametric ranking distributions (e.g. the Plackett-Luce or the Benter model), which enjoyed extensions to incorporate unit-related covariates into the analysis (for instance Gormley and Murphy, 2008a,b).

2 A mixture of experts model extension of the MMS

Let $\mathbf{r} = (r_1, \dots, r_n)$ be a ranking of n items, where $r_i \in \{1, \dots, n\}$ is the rank of item i . A ranking is a permutation of the first n integers and, thus, a point in the permutation space $\mathcal{P}_n$. Under the MMS, the probability of observing a ranking $\mathbf{r}$ is

$$P(\mathbf{r}|\boldsymbol{\rho}, \theta) = \frac{\exp\{-\theta d_S(\mathbf{r}, \boldsymbol{\rho})\}}{Z(\theta)} = \frac{\exp\{-2\theta (c_n - \boldsymbol{\rho}'\mathbf{r})\}}{Z(\theta)},$$

where $\boldsymbol{\rho} \in \mathcal{P}_n$ is the modal ranking, a.k.a. *consensus ranking*, $\theta \in \mathbb{R}_0^+$ is the concentration parameter, and $d_S(\cdot, \cdot)$ is the Spearman distance. Moreover, $c_n = n(n+1)(2n+1)/6$, and $Z(\theta) = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-2\theta (c_n - \mathbf{e}'\mathbf{r})}$ is the normalizing constant or *partition function*, with $\mathbf{e} = (1, 2, \dots, n)$ denoting the identity permutation.

In the mixture model extension of the MMS, it is assumed that the population can be modeled as a finite collection of MMS sub-populations. Suppose that a population consists of G sub-populations and that the probability of belonging to sub-population g is ω_g (a.k.a. mixing weight). The probability of observing $\mathbf{r}$ is therefore given by a convex combination of MMS probabilities

$$P(\mathbf{r}|\boldsymbol{\rho}, \boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{g=1}^G \omega_g P(\mathbf{r}|\boldsymbol{\rho}_g, \theta_g) = \sum_{g=1}^G \omega_g \frac{\exp\{-2\theta_g (c_n - \boldsymbol{\rho}_g'\mathbf{r})\}}{Z(\theta_g)}, \quad (1)$$

where $(\boldsymbol{\rho}_g, \theta_g, \boldsymbol{\omega}_g)$ are the parameters of the MMS characterizing the g -th component. The proposed MoE extension consists in letting the mixing weights depend on unit-related covariates via a generalized linear model (GLM).

In particular, let $\underline{\mathbf{r}} = (\mathbf{r}_1, \dots, \mathbf{r}_i, \dots, \mathbf{r}_N)$ be a random sample of size N drawn from a mixture of MMS and assume to have L covariates at disposal, collected in the $N \times (L+1)$ design matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N]^T$, where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iL})^T$ is the individual covariate profile for sample unit i . Under the MoE, the conditional probability of observing a generic ranking $\mathbf{r}_i$, given its associated explanatory variables $\mathbf{x}_i$, is given by (1) with subject-specific mixture weights depending on covariates, that is, $\omega_g(\mathbf{x}_i) := \omega_{ig}$ and $\sum_{g=1}^G \omega_{ig} = 1$ for $i = 1, \dots, N$. More specifically, the individual group membership probabilities of the i -th respondent, $\boldsymbol{\omega}_i = (\omega_{i1}, \dots, \omega_{iG})$ are modelled through a multinomial logistic regression function of the covariates as follows

$$\log \frac{\omega_{ig}}{\omega_{i1}} = \mathbf{x}_i^T \boldsymbol{\beta}_g \quad \text{equivalently} \quad \omega_{ig} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_g)}{\sum_{g'=1}^G \exp(\mathbf{x}_i^T \boldsymbol{\beta}_{g'})}, \quad \text{for } g = 2, \dots, G. \quad (2)$$

In order to fit this model, we adapt the EM algorithm of Crispino et al. (2023) by introducing an additional step for the estimation of the regression coefficients $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G)$ of the GLMs. This latter step is an adaptation of the Minorization-Maximization algorithm (Lange et al., 2000) used in the MoE of Benter models by Gormley and Murphy (2008b).

3 Illustrative applications

Simulation study

We evaluated the performance of the mixture of experts of Mallows model with Spearman distance (MoEMMS henceforth) on simulated data under different data generating processes (DGPs).

First we tested the ability of our fitting procedure: i) to select the right model used in the DGP, that is to correctly identify both the combination of covariates of the GLM model and the number of clusters; ii) to estimate the coefficients of the GLM model; iii) to correctly classify observations into the clusters found. As a matter of fact, the procedure correctly selects the model with the true number of mixture components and the right covariates in all trials. As expected, as the true concentration parameter decreases (that is, the level of difficulty of the simulation setting increases), the performance of the MoEMMS in terms of classifying the observations tends to decrease and the estimation of the coefficients becomes more variable and farther from the true value of the parameter. Moreover, the addition of covariates in the GLM model helps to better classify the observations in the groups, as desired.

Finally, we compared the performance of our model with the simple mixture of MMS (mi.xMMS henceforth). In particular, i) when the true DGP does not include

respondents' covariates, the MoEMMS and the simple mixMMS perform similarly; ii) when the true DGP includes respondents covariates the MoEMMS outperforms the simple mixMMS .

Real data example

We employ a dataset concerning preferences for sports which was obtained from a survey conducted in 2017 as part of the Big Data Analytics project in Sports of the University of Brescia, to investigate sport preferences (Simone and Iannario, 2018). The dataset is composed of $N = 647$ full rankings of $n = 8$ sports ordered by the respondents according to their personal preferences. The list of the sports is: Soccer, Swimming, Volleyball, Cycling, Basketball, Boxing, Tennis and Jogging. The data include many covariates information, including sex and age, which we consider in this illustrative example.

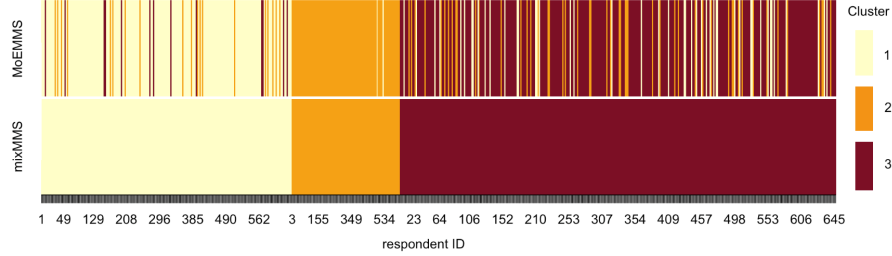
Table 1: Cross tabulations of the dummy covariates and the cluster memberships obtained by fitting the mixMMS .

(a) Gender				(b) Age			
Cluster	Gender		Total	Cluster	Age		Total
	M	F			≤ 25	> 25	
1	107 (52%)	97 (48%)	204 (100%)	1	123 (60%)	81 (40%)	204 (100%)
2	50 (57%)	38 (43%)	88 (100%)	2	41 (47%)	47 (53%)	88 (100%)
3	110 (31%)	245 (69%)	355 (100%)	3	185 (52%)	170 (48%)	355 (100%)
Total	267 (41%)	380 (59%)	647 (100%)	Total	349 (54%)	298 (46%)	647 (100%)

Table 1 (a) shows a cross tabulation of the gender dummy and the cluster memberships obtained by fitting the mixMMS model with $G = 3$ clusters on the data. It is clear that cluster 3 is mostly populated by females, while clusters 1 and 2 do not seem to be particularly associated with gender. Similarly, Table 1 (b) reports the cross tabulation of the age dummy and the cluster memberships, from which we note that cluster 1 is populated by younger respondents than clusters 2 and 3. However, these relationships and the associated uncertainties are not incorporated in the fitted model.

The MoEMMS assumes instead that the mixing weights depend on the covariates through the multinomial logit model of equation (2). Therefore, fitting such model on the data provides not only cluster memberships with the related Mallows parameters, that are similar to what found by fitting the mixMMS (see Fig. 1), but also the estimates of the coefficients of the multinomial logit, along with their standard errors, based on which we can check their statistical significance.

The odds ratios of the estimated coefficients, and their associated p-values (in parenthesis), are reported in Table 2. The parameters associated with cluster 2 are used as the reference parameters, that is, $\beta_2 = (\beta_{2,\text{Inter.}}, \beta_{2,\text{Female}}, \beta_{2,>25}) = (0, 0, 0)$.

Fig. 1: Cluster assignments estimated by the two models: `mixMMS` in the lower panel, `MoEMMS` in the upper panel.Table 2: Odds ratios of the estimated coefficients of the GLM, and their associated p-values (in parenthesis) estimated by the `MoEMMS`.

	Cluster 1	Cluster 3
Intercept	2.07 (0.003)	0.66 (0.108)
Female	0.88 (0.674)	4.52 (0.000)
>25	0.52 (0.024)	0.94 (0.819)

We notice that the odds of a female belonging to cluster 3 are around 4.5 times greater than the odds for cluster 2 and the coefficient is significant. The 3 consensus rankings estimated by the `MoEMMS`, reported in Table 3, show that cluster 3 is characterized by a ranking where Soccer is ranked last, while Jogging is the top-preferred sport. Cluster 1, instead, ranks Soccer in the top position. This appears to be an intuitive result, since, on average, Soccer is played more by males than females, at least in Italy.

By looking at the significant odds associated with the age covariate, we note that the odds of *older* people (with age > 25) to belong to cluster 1 are half the odds for being members of cluster 2, that is, it is more likely that younger people prefer Basket and Boxing (respectively in the second and third positions for cluster 1, see Table 3) to Swimming (ranked at the bottom) as opposed to *older* people who put Swimming in the second position and Basket and Boxing respectively in the fifth and the bottom positions (cluster 2).

We can conclude that, when there are informative covariates, the `MoEMMS` has the ability to simultaneously model and estimate the cluster structure and its association with the available covariates.

Table 3: Consensus rankings in the 3 clusters estimated by the `MoEMMS`.

	1	2	3	4	5	6	7	8
Cluster 1	Soccer	Basket	Boxing	Volley	Tennis	Jogging	Cycling	Swimming
Cluster 2	Soccer	Swimming	Volley	Cycling	Basket	Jogging	Tennis	Boxing
Cluster 3	Jogging	Swimming	Volley	Tennis	Boxing	Basket	Cycling	Soccer

4 Conclusion

This paper proposes a new model extension of the MMS that handles unit-related covariates. The implementation of the model exploits analytical simplifications of the MMS found in Crispino et al. (2023) and its extension to the finite mixture framework which allows to clustering assessors into groups according to their preferences. The inclusion of covariates in this setting is performed via a mixture of experts model (MoE) that is a mixture model whose mixing proportions depend on covariates through a generalized linear model. The underlying assumption is that even the respondents' characteristics can contribute to group formation, not only their preferences as in the simple mixture MMS.

Maximum likelihood estimates of model parameters are achieved via a hybrid of the EM and MM algorithms, the latter embedded for the covariate coefficients.

The model presented in this work has been implemented on full ranking data only. We aim at extending the entire procedure to partial ranking data, in order to be applied to many more real data situations.

