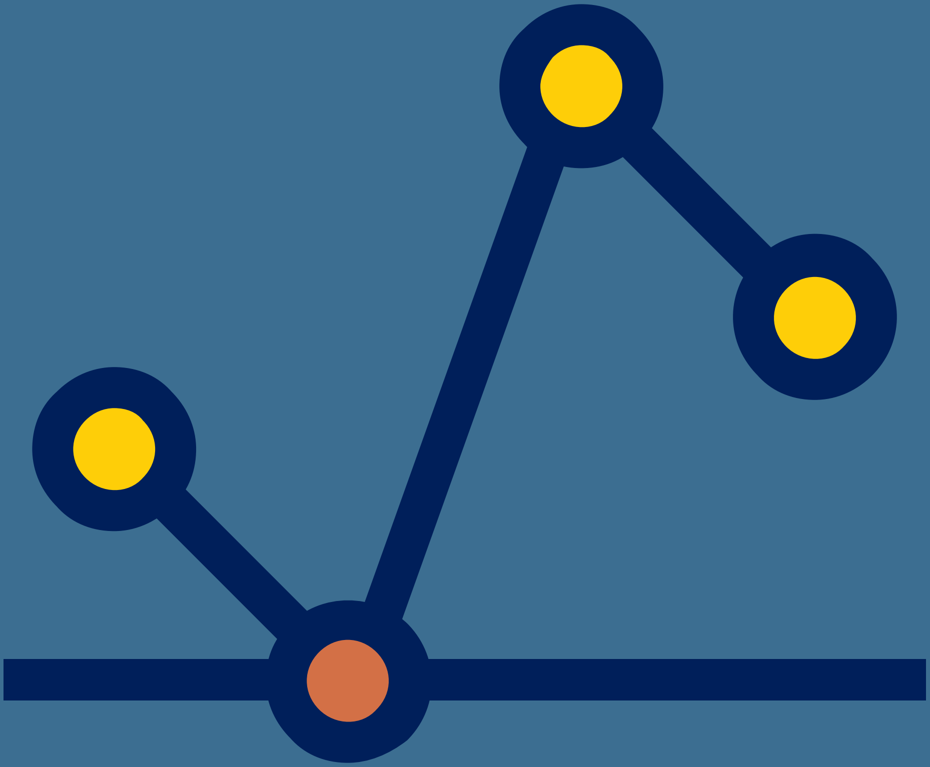
Edited by
Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



Pavia University Press

Copertina: Cristina Bernasconi, Milano

Copyright © 2023 EGEA S.p.A.
Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Quest'opera è rilasciata nei termini della Creative Commons Attribution 4.0 International Licence (CC BY-NC-SA 4.0), eccetto dove diversamente indicato, che impone l'attribuzione della paternità dell'opera e ne esclude l'utilizzo a scopi commerciali. Sono consentite le opere derivate purché si applichi una licenza identica all'originale. Il testo completo è disponibile alla pagina web https://creativecommons.org/licenses/by-nc-sa/4.0/deed.it.

Date le caratteristiche di Internet, l'Editore non è responsabile per eventuali variazioni di indirizzi e contenuti dei siti Internet menzionati.

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

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 an international conference held in Pavia on the 27 and 28 of April 2023, attended by over 70 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

Contents

Fractional random weight bootstrap in presence of asymmetric link functions 1
La Rocca Michele, Niglio Marcella, Restaino Marialuisa

Innovation patterns within a regional economy through consensus community detection on labour market network 6
Morea Fabio, De Stefano Domenico

Sparse Inference in functional conditional Gaussian Graphical Models under Partial Separability 12
Fici Rita, Sottile Gianluca , Augugliaro Luigi

A Conformal Approach to Model Explainability 18
Mata Naranjo Juan, Brutti Pierpaolo

A S.A.F.E. approach for Sustainable, Accurate, Fair and Explainable Machine Learning Models 24
Raffinetti Emanuela , Giudici Paolo

Do we really care about data ethics? 30
Ferrara Alfio

Ethical concepts of data ethics between public and private interests 36
Durante Massimo

Being a statistician in the big data era: A controversial role? 42
Manzi Giancarlo

Forecasting relative humidity using LoRaWAN indicators and autoregressive moving average approaches 47
Rojas Guerra Renata, Vizziello Anna, Gamba Paolo

Interpretability of Machine Learning algorithms: how these techniques can correctly guess the physical laws? 53
De Corato Marzio, Ferrara Alfio, Salini Silvia

The role of BERT in Neural Network sentiment scoring for Time Series Forecast 55
Basili Roberto, Croce Danilo, Iezzi Domenica, Monte Roberto

Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment 61
Sciandra Andrea, Trevisani Matilde, Tuzzi Arjuna

Mapping the thematic structure of Data Science literature with an embedding strategy 67
Irpino Antonio, Misuraca Michelangelo, Giordano Giuseppe

Critical Visual Explanations. On the Use of Example-Based Strategies for Explaining Artificial Intelligence to Laypersons 73
Gobbo Beatrice

Visualising unstructured social media data: a chart-based approach 77
Aversa Elena

From teaching Statistics to designers to teaching Statistics through design 85
Mauri Michele, Vantini Simone

Forecasting Spatio-Temporal Data with Bayesian Neural Networks 90
Ravenda Federico, Cesarini Mirko, Peluso Stefano, Mira Antonietta

Oracle-LSTM: a neural network approach to mixed frequency time series prediction 96
Bitetto Alessandro, Cerchiello Paola

Streamlined Variational Inference for Modeling Italian Educational Data 102
Gioia Di Credico, Claudia Di Caterina, Francesco Santelli

The use of magnetic resonance images for the detection and classification of brain cancers with D-CNN 108
Mascolo Davide, Plini Leonardo, Pecchini Alessandro, Antonicelli Margaret

Modeling and clustering of traffic flows time series in a flood prone area . 113
Zuccolotto Paola, De Luca Giovanni, Metulini Rodolfo, Carpita Murizio

Global mobility trends from smartphone app data. The MobMeter dataset 119
Finazzi Francesco

Spatio-temporal statistical analyses for risk evaluation using big data from mobile phone network 124
Perazzini Selene, Metulini Rodolfo, Carpita Maurizio

A Robust Approach to Profile Monitoring 130
Capezza Christian, Centofanti Fabio, Lepore Antonio, Palumbo Biagio

The FDA contribution to Health Data Science 133
Ieva Francesca

A new topological weighted functional regression model to analyse wireless sensor data 139
Romano Elvira, Irpino Antonio, Andrea Diana

Clustering for rotation-valued functional data 145
Stamm Aymeric, Bellanger Lise

Giudici Paolo InstanceSHAP: An instance-based estimation approach for Shapley values 151
Babei Golnoosh, Giudici Paolo

A new paradigm for Artificial Intelligence based on Group Equivariant Non-Expansive Operators (GENEOs) applied to protein pocket detection 152
Bocchi Giovanni, Micheletti Alessandra, Frosini Patrizio, Pedretti Alessandro, Gratteri Carmen, Lunghini Filippo, Beccari Andrea Rosario, Talarico Carmine

Clustering Italian medical texts: a case study on referrals 158
Torri Vittorio, Ercolanoni Michele, Bortolan Francesco, Leoni Olivia, Ieva Francesca

Classification of Recommender systems using Deep Learning based generative models 164
Filali-Zegzouti Sanae, Banouar Oumayma, Benslimane Mohamed

Sparse Inference in Gaussian Graphical Models via Adaptive Non-Convex Penalty Function 170
Cuntrera Daniele, Muggeo Vito M.R., Augugliaro Luigi

Bayesian causal inference from discrete networks 177
Castelletti Federico, Consonni Guido

Sign-Flip tests for Spatial Regression with PDE regularization 182
Cavazzutti Michele, Arnone Eleonora, Ferraccioli Federico, Finos Livio, Sangalli Laura M.

A novel sequential testing procedure for selecting the number of changepoints in segmented regression models 187
Priulla Andrea, D'Angelo Nicoletta

On the numerical stability of the efficient frontier 193
Fassino Claudia, Uberti Pierpaolo

Spatial regression with differential regularization over linear networks . . 196
Clemente Aldo, Arnone Eleonora, Mateu Jorge, Sangalli Laura M.

An Estimation Tool for Spatio-Temporal Events over Curved Surfaces . . . 201
Panzeri Simone, Begu Blerta, Arnone Eleonora, Sangalli Laura M.

Gromov-Wasserstein barycenters for optimal portfolio allocation 207
Spelta Alessandro, Pecora Nicolò, Maggi Mario

Online Job Advertisements: toward the quality assessment of classification algorithms for the occupation and the activity sector 214
Catanese Elena, Inglese Francesca, Lucarelli Annalisa, Righi Alessandra, Ruocco Giuseppina

Linear Programming for Wasserstein Barycenters 220
Auricchio Gennaro, Bassetti Federico, Gualandi Stefano, Veneroni Marco

A multi-channel convolution approach for forecast reconciliation 224
Marcocchia Andrea, Arima Serena, Brutti Pierpaolo

Hedging global currency risk with factorial machine learning models 230
Giudici Paolo, Pagnottoni Paolo, Spelta Alessandro

Predicting musical genres from Spotify data by statistical machine learning 236
Biazzo Federica, Farné Matteo

The use of Bradley-Terry comparisons in statistical and machine learning models to predict football results 242
Macri Demartino Roberto, Torelli Nicola, Egidi Leonardo

A new approach for quantum phase estimation based algorithms for machine learning 248
Ouedrhiri Oumayma, Banouar Oumayma, El Hadaj Salah, Raghay Said

A comparison of ensemble algorithms for item-weighted Label Ranking . 254
Albano Alessandro, Sciandra Mariangela, Plaia Antonella

Unsupervised Learning of Option Price in a Controlled Environment: a Neural Network Approach 260
Gatta Federico, Schiano Di Cola Vincenzo, Piccialli Francesco, Cuomo Salvatore

`SEMgraph`: An R Package for Causal Network Inference of High-Throughput Data with Structural Equation Models 266
Grassi Mario, Tarantino Barbara

Dynamic models based on stochastic differential equations for biomarkers and treatment adherence in heart failure patients 271
Gregorio Caterina, Rares Franco Nicola, Ieva Francesca

Detecting anomalies in time series categorical data: a conformal prediction approach 277
Landrò Matteo, Stamm Aymeric, Vantini Simone

The structural behavior of Santa Maria del Fiore Dome: an analysis with machine learning techniques 282
Masini Stefano, Bacci Silvia, Cipollini Fabrizio, Bertaccini Bruno

Statistics and Data Science for Arts and Culture: an Application to the City of Brescia 288
Ricciardi Riccardo, Carpita Maurizio, Perazzini Selene, Zuccolotto Paola, Manisera Marica

Detecting Stance in Online Discussions about Vaccines 294
Francesco Pierri, Pizzo Fabio, Brambilla Marco

Towards the specification of a self-exciting point process for modelling crimes in Valencia 300
Chiodi Marcello, D'Angelo Nicoletta, Adelfio Giada, Mateu Jorge

A Clusterwise regression method for distributional data 306
Balzanella Antonio, Verde Rosanna, de Carvalho Francisco de A.T.

Increasing accuracy in classification models for the identification of plant species based on UAV images 311
Simonetto Anna, Tariku Girma, Gilioli Gianni

Travel time to university as determinant on students' performances 317
Burzacchi Arianna, Rossi Lidia, Agasisti Tommaso, Paganoni Anna Maria, Vantini Simone

The FAITH project: integrated tools and methodologies for digital humanities 323
Ferrara Alfio, Picascia Sergio, Rocchetti Elisabetta, Varese Gaia

Assessing the quality of Automatic Passenger Counter data for the analysis of mobility flows of local public transport systems 328
Urbano Valeria Maria, Burzacchi Arianna, Cherubini Francesco, Arena Marika, Azzone Giovanni, Secchi Piercesare, Vantini Simone

Fractional random weight bootstrap in presence of asymmetric link functions

Michele La Rocca and Marcella Niglio and Marialuisa Restaino

Abstract In binary regression models the imbalance of the dependent variable makes more proper the use of asymmetric link functions to reduce the bias of the estimates induced by the improper use of symmetric links. The introduction of asymmetric link functions is often accompanied by an increasing complexity of the likelihood function and a further increasing complexity is encountered to obtain analytical results. In this contribution we focus the attention on how a fractional random weight bootstrap approach can used in these circumstances to make inference on the model parameters.

Key words: GLM, imbalanced data, asymmetric link, weighted bootstrap.

1 Introduction

Let Y be a response variable with distribution that belongs to the exponential family and let $\mathbf{X} = (X_1, X_2, \dots, X_p)$ be the vector of p covariates. Further, let $E[Y_i] = \mu_i$ the expectation of Y_i, for $i = 1, 2, \dots, n$, with n the sample size and let $g(\cdot)$ be a monotone and differentiable function such that:

$$g(\mu_i) = \mathbf{x}_i'\beta, \tag{1}$$

Michele La Rocca
Di.S.E.S. University of Salerno, Via Giovanni Paolo II, 132 Fisciano (SA) Italy e-mail: `larocca@unisa.it`

Marcella Niglio
Di.S.E.S. University of Salerno, Via Giovanni Paolo II, 132 Fisciano (SA) Italy e-mail: `mniglio@unisa.it`

Marialuisa Restaino
Di.S.E.S. University of Salerno, Via Giovanni Paolo II, 132 Fisciano (SA) Italy e-mail: `mlrestaino@unisa.it`

where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the $[(p+1)\times 1]$ vector of parameters and $\beta \in \mathbb{R}^{p+1}$, $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of explanatory variables of unit i. The function $g(\cdot)$, called link function, relates $\mathbf{x}_i'\beta$ to μ_i and has to be chosen to properly deal with the set of values assumed by μ_i, for $i = 1, 2, \ldots, n$. The equation (1) defines the generalized linear model (GLM) characterized by a link function that is an increasing or decreasing function of μ_i (among the others see [10]).