References

- Crispino, M., Mollica, C., Astuti, V., and Tardella, L. (2023). Efficient and accurate inference for mixtures of mallows models with spearman distance. *Statistics and Computing*, 33(5).
- Gormley, I. C. and Frühwirth-Schnatter, S. (2019). Mixtures of experts models. In *Frühwirth-Schnatter, S. Celeux, G., Robert, CP (eds.). Handbook of Mixture Analysis*. CRC Press.
- Gormley, I. C. and Murphy, T. B. (2008a). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027.
- Gormley, I. C. and Murphy, T. B. (2008b). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2(4):1452–1477.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.
- Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214.
- Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20.
- Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika*, 44(1/2):114–130.
- Simone, R. and Iannario, M. (2018). Analysing sport data with clusters of opposite preferences. *Statistical Modelling*, 18(5-6):505–524.

Value-Based Predictors of Voting Intentions: An Empirical and Explainable approach

Abstract This paper enhances the predictive understanding of Italian voting intentions by integrating demographic and value-based variables. While demographic factors are established predictors, our emphasis lies in examining the impact of specific values, such as attitudes towards globalization and drug legalization. The analysis is based on data from the annual SWG and Racheal Monitoring Survey, which includes responses from 1,500 Italian adults over the years 2017-2019. Beyond traditional demographic predictors, the study extends its focus to value-based variables. By combining these two dimensions, we seek to elucidate the factors influencing Italian voting intentions. This contributes to the broader discussion on political prediction modelling, employing Tree-ensemble methods and Explainable Artificial Intelligence.

Key words: Vote intention, XAI, Shap, Applied Machine Learning, Survey data

1 Introduction

This paper aims to leverage both demographic and value-based variables to enhance the predictive understanding of voting intentions. While the predictive value of demographic variables is well-established in the literature[12][10] our objective is to investigate whether certain value-based variables (such as attitudes towards globalization, legalization of soft drugs, Catholic teachings, etc.) can contribute to the prediction of voting intentions in the Italian context. The data under scrutiny were

Luca Pennella

Department of Economics, Business, Mathematics and Statistics, Università di Trieste, Piazzale Europa 1, 34127 Trieste , e-mail: luca.pennella@phd.units.it

Amin Gino Fabbrucci Barbagli

Department of Economics, Business, Mathematics and Statistics, Università di Trieste, Piazzale Europa 1, 34127 Trieste e-mail: amingino.fabbruccibarbagli@phd.units.it

collected as part of the annual SWG [16] and Rachael [17] Monitoring survey [11] - it is a coordinated system that analyses the country and its citizens, based on more than 60 thousand interviews per year and divided into annual and weekly surveys. The annual survey is based on a sample of 1,500 adults residing in Italy, stratified by region and urbanization level, and distributed according to age and gender quotas. A weekly survey integrates the longitudinal themes with current events and trends.

Rachael, established in 2020 through a collaboration between SWG, a leading Italian market and opinion research company, the University of Trieste, and SISSA, integrates traditional statistical research with big data and data science expertise.

This study not only explores the conventional demographic predictors of voting behaviour but extends the analysis to encompass a large spectrum of value-based variables. By combining these dimensions, we aim to analyze and give a comprehensive understanding of the factors influencing voting intentions in Italy, thereby contributing to the broader discourse on political prediction modelling [7, 8, 9] opening the floor to further developments and new applications such as the study of the possible role and influences of social network platforms [13] or the application of an algorithm for the sentiment analysis as a possible way to predict voting intentions [4]. The recent application applies Social Network tools to assess e.g. homophily and polarization [2]. An overview of the workflow can be obtained by observing Figure 1. The initial step involved data cleaning, target variable management, and encoding of various features. Subsequently, we conducted the training of two distinct classifiers, namely, Random Forest (RF) [5] and the Light Gradient Boosting Machine Classifier (LGBMC) [6]. The evaluation of classifier performance was performed using relevant metrics. Once the optimal classifier was identified, we employed the Recursive *Feature Elimination algorithm (RFE)* to select the least important features. Having identified the best classifier-features pair, we utilized eXplainable Artificial Intelligence (XAI) techniques to interpret and understand the most relevant variables for each class of the target variable. A multi-step approach not only facilitated model training and selection but also allowed for a deeper understanding of the underlying patterns and key features driving the predictive performance. The use of techniques such as RFE and XAI enhances the interpretability and transparency of the model, providing insights into the decision-making process and contributing to the overall trustworthiness of the predictive model.

2 Data

The examined data were gathered as part of the annual SWG and Rachael Monitoring survey. An over-quantitative survey is based on a sample of 1,500 adults residing in Italy is stratified by region and urbanization level and distributed according to age and gender. Conducted annually, our focus spans the biennium 2017-2019, employing online interviews administered through the CAWI method. Sampling units were randomly selected from the proprietary panel using predefined randomization criteria. The Data Preparation pipeline involves selecting only demographic and value-

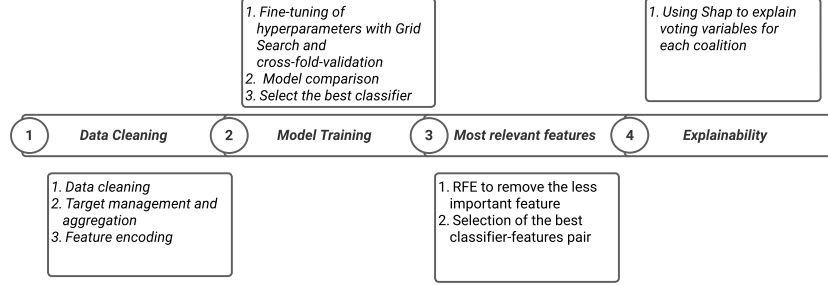


Fig. 1 Flow diagram of the analysis process

related variables; the initial dataset comprises 4,504 users and 116 variables. As the data originate from surveys, encoding is necessary for all predictive variables of interest. To be more adherent with the Italian political landscape at that time, we grouped various parties into the three main coalitions that emerged during the 2018 elections[18]: a center-left coalition that was guided by Matteo Renzi, a center-right led by Silvio Berlusconi and Matteo Renzi, and the Movimento 5 Stelle (M5S) [15] guided by Luigi di Maio that did not merge into any coalition.

3 Methods and Results

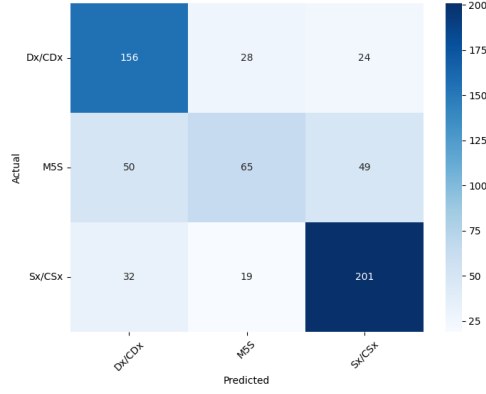
We evaluated the performance of two classifiers: RF and the LGBMC.

To optimize each classifier’s performance, a grid search procedure with cross-validation was employed to fine-tune the hyperparameters. We employ Accuracy as an optimization metric. The dataset is divided into an 80% training set and a 20% test set, with stratification based on the target variable. The cross-validation process encompasses dividing the training data into 5 folds. The model (i.e., classifier and corresponding hyperparameters) undergoes training and testing 5 times, using each fold as a validation set. We compute the average score for the metric of interest over these 5 tests. The classifier exhibiting the highest mean score was selected as the optimal classifier. Following the selection of optimal hyperparameter values for each dataset and classifier, we compute the standard metrics *Accuracy*, *F1*, *Precision*, and *Recall* on the test set. Table 1 presents a summary of the performance of the classifiers on the dataset

The previous experiment shows us that the LGBMC outperforms the RF. Thus, we consider the problem of determining if there exists a subset of the features that improve the quality of the classifier. To address such a problem, our idea is to consider the number of features as a further hyperparameter to be optimized. In practice, we perform a grid search procedure with cross-validation to optimize the Accuracy

Table 1 A comparison of the result of the grid search of two classifiers

Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.67	0.66	0.67	0.64
LGBM	0.69	0.69	0.69	0.68

**Fig. 2** Confusion matrices of the best classifier on the subset

and the number of features. To tune this last hyperparameter, we start considering all the features in the dataset, perform a grid search and 5-fold cross-validation procedure with the other hyperparameter of the LGBMC, and obtain the best-performing hyperparameter configuration. As for the previous experiments, we split the dataset into an 80% training set and a 20% test set. Then, we adopt the RFE algorithm to remove the less important feature. As a result, we obtain the highest mean Accuracy with a subset of the dataset, which consists of 35 features. The LGBMC trained with the RFE algorithm performs comparably to previous solutions but exploits only 35 features against the 116 used by datasets considering all the features, see Table 2

Table 2 Results of the LGBM classifier trained with the RFE algorithm.

	Accuracy	Precision	Recall	F1
Best classifier features pair	0.68	0.67	0.68	0.67

Additionally, we examine the confusion matrices(see Figure 2) of the model that summarize the number of correctly classified samples.

Notable performances are observed in terms of Recall and F1 for the *Dx/CDx* class, achieving 0.75 and 0.70, respectively, and for the *Sx/CSx* class with 0.80 and

0.76. However, less satisfactory results are noted for the *M5S* class, attaining a precision of 0.58.

To gain insights into the classifier behaviour on the dataset, we adopt the XAI library Shap¹ by Lundberg et al. [14]. This library allows the visual investigation of the impact of the most important features on the classification. Figure 3 shows the beeswarm plot of the feature importance based on Shap values for the test set. The plot provides a concise summary of how the top features in the dataset influence the model output. Dot colours, reflecting original feature values (blue for lower, red for higher), reveal each feature’s impact on prediction outcomes. We can determine each feature’s positive or negative effect on the prediction outcome by analyzing the plot.

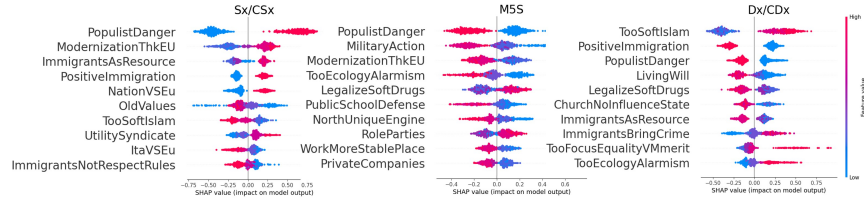


Fig. 3 Comparison of the top 10 features resulted from the application Shap values for each class.

The figure 3 illustrates the top 10 most relevant features for each class. Examining the graph for the *Dx/CDx* class, we observe that individuals in this class tend to assign high values to variables such as *TooSoftIslam*, *ImmigrantsBringCrime*, and *TooEcologyAlarmism* while assigning low values to variables like *PositiveImmigration*, *PopulistDanger*, *LegalizeSoftDrugs*, and *ImmigrantsAsResource*. Observing the graph for the *M5S* class, we observe that individuals in this class tend to assign high values to variables such as *LegalizeSoftDrugs* and *RoleParties* while assigning low values to other variables like *PopulistDanger*. Examining the graph for the *Sx/CSx* class, we observe that individuals in this class tend to assign high values to variables such as *PopulistDanger*, *ModernizationThkEU*, *PositiveImmigration* and *UtilitySyndicate* while assigning low values to variables to other variables like *OldValues*, *TooSoftIslam*, *ItaVSEu*, and *ImmigrantsNotRespectRules*.