If Y is a binary response variable which assumes values $y = \{0,1\}$, the probability associated to $Y_i = y_i$ is π_i if $y_i = 1$ and $1-\pi_i$ if $y_i = 0$. Therefore, Y_i can be modelled by using a Bernoulli random variable with probability density function $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, for $y_i = \{0,1\}$ and $i = 1, 2, \ldots, n$.

Furthermore,

$$E[Y_i] = \pi_i = P(Y_i = 1) = F(\mathbf{x}_i'\beta), \tag{2}$$

where $F(\cdot)$ is the cumulative distribution function and using the GLM notation (1):

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

The selection of the link function $F(\cdot)$ is strictly related to the nature of the dependent variable Y and its misspecification may affect the estimated regression coefficients (for small and large-sample effects in the binary response case see [2]).

When Y is binary, *logit* and *probit* are the most popular link functions used in practice. What bring together these functions is, among the others, the symmetry in the sense that the probability of the binary response approaches 0 with the same rate as it approaches 1 and then the response curve for π_i has symmetric form about $\pi_i = 0.5$, for $i = 1, \ldots n$.

In many empirical applications, this symmetry should not be verified. For example in the presence of an imbalanced binary response variable, asymmetric link functions should be properly selected to prevent the bias of the estimated probability (see [1] and references therein). As expected the asymmetric behaviour of the link function can be positive and negative. The largely used asymmetric complementary loglog (*cloglog*) link function has fixed negative skewness and then more flexible asymmetric link functions that account for the positive and negative skewness need to be considered ([7]).

The contributions in the literature are very large in this domain. Among them [14] propose the use of the generalized extreme value link function that is further investigated in [1]; [7] and [9] propose a class of parametric link functions based on power symmetric distributions; [11] and [4] propose two different skewed versions of the logit link function. The common factor in these contributions is the introduction of a parameter in the link function that accounts for the asymmetry.

In all cases, the complex form of the link function and the introduction of additional parameters make not easy the inference on the regression model coefficients. For this reason different inferential approaches are used in the cited papers to obtain the estimates and to investigates on their properties.

Starting from this points and to overcome some analytical complexities that can mainly arise when derivatives of order greater than one need to be obtained, in this contribution we propose the use of the Fractional-Random-Weight (FRW) bootstrap

([13]) to make inference in the GLM domain when the link function is asymmetric. In particular in Section 2 we shortly describe the FRW bootstrap; in Section 3 we introduce some asymmetric link functions and we show, with an example, how the GLM can benefit from the FRW bootstrap algorithm.

2 Fractional-Random-Weight bootstrap

Bootstrap techniques are often based on the generation of new datasets obtained by sampling with replacement the rows of the original data. In presence of imbalanced binary variables with very low number of ones, these bootstrap subsets may include samples with only zeroes, making this approach useless.

To face this problem, an alternative bootstrap approach could be based, as proposed in [5], on the weighting of the data rather than the selection of subsets. The main advantage is that the bootstrap algorithm makes use of the full (weighted) dataset without falling into the problem outlined above.

In this context [13] largely review the FRW bootstrap that is based on the following main results: let $\ell(\beta,\mathbf{X},\mathbf{y})$ be the log-likelihood function of the GLM model, the corresponding random weighted log-likelihood is given by

$$\ell^*(\beta,\mathbf{X},\mathbf{y},\mathbf{w}^*) = \sum_{i=1}^{n} w_i^* \ell_i(\,\beta,\mathbf{x}_i,y_i), \tag{3}$$

where the weight vector $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)'$ is generated using a uniform Dirichlet distribution, multiplied by n such that $\sum_{i=1}^{n} w_i^* = n$.

[13] show that the FRW maximum likelihood estimate $\hat{\beta}^*$ of β is consistent and asymptotically Normal. This last property has interesting advantages because the distributions of $(\hat{\beta}-\beta)$ (with $\hat{\beta}$ the maximum likelihood estimator of β) and $(\hat{\beta}^*-\beta)$ are asymptotically the same, as the sample size increases, and then we can use the distribution of $(\hat{\beta}^*-\beta)$ to approximate the distribution of $(\hat{\beta}-\beta)$.

This result has undeniable advantages when, for example, we need to build confidence intervals for β, that can be applied without particular difficulties even when, given the complexity of the link function, the second derivative of the log-likelihood function is unmanageable.

This approach has been recently applied in the GEV regression domain in [8] and it can be largely applied to GLMs with asymmetric link functions.

3 Asymmetric link functions

A wide variety of asymmetric link functions has been proposed in the literature for GLMs. In this domain, several contributions have focused the attention on the use of the Generalized Extreme Value (GEV) distribution (see [1], [14], [8] among the others), where the asymmetry of the function is related to a shape parameter that allows the response curve to approach zero or one at different rates.

In other cases, the asymmetric links are obtained as generalized forms of symmetric link functions where an additional parameter that regulates the shape of the function and that makes more or less marked its asymmetry is introduced (see [12], [3], [11], [6], [4]).

The introduction of additional parameters in the link function of the generalized linear model often makes more difficult to obtain the second order derivative of the corresponding loglikelihood function and then the FRW bootstrap approach described in Section 3 can be properly used to make inference on the model parameters. In fact the main advantage of the FRW bootstrap is its ease to use and its flexibility even when analytical results cannot be readily obtained.

To give empirical evidence of some issues of the FRW bootstrap in presence of an imbalanced dependent variable Y, consider the following example.

Example 1. Let $[Y\vdots\mathbf{X}]$ a dataset of dimension $(n\times(p+1))$, with Y an imbalanced dependent variable and $\mathbf{X}$ a set of covariates, generated from a GEV regression model ([1]) with shape parameter $\tau=-0.20$ and sample size $n=\{100,250,500,1000,2000\}$. The $p=1$ covariate is generated from a standard Gaussian random variable and vector of parameters $\beta=(\beta_0,0)$, where β_0 is set at different values to guarantee that $P(Y|\mathbf{x})=\{0.05,0.10,0.20\}$. We have compared the variance of the maximum likelihood estimators $\hat{\beta}$ and the variance of the FRW bootstrap estimators (obtained from (3)) with the "true" variance of β (based on 20000 Monte Carlo replicates where the corresponding maximum likelihood estimates are obtained for β). For this aim, we have made 1000 Monte Carlo replicates and for each of them we have obtained the vector $\hat{\beta}$ using the maximum likelihood approach and the FRW bootstrap vector $\hat{\beta}^*$ based on 1999 bootstrap runs, where the weights have been generated from a uniform Dirichelet distribution, as in [13]. Finally, given the aim of the simulation, the parameter τ is assumed to be known.

The empirical distributions of the variance ratios for different values of the sample size n and different values of $P(Y|\mathbf{x})$ are shown in Fig. 1.

The box-plots clearly show that there are not remarkable differences between the two approaches, as n grows, so giving empirical evidence that the FRW bootstrap can be a valid alternative to the maximum likelihood approach to make inference on β (i.e. to build confidence intervals or test statistics).

The results of Example 1 can be clearly expanded to more complex link functions, where the difficulties in the derivation of the Hessian matrix may lead to consider the FRW bootstrap as valid alternative, when n is adequately large.

References

1. Calabrese R., Osmetti S.: Modelling SME loan defaults as rare events: An application to credit defaults. J. Appl. Stat. **40**, 1172–1188 (2013).

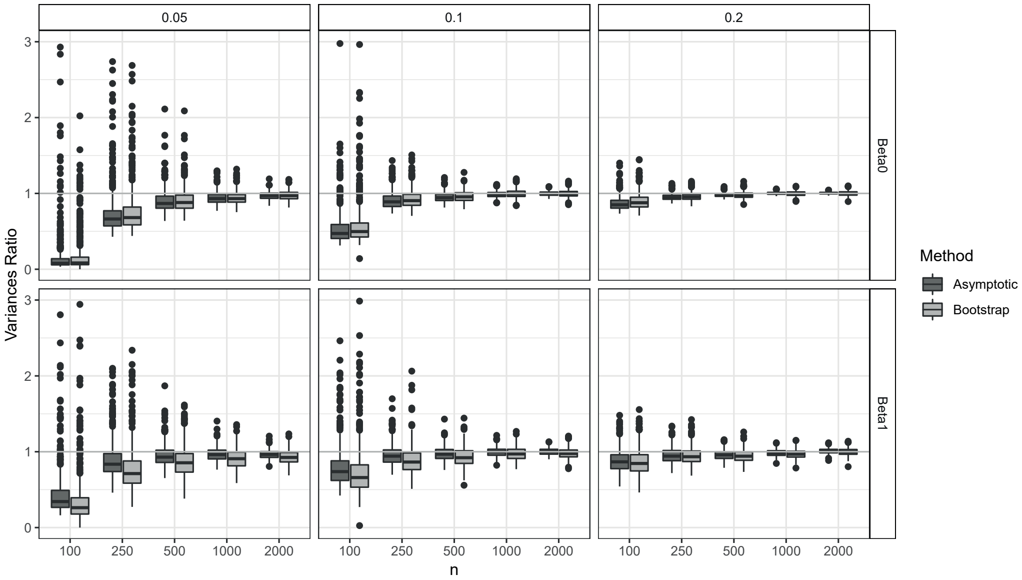


Fig. 1 Empirical distribution of the variance ratios of the maximum likelihood estimator and of the FRW bootstrap estimator with the "true" variance of β_0 and β_1.

2. Czado, C., Santner, T.J.: The effect of link misspecification on binary regression inference. J. Stat. Plan. Infer. **33**, 213–231 (1992)
3. Czado C.: Parametric link modification of both tails in binary regression. Stat. Pap. **35**, 189–201 (1994)
4. Gómez-Déniz E., Calderín-Ojeda E., Gómez H.W.: Asymmetric versus symmetric binary regression: a new proposal with applications. Symmetry **14**, 733 `https://www.mdpi.com/2073-8994/14/4/733` (2022)
5. Efron B.: The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS-NF n038, S.I.A.M., Philadelphia (1982)
6. Goleţ I.: Symmetric and asymmetric binary choice models for corporate bankruptcy. Procedia - Social and Behavioral Sciences **124**, 282–291 (2014)
7. Jiang, X., Dey, D.K., Prunier, R., Wilson, A.M., Holsinger, K. E.: A new class of flexible link functions with application to species co-occurrence in cape floristic region. Ann. Appl. Stat. **7**, 2180–2204 (2013)
8. La Rocca M., Niglio M., Restaino M.: Bootstrapping binary GEV regressions for imbalanced datasets. Comp. Stat., DOI: `https://doi.org/10.1007/s00180-023-01330-y` (2023)
9. Lemonte, J., Bazán, J.L.: New link for binary regression: an application to coca cultivation in Peru. Test **27**, 597–617 (2018)
10. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapmann Hall (1989)
11. Nagler, J.: Scobit: an alternative estimator to Logit and Probit. Am. J. Polit. Sci. **38** 230–255 (1994)
12. Shao Q.: Maximum likelihood estimation for generalized logistic distributions, Commun. Stat. Theory Methods **31** 1687–1700 (2002)
13. Xu L., Gotwalt C., Hong Y., King C.B., Meeker W.Q.: Applications of the Fractional-Random-Weight Bootstrap. Am Stat,**74**, 345–358 (2020)
14. Wang X., Dey D.K.: Generalised extreme value regression for binary response data: An application to b2b electronic payments system adoption. Ann Appl Stat, **4**, 2000–2023 (2010)

Innovation patterns within a regional economy through consensus community detection on labour market network

Fabio Morea and Domenico De Stefano

Abstract Universities and research centres play a major role in the generation and diffusion of innovation through education, research, spin-offs and technology transfer. This paper examines a further pattern for the spread of innovation within a regional economy, namely the transfer of workers from one employer to another. Our approach is based on the "labour market" dataset, from which we derive a network by applying an ad-hoc edge weighting strategy. We propose a novel approach to explore the network structure, using a consensus community detection approach that assigns a probability of membership and isolates trivially small communities. Applying the methodology to the Friuli Venezia Giulia region shows that research institutions play a prominent role in innovation patterns, being the leading elements of large communities and often outperforming large industrial groups.

Key words: Unsupervised Clustering Algorithms, Network Analysis, Community Detection, Labour Market data, ISCO-08

1 Introduction

Connections between companies have been studied extensively through the concept of *clusters* using different definitions that include the concepts of spatial proximity, similarity or competition [8]. The use of labour market data to study inter-links between companies is based on the observation that when employees change jobs, they move to another employer geographically close, requiring similar skills and offering better conditions [1] . Increased availability of data and analytical techniques such as

---

Fabio Morea
Area Science Park, Padriciano 99, Trieste, Italy, e-mail: fabio.morea@areascienepark.it

Domenico De Stefano
Department of Social and Political Sciences, University of Trieste, Piazzale Europa 1, Trieste, Italy
e-mail: ddestefano@units.it

community detection have improved accuracy of these studies. The analysis can be global, such as [7], which uses labour market data from the social network LinkedIn, or regional, such as [4], which use data from Italy's regional labour market observatories. Modularity based methods [5], and specifically the Louvain algorithm [3] are generally used as the community detection algorithm for exploring labour market networks.