4 Conclusion and future work

This work shows how value-based variables are useful in classifying voting intentions. Additionally, thanks to XAI we can detect the most relevant features for each party and category. Our primary objective is to enhance the model’s performance by expanding the dataset with new records from recent years and incorporating additional features related to social networks and leisure activities. Next, we aim to

¹ <https://shap.readthedocs.io/en/latest/index.html>

extend the XAI aspect of our work. Specifically, we plan to employ additional techniques and meticulously analyze the characteristics of misclassified instances. As a further direction, we will study how to make our classifier robust against smart Ponzi contracts that conceal their fraudulent nature inside apparently harmless contracts. As a further direction, we will explore diverse fields of application for Rachael and SWG survey data, encompassing areas such as proactive healthcare, market research, and business surveys. Data will be enriched by using different data sources trying to add more variables and understanding the best predictors. Furthermore, we will apply Social Networks Analysis tools, assessing e.g. homophily looking for common characteristics between voters.

References

1. Broy, M.: Software engineering — from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) *Software Pioneers*, pp. 10-13. Springer, Heidelberg (2002)
2. Esteve-Del-Valle, Marc. "Homophily and Polarization in Twitter Political Networks: A Cross-Country Analysis." *Media and Communication* [Online], 10.2 (2022): 81-92
3. Andreas Graefe, Accuracy of Vote Expectation Surveys in Forecasting Elections, *Public Opinion Quarterly*, Volume 78, Issue S1, 2014, Pages 204–232
4. Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. *Social Science Computer Review*, 33(1), 3-20.
5. Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32
6. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30 (2017).
7. De Stefano, Domenico, Francesco Pauli, and Nicola Torelli. "Preelectoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections." *The Annals of Applied Statistics* 16.1 (2022): 460-476.
8. Ford, Robert, et al. "From polls to votes to seats: forecasting the 2015 British general election." *Electoral Studies* 41 (2016): 244-249.
9. Ganser, Christian, and Patrick Riordan. "Vote expectations at the next level. Trying to predict vote shares in the 2013 German federal election by polling expectations." *Electoral Studies* 40 (2015): 115-126.
10. D'Souza, Gypsyamber et al. "Moderate predictive value of demographic and behavioral characteristics for a diagnosis of HPV16-positive and HPV16-negative head and neck cancer." *Oral oncology* vol. 46,2 (2010): 100-4.
11. Observatory of Italian public opinion provided by SWG, website: <https://www.swg.it/osservatorio>. Accessed: 18/01/2024
12. Islam, Towhidul, et al. "The usefulness of socio-demographic variables in predicting purchase decisions: Evidence from machine learning procedures." *Journal of Business Research* 151 (2022): 324-338.
13. Leiter, Debra, et al. "Social networks and citizen election forecasting: The more friends the better." *International Journal of Forecasting* 34.2 (2018): 235-248.
14. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
15. Chiamonte, A. et al., "Populist Success in a Hung Parliament: The 2018 General Election in Italy", *South European Society and Politics*, 23:4,(2018), 479-501
16. SWG, website: <https://www.swg.it/>. Accessed: 26/01/2024
17. Rachael, website: <https://rachael.swg.it/>. Accessed: 26/01/2024
18. <https://www.senato.it/leg/18/Elettorale/riepilogo.htm> Accessed: 28/02/2024

A dynamic version of the Massey's rating system with an application in basketball

Paolo Vidoni and Enrico Bozzo

Abstract This paper proposes a flexible, dynamic extension of the popular Massey's method for rating players and teams involved in sports competitions. The original Massey's approach is static since the computation of a team rating is based on the strength of the opponent teams evaluated at the current time. The proposed dynamic extension updates the team rating considering the strength of the opponent teams evaluated at the time when the matches were played. An application of the new rating procedure to the Euroleague Basketball 2018-2019 is presented.

Key words: basketball, match prediction, ranking, rating, sport analytics

1 A brief introduction to the Massey's method

The main idea underlying the method proposed by Massey [4] in his honors thesis, and shared by most models and rating methods, is that the match outcome between two teams depends on the team strengths only through their difference. Let us consider a tournament involving $I \geq 2$ teams and $N \geq 1$ games. If y_n , $n = 1, \dots, N$, is an outcome of match n played by teams i_n, j_n , then

$$y_n \approx g(r_{i_n} - r_{j_n}),$$

where r_{i_n} and r_{j_n} are the ratings of teams i_n and j_n . Although there are various possibilities, here and in what follows the outcome y_n is the final score difference

Paolo Vidoni
Department of Economics and Statistics - University of Udine, via Tomadini 30/A, I-33100 Udine, Italy, e-mail: paolo.vidoni@uniud.it

Enrico Bozzo
Department of Mathematics, Computer Science and Physics - University of Udine, via delle Scienze 206, I-33100 Udine, Italy, e-mail: enrico.bozzo@uniud.it

in the game from the perspective of team i_n and, usually, $g(x) = x$. In this case, a simple approach for computing the team ratings after N matches is to minimize the quadratic objective function

$$\sum_{n=1}^N \{y_n - (r_{i_n} - r_{j_n})\}^2. \quad (1)$$

It is immediate to see that (1) can be viewed as the sum of the squared residual for a linear regression model with a suitable design matrix, however the associated linear system is overdetermined due to linear dependence. A standard approach to address this problem is to impose a linear constraint on the team ratings; following [4], we assume the sum constraint so that $\sum_{i=1}^I r_i = 0$.

It is interesting to note that Massey's ratings are interdependent and that the rating of the team i after N matches can be written as the sum of the mean rating of the teams matched by i and the mean of the score differences from the perspective of i ; that is,

$$r_i = \frac{1}{m_i} \sum_{j=1}^I A_{ij} r_j + \frac{p_i}{m_i}, \quad (2)$$

where m_i is number of games played by i , A_{ij} is the number of matches between i and j and p_i is the sum of score differences from the perspective of i . In this case, the rating of the opponent teams is evaluated at the end of the sequence of matches.

An effective description of Massey's method, with the introduction of an appropriate statistical model for score differences, can be found in [2].

2 A temporized extension of the Massey's method

A first proposal for obtaining time-dependent ratings, i.e. to consider the strength of the opponent teams at the time of the matches, can be found in [1] and it is based on a suitable generalization of formula (2). In this case, the rating of team i , after the match played at time $t \in \{1, 2, \dots, T\}$, is defined as

$$r_i(t) = \frac{1}{m_i(t)} \sum_{k=1}^{m_i(t)} \{r_{j_k}(t_{k-1}) + y_i(t_k)\}, \quad (3)$$

where $m_i(t)$ is the number of games played by i until time t , $j_k, k = 1, \dots, m_i(t)$ are all the teams matched by i until time t , at the timestamps $t_1, \dots, t_{m_i(t)} = t$, and $y_i(t_k)$ is the score difference from the perspective of i in the match of time t_k . The initial ratings $r_i(t_0) = r_i(0)$, $i = 1, \dots, I$, are usually set as equal to zero; alternatively, they can be estimated or specified as the team ratings computed at the end of the previous season. With simple algebra, it is quite easy to obtain the iterative formula

$$r_i(t) = r_i(t^-) + \frac{1}{m_i(t)} [y_i(t) - \{r_i(t^-) - r_j(t^-)\}],$$

where j is the team faced by i at time t and $r_i(t^-)$ and $r_j(t^-)$ are the ratings of i and j before the match. Then, the rating of i at time t can be iteratively defined as an update of the rating before time t , which involves the difference between the realized score difference $y_i(t)$ and the expected score difference as described by the ratings difference $r_i(t^-) - r_j(t^-)$. Notice that this time-aware version of the Massey method has the advantage of considering the rating of the opponent teams at the time of the matches. However, it suffers from an *excess of stability*, since as the number of games $m_i(t)$ increases the rating becomes more and more stable, and a new result can only slightly modify the current reputation of the team.

A more general dynamic extension of Massey's method can be specified by considering a suitable weighted mean, instead of the simple mean, in equation (3). Therefore,

$$r_i(t) = \sum_{k=1}^{m_i(t)} \omega_{m_i(t),k} \{r_{j_k}(t_{k-1}) + y_i(t_k)\}, \quad t \in \{1, 2, \dots, T\}, \quad (4)$$

with $\omega_{m_i(t),k}$, $k = 1, \dots, m_i(t)$, a system of non-negative (normalized) weights. With an appropriate choice of the weights, it is possible to calibrate the influence that the most recent results have in defining the present team rating. Under this respect, an interesting option is to set $\omega_{1,1} = 1$ and, for $m > 1$, $\omega_{m,1} = (1 - \gamma)^{m-1}$, $\omega_{m,k} = (1 - \gamma)^{m-k} \gamma$, $k = 2, \dots, m-1$, $\omega_{m,m} = \gamma$, with $\gamma \in (0, 1)$. In this case, the average assigns weights that decrease exponentially as the matches come from further in the past and it can be interpreted as a simple exponential smoothing procedure, giving the following iterative relation

$$r_i(t) = r_i(t^-) + \gamma[y_i(t) - \{r_i(t^-) - r_j(t^-)\}],$$

with smoothing parameter γ .

To account for the uncertainty associated with the rating values, and then assess the significance of the differences in teams' abilities, and to specify a procedure to predict match results, we introduce a suitable statistical model for the match outcomes y_n , $n = 1, \dots, N$. This model can be viewed as the natural stochastic data-generating process that includes the new dynamic rating methodology just defined. More precisely, we consider the innovations state space model defined by the following measurement and state equations

$$\begin{aligned} y_i(t) &= r_i(t^-) - r_j(t^-) + \varepsilon_i(t) \\ r_i(t) &= r_i(t^-) + \gamma \varepsilon_i(t), \end{aligned}$$

where $\varepsilon_i(t) \sim N(\delta, \sigma^2)$, $t = 1, \dots, T$, $i = 1, \dots, I$, is a sequence of uncorrelated Gaussian error terms; δ describes the potential home field advantage of team i (see, e.g., [2]). An accurate description of this type of model can be found in [3]. To improve the flexibility of the model, we may consider a convenient transformation of the ratings difference, in particular $g\{r_i(t^-) - r_j(t^-); \alpha, \beta\} = \alpha + \beta\{r_i(t^-) - r_j(t^-)\}$. In this framework, the estimates of the unknown model parameters, with the asso-

ciated estimated standard errors, can be obtained quite easily using the maximum likelihood approach, since the likelihood function has a simple explicit expression.

Finally, we emphasize that this model can be generalized by considering a specification for the match outcomes y_n , $n = 1, \dots, N$, different from the score difference and by defining offensive and defensive rating instead of a single general rating for each team. A further, useful improvement requires the inclusion of suitable team- and game-specific covariate information, such as injuries, coaching changes, weather conditions, time of the day, participation in other competitions, etc.

3 An application to the Euroleague Basketball 2018-2019

As a simple application, we consider the match results of the 2018/2019 regular season of the Turkish Airlines Euroleague Championship. The regular season of the Euroleague Championship is a round-robin tournament where sixteen teams play each other home and away and the top eight teams advance to the playoffs. Then, the total number of matches is 240, since each team competes twice against other teams in the league.