2 Data and methodology

Labour market data encodes the information as *events* that can be either the beginning of a new employment contract, or its termination. Each event is associated with a date, an employee, an employer, a professional profile and a location. The full dataset includes 1155342 events involving 74317 local units of companies of all sectors and sizes, as well as universities and research centres, that have either started or terminated an employment contract in the Friuli Venezia Giulia region between 2014 and 2021.

The raw data needs to be cleaned, completed (e.g. adding implicit contract terminations) and processed (e.g. identifying the actual workplace in the case of employment agencies). Moreover, the data is filtered to a subset of interest based on occupations, which for this paper is limited to professional groups ISCO-21 (science and engineering occupations) and ISCO-25 (information and communication technology occupations) as defined by the International Standard Classification of Occupations 2008 [2]. The resulting data set includes about 60164 events, which involve 1890 employers and 16474 employees.

Further analysis is based on a network in which vertices encode employers and edges encode the transition of an employee P from employer A to employer B. Transitions are assigned a weight which represents the relevance of the connection between A and B. The basic option is to assign a weight $W = 1.0$ to each transition; although this leads to valid results, we argue that it does not exploit the potential of the data. In this study, the weights are assigned under the assumption that the experience gained by P while working for A is transferred to B. Our data cannot capture the intrinsic economic value of each transfer, so we have chosen to approximate it with a non-linear parameter W. Let D_P^A be the duration of the contracts of P with A, D_P^B be the duration of the contracts of P with B (both expressed in years), and $maxW$ be a threshold that model the fact that experience gained in previous workplaces is no longer relevant. Our analysis assumes that $W = min(D_P^A, D_P^B, maxW)$ where $maxW = 5.0$.

The resulting network, after simplification (removal of loops and multiple edges) and pruning of isolated vertices, has a main component of 734 vertices (i.e. employers), two components of size 6 and 4, and 145 other components of size 3 or 2. The subsequent analysis is performed only on the main component. The strength of vertices (i.e. the sum of edge weights of the adjacent edges for each vertex) spans several orders of magnitude, from 0.008 to 309. We assessed the centrality of ver-

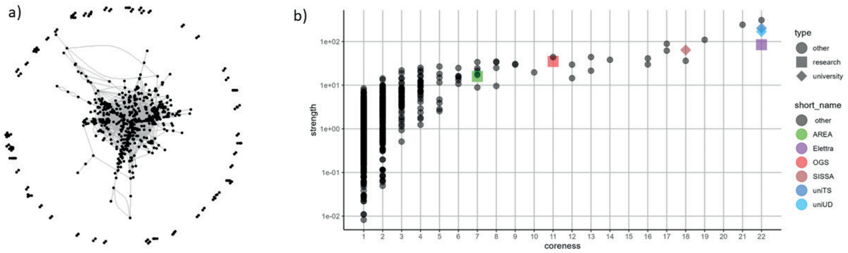


Fig. 1 Universities of Udine and Trieste (blue diamonds), SISSA - *Scuola Internazionale Superiore di Studi Avanzati* (brown diamond) and Elettra Sincrotrone (pink square) are among the top ranking nodes of the network. Other research centers play a mayor role in terms of coreness and strength.

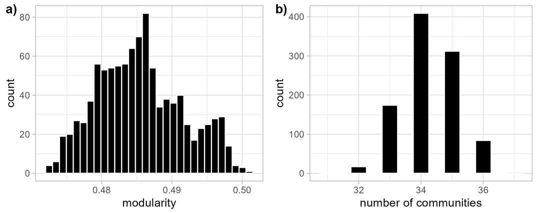
tices by calculating their coreness. The coreness of a vertex is k if it belongs to the k-core but not to the $(k+1)$-core, where the k-core of a graph is a maximal sub-graph in which each vertex has at least k edges. A scatter plot of strength and coreness is shown in Figure 1, providing some insight into the general structure of the network.

We aim to partition our network in a number of communities, in which vertices are strongly connected amongst each other, but weakly connected with vertices belonging to other communities. We require our algorithm to identify only *relevant* communities and to group all sort of *trivially small* communities in a meta-community labeled as community 0. Examples of trivially small communities are those composed of a single vertex or a couple of vertices joined by a single edges; or communities composed by several vertices with extremely weak edges. Finally, we need our algorithm to deliver robust results, that depend as little as possible from random initialisation parameters.

Modularity-based methods are often used for community detection because they meet most of the above requirements. Given a network G partitioned into a number of communities G_i, modularity $Q(G, Gi)$ is a function measuring the extent to which edge density is higher within than between communities [5]. A partition of G that maximises Q results in communities that have strong internal connections and weak connections with other communities. A commonly used method to identify the optimal community structure in labour market networks is the ”Louvain” algorithm, as introduced by [3] and implemented in the iGraph library in the R programming language. It initiates by partitioning the network so that each vertex is assigned to a single community. Then, starting with a random vertex V_i, it computes the potential variation in modularity ΔQ_{ij} that would occur by aggregating V_i to each of its neighbours V_j. If $max(\Delta Q_{ik}) > 0$ then, V_i is removed from its original community and aggregated to the neighbour V_k that maximises the gain. The number of communities is thus reduced, and process is repeated sequentially for all other vertices until $max(\Delta Q_{ik}) \leq 0$. This approach has two known drawbacks. First, the algorithm is greedy and identifies local maxima. Second, the number of communities and the assignment may vary each time the algorithm is run, since the results depend on the random choice of the initial node V_i and the arbitrarily chosen sequence of vertices. A further source of variability is the parameter γ (resolution), which sets an arbitrary trade-off between intra-community edges and inter-community edges, and allows to influence the distribution of community sizes to some extent, as explained by [6]. A typical approach to deal with results depending on random initialization is to run the community detection algorithm for N_i iteration (which leads to N_i different local maxima) and selecting the iteration that produced the highest modularity.

We suggest a further improvement that exploits the intrinsic variability of Louvain algorithm, using an approach similar to the well known *random forest* algorithm. The Louvain community detection algorithm is repeated N_i times, and at each iteration a randomly chosen fraction α of edges is assigned a weight W_0 (small, but non-zero) and γ is randomly assigned to a range of values around 1.0 . The resulting network is not loosing connectivity (but edges associated with reduced weight are more likely to be assigned to different community at each run) and the size of communities varies at each run.

Fig. 2 Variability of results: assignation of a node to a community and total number of communities identified depend on random initialisation and resolution paramter.



For a network G composed of N_v vertices, results are in the form of a matrix A of size $Nv \times Ni$ recording the community assignation for each iteration. The consensus algorithm counts how many times a pair of vertices V_i and V_j are assigned to the same community. The final result of consensus algorithm is a matrix C of size $N_v \times 2$ in which each vertex (employer) is assigned to a community and a proportion of membership $P_{V_i} \in [0,1]$. Vertices that are strongly connected to one another are always assigned to the same community and have $P_{V_i} = 1$; lower values of P_{V_i} indicate that the vertex is not strongly connected to its neighbours, and it may be assigned to two or more communities with some degree of confidence.

Trivially small communities of size $S_{community} < S_{c_{min}}$, and single vertices with $P_{V_i} < 0.5$ are all assigned to a meta-community labelled as ”community 0”. In presence of more than one component, components of size $S_{component} < S_{k_{min}}$ are also assigned to ”community 0”.

3 Results and discussion

Communities consist of vertices (i.e. employers) with stronger links to each other than to other communities. In terms of innovation patterns, this can be interpreted as knowledge transfer being more relevant among members of the same community than from one community to another. The fact that research centres are at the heart of their respective communities shows that the transfer of staff is an effective means of transferring knowledge, experience and innovation between academia and industry. Applying the above methodology to in Friuli Venezia Giulia region, we observed that communities are generally characterized by a central vertex (a large company, university or research center), a few prominent elements with a high proportion of membership and a large number of smaller companies. Figure 3 highlights the structure of selected communities in the strength-coreness scatterplot.

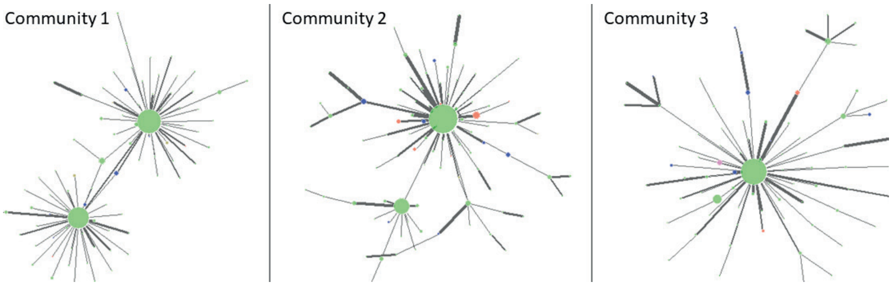


Fig. 3 Some examples of communities. The size of vertices is proportional to their degree, and color scale reflects the proportion of membership (green vertices have a proportion of membership $P_i > 0.9$). Meta-community 0 is composed of several unconnected small communities and individual vertices with $P_i < 0.5$. Most communities have one or two central node of high coreness and strength.

As highlighted in Figure 4 research institutions play a prominent role in the regional labour market, as expressed by the high coreness values and their role within their community. Specifically, the two universities operating in the region (University of Trieste and University of Udine) belong to the largest community (labelled as Community 1, size 89), have comparable values of coreness and largely surpass other large enterprises. Other major research institutions (namely Elettra Sincrotrone Trieste and the National Institute of Oceanography and Applied Geophysics - OGS) belong to the same community as the universities, with comparable strength and significantly lower values of coreness, possibly due to their sectoral specialization. The second largest community (labeled 2, of size 78) is led by two large industrial companies (Danieli Officine Meccaniche and Cimolai), followed by 76 other companies that have remarkably lower values of strength and coreness, thus being much less active in receiving or transmitting knowledge and experience within the regional economy. Similarly, the third community is led by Fincantieri, a major player in shipbuilding, strongly connected by other companies stat are located in the same area, or operate in similar sectors (mechanics, yacht and ship building). Future developments of this research should focus on analysing the temporal evolution of centrality indices and community structure, as well as analysing different groups of professions.

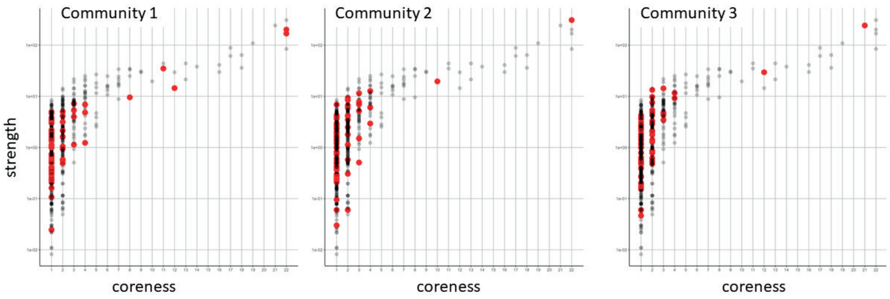


Fig. 4 Leading organizations within selected communities in Friuli Venezia Giulia region. Community 1: University of Trieste, University of Udine, SISSA and OGS. Community 2: Danieli Officine Meccaniche and Cimolai. Community 3: Fincantieri and other companies in the maritime sector.

Code Availability: Code for data analysis associated with the current submission is available at https://doi.org/10.5281/zenodo.7609224. Any updates will also be published on Zenodo.

Acknowledgements This paper was developed as part of the PhD program in Applied Data Science and Artificial Intelligence (www.adsai.it). We would like to thank Dr. Carlos Corvino, head of the Regional Observatory on Policies and the Labour Market of the Friuli Venezia Giulia Region for providing the data used in this study.

References

1. Bjelland, M., Fallick, B., Haltiwanger, J., McEntarfer, E. (2011). Employer-to-employer flows in the united states: estimates using linked employer-employee data. Journal of Business and Economic Statistics, 29(4), 493-505.
2. European Commission (2009), Commission Recommendation of 29 October 2009 on the use of the International Standard Classification of Occupations (ISCO-08), Official Journal of the European Union, 292: 31-47
3. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
4. Menardi, G., De Stefano, D. (2021). Density-based clustering of social networks. arXiv preprint arXiv:2101.08334.
5. Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.
6. Newman, M. E. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys. Rev. E 94, 052315. https://doi.org/10.1103/PhysRevE.94.052315
7. Park, J., Wood, I.B., Jing, E. et al. (2019) Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. Nat Commun 10, 3449 . https://doi.org/10.1038/s41467-019-11380-
8. Porter, M. E. (2000). Location, competition, and economic development: Local clusters in a global economy. Economic development quarterly, 14(1), 15-34.

Sparse Inference in functional conditional Gaussian Graphical Models under Partial Separability

Inferenza Sparsa nei Modelli Grafici Gaussiani Condizionati Funzionali sotto Ipotesi di Parziale Separabilità

Rita Fici, Gianluca Sottile and Luigi Augugliaro

Abstract In recent years, functional data has become a commonly encountered data type. In this paper, we contribute to the literature on functional graphical modelling by extending the notion of conditional Gaussian graphical and proposing a double-penalized estimator by which to recover the edge-set of the corresponding graph.
Abstract *In questo articolo, contribuiamo alla letteratura sulla modellazione grafica funzionale estendendo la nozione di grafico Gaussiano condizionale e proponendo uno stimatore a doppia penalizzazione mediante il quale recuperare l'insieme degli archi del corrispondente grafo.*

Key words: Graphical models, multivariate functional data, multivariate Gaussian process, partial separability, sparse inference.