The original Massey method (M) and the two time-aware versions, that one based on the simple mean (3) (Msimple) and that one based on the weighted mean (4) with simple exponential smoothing weights (Mses), are applied to the results of the tournament. The estimated ratings, with the associated ranking, are reported in Table 1. The estimates for the unknown model parameters, namely the home field advantage

Table 1 Teams ranked according to the official score at the end of the season, with the games won and the point difference, and the ratings (with the associated ranks in parentheses) obtained using the original Massey method (M) and the two temporalized methods (Msimple and Mses).

Team	Won	Difference	M	Msimple	Mses
Fenerbahce Istanbul	25	267	8.03 (1)	85.73 (1)	86.03 (1)
CSKA Moscow	24	193	6.09 (3)	65.92 (3)	63.34 (3)
Real Madrid	22	236	7.37 (2)	70.25 (2)	73.28 (2)
Anadolu Efes Istanbul	20	156	4.87 (4)	62.26 (4)	62.18 (4)
FC Barcelona	18	76	2.47 (5)	28.901 (9)	26.06 (9)
Panathinaikos Athens	16	37	1.03 (8)	47.25 (5)	47.61 (5)
Baskonia Vitoria-Gasteiz	15	71	2.09 (6)	35.72 (7)	32.93 (8)
Zalgiris Kaunas	15	37	1.12 (7)	40.11 (6)	43.03 (6)
Olympiacos Piraeus	15	25	0.91 (10)	-14.39 (11)	-15.45 (11)
Maccabi Tel Aviv	14	30	0.94 (9)	33.99 (8)	33.27 (7)
FC Bayern Munich	14	-56	-1.62 (12)	-32.35 (12)	-29.40 (12)
Olimpia Milan	14	1	0.09 (11)	-2.89 (10)	-2.53 (10)
Khimki Moscow Region	9	-116	-3.44 (13)	-61.51 (13)	-62.58 (13)
Gran Canaria	8	-299	-9.41 (14)	-117.60 (15)	-116.35 (15)
Buducnost Podgorica	6	-320	-10.00 (15)	-101.69 (14)	-101.13 (14)
Darussafaka Istanbul	5	-338	-10.56 (16)	-139.72 (16)	-140.28 (16)

δ , the variance of the error terms σ^2 , and (for Msimple and Mses) α and β , are obtained using the maximum likelihood approach. Note that the three methods produce the same ranking for the four best teams, and this corresponds to the order that we have considering the difference in points. The two time-dependent procedures give similar results and they are both affected by any variations in the quality of the teams' performances that may occur, in particular, at the end of the season. In this respect, the Massey method is more rigid and stable, which can be a disadvantage in some cases. The following Figure 1 describes the pattern of the ratings of Fenerbahce Istanbul, CSKA Moscow, FC Barcelona, Maccabi Tel Aviv and Baskonia Vitoria-Gasteiz from day 10 to the end of the regular season, by considering only the time-dependent approaches.

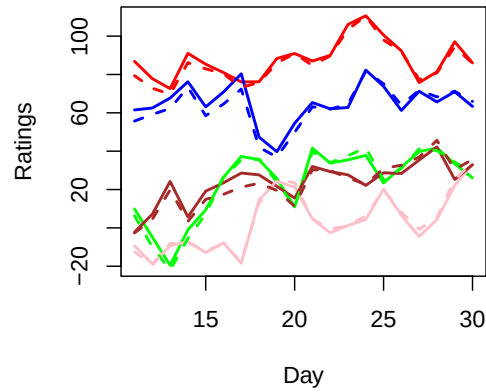


Fig. 1 Pattern of the ratings of Fenerbahce Istanbul (red), CSKA Moscow (blue), FC Barcelona (green), Maccabi Tel Aviv (pink), Baskonia Vitoria-Gasteiz (brown) obtained using Msimple (dashed line) and Mses (solid line).

Finally, we discuss the predictive ability of the statistical models based on the three rating methods. In particular, we evaluate both the accuracy (understood as the percentage of correct predictions for the win, both in general and for the home and the away team) and the mean squared prediction error, concerning the prediction of score differences. We consider a sequential procedure, so that at each day of the season, the model parameters are estimated using the results of the matches played up to that time. The findings are reported in Table 2 and show that, regarding the predictive accuracy, the behaviour of all three methods is quite good, with a slight prevalence for Massey's method. On the other hand, the temporalized method Mses has the best performance regarding the prediction of point differences. It must be said that the present analysis refers to a specific sports tournament, which does not present interesting dynamics, such as upsets, significant variations in performance, etc.

Table 2 Percentage of correct predictions for the win, in general (AccW) and for the home (AccWH) and the away team (AccWA), and mean squared prediction error (MSPE) for the point difference, using models based on the Massey (M) and on the two temporalized (Msimple and Mses) methods.

Method	AccW	AccWH	AccWA	MSPE
M	0.750	0.855	0.520	199.77
Msimple	0.744	0.864	0.480	189.45
Mses	0.744	0.864	0.480	189.05

4 Final remarks

In this paper, we introduced a new temporalized rating method. This approach adequately takes into account the fact that teams' capabilities change over time. In addition, the associated statistical model is flexible and easily extensible to improve its predictive ability. Sports organizations or analysts could effectively use this dynamic method to achieve a less rigid, more realistic, and up-to-date assessment of the current strength of teams engaged in sports competitions. In addition, the associated statistical model, even in its simplest version, has demonstrated good predictive accuracy and thus can be used in the context of sports event betting.

Acknowledgements This research was partially supported by the Departmental Strategic Plan of the University of Udine, Department of Economics and Statistics (2022-2025), and by the Research Project PRIN 2022, granted by European Union – Next Generation EU, "Statistical Models and AlgoriThms in sports (SMARTsports). Applications in professional and amateur contexts, with able-bodied and disabled athletes", project nr. 2022R74PLE, CUP: G53D23001870006.

References

1. Franceschet, M., Bozzo, E., Vidoni, P.: The Temporalized Massey's Method. *Journal of Quantitative Analysis in Sports* **13**, 37–48. (2017)
2. Glickman, M.E., Stern, H.S.: Estimating Team Strength in the NFL. In: Albert, J., Glickman, M.E., Swartz, T.B., Koning, R.H. (eds.) *Handbook of Statistical Methods and Analyses in Sports*, pp. 113–136. CRC Press, Boca Raton. (2017)
3. Hyndman, R., Koehler, A., Ord, K., Snyder, R.: *Forecasting with Exponential Smoothing*. Springer Berlin, Heidelberg. (2008)
4. Massey, K.: *Statistical models applied to the rating of sports teams*. Technical Report, Bluefield College. (1997)

19. Contributed - Advances in text mining

- 19.1 - Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?
- 19.2 - A Fuzzy Topic Modeling approach to legal corpora
- 19.3 - EmurStat: a digital tool for statistical analysis of emur flow
- 19.4 - Graph Neural Networks for clustering medical documents

Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?

Andrea Sciandra and Arjuna Tuzzi

Abstract: With reference to a large corpus of 76 Italian contemporary popular mystery novels by 16 different authors, this study aims to assess the performance of large language models in an authorship attribution test. The results obtained through both transformers and correspondence analysis vector representations are compared and contrast in machine learning classification tasks. Although in previous works transformers have been shown to perform better than other alternatives, in this case, correspondence analysis wins the challenge. Results support the hypothesis that specialized large corpora require tailor-made representations.

Riassunto: Attraverso l'analisi di un corpus di 76 gialli contemporanei italiani di 16 diversi autori, questo studio intende valutare le *performance* dei *large language models* in una prova di attribuzione d'autore. Vengono confrontati in diverse classificazioni basate su *machine learning* i risultati ottenuti adottando rappresentazioni vettoriali con i *transformers* e con l'analisi delle corrispondenze. Sebbene in lavori precedenti i *transformers* si siano dimostrati migliori di altre alternative, in questo caso è l'analisi delle corrispondenze a vincere la sfida. I risultati confermano l'ipotesi che grandi corpora specializzati richiedano soluzioni su misura.

Key words: authorship attribution, machine learning, correspondence analysis, transformers, popular Italian mystery novels

¹ Andrea Sciandra, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA); email: andrea.sciandra@unipd.it

Arjuna Tuzzi, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA); email: arjuna.tuzzi@unipd.it

Introduction

The emergence of large language models (LLM) and the availability of transformers (Wolf et al., 2020) for vector representation of words and texts have significantly changed the landscape of text mining. They have proven particularly innovative in numerous applications related to text classification tasks. However, the question to determine whether LLMs are the ultimate solution to all problems persists, particularly because the black-box nature of this technology significantly constrains the ability to achieve explainable models. This study aims to assess the performance of LLMs in an authorship attribution task, comparing and contrasting the results with those obtained through correspondence analysis (CA, Greenacre, 1984; Murtagh, 2005) as suggested by Lebart (1997).

The corpus includes 76 Italian contemporary popular mystery novels by 16 different authors. Each author contributes a minimum of three to a maximum of six novels. It is a very homogeneous corpus by text-genre, created with the support of the publishing house, which lively contributed to the choice of the most significant authors and works. On the basis of simple tokenization that takes into account only alphabetic sequences (all other signs were considered separators, and numbers were excluded) and simple normalization that only transformed the words into lowercase letters, the corpus has a total size of just under five million words (4,845,966 word tokens) and expresses a vocabulary of just over 100 thousand different words (109,643 word types). The length of the novels ranges from a minimum of 22,402 to a maximum of 123,457 words, thus the length of the works is quite variable, as is the contribution to the corpus of each author (Fig. 1). Given that the type-token ratio is equal to 2.2% (average frequency of word types 44.2) and the percentage of *hapax legomena* is 38.1% of the vocabulary, the corpus shows a desirable high degree of redundancy.

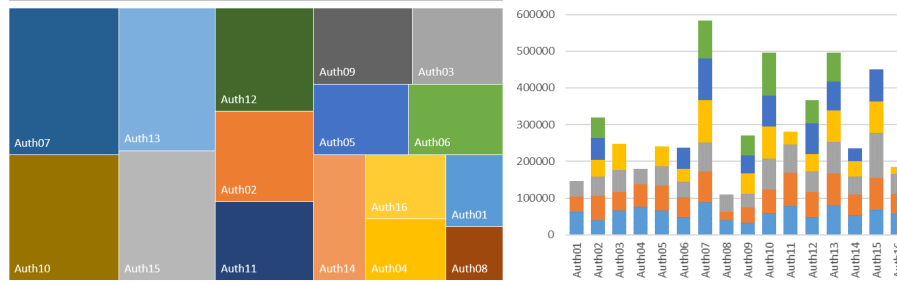


Figure 1: Dimensions in word tokens by author (left) and by author and novel (right).

As it is homogeneous in terms of text genre and includes novels of undisputed authorship, our corpus represents a good basis for testing the ability of a classification pipeline for an authorship attribution task. This study aims at assessing the performance of the same classification methods based on machine learning but with different ways of representing the features: we compete between vectorization based

on modern transformers and that provided by traditional correspondence analysis. All texts are partitioned into 9526 equal-size text chunks (500 words, since we used a LLM model that can process texts of the maximal length of 512 tokens), and performance is measured based on the ability to attribute each chunk of the test set to the right author.

Methods

We selected word types with number of occurrences ≥ 76 (one occurrence per novel on average) and started from the (5061 x 9526) Term-Document Matrix (TDM) with (relative) frequencies as weights. After calculating the chi-square distance and Singular Value Decomposition (SVD), each text (word) becomes a column (row) vector in the 5060-dimensional space of CA orthogonal axes. We can take the first dimensions of these vectors as representations of texts (words) to submit to ML classification models. We chose to select the first 76 axes, equal to the number of texts available. The explained variance for the first 76 axes is about 13%.

Next, we obtained a representation of texts using transformers, specifically by exploiting the BERT Base Italian Cased LLM pre-trained vectors (<https://huggingface.co/dbmdz/bert-base-italian-cased>). BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) provides contextualised embeddings that consider the context in which a word is used. Therefore, they dynamically produce word representations informed by the surrounding words. In this study, we concatenated the 12 BERT layers that represent the same token and used the mean to aggregate the embeddings from different tokens to represent a text. As the BERT-base mapping involves 768 vectors, to reduce dimensionality we applied a principal component analysis, extracting the first 76 components. These components account for approximately 82% of the cumulative proportion of variance of embeddings obtained via LLM.

Regarding ML-based classification methods, we chose two models: Random Forests (RF; Breiman, 2001), and Extreme Gradient Boosting (XGBoost; Chen et al., 2019) to provide two informative validations through bagging and boosting techniques. The training set was generated by randomly selecting 80% of the chunks of each author, while the remaining chunks were used for the test set. Each model involved a 5-fold cross-validation, repeated 30 times. The evaluation of the results included the analysis of accuracy, unweighted Kappa statistic, and a one-sided test to verify if the accuracy is greater than the "no information rate" (the largest class percentage in the data: 12.07%; Kuhn, 2008).

Discussion and conclusions

The results of the models for the test set are presented in Table 1.

The models using CA consistently outperformed those using LLM. The best model overall is XGBoost with features derived from CA, achieving an accuracy of approximately 99%. The Kappa statistic confirms that the XGBoost model with CA coordinates is the most accurate. All models are significant according to the one-sided test, as the accuracy is higher than the no information rate (NIR). These results indicate that CA features demonstrate the highest classification capacity in this particular corpus.

The heatmaps (confusion matrices) in Figure 2 compare the classification results of the XGBoost-CA and XGBoost-LLM models in terms of the probability of attributing chunks of the test set to the authors of the corpus. The red-coloured chunks represent correct author attributions with probabilities close to 100%, highlighting the superior performance of the CA model.

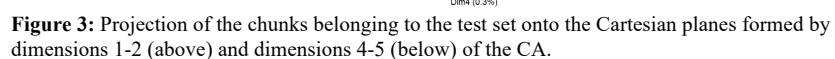
The ability of the CA to discriminate the authors of the novels is evident by projecting the position of the chunks belonging to the test set onto a Cartesian plane (Dim. 1 – Dim. 2, and Dim. 4 – Dim. 5 in this example). It can be observed that the first dimension isolates author number 2, while the fifth dimension isolates authors number 9 and 14 (Fig. 3).

Although in previous works (Jones et al., 2022; Ai et al., 2022; Rodella et al., forthcoming), transformers have been shown to perform better than other alternatives, in this case, CA wins the challenge. Our hypothesis is that the performance of transformers depends on the specific application context. In particular, we are convinced that transformers, born from the training of LLMs on general, multilingual, multifaced, multigenre databases, offer a very good representation for a text classification in general, but perform worse when the language is not English, when texts are large and complex, when the corpus represents a special language and a specific genre. In the latter cases, methods that arise from the analysis of the textual data of the corpus itself could be more powerful, as they exploit endogenous information from the training set. Our hypothesis requires better testing on different corpora and with different ML methods.

It is important to note that some LLMs do not explicitly state the material on which they were trained, which can pose a challenge for this type of research. Given the vast array of open-source tools, resources, and code available online, we can observe that there is not a commensurate availability of discussion, documentation, and scientific publications. It is very challenging to choose among so many opportunities without any criteria to evaluate which ones are the best, for what purposes, and for what contexts of use.

Table 1: Comparison of author classification performance: LLM and CA (Diagnostics: Accuracy, unweighted Kappa statistic, and a one-sided test p -value to verify if the accuracy is greater than NIR).

<i>ML Model</i>	<i>Accuracy</i>	<i>Kappa</i>	<i>P-Value (Acc > NIR)</i>
LLM - RF	0.7608	0.7395	< 2.2e-16
LLM - XGBoost	0.8667	0.8557	< 2.2e-16
CA - RF	0.9831	0.9812	< 2.2e-16
CA - XGBoost	0.9863	0.9852	< 2.2e-16



References

1. Ai, B., Wang, Y., Tan, Y., & Tan, S. (2022). *Whodunit? Learning to Contrast for Authorship Attribution* (arXiv:2209.11887). arXiv. <http://arxiv.org/abs/2209.11887>
2. Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
3. Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2019). *xgboost: Extreme gradient boosting*. R package version 0.81.0.1, 1–4
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
5. Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
6. Jones, K., Nurse, J. R. C., & Li, S. (2022). Are You Robert or RoBERTa? Deceiving Online Authorship Attribution Models Using Neural Text Generators. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 429–440. <https://doi.org/10.1609/icwsm.v16i1.19304>
7. Kuhn, M. (2008), “Building predictive models in R using the caret package” *J. of Statistical Softw.*, (doi:10.18637/jss.v028.i05).
8. Lebart, L. (1997). Correspondence analysis, discrimination, and neural networks. In *Data Science, Classification, and Related Methods*. Hayashi C., Ohsumi N., Yajima K., Tanaka Y., Bock H.- H. and Baba Y. (eds), Springer, Berlin, 423-430.
9. Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. CRC Press.
10. Rodella, I., Sciandra, A., & Tuzzi, A. (forthcoming). Analysis of Marie Skłodowska-Curie Actions (MSCA) evaluations and models for predicting the success of proposals, *JADT - 17th International Conference on Statistical Analysis of Textual Data*.
11. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

A Fuzzy Topic Modeling Approach to legal corpora

Antonio Calcagni, Arjuna Tuzzi

Abstract This study investigates the application of Fuzzy Latent Semantic Analysis (fLSA) in analyzing expenditure chapters within legal texts, using Italy’s budget law 178/2020 as a case study. Faced with challenges in legal studies, such as specialized language and heterogeneity, fLSA combines Latent Semantic Analysis (LSA) with dimensionality reduction and soft clustering. Results show comparable performance with the widely used Latent Dirichlet Allocation (LDA) in identifying coherent expenditure chapters, with fLSA showing a tendency to retrieve more distinctive and exclusive topics.

Key words: topic modeling, fuzzy modeling, lda, topic coherence

1 Introduction

Topic models are nowadays one of the most relevant techniques for uncovering latent structures in large textual datasets, spanning various domains such as social sciences [9], digital humanities [7], scientometrics [8], and legal studies [6]. Particularly challenging is their application in legal studies, where specialized language and heterogeneity pose difficulties [3, 4]. This complexity is amplified when dealing with short legal corpora, as in the case of budget laws, which serve as critical guides for a country’s finances. In this contribution, we adopt a fuzzy topic modeling approach (fLSA) to analyze the Italian budget law 178/2020. This technique combines the strengths of Latent Semantic Analysis (LSA) with dimensionality reduction and soft clustering to handle sparsity and high-dimensionality challenges in word tokens

Antonio Calcagni,
University of Padova, e-mail: antonio.calcagni@unipd.it
GNCS, National Institute of Advanced Mathematics (INdAM)

Arjuna Tuzzi,
University of Padova, e-mail: arjuna.tuzzi@unipd.it

and features. The focus is on retrieving -which is quite challenging in legal corpora- as well as assessing the capability of fLSA in detecting highly coherent expenditure chapters (i.e., high-quality topics), especially in comparison with widely used techniques such as Latent Dirichlet Allocation (LDA). All the materials like algorithms and datasets used throughout this contribution are available to download at https://github.com/antcalcagni/flsa_legalCorpora.

2 Fuzzy Latent Semantic Analysis

In the analysis of text data, a corpus contains n documents of N word-tokens. The i -th document is represented as a vector of N_i words $\mathbf{w}_i = \{w_1, \dots, w_{N_i}\}$ such that $\sum_{i=1}^n N_i = N$. To represent the joint information of documents and lexicon, the document-term matrix $\mathbf{X}_{n \times J}$ is usually constructed (J is the number of word-types), which constitutes the input for the topic analysis. Unlike other LDA-based methods [1], the fLSA algorithm computes the basic topic modeling matrices - namely, $\mathbf{P}_{D|T}$ (probability of documents per topic) and $\mathbf{P}_{W|T}$ (probability of words per topic) - by means of a deterministic algorithm that uses both SVD and fuzzy c-means in its iterative steps [5]. Given a (possibly locally-globally weighted) document-term matrix $\mathbf{X}_{n \times J}$, the following computations are then required for K topics:

$$\mathbf{P}_{D|T_{(n \times K)}} = (\mathbf{\Xi}_{(n \times K)} \circ \mathbf{p}_{D_{n \times 1}} \mathbf{1}_K^T) \text{diag}(\mathbf{1}_n^T (\mathbf{\Xi}_{(n \times K)} \circ \mathbf{p}_{D_{n \times 1}}) \mathbf{1}_K^T)^{-1} \quad (1)$$

$$\mathbf{P}_{W|T_{(J \times K)}} = (\text{diag}(\mathbf{1}_n^T \mathbf{X}_{n \times J})^{-1} \mathbf{X}_{n \times J}^T) \mathbf{P}_{D|T_{n \times K}} \quad (2)$$

where $\mathbf{p}_D = (\mathbf{1}_n^T \mathbf{X})(\mathbf{I} s)^{-1}$, with $s = (\mathbf{1}_n^T \mathbf{X} \mathbf{1}_K)$, is the vector of document probability, $\mathbf{\Xi}$ is the matrix of fuzzy membership degrees computed via fuzzy c-means algorithm on the matrix $\mathbf{U}_{n \times K}$ of the first K left singular vectors of $\mathbf{X}$, $\text{diag} : \mathbf{R}^M \rightarrow \mathbf{R}^{M \times M}$ is the usual linear operator, whereas $\circ$ denotes the Hadamard product. Put this way, the fLSA generalizes the standard LSA by using the fuzzy c-means (or any other fuzzy clustering technique such as fuzzy k-medoids) to find the unobserved topics through the set of documents. Although it misses a consistent stochastic framework to compute $\mathbf{P}_{D|T}$ and $\mathbf{P}_{W|T}$, the technique has shown optimal results in terms of classification, document clustering, and document redundancy, if compared to other commonly used methods for topic modeling [5].

3 Application: The case of Italian budget law 178/2020

Data and preprocessing

The corpus being analyzed consists of $n = 1162$ clauses referring to the Italian budget law 178/2020.¹ The entire corpus conveys $N = 52223$ word-tokens and $J = 6207$ word-types, which highlights a good level of redundancy (TTR = 11.80%, hapax = 31.20%). The minimum and maximum lengths are 3.00 and 648.00 word-tokens, respectively, with a Q3-Q1 range of 35 word-tokens. The raw corpus has been preprocessed to achieve a standardized format. Hence, all uppercase letters have been converted to lowercase, punctuation marks and symbols have been removed, and a first subset of multiword expressions (e.g., *partita iva*) have been identified and transformed (e.g., *partita.iva*) by matching with a predefined list. Additionally, to reduce unnecessary word redundancy, all words referring to other laws (e.g., *d.l. 30 giugno 1998*) have been considered a single word (e.g., *dl.30.6.1998*), while a second subset of multiword expressions have been identified and created by considering 5-to-2-grams with highest pointwise mutual information (PMI). With the same aim, commonly used modal verbs (e.g., *essere*, *avere*, *dovere*, *potere*) as well as words with less than nine occurrences have been filtered out. As a result, the final corpus has $N = 32878$ word-tokens and $J = 1392$ word-types, with TTR = 6.90%. Finally, the document-term-matrix has been computed and the common term-frequency/inverse-document-frequency (TF-IDF) schema has been adopted to locally/globally weight the ensuing frequency matrix.