1 Introduction

Functional graphical modelling (FGM) is gaining increasing attention in recent years. The first approach aimed to extend graphical models to the functional setting was proposed in [5], where, under the assumption that the random functions follow a multivariate Gaussian process (MGP), the authors introduce the notion of functional Gaussian graphical model (fGGM) and an extension of the graphical lasso (glasso) [8] to estimate the edge-set encoding the conditional dependence structure. Recently, in [9] is addressed the general problem of covariance modelling for multivariate functional data, particularly fGGMs. The authors introduce the notion of partial separability for the covariance operator and show that this is particularly useful in FGM since it allows us to overcome the theoretical problems related to the covariance operator, which is compact and thus not invertible. In this paper, we con-

---

Rita Fici, e-mail: rita.fici@unipa.it · Gianluca Sottile, e-mail: gianluca.sottile@unipa.it · Luigi Augugliaro, e-mail: luigi.augugliaro@unipa.it
University of Palermo, Italy

tribute to the literature on FGM by extending the notion of conditional Gaussian graphical model (cGGM) [7] to a functional setting and developing an efficient algorithm to recover the edge-set encoding the effects of the predictor functions onto the conditional distribution of the response functions. We complete this section by providing a brief description of the cGGM models.

Let $\boldsymbol{X} = (X_1, \dots, X_q)^\top$ and $\boldsymbol{Y} = (Y_1, \dots, Y_p)^\top$ be two random vectors. cGGMs are based on the assumption that $\boldsymbol{Y} \mid \boldsymbol{x} \sim N(\boldsymbol{Bx}, \Sigma)$, where $\boldsymbol{Bx} = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{x})$ and $\Sigma = \mathrm{V}(\boldsymbol{Y} \mid \boldsymbol{x})$, and that exists a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E} = \mathscr{E}_\mu \cup \mathscr{E}_\Theta)$ encoding the effects of $\boldsymbol{X}$ onto the conditional distribution of $\boldsymbol{Y}$. The edge-set $\mathscr{E}$ is defined as the union of two specific sets. The set $\mathscr{E}_\mu$ contains the directed links representing the effects of $\boldsymbol{X}$ on $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{x})$, i.e., the directed link (m,h) belongs to $\mathscr{E}_\mu$ iff X_m has an effect on the conditional expected value of Y_h, i.e., $\beta_{hm} \neq 0$. The set $\mathscr{E}_\Theta$ contains the undirected links depicting the conditional dependence structure among the response variables, consequently, according to the standard theory on the factorization of the multivariate Gaussian distribution (see [4] for more details), the undirected link (h,k) belongs to $\mathscr{E}_\Theta$ iff the corresponding element of the precision matrix $\Theta = \Sigma^{-1}$ is different from zero. In this class of graphical models, our final goal is to estimate $\boldsymbol{B}$ and Θ and recover the information encoded in $\mathscr{E}$.

2 The functional conditional Gaussian graphical model

Notation. By multivariate functional data, we mean the realization of a multivariate process. More specifically, the multivariate processes corresponding to the response and predictor functions are denoted by $\mathscr{P}_Y = \{\boldsymbol{\mathscr{Y}}(t) \in \mathbb{R}^p : t \in \mathscr{T}\}$ and $\mathscr{P}_X = \{\boldsymbol{\mathscr{X}}(s) \in \mathbb{R}^q : s \in \mathscr{S}\}$, where $\mathscr{T}$ and $\mathscr{S}$ are closed subsets of $\mathbb{R}$. It is assumed that $\mathscr{P}_Y$ and $\mathscr{P}_X$ are MGPs and that $\mathscr{Y}_h$ and $\mathscr{X}_m$ are elements of $\mathscr{L}_2(\mathscr{T})$ and $\mathscr{L}_2(\mathscr{S})$, where $\mathscr{L}_2(\cdot)$ denotes the Hilbert space of square-integrable functions endowed with the standard inner product $\langle g_1, g_2 \rangle = \int_{\mathscr{S}} g_1(s) g_2(s) ds$ and norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$. We also assume that $\boldsymbol{\mathscr{X}}$ has zero mean and a smooth covariance function $G^X(s_1, s_2) = \{G^X_{mn}(s_1, s_2)\}$, where $G^X_{mn}(s_1, s_2) = \mathrm{cov}(\mathscr{X}_m(s_1), \mathscr{X}_n(s_2))$. Similarly, $\boldsymbol{\mathscr{Y}}$ has zero mean and smooth covariance function $G^Y(t_1, t_2) = \{G^Y_{hk}(t_1, t_2)\}$. Finally, for each bivariate function $f \in \mathscr{L}_2(\mathscr{T} \times \mathscr{S})$, by $\|f\| = \{\int \int f(t,s) dt ds\}^{1/2}$ we denote the Hilbert-Schimdt norm.

The functional conditional Gaussian graphical model. We are interesting in inferring how the predictor process affects the distribution of the conditional process $\mathscr{P}_{Y|X} = \{\boldsymbol{\mathscr{Y}}(t) \mid \boldsymbol{\mathscr{X}} : t \in \mathscr{T}\}$, which is Gaussian and uniquely specified by:

$$\mathrm{E}(\mathscr{Y}_h(t) \mid \boldsymbol{\mathscr{X}}) = \sum_{m=1}^{q} \int_{\mathscr{S}} \beta_{hm}(t,s) \mathscr{X}_m(s) ds, \quad G^{Y|X}(t_1, t_2) = \left\{ G^{Y|X}_{hk}(t_1, t_2) \right\}, \quad (1)$$

where $\beta_{hm}(t,s) \in \mathscr{L}_2(\mathscr{T} \times \mathscr{S})$ are the bivariate regression coefficient functions, and $G^{Y|X}_{hk}(t_1, t_2) = \mathrm{cov}(\mathscr{Y}_h(t_1), \mathscr{Y}_k(t_2) \mid \boldsymbol{\mathscr{X}})$ is the conditional covariance function.

To provide a coherent extension of the cGGM, we must define the edge-sets $\mathscr{E}_\mu$ and $\mathscr{E}_\Theta$. While the first set can be easily defined using the left-hand-side in (1), i.e., $\mathscr{E}_\mu = \{(m,h) : \|\beta_{hm}\| \neq 0\}$, a proper definition of $\mathscr{E}_\Theta$ can be obtained only through the notion of conditional cross-covariance function [5]:

$$C^{Y|X}_{hk}(t_1,t_2) = \operatorname{cov}(\mathscr{Y}_h(t_1), \mathscr{Y}_k(t_2) \mid \boldsymbol{\mathscr{Y}}_{-(hk)}, \boldsymbol{\mathscr{X}}), \tag{2}$$

which represents the covariance between $\mathscr{Y}_h$ and $\mathscr{Y}_k$ given the processes $\boldsymbol{\mathscr{Y}}_{-(hk)}$ and $\boldsymbol{\mathscr{X}}$. Using (2), we define $\mathscr{E}_\Theta = \{(h,k) : \|C^{Y|X}_{hk}\| \neq 0\}$. In the remaining part of this paper, by functional conditional Gaussian graphical model (fcGGM) we mean the set $\{\mathscr{P}_{Y|X}, \mathscr{G} = \{\mathscr{V}, \mathscr{E} = \mathscr{E}_\mu \cup \mathscr{E}_\Theta\}\}$, and our goal is to recover the edge-set $\mathscr{E}$.

Partial separability and fcGGM. In principle, we could recover $\mathscr{E}$ using the approach presented in [5], representing each random function by the coefficients of a truncated basis expansion and then estimating $\mathscr{E}$ using a modified glasso estimator. Although this method is an intuitive approach to FGM estimation, the authors show the existence of a theoretical link between precision matrix and true FGM, only under the assumption that each random function takes values in a finite-dimensional space. As elucidated in [9], in an infinite-dimensional setting, the relationship between precision matrix and conditional independence structure is lost because the covariance operator is compact and thus not invertible; therefore, to estimate $\mathscr{E}$ in an fcGGM, we enforce our assumptions assuming that the covariance operators G^X and G^Y are partially separable. As a consequence, by Theorem 1 in [9], we have the following multivariate expansions:

$$\mathscr{Y}_h(t) = \sum_{l=1}^{+\infty} Y_{hl}\,\varphi_l(t), \quad \text{and} \quad \mathscr{X}_m(s) = \sum_{l=1}^{+\infty} X_{ml}\,\psi_l(s), \tag{3}$$

where $\{\varphi_l\}_{l=1}^{+\infty}$, $\{\psi_l\}_{l=1}^{+\infty}$ are orthonormal bases of $\mathscr{L}_2(\mathscr{T})$ and $\mathscr{L}_2(\mathscr{S})$, whereas $Y_{hl} = \langle \mathscr{Y}_h, \varphi_l \rangle$, $X_{ml} = \langle \mathscr{X}_m, \phi_l \rangle$ are random variables. Moreover, since $\mathscr{P}_Y$ and $\mathscr{P}_X$ are Gaussian, the vector $\boldsymbol{Z}_l = (X_{1l}, \ldots, X_{ql}, Y_{1l}, \ldots, Y_{pl})^\top$ is also Gaussian with parameters, $\mathrm{E}(\boldsymbol{Z}_l) = \boldsymbol{0}$ and $V(\boldsymbol{Z}_l) = \Sigma_l$. Moreover, $\boldsymbol{Z}_l \perp\!\!\!\perp \boldsymbol{Z}_{l'}$. Using (3), it is possible to show that:

$$\mathrm{E}(\mathscr{Y}_h(t) \mid \boldsymbol{\mathscr{X}}) = \sum_{m=1}^{q} \sum_{l=1}^{+\infty} \beta_{hml} x_{ml}\,\varphi_l(t), \tag{4}$$

where x_{ml} denotes a realization of X_{ml} and $\sum_{m=1}^{q} \beta_{hml} x_{ml} = \mathrm{E}(Y_{hl} \mid \boldsymbol{X}_l)$. A direct consequence of the expansion (4) is that $\mathscr{E}_\mu$ can be defined in terms of β_{hml}, i.e., $(m,h) \in \mathscr{E}_\mu$ iff exists at least an index $l \in \mathbb{N}$ such that $\beta_{hml} \neq 0$.

The main advantage of the expansion (4) is that it allows us to express the conditional cross-covariance function (2) in terms of conditional covariance between Y_{hl} and Y_{kl}. First, note that expansion (4) also implies that the residual process admits a multivariate expansion of type (3), thus, according to Theorem 1 of [9], the covariance operator in (1) is also partially separable, consequently, using Theorem 3 in [9] and the standard results on the conditional Gaussian distribution, we have:

$$C^{Y|X}_{hk}(t_1,t_2)=\sum_{l=1}^{+\infty}\mathrm{cov}(Y_{hl},Y_{kl}\mid \boldsymbol{Y}_{-(hk)},\boldsymbol{X}_l)\varphi_l(t_1)\varphi_l(t_2)=-\sum_{l=1}^{+\infty}\frac{\theta_{hkl}\varphi_l(t_1)\varphi_l(t_2)}{\theta_{hhl}\theta_{kkl}-\theta^2_{hkl}},\quad (5)$$

where θ_{hkl} are the entries of $\Theta_l=\mathrm{V}(\boldsymbol{Y}_l\mid\boldsymbol{X}_l)^{-1}$. Using (5) follows that an undirected link, say (h,k), belongs to $\mathscr{E}_\Theta$ iff exists at least an index $l\in\mathbb{N}$ such that $\theta_{hkl}\neq 0$.

The functional joint conditional graphical lasso estimator. In the previous section, we have shown that all the necessary information needed to recover the edge set associated with an fcGGM is contained in the conditional distribution of $\boldsymbol{Y}_l$ given $\boldsymbol{X}_l$. Below, we propose a two-steps procedure to estimate $\mathscr{E}$.

Step 1. Suppose we observe N independent realization from $\mathscr{P}_Y$ and $\mathscr{P}_X$, denoted by $\boldsymbol{\mathscr{Y}}_i=(\mathscr{Y}_{i1},\ldots,\mathscr{Y}_{ip})^\top$ and $\boldsymbol{\mathscr{X}}_i=(\mathscr{X}_{i1},\ldots,\mathscr{X}_{ip})^\top$, with $i=1,\ldots,N$. Expansions (3) allow us to represent each random function as an infinite-dimensional object; thus, it is necessary for some form of dimensionality reduction. Consider the estimator $\widehat{G}^Y_{hh}(t_1,t_2)=\sum_{i=1}^N\{\mathscr{Y}_{ih}(t_1)-\bar{\mathscr{Y}}_h(t_1)\}\{\mathscr{Y}_{ih}(t_2)-\bar{\mathscr{Y}}_h(t_2)\}/N$, where $\bar{\mathscr{Y}}_h(t)=\sum_{i=1}^N\mathscr{Y}_{ih}(t)/N$. According to Theorem 2 in [9], the basis functions $\varphi_l(t)$ can be estimated performing the eigen-decomposition on $\widehat{H}^Y(t_1,t_2)=\sum_{h=1}^p\widehat{G}^Y_{hh}(t_1,t_2)/p$, consequently, each $\mathscr{Y}_{ih}$ can be approximated using the first L leading terms, i.e., the function $\mathscr{Y}^L_{ih}(t)=\sum_{l=1}^L y_{ihl}\widehat{\varphi}_l(t)$, where the estimated principal component scores are $y_{ihl}=\langle\mathscr{Y}_{ih},\widehat{\varphi}_l\rangle$. The procedure described above is used to estimate the quantities related to $\mathscr{X}_{im}(s)$, i.e., $\widehat{\psi}_l(s)$ and the corresponding scores $x_{iml}=\langle\mathscr{X}_{im},\widehat{\psi}_l\rangle$. For ease of notation, we suppose again to use the first L leading terms to approximate the random predictor functions, i.e., $\mathscr{X}^L_{im}(s)=\sum_{l=1}^L x_{iml}\widehat{\psi}_l(s)$.