Methods and procedures

The performance of fLSA method to retrieve highly coherent topics has been assessed against the LDA method in a repeated 10-fold cross validation schema (with 25 random repetitions).² Two commonly used coherence metrics, namely UMass-like (i.e., `mean_logratio`) and UCI-like (i.e., `mean_pmi`) metrics, as implemented in the R library `text2vec`, have been chosen for the between-method comparison. The number K of topics has been varied from 2 to 45. Additional coherence or topic-quality measures are available as Supplementary Materials in the repository indicated in Section 1.

¹ In this context, each document represents a single clause, as the budget law is structured based on the use of a single article.

² The validation process involves partitioning the dataset into training and testing sets during each cross-validation iteration. The model is trained on the training set to generate the word-topic probability matrix $\mathbf{P}_{W|T}$, which is then used to compute coherence on the test dataset. In training, the thirty most significant words for each topic are identified, and in testing, these words are used to calculate coherence on the remaining dataset.

Results

Figure 1 shows the coherence metrics averaged over the 10 folds for both methods. Although the criteria indicate different optimal numbers of topics for fLSA and LDA — specifically, $K^\dagger \in \{11, 12\}$ in the first case and $K^\dagger \in \{7, 14\}$ in the second — both highlight a slightly higher coherence for LDA compared to fLSA.³ Interestingly, the methods differ in the topic construction, with fLSA producing a less uncertain marginal topic distribution than LSA (Figure 2). Moreover, fLSA detects fewer exclusive topics than LDA, resulting in a higher level of distinctiveness. Based on Frex top words⁴, the topics produced by both methods span from labor policies to health policies related to COVID-19, covering social welfare as well. However, if we restrict the focus only to those topics referring to covid-19 pandemic, we notice that fLSA identifies only two ($k \in \{5, 11\}$), which have the largest probability to occur and are highly exclusive. By contrast, LDA identifies five topics of this type ($k \in \{3, 5, 7, 11, 14\}$), which are hard to distinguish from the others. Figure 3 shows the the topic distances in an MDS projection for both methods. Interestingly, the solution provided by fLSA suggests that topics can be aggregated into three clusters (blue dashed rectangles), while no clear hierarchical structure emerges for the LDA method. To see whether the fLSA-based analysis would benefit from a three topics solution, the algorithm has been run by setting $K = 3$. The new solution conveys topics with higher average coherence compared to the case where $K = 12$. These topics are almost uniformly distributed over the clauses ($p_{K=1} = 0.415$, $p_{K=2} = 0.361$, $p_{K=3} = 0.223$) but exhibit different levels of exclusivity ($ER_{K=1} = 0.238$, $ER_{K=2} = 0.578$, $ER_{K=3} = 0.182$). Figure 4 depicts the FREX-based top words for the three topics alongside the word-topic proportion (colored bars). Unlike the previous solution, the latter shows a coarse classification of expenditure chapters into three general macro-areas of economic activity.

4 Conclusions

In this contribution, we have explored the feasibility of applying topic modeling to specialized legal texts, particularly in the analysis of expenditure chapters within budget laws. Specifically, Fuzzy Latent Semantic Analysis (fLSA) has been employed on Italy’s budget law 178/2020. Our focus has been on assessing the effectiveness of fLSA in identifying coherent expenditure chapters compared to the commonly used Latent Dirichlet Allocation (LDA). The results indicate that fLSA has been capable of retrieving coherent and informative topics to the same extent as LDA. However, they differ in terms of marginal probability and exclusivity, with topics tending to cluster when projected onto a lower MDS-based space. Further investigations are needed to evaluate the extent to which fLSA differs from more

³ For the analysis, the upper bound of $K^\dagger$ has been selected for both methods.

⁴ FREX-based top words for each topic and both methods are available in the repository indicated in Section 1.

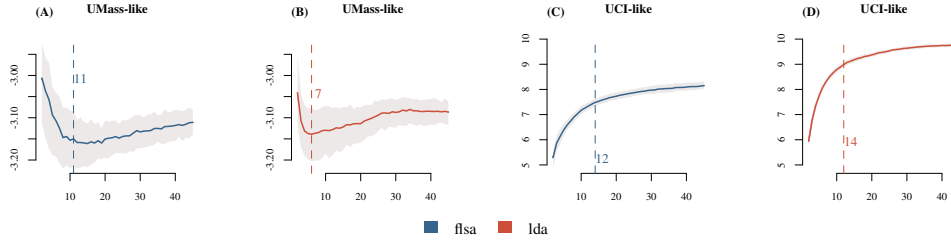


Fig. 1 Average Coherence Scores for fLSA (blue curve), and LDA (red curve) as a function of the number of topics. Note that dashed vertical lines indicate the elbow of the curves, gray areas are the IQR ranges (i.e., $q_{0.75} - q_{0.25}$ tolerance intervals), and numbers indicate the corresponding elbow points (i.e., the optimal number of topics). All metrics have been computed using the top thirty topic words.

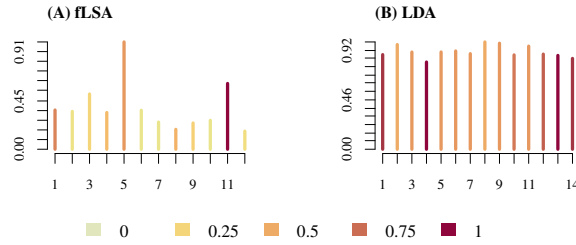


Fig. 2 Marginal topic distributions (bars) alongside the exclusive term ratio (gradient color). Note that frequencies and exclusivity index have been normalized for the sake of comparison.

specialized methods like the Correlated Topic Model (CTM). Similarly, additional studies would be needed to assess how fuzzy membership degrees $\mathcal{E}$ can be integrated to compute $\mathbf{P}_{D|T}$. While the current implementation has used a linear aggregator, other solutions can be explored, such as those based on generalized Bayes rules [2].

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of machine Learning research* **3**(Jan), 993–1022 (2003)
2. Coletti, G., Gervasi, O., Tasso, S., Vantaggi, B.: Generalized bayesian inference in a fuzzy context: From theory to a virtual reality application. *Computational Statistics & Data Analysis* **56**(4), 967–980 (2012)
3. Cortelazzo, M.: La lingua delle leggi italiane. In: M.E. Piemontese, et al. (eds.) *Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile*, pp. 110–122. Carocci (2023)
4. Garavelli, B.M.: *Le parole e la giustizia*. Einaudi (2001)
5. Karami, A., Gangopadhyay, A., Zhou, B., Kharrazi, H.: Fuzzy approach topic discovery in health and medical corpora. *International Journal of Fuzzy Systems* **20**, 1334–1345 (2018)

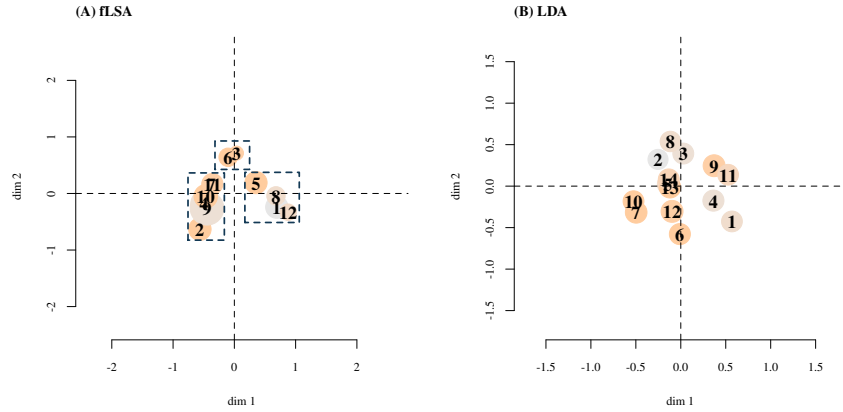


Fig. 3 Inter-topic distance via MDS-based plot for both fLSA and LDA methods. Note that each circle represents a topic, the circle radius represents the marginal probability of the topic (in log scale), the circle color represent the UCI-like coherence (the stronger the saturation of orange, the higher the coherence), whereas the blue rectangles indicate topic clusters extracted via Ward-based hierarchical clustering on the matrix of inter-topic distances.

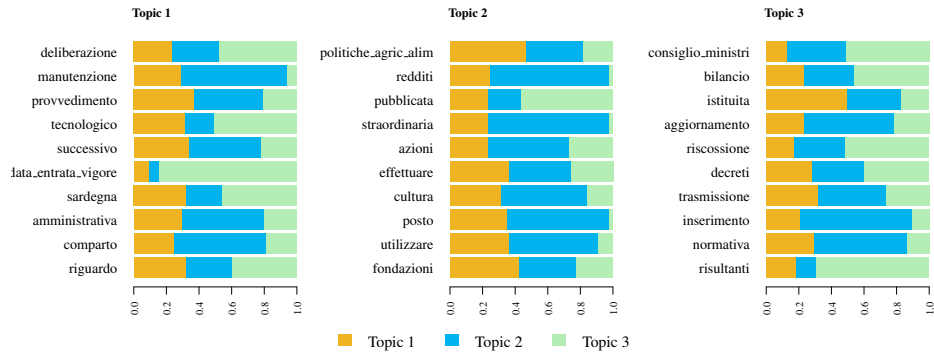


Fig. 4 Highest Frex topic words alongside the proportion of occurrence for each estimated topic (fLSA).

6. Luz De Araujo, P.H., De Campos, T.: Topic modelling brazilian supreme court lawsuits. In: Legal Knowledge and Information Systems, pp. 113–122. IOS Press (2020)
7. Schröter, J., Du, K.: Validating topic modeling as a method of analyzing sujet and theme. *Journal of Computational Literary Studies* **1**(1) (2022)
8. Tuzzi, A., et al.: Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Springer (2018)
9. Valdez, D., Pickett, A.C., Goodson, P.: Topic modeling: latent semantic analysis for the social sciences. *Social Science Quarterly* **99**(5), 1665–1679 (2018)

EmurStat: a digital tool for statistical analysis of emur flow

Simone Paesano, Maria Gabriella Grassia, Marina Marino, Dario Sacco, Rocco Mazza

Abstract New Public Management (NPM) emphasizes the use of market-based techniques to improve efficiency and effectiveness in public service delivery. This approach seeks to promote accountability and performance measurement. Key performance indicators describe the performance of processes that characterize a specific workflow. One of the concepts that has emerged in the last decade is Precision Public Health, which integrates traditional determinants of health with new approaches such as data science and health economics. Moreover, visualizations help to understand social determinants of health and public health indicators. This paper aims to present a useful application for data visualization, processing, and analysis for understanding and evaluating the performance of services provided by emergency rooms, through the lunge on a specific case.

Key words: Data visualization, Emergency room.

Simone Paesano

University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: simone.paesano@unina.it

Maria Gabriella Grassia

University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: mgrassia@unina.it

Marina Marino

University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: marina.marino@unina.it

Dario Sacco

University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: davide.sacco@unina.it

Rocco Mazza

University of Bari Aldo Moro, Department of Political Science, Piazza Cesare Battisti, Bari e-mail: rocco.mazza@uniba.it

1 Introduction

The New Public Management (NPM) represents a managerial approach applied in government agencies at both local and national levels to enhance the efficiency and effectiveness of public service organizations. Essentially, NPM entails the incorporation of principles and practices from the private sector into the public sector (Gomes & Mendes, 2022). Pollitt's taxonomy (1997) delineates eight interrelated components that encapsulate the objectives of NPM: cost reduction, budget constraints, transparent resource allocation, organizational disaggregation into separate agencies, decentralization of management authority within public agencies, separation of service provision from procurement, introduction of market-like mechanisms, and implementation of performance management incorporating goals, indicators, and results. NPM also advocates a shift in public employment practices, moving from tenure-based, nationally standardized pay and conditions to group contracts, performance-related pay, and local determination of pay and working conditions. The implementation of New Public Management (NPM) in healthcare has led to a significant emphasis on digitization processes and measurement of Key Performance Indicators (KPIs) (McGee-Lennon et al., 2017). There has been a growing momentum in digital health initiatives, with a focus on integrating digital technologies to improve health care delivery and improve patient outcomes (McGee-Lennon et al., 2017). Longitudinal qualitative evaluations have emphasized the importance of normalization process theory (NPT) and a comprehensive approach to framework analysis in assessing preparedness to deliver digital health care at scale (McGee-Lennon et al., 2017). In addition, the governance of digital transformation in health services has been recognized as a critical issue that requires in-depth evaluation and monitoring (Barros et al., 2019). In addition, the role of KPIs in digital health has been highlighted, with a call for the development of digital public health interventions and an assessment of the current state of knowledge in digital health from a management perspective (Wienert et al., 2022). This reflects a growing recognition of the need for robust measurement tools and frameworks to assess the impact and effectiveness of digital health initiatives. In addition, ethical issues related to digital health, such as the development of digital twins for personalized health services, have emerged, underscoring the complexities and considerations involved in the digitization of health services (Huang et al., 2022). In conclusion, the application of NPM in healthcare, particularly in the context of digitization processes and the importance of KPIs, is a multifaceted and evolving area. It requires a comprehensive understanding of preparedness, governance, ethical considerations, and the development of measurement tools to ensure the proper implementation and maintenance of digital health initiatives.

2 The guidelines on which the application is based.

The guidelines described by the Italian Ministry of Health can be considered as benchmarks for medical conduct (Mendogni et al., 2019). These guidelines have been instrumental in shaping various aspects of healthcare practices, including rehabilitation activities, pain management, clinical practice, and public health interventions. The Italian Ministry of Health has been actively involved in supporting projects, issuing national recommendations, and developing safety protocols based on these guidelines (Vitale et al., 2022). Specifically, this application was developed on three guidelines produced by the Ministry of Health, in the social health field with a focus on emergency room management. These are the national guidelines on intra-hospital triage; national guidelines on short intensive observation - obi and national guidelines on the development of the overcrowding management plan in emergency rooms. In addition, a session dedicated exclusively to socio-demographic analysis was produced. Before proceeding, it is beneficial to provide a brief definition of the focus of these guidelines. Triage is the initial point of reception for individuals arriving at the Emergency Department. The purpose of Triage is to oversee the management of all individuals accessing the Emergency Department, with particular attention to those in critical conditions. This involves assessing data and clinical indicators that characterize situations of risk, complications, and/or undesirable effects of ongoing treatments. This evaluation is conducted through clinical observation, the utilization of suitable scales/instruments, and appropriate monitoring systems. Triage services must be consistently available on a 24-hour basis.

2.1 Triage

Triage, is the initial point of reception for individuals arriving at the Emergency Department. The purpose of Triage is to oversee the management of all individuals accessing the Emergency Department, with particular attention to those in critical conditions. This involves assessing data and clinical indicators that characterize situations of risk, complications, and/or undesirable effects of ongoing treatments. This evaluation is conducted through clinical observation, the utilization of suitable scales/instruments, and appropriate monitoring systems. Triage services must be consistently available on a 24-hour basis. From a methodological standpoint, the Triage process unfolds in four phases: 1) Immediate Assessment Phase (at the entrance): This involves a swift observation of the individual's general appearance to identify those with care needs requiring immediate intervention. 2) Subjective and Objective Assessment Phase: a. Subjective assessment: Conducted through targeted interviews (focused medical history). b. Objective assessment: Carried out through the observation of clinical signs, vital parameter measurements, and analysis of available clinical documentation. 3) Triage Decision Phase: This entails assigning a priority code, implementing necessary care measures, and potentially activating diagnostic and therapeutic pathways. 4) Reassessment Phase: This phase involves

confirming or modifying the assigned priority code for patients in waiting, ensuring ongoing accuracy in the prioritization of cases. The statistical analyses included within the application are based on these methodological principles. The dimensions considered by the application are as follows: Mode of arrival; Absolute frequencies triage Igresso; Appropriateness triage code; Appropriateness intake times by triage code; Average waiting time and standard deviation grouped by code; Appropriateness by emergency department permanent times; Absolute frequencies type of trauma; Diagnosis code; Outcome pathway in emergency department (including patients in OBI).

2.2 *Obi*

Observation and Brief Intervention (OBI) is a management modality for emergency patients presenting with acute clinical conditions that are either high critical but low developmental risk, or low critical but with potential developmental risk. These cases have a high probability of reversibility and require an urgent and nondeferrable diagnostic and therapeutic process that cannot be managed in alternative healthcare settings. OBI is distinguished by its high intensity of care, involving significant medical and nursing staff commitment, diagnostic investigations, clinical monitoring, and strategic therapeutic planning, and is administered within a specifically defined and restricted time frame. Its goal is to ascertain the most appropriate level of care. The functions of the OBI include: - Clinical observation - Provision of short-term therapy for moderately complex conditions - Facilitation of diagnostic and therapeutic insights aimed at appropriate admission or discharge. The expected consequences are improved appropriateness of hospital admissions and safer discharge of patients from the emergency department. The three dimensions considered by the application are as follows: (i) Need for route in OBI. (ii) Appropriateness of OBI length of stay. (iii) Average time of admission to OBI and Standard deviation.

2.3 *Overcrowding management plan in emergency rooms*

Regarding these guidelines, the focus is on the concept of overcrowding, which serves as a reference for the other two discussed guidelines. Specifically, overcrowding is the situation where the normal functioning of Emergency Departments is constrained by the imbalance between healthcare demand (number of patients waiting and in care) and available logistical, instrumental, and professional resources. When overcrowding is persistent, it is generally due to the under-sizing of Emergency Departments about actual access, stemming from a real lack of structural, technological, human, and/or organizational resources. In this direction, the application shows a general data section where the total number of patients present within the emergency department in a single day is calculated.

3 Application workflow

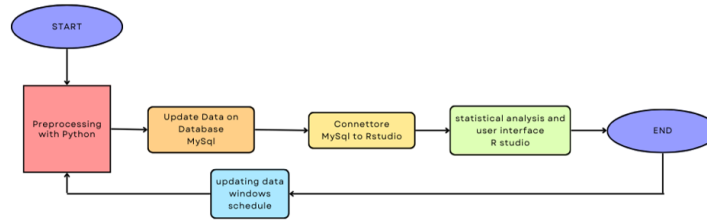


Fig. 1 Work flow aplication

Data mining, processing, and analysis play a key role in understanding and evaluating the performance of health services. These actions help discern patterns and trends and provide support for clinical and managerial decision-making (Vitale et al., 2022). This technical exposition outlines the complexities of the process of extracting and transforming health data. The main programming languages used are Python and R, while the database on which the application relies is MySQL. We used the Python programming language to facilitate the implementation of the data extraction process. This language boasts specialized libraries for adept data acquisition and processing, enabling access to diverse data sources such as CSV files, APIs, or databases. In healthcare, data extraction sources encompass electronic records, hospital management systems, and other health information systems. The specific case involves the extraction of files from Elipse based on The Urgent Emergency Information Flow (also called, Emur Flow), formatted as table format. After extraction, the data undergoes a series of transformations through Python. These encompass data cleaning, aggregation, and the generation of new variables. These operations are crucial for priming the data for subsequent analyses, ensuring consistency and quality. Following transformations, the data is securely stored in the MySQL database. Renowned for its scalability and performance, MySQL serves as a relational database management system, facilitating structured data retention and facile access for subsequent analyses. More intricate transformations, such as statistical analyses, are executed using the R programming language. R, widely acknowledged in statistics and data analysis, offers an array of specialized packages and functionalities. A monthly recall function, developed using Windows Scheduler, ensures the automatic loading of data into the MySQL Database.

4 Conclusion

The extraction and processing of healthcare data through Python, R, and MySQL yield indispensable insights for healthcare management and performance monitor-

ing. These tools enable the analysis of diverse healthcare datasets, contributing to a comprehensive understanding of healthcare systems and patient pathways. Moreover, the incorporation of standardized guidelines, particularly those from the Ministry of Health, safeguards the precision and consistency of statistical analyses (Mulugeta et al., 2019). By adhering to established guidelines, healthcare data analysts ensure the reliability and validity of their findings, thereby enhancing the quality of healthcare data analysis. Furthermore, the development of interactive dashboards facilitates effective data visualization, streamlining the understanding and interpretation of results (Radbron et al., 2021). Interactive dashboards provide a user-friendly interface for healthcare professionals and decision-makers to explore and comprehend complex healthcare data, thereby supporting informed decision-making and strategic planning. This integrative approach advances comprehension of the healthcare system, identifies avenues for enhancement, and buttresses data-driven decisions to uphold the quality and efficacy of healthcare services (Coates Mickan, 2020). In conclusion, the utilization of Data science tools for healthcare data analysis, in conjunction with adherence to standardized guidelines and the development of interactive dashboards, represents a comprehensive and robust approach to advancing healthcare management and performance monitoring.

References

1. Bourke, J. G. (2023). Scatalogic rites of all nations. BoD–Books on Demand.
2. Coates, D., Mickan, S. (2020). Challenges and enablers of the embedded researcher model. *Journal of health organization and management*, 34(7), 743-764.
3. Gomes, P., Mendes, S. M. (2023). Organizational context, use of performance management practices and their effects on organizational performance: an empirical look at these interrelationships. *International Journal of Productivity and Performance Management*, 72(8), 2467-2495.
4. Mulugeta, H., Wagnew, F., Dessie, G., Biresaw, H., Habtewold, T. D. (2019). Patient satisfaction with nursing care in Ethiopia: a systematic review and meta-analysis. *BMC Nursing*, 18, 1-12.
5. Pimentel-Nunes, P., Libânio, D., Marcos-Pinto, R., Areia, M., Leja, M., Esposito, G., ... Dinis-Ribeiro, M. (2019). Management of epithelial precancerous conditions and lesions in the stomach (maps II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy*, 51(04), 365-388.
6. Pollitt, C., Dan, S. (2011). The impacts of the New Public Management in Europe: A meta-analysis.
7. Radbron, E., McCance, T., Middleton, R., Wilson, V. (2021). Maintaining momentum in action research. *Nurse researcher*, 29(3).
8. Tosi, D., Nosotti, M., Bonitta, G., Mazzucco, A., Righi, I., Mendogni, P., ... Crisci, R. (2019). Uniportal and three-portal video-assisted thoracic surgery lobectomy: analysis of the Italian video-assisted thoracic surgery group database. *Interactive cardiovascular and thoracic surgery*, 29(5), 714-721.
9. Visco, V., Vitale, C., Rispoli, A., Izzo, C., Virtuoso, N., Ferruzzi, G. J., ... Ciccarelli, M. (2022). Post-COVID-19 syndrome: involvement and interactions between respiratory, cardiovascular, and nervous systems. *Journal of Clinical Medicine*, 11(3), 524.