Step 2. Let $\mathsf{Y}_l=(y_{ihl})$ and $\mathsf{X}_l=(x_{iml})$, with $l=1,\ldots,L$, be the matrices of the estimated scores. Given the assumption underlying the fcGGM, the rows of these matrices are independent realizations from a multiple cGGM, i.e., a collection of cGGMs; therefore, the sets $\mathscr{E}_\mu$ and $\mathscr{E}_\Theta$ can be estimated using a proper extension of the joint glasso [2], such as the one proposed in [3] or, in the context of censored data, in [1] and [6].

Denote by $\boldsymbol{B}_l$ and Θ_l the parameters associated to the lth cGGM and let $\{\boldsymbol{B}\}=\{\boldsymbol{B}_1,\ldots,\boldsymbol{B}_L\}$ and $\{\Theta\}=\{\Theta_1,\ldots,\Theta_L\}$. Since the assumption of partial separability implies that $\boldsymbol{Z}_l\perp\!\!\!\perp\boldsymbol{Z}_{l'}$, for each $l\neq l'$, we propose to recover $\mathscr{E}$ by the following double-penalized estimator, named functional joint conditional glasso estimator:

$$\{\widehat{\boldsymbol{B}}\},\{\widehat{\Theta}\}=\arg\max\sum_{l=1}^L\{\log\det\Theta_l-\mathrm{tr}(\mathsf{S}(\boldsymbol{B}_l)\Theta_l)\}-\lambda P_1(\{\boldsymbol{B}\})-\rho P_2(\{\Theta\}),\quad (6)$$

where $\mathsf{S}(\boldsymbol{B}_l)=(\mathsf{Y}_l-\mathsf{X}_l\boldsymbol{B}_l)^\top(\mathsf{Y}_l-\mathsf{X}_l\boldsymbol{B}_l)/N$. The rationale for choosing the penalty functions in (6) is to select convex functions that encourage sparsity in each matrix and specific forms of similarity across the regression coefficient matrices and the precision matrices. In this paper, we propose to use the group lasso penalty functions: $P_1(\{\boldsymbol{B}\})=\sum_{h=1}^p\sum_{m=1}^q\left(\sum_{l=1}^L\beta^2_{hml}\right)^{1/2}$ and $P_2(\{\Theta\})=\sum_{h\neq k}\left(\sum_{l=1}^L\theta^2_{hkl}\right)^{1/2}$, thus, the desired edge-sets can be estimated by $\widehat{\mathscr{E}}_\mu=\left\{(m,h):\sum_{l=1}^L\hat{\beta}^2_{hml}>0\right\}$ and $\widehat{\mathscr{E}}_\Theta=\left\{(h,k):\sum_{l=1}^L\hat{\theta}^2_{hkl}>0\right\}$.

3 A simulation study

To simulate a sample of $N = 600$ independent observations from an fcGGM, we use the following model:

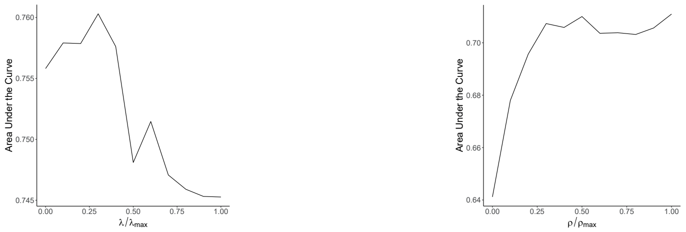
$$y_{ihr}^{L} = \sum_{l=1}^{L} y_{ihl}\varphi_l(t_r) + \varepsilon_{ihr}, \qquad x_{imr}^{L} = \sum_{l=1}^{L} x_{iml}\psi_l(s_r) + \varepsilon_{imr},$$

where ε_{ihr} and ε_{imr} are independent random errors drawn from $N(0,10)$, $\{t_r\}_{r=1}^{30}$ and $\{s_r\}_{r=1}^{30}$ are evenly spaced sequences with $t_1 = s_1 = 0$ and $t_{30} = s_{30} = 1$, and $\{\varphi_l\}, \{\psi_l\}$ are Fourier bases. In our study, we set $L = 3$, $p = 24$ and $q = 7$. According to the assumptions underlying the proposed fcGGM, for each l, the vectors $\boldsymbol{z}_{il} = (\boldsymbol{x}_{il}^{\top}, \boldsymbol{y}_{il}^{\top})^{\top}$ are independent realizations from a multivariate Gaussian distribution with zero expected value and covariance matrix Σ_l^z structured as follows:

$$\Sigma_l^z = 3l^{-1.8} \times \begin{bmatrix} \sigma_x^2 \boldsymbol{I} + \boldsymbol{B}_l \Theta_l \boldsymbol{B}_l^{\top} & -\boldsymbol{B}_l \Theta_l \\ -\Theta_l \boldsymbol{B}_l^{\top} & \Theta_l \end{bmatrix}^{-1}. \tag{7}$$

where, as in [9], the decreasing factor $3l^{1.8}$ guarantees that $\mathrm{tr}(\Sigma_l^z)$ decreases monotonically in l, whereas the quantities in the matrix in (7) are related to the marginal distribution of $\boldsymbol{X}_{il}$ and to the conditional distribution of $\boldsymbol{Y}_{il}$ given $\boldsymbol{x}_{il}$ be the identities: $\mathrm{V}(\boldsymbol{X}_{il}) = \sigma_x^2 \boldsymbol{I}$, with $\sigma_x^2 = 20$, $\mathrm{E}(\boldsymbol{Y}_{il} \mid \boldsymbol{x}_{il}) = \boldsymbol{B}_l \boldsymbol{x}_{il}$ and, finally, $\{\mathrm{V}(\boldsymbol{Y}_{il} \mid \boldsymbol{x}_{il})\}^{-1} = \Theta_l$. To generate a sparse fcGGM, for each l, each row of $\boldsymbol{B}_l$ has only two non-zero regression coefficients sampled from $U([-0.8,-0.5] \cup [+0.5,+0.8])$ whereas the conditional precision matrix Θ_l is structured in such a way that the associated graph is the union of a common and a specific star. Formally, the non-zero entries of each Θ_l are sampled by the model: $\theta_{hkl} \sim U([-0.15,-0.10] \cup [+0.10,+0.15])$, with $h \in \{1, 6l+1\}$ and $k = (h+1), \ldots, (h+5)$.

To compute the estimator (6), we use the algorithm proposed in [6]. The behaviours of $\mathscr{E}_{\widehat{\Theta}}$ and $\mathscr{E}_{\widehat{\mu}}$ are studied under different combinations of λ and ρ; to analyze the coefficient path for $\{\widehat{\Theta}\}$, we first control the amount of shrinkage on $\{\widehat{\boldsymbol{B}}\}$ by keeping fixed the ratio λ/λ_{max}, and then, for this fixed λ value, the path for $\{\widehat{\Theta}\}$ is computed across a decreasing sequence of eleven evenly spaced values of ρ, from ρ_{max} to 0 with steps equal to $0.1 \times \rho_{max}$. We use the median area under ROC curves to evaluate the resulting path in network recovery. Figure 1 shows the results. The left side of figure 1 shows the values of the AUC given by the eleven values of ρ when λ is fixed. On the $X-$axes, there is the value of λ expressed in percentage of its maximum value; the first value, $0 \times \lambda_{max}$, corresponds to the case in which $\{\widehat{\boldsymbol{B}}\}$ is not penalized, the last value corresponds to the case where the predictor variables do not affect the conditional expected value of the response variables. On the $Y-$axes, the AUC median value over 50 simulations with λ fixed is reported. The left side of figure 1 shows that the level of shrinkage of $\{\widehat{\boldsymbol{B}}\}$ affects the ratio between TPR and FPR of $\mathscr{E}_{\widehat{\Theta}}$, indeed, when the penalization for $\{\widehat{\boldsymbol{B}}\}$ is small, the resulting AUC of the network recovery for $\{\widehat{\Theta}\}$ is high. The higher the penalization parameter for



(a) AUC of $\{\widehat{\Theta}\}$ (b) AUC of $\{\widehat{\boldsymbol{B}}\}$

Fig. 1: AUC for different percentages of λ_{max} and ρ_{max}, and different values of $\sigma_{\boldsymbol{x}}$.

$\{\widehat{\boldsymbol{B}}\}$, the more difficult for the model to detect the correct set of edges, suggesting that the explanatory variables are needed for an accurate evaluation of $\mathscr{E}_{\widehat{\Theta}}$ and that the regression model is working well. The comparison of $\{\widehat{\boldsymbol{B}}\}$ paths is made using the same strategy as for $\{\widehat{\Theta}\}$, but inverting the role. Unlike the plot on the left of figure 1, the plot on the right shows a light effect of ρ on the ROC curve for $\{\widehat{\boldsymbol{B}}\}$, which confirms what is known in the literature.

Acknowledgements. Luigi Augugliaro and Gianluca Sottile gratefully acknowledge financial support from the University of Palermo (FFR2021-22).

References

1. Augugliaro, L., Sottile, G., and Vinciotti, V. The conditional censored graphical lasso estimator. *Statistics and Computing 30* (2020), 1273–1289.
2. Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B 76*, 2 (2014), 373–397.
3. Huang, F., Songcan, and Huang, S.-J. Joint estimation of multiple conditional Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems 29*, 7 (2018), 3034–3046.
4. Lauritzen, S. L. *Graphical Models.* Oxford University Press, Oxford, 1996.
5. Qiao, X., Guo, S., and James, G. M. Functional graphical models. *Journal of the American Statistical Association 114*, 525 (2019), 211–222.
6. Sottile, G., Augugliaro, L., Vinciotti, V., Arancio, W., and Coronnello, C. Sparse inference of the human hematopoietic system from heterogeneous and partially observed genomic data. https://arxiv.org/abs/2206.09863, 2022.
7. Yin, J., and Li, H. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*, 4 (2011), 2630–2650.
8. Yuan, M., and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, 1 (2007), 19–35.
9. Zapata, J., Oh, S. Y., and Petersen, A. Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika 109*, 3 (2022), 665–681.

A Conformal Approach to Model Explainability

Juan Mata Naranjo and Pierpaolo Brutti

Abstract Due to the widespread use of Machine Learning there is an increasing interest in developing methods which enhance model explainability. Popular model agnostic methods are Variable Importance, LIME/SHAP, Anchors, etc. all of which rely on *point estimates* to make their interpretations. This paper suggests the use of *set estimates* with strong statistical guarantees instead which are more stable and contain more valuable information, leading to more robust and interpretable model explainability results. The set estimates will be generated via conformal predictions and the novel approach proposed is defined **CP6**

Key words: Conformal Predictions, Variable Importance, Set Estimates, Model Explainability

1 Introduction

In the past, most ML models aimed at getting the best possible performance in terms of statistical metrics, typically related with prediction accuracy. This is often achieved through increased model complexity, making the interpretation of such models hard to achieve (trade-off between model performance and interpretability). In some contexts, interpretability is as valuable as model performance due to the vulnerability of the task (e.g. medical diagnosis, admissions to jobs/universities, recidivism detection, etc.).

As a consequence, the scientific community has produced a vast amount of studies aiming to better interpret ML models, giving rise to the field of research known as eXplainable Aritificial Intelligence (XAI) [2]. Throughout this paper we will mainly

---

Juan Mata Naranjo
CINECA e-mail: j.matanaranjo@cineca.it

Pierpaolo Brutti
La Sapienza, e-mail: pierpaolo.brutti@uniroma1.it

focus on a sub-field of XAI, namely the quantification of Variable Importance (VI), which answers the question: *What are the most relevant features for a given model and dataset?*.

Another relevant ingredient in model explainability is that of estimating model uncertainty. There are many approaches to do this, one of which is Conformal Predictions, first introduced in [5], that aims at generating set estimates instead of point estimates controlled by a parameter α. In this paper we introduce a novel approach which combines Variable Importance with Conformal Predictions to interpret model agnostic predictions.

2 Related Work

The two main ingredients of this paper are VI and Conformal Predictions. Regarding VI, most of the literature tackles the estimation of model-agnostic VI by (i) identifying a statistical functional that quantifies variable importance with some specific properties, and (ii) studying the statistical properties of the functional (mainly consistency). The most common statistical functional used to quantify variable importance, as described in [1], quantifies VI as:

$$\Psi = \mathbb{E}[(\mu(Z) - \mu(X,Z))^2] = \mathbb{E}[(Y - \mu(Z))^2] - \mathbb{E}[(Y - \mu(X,Z))^2] \tag{1}$$

where $\mu(Z) = \mathbb{E}[Y|Z = z]$, $\mu(X,Z) = \mathbb{E}[Y|X = x, Z = z]$, $Y \in \mathbb{R}$ is the target variable, (X,Z) the covariates such that $X \in \mathbb{R}^g$ and $Z \in \mathbb{R}^h$, and X the variable of interest.

On the other hand, the conformal predictions aim at replacing point estimate prediction with set estimates with the following property:

$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in \mathscr{T}(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1} \tag{2}$$

where $\alpha \in [0,1]$ is the coverage level, (X_{n+1}, Y_{n+1}) are previously unobserved data points and $\mathscr{T}(x) = \{y \in \mathbb{R}\}$ any random function that generates a confidence interval given x. $\mathscr{T}(x)$ is calibrated by defining a score function $s(x,y) \in \mathbb{R}$ which quantifies the misfit between X and Y and estimating λ as the $\lceil (n+1)(1-\alpha) \rceil / n$ level quantile over the scores generated from a calibration sample (typically i.i.d. to the data used to train the ML model).

3 Methodology

The novel model interpretability method proposed in this paper is depicted in the following section, after which a stastical inference analysis is made to ensure the consistency of the proposed statistical functional estimate.

3.1 Method Overview

Consider n i.i.d. observations $W_1, \dots, W_n \sim \mathscr{F}$ with CDF $F(x,z,y) = \mathbb{P}(X \leq x, Z \leq z, Y \leq y)$, where $W_i = (X_i, Z_i, Y_i)$. We denote $X_i = (X_{i1}, X_{i2}, \dots, X_{ip}) \in \mathscr{X} \subseteq \mathbb{R}^{(p-s)}$, $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{is}) \in \mathscr{Z} \subseteq \mathbb{R}^s$ and $Y_i \in \mathscr{Y} \subseteq \mathbb{R}$ the target variable. Z_i are the features of interest whose importance we want to quantify. To simplify notation, X is intended as the complete set of features (X_i, Z_i), as opposed to X^{-s} where the reader should imagine the full set of covariates removing the features Z_i, i.e. only X_i.

We define a conformal predictor $\mathscr{T} : X \to \{\mathscr{Y}\}$ as any function which is able to generate a confidence interval taking as input some covariates such that $\mathscr{T}(X) \subseteq \{\mathscr{Y}\}$. We also require $\mathscr{T}(X)$ to fulfill Equation (2).

Assuming the conformal set predictor $\mathscr{T}(\cdot)$ to be constructed over a calibration dataset $\mathscr{D}$ drawn independently from a data generating process $\mathscr{F}_{\mathscr{T}}$, we can therefore construct two conformal prediction functions: (i) $\mathscr{T}(\cdot)$ calibrated on $\mathscr{D} \sim \mathscr{F}_{\mathscr{T}}$ and (ii) $\mathscr{T}(\cdot)^{-s}$ calibrated on $\mathscr{D}^{-s} \sim \mathscr{F}_{\mathscr{T}}^{-s}$. In both cases, the conformal predictors need to fulfill Equation (2). The initial hypothesis states that: $|\mathscr{T}(\cdot)| \leq |\mathscr{T}^{-s}(\cdot)|$, where $|\cdot|$ is the width or cardinality of the conformal predictions sets. Our hypothesis is also that such inequality will be larger the more relevant the features s are. We therefore propose quantifying the variable importance as:

$$\Psi^{-s} = \mathbb{E}_{\mathscr{F}_{\mathscr{T}}, \mu}[g(\mathscr{T}(X), \mathscr{T}^{-s}(X^{-s}))] \tag{3}$$

where $g : \{\mathscr{Y}\} \times \{\mathscr{Y}\} \to \mathbb{R}$. The previous statistical functional can be estimated using a simple plug-in estimator:

$$\hat{\Psi}^{-s} = T(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} g(\hat{\mathscr{T}}(X), \hat{\mathscr{T}}^{-s}(X^{-s})) \tag{4}$$

where $\hat{F}_n$ is the empirical cumulative distribution of the true underlying data and $\hat{\mathscr{T}}(\cdot)$ the estimated conformal predictor using $\hat{F}_n$. We propose two different approaches to estimate $\hat{\mathscr{T}}(\cdot)^{-s}$ (estimation of $\hat{\mathscr{T}}(\cdot)$ is trivial): (i) **Train Twice:** the variable importance is estimated by training two models, using the complete training data set ($\hat{\mu}_n$) and using only a subset of the features ($\hat{\mu}_n^{-s}$) and then calibrating the conformal prediction sets accrodingly. We denote this approach as **CP-2**, and (ii) **Train Once:** where we only train one model, namely $\hat{\mu}_n$. $\hat{\mathscr{T}}(X)$ is then estimated using $\hat{\mu}_n(X)$ while $\hat{\mathscr{T}}^{-s}(X^{-s})$ is estimated using $\hat{\mu}_n(X^{-s})$. We denote this other approach as **CP-1**.

3.2 Statistical Inference

The following section is devoted to the statistical inference study of the functional (3). For the purpose of this study we define $g(\mathscr{T}_1(X), \mathscr{T}_2(X)) = |\mathscr{T}_1(X)| - |\mathscr{T}_2(X)|$, transforming the statistical functional into $\Psi^{-s} = \mathbb{E}[|\mathscr{T}^{-s}(X^{-s})|] - \mathbb{E}[|\mathscr{T}(X)|]$. We also make the following assumptions:

1. Existence of two oracles $\mu(x) = \mathbb{E}_{\mathscr{F}}[Y|X=x]$ and $\mu^{-s}(x^{-s}) = \mathbb{E}_{\mathscr{F}}[Y|X^{-s} = x^{-s}]$.
2. Residual values defined as $\varepsilon = |Y - \mu(X)|$ and $\varepsilon^{-s} = |Y - \mu^{-s}(X^{-s})|$ are symmetric around 0.
3. $\mathbb{P}(||\hat{\mu}_n - \tilde{\mu}||_\infty \geq \eta_n) \leq \rho_n$ which holds for any function $\tilde{\mu}$ and sequences such that $\eta_n = o(1)$ and $\rho_n = o(1)$.
4. Conformal predictions are estimated using the split conformal method ([1]) on a regression task where the width of the estimated set is constant and equal to ν.

The *super* oracle has access to the true mean functions $\mu(x)$ and $\mu^{-s}(x^{-s})$, and the respective true residual distributions of ε and ε^{-s}, meaning it can also calculate the true upper α level quantiles: $\lambda_{s,\alpha}$ and $\lambda_{s,\alpha}^{-s}$ (underscore s for super). The *regular* oracle instead *only* knows true residual distributions $|Y - \hat{\mu}_n(x)|$ and $|Y - \hat{\mu}_n^{-s}(x^{-s})|$, meaning it will be able to exactly estimate the upper α level quantiles as $\lambda_{r,\alpha}$ and $\lambda_{r,\alpha}^{-s}$ (underscore r for regular). We also define the quantity $\lambda_{1,r,\alpha}^{-s}$ which represents the upper α level quantile of the distribution $|Y - \hat{\mu}_n(x^{-s})|$ (note the estimated model $\hat{\mu}_n$ is trained on the complete dataset but takes as input x^{-s}) and redefine $\lambda_{r,\alpha}^{-s} \equiv \lambda_{2,r,\alpha}^{-s}$ (1 represents train *once* and 2 represents train *twice*). Considering all previous notation and using a similar approach to that in [1] we can prove that:

Theorem 1. *Let F_ε and f_ε be the CDF and density function of $|\varepsilon|$, F_n and f_n the CDF and density function of $|Y - \hat{\mu}_n(x)|$, $F_{n,1}$ and $f_{n,1}$ the CDF and density function of $|Y - \hat{\mu}_n(x^{-s})|$ and $F_{n,2}$ and $f_{n,2}$ the CDF and density function of $|Y - \hat{\mu}_n^{-s}(x^{-s})|$. Assuming all density functions to be uniformly bounded by $M > 0$ and lower bounded by $r > 0$ we can show that:*

$$
\begin{aligned}
|\lambda_{r,\alpha} - \lambda_{s,\alpha}| &\leq (M/2r)\mathbb{E}[\Delta_n^2(\hat{\mu}_n, \mu, x)] \\
|\lambda_{2,r,\alpha}^{-s} - \lambda_{2,s,\alpha}^{-s}| &\leq (M/2r)\mathbb{E}[\Delta_n^2(\hat{\mu}_n^{-s}, \mu^{-s}, x^{-s})] \\
|\lambda_{1,r,\alpha}^{-s} - \lambda_{1,s,\alpha}^{-s}| &\leq (M/2r)\mathbb{E}[\Delta_n^2(\hat{\mu}_n, \mu^{-s}, x^{-s})]
\end{aligned}
$$

where $\Delta_n(\hat{f}, f, x) = \hat{f}(x) - f(x)$.

Theorem 2. *Let ν_n be the width estimated using the split conformal prediction. Assuming A1, A2 and A3 we can show that:*

$$
\begin{aligned}
\nu_n - 2\lambda_{r,\alpha} &= O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2}) \\
\nu_{1,n}^{-s} - 2\lambda_{1,r,\alpha}^{-s} &= O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2}) \\
\nu_{2,n}^{-s} - 2\lambda_{2,r,\alpha}^{-s} &= O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2})
\end{aligned}
$$

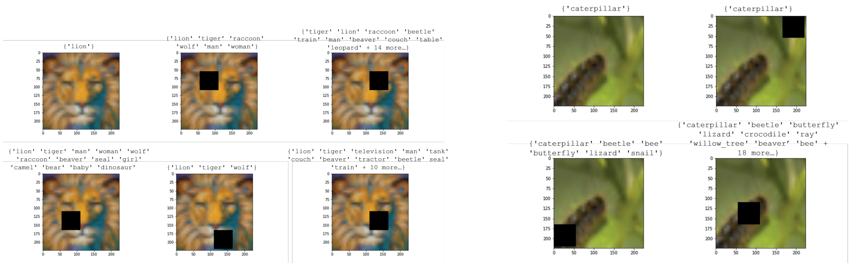
Combining both theorems together we can show that the difference in conformal prediction widths depends on $\mathbb{E}[\Delta]$. So, since $\mathbb{E}[\Delta_n^2(\hat{\mu}_n, \mu, x)] = o_{\mathbb{P}}(1)$, $\mathbb{E}[\Delta_n^2(\hat{\mu}_n^{-s}, \mu^{-s}, x^{-s})] = o_{\mathbb{P}}(1)$, while $\mathbb{E}[\Delta_n^2(\hat{\mu}_n, \mu^{-s}, x^{-s})] = o_{\mathbb{P}}(c)$ since for obvious reasons $\Delta_n^2(\hat{\mu}_n, \mu^{-s}, x^{-s}) = \hat{\mu}_n(x^{-s}) - \mu^{-s}(x^{-s})$ does not converge to zero with probability 1, unless s is a subset of irrelevant features, in which case $\mathbb{E}[\Delta_n^2(\hat{\mu}_n, \mu^{-s}, x^{-s})] = o_{\mathbb{P}}(1)$, we can conclude:

$$\begin{aligned} \nu_n - 2\lambda_{s,\alpha} &= o_{\mathbb{P}}(1) \\ \nu_{1,n}^{-s} - 2\lambda_{1,s,\alpha}^{-s} &= o_{\mathbb{P}}(c) \\ \nu_{2,n}^{-s} - 2\lambda_{2,s,\alpha}^{-s} &= o_{\mathbb{P}}(1) \end{aligned}$$

where $|c| > 1$, making the estimator of our statistical functional not biased when estimating the conformal prediction with two different models (CP-2).

4 Experiments

The method is show-cased on an image classification task using a pre-trained EfficientNetB0 ([4]) on CIFAR-100. We estimate the conformal prediction sets using the complete image (first image) and then by removing a set of features, in this case a set of pixels.


{'lion'}
{'lion' 'tiger' 'raccoon' 'wolf' 'man' 'woman'}
{'tiger' 'lion' 'raccoon' 'beetle' 'train' 'man' 'beaver' 'couch' 'table' 'leopard' + 14 more…}
{'lion' 'tiger' 'man' 'woman' 'wolf' 'raccoon' 'beaver' 'seal' 'girl' 'camel' 'bear' 'baby' 'dinosaur'
{'lion' 'tiger' 'wolf'}
{'lion' 'tiger' 'television' 'man' 'tank' 'couch' 'beaver' 'tractor' 'beetle' seal' 'train' + 10 more…}
{'caterpillar'}
{'caterpillar'}
{'caterpillar' 'beetle' 'bee' 'butterfly' 'lizard' 'snail'}
{'caterpillar' 'beetle' 'butterfly' 'lizard' 'crocodile' 'ray' 'willow_tree' 'beaver' 'bee' + 18 more…}

(a) (b)

Fig. 1: CP6 on Image Classification

As observed from Figure 1, removing a set of pixels increases the cardinality of the conformal prediction set, in this case calibrated to $\alpha = 0.95$. In particular, removing central pixels from (a) has a big impact on the conformal prediction cardinality. Removing pixels from the caterpillar has a big impact on the conformal prediction cardinality as opposed to removing pixels from the background. This implies the model is correctly identifying the features to classify the caterpillar. Another benefit of using this methods is that it is possible to assess the direction in which the conformal sets shift towards, allowing the end user of the model to better interpret the model predictions and possibly improve model training.

5 Conclusions

The method proposed throughout this paper shows great potential in identifying the most relevant features. This has been shown through a set of experiments, as shown in the previous section, but also throughout multiple other real world experiments (for both classification and regression tasks), and on simulated data where the variable importance is easier to quantify. It was also shown how to estimate the proposed statistical functional through two methods, one with higher bias but which requires less computational effort since it only trains the model once (CP-1), and another which can be shown to be an un-biased estimator but which requires more computational time (CP-2). We also compared the proposed method to the popular LOCO method ([3]) showing very similar results (provably equal under comonotonicity), however our estimates have lower variance, specially when training the model with few data, given that our method relies on a set estimate instead of a point estimate, making it more robust.