Graph Neural Networks for clustering medical documents

Vittorio Torri, Francesca Ieva

Abstract Clustering is one of the most challenging tasks in the field of Natural Language Processing, due to the high dimensionality of textual data. Different types of document embeddings have been proposed in the past, often based on the transformer neural network architecture. In this work, we propose to exploit a graph-based representation combining it with the recent advancements in the field of graph neural networks. While graph neural networks achieved promising results in document classification, their potential for document clustering has not been explored yet. In particular, we propose an application in the medical domain, where document clustering is of paramount importance due to the large amount of information present in medical documents and the difficulties in labelling them.

Key words: Natural Language Processing, Clustering, Graph Neural Networks, Graph Embeddings, Medical Documents

1 Introduction

Document clustering is one of the most relevant tasks in the field of Natural Language Processing (NLP). The field has seen significant advancements in recent years, particularly with the development of the Transformer architecture and the latest large language models (LLMs) like ChatGPT. Nevertheless, there is still a

Vittorio Torri
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: vittorio.torri@polimi.it

Francesca Ieva
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: francesca.ieva@polimi.it
HDS – Health Data Science Centre, Human Technopole, Milan, Italy

need for tailored models oriented to specific tasks in specific domains, for multiple reasons:

- performances of current LLMs have still margins of improvement and are not always state-of-the-art [5]
- the cost of their use and/or deployment is often unsustainable
- in specific domains, like medicine, where sensitive data have to be processed, it is not legally possible to transmit them to commercial companies

In this work, we propose a new pipeline for document clustering that exploits a graph-based representation of the data that is subsequently embedded with a graph-neural network (GNN) autoencoder. To the best of our knowledge, this is the first work to propose the use of graph neural networks for document clustering.

We apply our pipeline to a case study in the medical domain, where there is a huge need of clustering textual data, due to the large amount of useful information that they store and the lack of labelled datasets [9].

The rest of the paper is organized as follows: in Section 2 we discuss the dataset and the clustering pipeline, in Section 3 we present the results and in Section 4 we summarize our findings.

2 Materials and methods

In this section we present the data for the case study and the clustering pipeline, discussing its main components.

2.1 Data

The data used in this study consists of documents from the Italian section of the E3C Corpus [7]. This dataset is composed of publicly available documents collected from a variety of sources, including case reports from medical journals and texts of exams for medical students. This corpus is multilingual, containing data in five different languages, each with three distinct subsets:

- Layer 1: documents manually annotated with respect to clinical concepts, events and temporal relations among events
- Layer 2: documents semi-automatically annotated with respect to the same entities of Layer 1
- Layer 3: unlabelled documents

In this work, we primarily focused on the documents present in Layer 1 and Layer 2 of the Italian section, consisting of 86 and 174 documents, respectively. This is necessary to assess the performance of our clustering pipeline, as we require ground truth labels. Clustering metrics that don't rely on labels evaluate clustering quality

within a specific embedding space. However, the document embedding itself is a crucial part of our clustering pipeline, requiring evaluation.

This dataset does not have direct labels for document classes/clusters and we had to derive them by exploiting the existing labels for clinical concept entities. In particular, Layer 1 and Layer 2 of the dataset are annotated with entities defined as *clinical concepts* and linked to the corresponding concepts in the UMLS ontology [1]. The UMLS ontology includes various types of relationships among its concepts and we exploited the *parent* relationship to derive the more general topics related to the annotated entities. We annotated each UMLS *parent* concept that we extracted from the annotated *clinical entities* with respect to the main subfields of medicine (e.g.: cardiology, oncology, gastroenterology). Consequently, we assigned each document a label corresponding to the most frequent subfield among the parent concepts of its entities. Considering the dataset’s relatively limited size and broad scope, our goal is to identify clusters corresponding to medical fields. These clusters typically contain documents with similar symptoms, exams, or procedures. In the annotated dataset we identified 7 clusters corresponding to different medical areas, covering 213 of the 260 documents. Although only documents from Layer 1 and Layer 2 can be used for the clustering and the evaluation, we also leveraged Layer 3 (consisting of 10209 documents) in the unsupervised training of the graph autoencoder, one of the components of our clustering pipeline.

2.2 Clustering pipeline

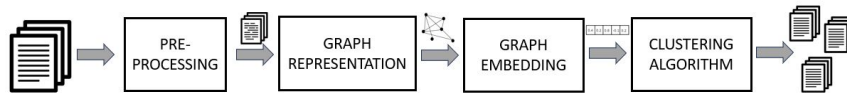


Fig. 1 Schema of the clustering pipeline

Our clustering pipeline is depicted in Figure 1 and consists of four components:

1. Preprocessing
2. Graph-based representation
3. Graph embedding
4. Clustering of graph embeddings

The following subsections detail these components.

2.2.1 Preprocessing

Preprocessing is a fundamental step in NLP pipelines. Given that our pipeline relies on a graph representation aiming to capture the lexical and semantic structure of documents, traditional techniques such as lemmatization and lower-casing were unnecessary. However, we conducted punctuation removal and expanded the most frequent medical term abbreviations present in the dataset.

2.2.2 Graph-based representation

Graph-based representations have a long history in NLP [8], experiencing renewed interest with the emergence of graph neural networks [11]. Various methods exist for representing both individual documents and entire document corpora as graphs. In this study, we propose a document-level graph representation based on dependency parsing, which encapsulates syntactic relationships between words independent of their textual distance. We utilized the Spacy *it_core_news_sm* dependency parsing model for the Italian language [2] to extract these word relationships. Node features were derived from part-of-speech tags, UMLS Semantic Types of words, and Word2Vec embeddings.

2.2.3 Graph embedding

To obtain an embedding vector representing the graph we propose a graph autoencoder architecture. It is derived by the GAE architecture proposed by [3] with the addition of a SAG pooling layer [6]. The basic GAE architecture applies two (or more) layers of Graph Convolution (as defined in [4]), reducing the dimensionality of the node features but not the number of nodes:

$$Z = (\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}) \text{ReLU}(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} X W_0) W_1 \quad (1)$$

where $\tilde{A} = A + I$, being A the adjacency matrix of the graph, X the input features, $\tilde{D}$ the degree matrix of $\tilde{A}$ and W the weights of the network.

Our textual data's high dimensionality depends not only on the features of its nodes (tokens) but also on their number. To address this, we incorporate the SAG Pooling layer, which retains only the top k nodes based on an attention score:

$$\text{att_scores} = \tanh(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} X \Theta) \quad (2)$$

where Θ the attention parameters. This graph autoencoder is trained using as loss function the binary cross-entropy on the edges of the reconstructed graph, with early stopping on a validation set. We used the Adam optimizer with $lr = 10^{-5}$. The size of the embedding representation is 50, chosen with a grid over number of nodes and length of node embeddings in $[5, 10, 20] \times [5, 10, 20]$.

2.2.4 Clustering

Any traditional clustering algorithm can be in principle applied to graph embeddings, being them dense numerical vectors. We evaluated KMeans, hierarchical agglomerative clustering and HDBScan, finding the best performances in the latter. This confirms previous results on clustering embeddings from neural networks [10].

3 Results

Results are measured in terms of precision, recall and F1-score, weighted averaged with respect to the cluster labels discussed in Section 2.1. In Table 1 we compare the pipeline we are currently proposing with a baseline pipeline based on a TF-IDF representation and with our previously proposed pipeline based on the Umberto model [10]. Our GNN-based pipeline shows improvement over previous models. The Umberto-based pipeline performs even worse than TF-IDF on this dataset. This may be attributed to the longer texts, which would necessitate a different pooling mechanism than the one used in [10]. In Table 2 we compare different clustering algorithm in our GNN-based pipeline, with results highlighting HDBScan as the best.

Table 1 Results of the different clustering pipelines

Pipeline	W. Precision	W. Recall	W. F1-Score
TF-IDF	0.2150	0.3240	0.2343
Umberto	0.0690	0.2626	0.1092
GNN	0.3027	0.3923	0.3277

Table 2 Comparison of different clustering algorithms in our GNN-based pipeline

Cluster. Alg.	W. Precision	W. Recall	W. F1-Score
KMeans	0.1659	0.2961	0.2066
Aggl. Clust.	0.2675	0.3240	0.2351
HDBScan	0.3027	0.3923	0.3277

We executed our pipeline on a Google Colab virtual machine with 8 Intel Xeon @ 2.2Ghz, 51 GB of RAM and an Nvidia T4 GPU. The graph construction required 30 minutes (7 sec/doc), mainly due to the delays in calling the UMLS APIs. The graph autoencoder training required 10 minutes and the subsequent clustering took negligible time. Considering these numbers, we consider our approach

scalable to larger dataset by setting up a local UMLS database instance that might severely reduce the time required for the graph construction.

4 Conclusions

This work introduces a novel pipeline for document clustering, leveraging Graph Neural Networks (GNNs), and applies it to a recent dataset within the medical domain. We also devised a semi-automatic mechanism to derive cluster labels for this dataset, facilitating the evaluation process.

The graph-based representation produced by GNNs combines the power of neural-network based embeddings with a pooling mechanism that reduces the number of nodes in a way that seems to be more effective on long texts with respect to the pooling mechanisms that are typically adopted for BERT-based models. The results compared to other existing methods make this research direction promising. Future work will focus on experimenting with different graph representations and autoencoder architectures.

References

1. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32.suppl_1, D267-D270 (2004)
2. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
3. Kipf, T. N., Welling, M.: Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016)
4. Kipf, T. N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations* (2017)
5. Kocof, J., et al.: ChatGPT: Jack of all trades, master of none. *Information Fusion*, 101861 (2023)
6. Lee, J., Inyeop, L., Jaewoo, K.: Self-attention graph pooling. *International conference on machine learning*. PMLR (2019)
7. Magnini, B., Altuna, B., Lavelli, A., Speranza, M., Zanolli, R.: The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy. Associazione Italiana di Linguistica Computazionale. (2020)
8. Sonawane, S. S., and Kulkarni, P.A.: Graph based representation and analysis of text document: A survey of techniques. *International Journal of computer applications* 96.19 (2014)
9. Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., Godtliebsen, F.: Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6), e1549. (2021)
10. Torri, V., Ercolanoni, M., Bortolan, F., Leoni, O., Ieva, F.: Clustering Italian medical texts: a case study on referrals. In *Proceedings of the Statistics and Data Science Conference* (pp. 158-163). Pavia University Press (2023)
11. Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, G., Long, B.: Graph neural networks for natural language processing: A survey. *Foundations and Trends in Machine Learning* 16.2, 119-328 (2023)