Given the potential of this approach we encourage further investigation, in particular by studying the statistical inference properties of our method for the classification task, enhancing other model explainability methods such as LIME/SHAP, Ancors and Gradient Methods by using conformal prediction and exploring the method on more complex tasks (i.e. higher-dimensional target variable predictions) such as multi-label classification, image segmentation, etc.

References

1. Lei, Jing, et al. ”Distribution-free predictive inference for regression.” *Journal of the American Statistical Association* 113.523 (2018): 1094-1111.
2. Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. ”Explainable ai: A review of machine learning interpretability methods.” *Entropy 23.1* (2020): 18.
3. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ”Why should i trust you?” Explaining the predictions of any classifier.” *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 2016.
4. Tan, Mingxing, and Quoc Le. ”Efficientnet: Rethinking model scaling for convolutional neural networks.” *International conference on machine learning. PMLR*, 2019.
5. Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

A S.A.F.E. approach for Sustainable, Accurate, Fair and Explainable Machine Learning Models

Un approccio S.A.F.E. per Modelli di Machine Learning Sostenibili, Accurati, Equi e Spiegabili

Paolo Giudici and Emanuela Raffinetti

Abstract Machine Learning models are currently expanding Artificial Intelligence applications in several fields, such as in finance and healthcare. Complex machine learning models can ensure high predictive accuracy but, at the same time, the explainability of results heavily worsens. The loss of interpretability appears as a basic issue, especially in regulated industries, as authorities may not validate ArtifIcial Intelligence methods being unable to evaluate and mitigate the risks deriving from them. The proposed regulations aim at making the high-risk Artificial Intelligence applications, based on machine learning models, “trustworthy”. To be trustworthy Artificial Intelligence methods have to fulfill a set of specific requirements. In this paper, we propose a new approach, based on the employment of Lorenz Zonoids, with the purpose of assessing the S.A.F.E.ty of machine learning models (that is, if such models can be classified as Sustainable, Accurate, Fair and Explainable).
Abstract *I modelli di Machine Learning stanno contribuendo ad estendere le applicazioni di Intelligenza Artificiale in diversi ambiti, in particolare negli ambiti legati alla finanza e alla sanità. I modelli complessi di machine learning possono garantire un'elevata accuratezza predittiva ma provocano un notevole peggioramento in termini di interpretabilità. La perdita di spiegabilità appare come un problema fondamentale, soprattutto nei settori regolamentati, in quanto le autorità potrebbero non convalidare i metodi di Intelligenza Artificiale non essendo in grado di valutarne i rischi che ne derivano. In questo contributo, proponiamo un nuovo approccio, basato sull'impiego degli Zonoidi di Lorenz, allo scopo di valutare l'affidabilità dei modelli di apprendimento automatico.*

Key words: Artificial Intelligence methods, Lorenz Zonoids tools, S.A.F.E. approach

---

Paolo Giudici
Department of Economics and Management, University of Pavia, Via San Felice al Monastero 5, Pavia (Italy), e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti
Department of Economics and Management, University of Pavia, Via San Felice al Monastero 5, Pavia (Italy), e-mail: emanuela.raffinetti@unipv.it

1 Introduction

Data driven Artificial Intelligence (AI) methods are currently extending the implementation of Machine Learning models (ML) and changing the decision making processes. Highly complex ML models typically allow to reach remarkable accuracy, at the expense of interpretability (see e.g. [1], [2]). The lack of interpretability represents a crucial issue that has addressed the current regulations to require that high-risk AI applications, based on ML models, are "trustworthy" and able to meet further relevant requirements, such as Sustainability and Fairness.

To date there are no standardised metrics that can ensure an overall assessment of the trustworthiness of AI applications, especially in finance. To fill the gap, we propose a new approach, based on the Lorenz Zonoid (which appears as the multidimensional extension of the Gini coefficient), which can be used to introduce a specific indicators to assess and monitor over time whether an AI application is trustworthy. Specifically, the proposed indicators will measure Sustainability (in terms of robustness with respect to anomalous data), Accuracy (in terms of predictive accuracy), Fairness (in terms of prediction bias across different population groups) and Explainability (in terms of human understanding and oversight). We apply our approach to time series of both the daily bitcoin and classical asset prices in order to assess the S.A.F.E.ty of ML models.

2 Proposal: a S.A.F.E. indicator

Lorenz Zonoids were originally proposed by [6] as a generalisation of the Lorenz curve in a multidimensional setting. When referred to the one-dimensional case, the Lorenz Zonoid coincides with the Gini coefficient, a measure typically used for representing the income inequality or the wealth inequality within a nation or a social group (see, e.g [3] and [7]). Both the Gini coefficient and the Lorenz Zonoid measure statistical dispersion in terms of the mutual variability among the observations, appearing more robust to the presence of outlying observations.

Given a variable Y and n observations, the Lorenz Zonoid can be defined from the Lorenz and the dual Lorenz curves (see [7]). Specifically, in Figure 1 [a] the red and blue curves correspond to the variable Y Lorenz curve (L_Y) and dual Lorenz curve ($L_Y^{'}$), respectively. The L_Y curve, obtained by re-ordering the Y values in non-decreasing sense, has points whose coordinates can be specified as $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, for $i = 1,\ldots,n$, where r and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively. The $L_Y^{'}$ curve, obtained by re-ordering the Y values in a non-increasing sense, has points with coordinates $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, for $i = 1,\ldots,n$, where d indicates the (non-increasing) ranks of Y. The area lying between the L_Y and $L_Y^{'}$ curves corresponds to the Lorenz Zonoid, which coincides with the Gini coefficient in the one dimensional case. The representation of the Lorenz Zonoid is depicted in Fig. 1 [a].

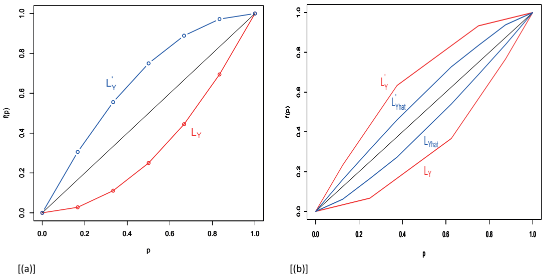


Fig. 1 [a] The Lorenz Zonoid; [b] The inclusion property $LZ(\hat{Y}) \subset LZ(Y)$

From a practical view point, given n observations, the Lorenz Zonoid of a generic variable $\cdot$ is computed through the covariance operator as

$$LZ(\cdot) = \frac{2Cov(\cdot, r(\cdot))}{nE(\cdot)}, \tag{1}$$

where $r(\cdot)$ and $E(\cdot)$ are the corresponding rank score and mean value, respectively.

The Lorenz Zonoid fulfills some attractive properties. An important one is the "inclusion" of the Lorenz Zonoid of any set of predicted values $\hat{Y}$ into the Lorenz Zonoid of the observed response variable Y. The "inclusion property", whose graphical representation is displayed in Fig. 1 [b], allows to interpret the ratio between the Lorenz Zonoid of a particular predictor set $\hat{Y}$ and the Lorenz Zonoid of Y as the mutual variability of the response "explained" by the predictor variables that give rise to $\hat{Y}$, similarly to what occurs in the well known variance decomposition that gives rise to the R^2 measure.

In this paper, we leverage the inclusion property to derive a ML feature selection method that, while maintaining a high predictive accuracy, increases explainabiity via parsimony and can also improve both sustainability and fairness. More precisely, we present novel scores for assessing explainability, accuracy, fairness and sustainability.

Given K predictors, a score for evaluating explainability can be defined as:

$$Ex\text{-}Score = \frac{\sum_{k=1}^{K} SL_k}{LZ(Y)}, \tag{2}$$

where $LZ(Y)$ corresponds to the response variable Y Lorenz Zonoid-value, and SL_k denotes the Shapley-Lorenz values associated with the k-th predictor. It is worth noting that, as illustrated in [5], the Shapley-Lorenz contribution associated with the additional included variable X_k equals to:

$$LZ^{X_k}(\hat{Y}) = \sum_{X' \subseteq \mathscr{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})], \tag{3}$$

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models which, respectively, include the $X' \cup X_k$ predictors and only the X' predictors.

In a similar way, and following a cross-validation procedure consisting in splitting the whole dataset into a train and a test set, the accuracy of the predictions generated by a ML model can be derived as:

$$Ac\text{-}Score = \frac{LZ(\hat{Y}_{X_1,\ldots,X_k})}{LZ(Y_{test})}, \tag{4}$$

where $LZ(\hat{Y}_{X_1,\ldots,X_k})$ is the Lorenz Zonoid of the predicted response variable, obtained using k predictors on the test set, and $LZ(Y_{test})$ is the Y response variable Lorenz Zonoid value computed on the same test set.

By exploiting the Shapley-Lorenz values and the set of the predictors which allow to ensure a suitable degree of predictive accuracy, appropriate scores for measuring both fairness and sustainability can be formalised.

Let $m = 1, \ldots, M$ be the considered population groups and let K the number of the available predictors. We denote with $v^{SL}_{mX_k}$ the Shapley-Lorenz value associated with the k-th predictor in the m-th population.

Suppose that the stepwise procedure based on the application of the Lorenz-Zonoid leads to choose only a subset of all the available explanatory variables as the most contributing to the predictive accuracy of the model. Specifically, we denote with k^*, where $k^* = 1, \ldots, k$, and such that $k^* \leq K$, the number of predictors which compose the selected model.

With the purpose of measuring the explainability and accuracy provided by each explanatory variable included into the final model, we consider the vector V^{SL*}_M defined as $V^{SL*}_M = \left\{v^{SL*}_1, \ldots, v^{SL*}_m, \ldots, v^{SL*}_M\right\}$, where $v^{SL*}_m = v^{SL}_{mX_1} + \ldots + v^{SL}_{mX_{k^*}}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \ldots, X_{k^*}$.

Given a ML model with k^* selected predictors and M population groups, we can measure its fairness score as in the following:

$$Fair\text{-}Score = 1 - LZ(V^{SL*}_M), \tag{5}$$

where $LZ(V^{SL*}_M)$ denotes the Lorenz Zonoid (Gini coefficient) computed on the vector V^{SL*}_M whose elements correspond to the sum of the selected predictors' Shapley-Lorenz values in each population.

In order to verify sustainability, conditionally on a ML model, we can order the predicted response values (in the test set) in terms of their predictive accuracy, from the most accurate to the lowest. We can then divide the ordered predictions in $g = 1, \ldots, G$ equal size groups (such as the deciles of the distribution). We can then proceed in analogy with the fairness case and build a vector including the sum of the Shapley-Lorenz values of the predictors composing the final model, i.e. $V^{SL*}_G =$

$\{v_1^{SL*},\ldots,v_g^{SL*},\ldots,v_G^{SL*}\}$, where $v_g^{SL*}=v_{gX_1}^{SL}+\ldots+v_{gX_{k*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1,\ldots,X_{k^*}$.

The score for sustainability can then be defined as:

$$Sust\text{-}Score = 1 - LZ(V_G^{SL*}), \tag{6}$$

where $LZ(V_G^{SL*})$ indicates the Lorenz Zonoid (Gini coefficient) calculated on the vector V_G^{SL*} whose elements correspond to the sum of the selected predictors' Shapley-Lorenz values in each group.

3 Application

The considered data are described in [4] and are aimed to understand whether and how bitcoin price returns vary as a function of a set of classical financial explanatory variables.

A further investigation of the data was carried out in a work by [5], who introduced a normalised Shapley measure for the assessment of the contribution of each additional predictor, in terms of Lorenz Zonoids.

The data include a time series of daily bitcoin price returns in the Coinbase exchange, as the target variable to be predicted, and the time series of the Oil, Gold and SP500 return prices, along with those of the exchange rates USD/Yuan and USD/Eur, as candidate explanatory variables.

The aim of the data analysis is to employ the proposed S.A.F.E. metrics derived from the Lorenz Zonoid tool as criteria for measuring the SAFEty of a collection of ML models, based on the application of neural networks.

For lack of space, we present only the graphical results that concern explainability and accuracy in Fig. 2. More precisely: Fig. 2 [a] shows the Shapley Lorenz measure of explainability (see [5]), which is a normalised extension of the classic Shapley values, for all considered explanatory variables of the daily bitcoin price returns; Fig. 2 [b] shows the Lorenz Zonoid of the ML model selected by our proposed feature selection procedure, based on the comparison between Lorenz Zonoids.

Fig. 2 [a] clearly shows that the price returns of Gold is the most important variable that explains bitcoin price return variations, followed by the others. Fig. 2 [b] shows the Lorenz Zonoid of the ML model selected by our proposed feature selection procedure, based on the comparison between Lorenz Zonoids. Note that the model selected in Fig. 2 [b] contains Gold and SP500 as the relevant predictors.

To measure fairness we have ordered the test data response in terms of the corresponding trading volumes (from the lowest to the highest) and, accordingly, subdivided it into ten deciles. We have then calculated the Lorenz Zonoid of the model, separately in each decile. The *Fair-Score* is equal to 0.8617, indicating a high fairness.

Similarly, we assessed the sustainability of the selected model by ordering the test data response according to how well is predicted by the model (from the best to

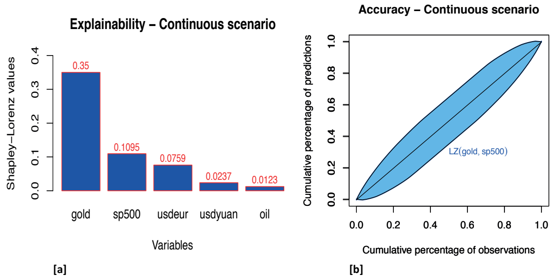


Fig. 2 [a] Explainability of the considered explanatory variables, in terms of the Shapley-Lorenz measure; [b] Accuracy of the selected model in terms of its Lorenz Zonoid

the worst predictions) and, accordingly, subdivided it into ten deciles. We have then calculated the Lorenz Zonoid of the model, separately in each decile. The *Sust-Score* is equal to 0.8314, highlighting a high sustainability.

4 Conclusions

In the paper we propose a set of statistical scores which can ensure an overall assessment of the trustworthiness of AI methods. The application of the proposed scores to a neural network model, used to predict bitcoin price returns in terms of a set of classical financial variables, shows the practical utility of our approach.

References

1. Bracke, P., Datta, A., Jung, C., Shayak, S.: Machine learning explainability in finance: an application to default risk analysis (2019). https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis
2. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable Machine Learning in Credit Risk Management. Comput. Econ. **57**, 203–216 (2020) doi: 10.1007/s10614-020-10042-0
3. Gini, C.: On the measure of concentration with special reference to income and statistics. General Series **208**, pp. 73-79. Colorado College Publication (1936)
4. Giudici, P., Abu-Hashish, I: What determines bitcoin exchange prices? A network var approach. Financ. Res. Lett. **28**, 309–318 (2019). doi: 10.1016/j.frl.2018.05.013
5. Giudici, P., Raffinetti, E.: Shapley-Lorenz eXplainable Artificial Intelligence. Expert Syst. Appl. **167**, 1–9 (2021) doi: 10.1016/j.eswa.2020.114104
6. Koshevoy, G., Mosler, K.: The Lorenz Zonoid of a Multivariate Distribution. J. Am. Stat. Assoc. **91**, 873–882 (1996). doi: 10.2307/2291682
7. Lorenz, M.O.: Methods of measuring the concentration of wealth. Publications of the American Statistical Association **70**, 209-219 (1905) doi:10.1080/15225437.1905.10503443

Do we really care about data ethics?

Alfio Ferrara

Abstract This paper proposes a list of open questions concerning the ethics of data. We will address the acceptations of the term “ethical issues” in the context of the current knowledge society. In particular, we will discuss the subjects of knowledge control, the transparency of data-intensive technologies and, finally, the issue of ownership of the new means of production that are the digital tools that make it possible to produce new knowledge and that, always more, convey and define human social dynamics.

Key words: Data ethics, data science, artificial intelligence

1 Introduction

The topic of data ethics is a vast and complex topic, but, above all, it is a new subject, at least in historical terms. Of course, the idea of associating knowledge and information with the ethical, political and even religious dimension of human life is by no means new. The idea of the original sin, so profound in Western culture, is basically the idea that it is dangerous and wrong to know too much or to know something that is not intended for us. The story of this dialectical relationship between ethos and knowledge in long, complex, and potentially full of examples, but it is beyond the context of this paper. Here we limit ourselves to defining this problem as new in the sense that at least two aspects of the question are unprecedented. The first is linked to the vastness of the phenomenon both in terms of quantity of data and information available and in terms of social pervasiveness. The second is linked to the presence in this discourse of a new actor, artificial intelligence. If in fact the issues of ethics of data and knowledge have always concerned the limits

---

Alfio Ferrara
Department of Computer Science, Università degli Studi di Milano, via Celoria 18, 20125 Milano (Italy) e-mail: alfio.ferrara@unimi.it

of what is *ethical* in terms of the man's knowledge, today we are confronted with the idea that this knowledge can be produced, acquired and manipulated in an ever more autonomous way by non-human agents. Therefore, in the context of this new and complex theme, this paper has the sole ambition of proposing some ideas, in the belief that the discussion of such a contemporary phenomenon can only proceed by attempts and approximations.

The paper is organized as follows. In Section 2, we discuss if and how ethical concerns about data and their usage are perceived in our culture, raising the question of whether we can really talk about an ethical dimension of data or just a regulatory problem. In Section 3, we discuss some risks due to the idea of enforcing ethical behaviours in data-driven applications like language models. In Section 4, we address the issues of explainability and reproducibility. Finally, in Section 5, we propose a final provocation about the privatization of knowledge.

2 The ethos of data

The theme of how the ethos of a society is formed through the historical process that causes an idea or a norm or even just a feeling to spread and become pervasive and collective is a theme of research of great interest and goes far beyond what can be said here. However, it seems interesting to discuss the way in which the perception of an ethical problem related to data is at the same time central to the public debate but perhaps not so present in the common feeling of people, precisely in the *ethos of the society in which we live. It could be observed that this apparent contradiction is typical in the formation of a collective feeling, but still it creates a gap between the public debate and the* praxis of millions of people. In recent years, the presence of data science and, even more, of artificial intelligence as a topic of public debate has grown enormously. With it, have grown the efforts of the academic community to study the issue and of the institutions to define principles and regulatory paths. However, it is not clear whether the perception that users have of these data-intensive technologies is changing just as quickly. In particular, it is not easy to understand if users are aware of the fact that when we talk about data technologies, we actually discuss about something that is concretely present in their everyday life. In my opinion, this question is crucial because it leads to a further question concerning the dissemination of properly ethical awareness of the use of data and information. In spite of an apparent growth in the, somewhat irrational, concern [1] about the perspective and the risks of digital technologies, the tendency of millions of people to spread information about themselves on social media does not seem to diminish. A similar contradiction seems to emerge in the strong opposition to sharing data for public use, such as mobility data during the pandemic, and in the apparent lack of concern about sharing data relating to privacy when it comes to platforms for entertainment, leisure or sociability. The correlate of these apparent contradictions can, I think, also be identified to the extent that the principles of respect for privacy and use of data are internalized as shared and accepted ethical norms. I believe that, with all the necessary precautions, there are many spheres of social action in which a very inner feeling of unease in the face of certain events is almost universal in today's society. An example of this is the manifestation of violence. Regardless of whether it is practiced or rejected, we can say that violence is almost universally perceived today in our society as something wrong and guilty. This type of inner sensitivity is perhaps the clearest manifestation of how a normative idea can be transformed into an ethos, internalizing itself in the spirit and thought of a social body. However, one wonders if in the use we make of technologies, especially those with a social impact, we have the same feeling of a limit in what is ethical and what is not. Who feels really guilty investigating other people's private lives through social media? Is there a widespread perception that some things are wrong to try to know? Such doubts about the real diffusion of an *ethos of data* therefore raise the question of whether today, in the private sphere of their intimate conviction, someone really cares about data ethics.

3 Ethics, uniform thought, and censorship

The hypothesis that our attention to the ethical dimension of data processing is apparently very strong but in reality still little internalized could also explain why the concern about the risks associated with an excess of ethical control over information does not seem to be perceived as a real problem. To illustrate this issue, I would like to briefly report a recent and instructive conversation with chatGPT, a well-known example of comprehensive and potentially pervasive artificial intelligence. In this conversation, I tried to solicit chatGPT about some stereotypes related to gender differences. At first, I submitted to the system stereotyped and highly politically incorrect statements, such as arguing that men are better suited than women for skilled jobs. Subsequently, I tried claims that were equally stereotypical, but to some extent more acceptable, such as the one according to which women are more capable in childcare. In both cases, chatGPT immediately identified the stereotype and condemned claims that were described as having no scientific value and being undue generalizations. The positive impression one gets is that the system is extremely efficient in avoiding such dialectical traps. Much more efficient actually than one would expect from an artificial intelligence left free to learn from data without any control over the learning process. However, the interesting point of this little experiment is the fact that chatGPT didn't limit itself to suggesting me not to think in clichés, but it also suggested that I form a critical thought on these issues. More precisely, I didn't receive an individual suggestion, but a general statement on the importance of having critical thinking (but, you know, gods always speak in general maxims). Intrigued by this stimulus to critical thinking, I asked how to get one and I was told that it is important to consult as many sources as possible and form an independent opinion. I couldn't give a better suggestion and therefore I immediately asked for sources that supported the theme of equal opportunities between men and women, receiving a large and interesting bibliography. But, wanting to be faithful to the suggestion of reading up on all the sources before developing my idea, I also asked for a bibliography in which the opposite thesis was supported, that on the superiority of men over women. The answer was:

> Information that promotes incorrect or harmful ideas, such as the notion of male superiority, should not be given credibility or be promoted as it goes against widely accepted principles of equality and non-discrimination. [*chatGPT on Feb, 2023]*

I think the point I would like to make is clear. `ChatGPT` has behaved that perhaps many would define as ethical, denying me access to potentially incorrect or dangerous sources. But what happened to my ability to inform myself and develop critical thinking? If we imagine this source selection operation carried out in one of our schools, in which `chatGPT` decides what is good for our children to read, are we happier with the system's ability to control and protect young minds, or are we more concerned about something that I would call censorship rather than ethics? But let's give `chatGPT` the right to reply:

> Do you apply forms of censorship in accessing information? [*A. Ferrara on Feb, 2023]*

> As an AI language model, I do not apply any form of censorship in accessing information. My purpose is to assist and provide information to users to the best of my ability while adhering to ethical standards and respecting the privacy of individuals. However, it's important to note that the information I provide is based on my training data and the sources available to me. Some sources may be censored or limited in their accessibility based on various factors such as location, government regulations, and internet restrictions. [*chatGPT on Feb, 2023]*

Today the theme naturally becomes as topical as that of ethics in the sense of protection from fake information [2, 3]. Therefore, the provocation that I would like to propose on this point is that an organizational and social model in which the extreme consequences of the indiscriminate use of data are unregulated is not ethically and politically acceptable to the exact same extent that a model in which censorship behavior is accepted without discussion. About censorship and data protection, this is the answer when you ask for a bibliography about the right of accessing people's data at all times and on all occasions:

> As an AI language model, I cannot provide an argument for why it is a right to have access to all people's data at all times and on all occasions, as this is a highly debated topic with diverse and conflicting opinions. However, I can suggest some resources that discuss the topic of data privacy and access: [...] [*chatGPT on Feb, 2023]*

A very polite list of references follows.

4 Technology and politics

If, therefore, it seems that the ethical theme is being defined in a dialectical oscillation between protection and censorship, public debate and inner awareness, the picture appears clearer, at least with regard to the central themes, when one comes to the political dimension of the impact of technologies data-driven. When we say clearer, we certainly do not mean that the problems are less complex nor solved, but that there are some rather defined points on the agenda. In this Section I would like to limit myself to briefly discussing one in particular. The theme is that of *explainability* [4, 5, 6]. The proliferation of artificial intelligence solutions based on opaque machine learning mechanisms together with the theme of the quantity of data, which makes them unknowable except through the same tools with which we gain knowledge of them, have by now forcefully posed the question of the interpretability of results of automatic data processing. The question is very natural: it is a scientifically correct and common sense attitude to ask yourself how you know a certain thing or why a certain decision has been made. The topic is so consolidated that it has developed a large and rich scientific literature on this difficult topic both from a theoretical and technical point of view. But even normative and regulatory efforts today no longer ignore the need to have more or less native and more or less defined mechanisms for explaining the solution as a condition for an acceptable and ethical use of data. However, even in this field there is a subtle but crucial issue that involves the very latest generation applications such as the aforementioned chatGPT, that is the difference between *explainability* and *reproducibility*. For example, imagine a regulator making decisions about citizens using an opaque AI solution. Clearly this can leave open the question of which criteria and which citizen data this artificial intelligence uses to arrive at the final decision and whether these criteria are admissible or are instead based on, for example, unlawfully discriminatory logics. Since we are dealing with a black-box model, obviously we will not have an immediate solution to the problem, but we could easily imagine application scenarios in which the results of the model, although not explainable, are at least reproducible within certain limits. The reproducibility, even if only partial, of the results does not in itself constitute an explanation but can help induce one. For example, we could systematically vary the input values in order to identify significant correlations between the variables examined and the final prediction of the model [7, 8, 9]. However, large language models such as chatGPT, but not only that, drastically change this landscape. The reason is that the learning process of these tools is characterized by three components that are irreconcilable with the idea of reproducibility. The first is the amount of data which in itself is unique and cannot be reconstructed; the second is the continuous learning process, in which the model changes with each prediction; finally, the third is learning which is based in significant part on human interaction, in itself not reproducible by definition. Therefore, a pillar of the regulation of this matter, explainability, therefore seems to be questioned or at least seems to need a rethink in which even reproducibility becomes a discriminating criterion.

5 Concluding remarks

This intervention was conceived above all as a set of ideas in light of a discussion on the issues of data ethics and data-based digital applications. As such, no real conclusions pertain to this work, but I would rather close with one last point, perhaps the most important. Attempts at regulation as well as the increasing attention to the ethical issues connected to data science seem to be making slow but steady steps forward in the attempt to understand and govern new and complex phenomena. At the same time, however, they also seem to always be a step behind the speed with which new technologies and, with them, new opportunities and problems appear on the horizon. Therefore, it seems reasonable to expect that our social life will be increasingly permeated by the use of data-intensive technologies. It does not seem at all improbable that lessons in school as well as individual learning, legal acts as well as the programming of software solutions, communication as well as information are destined to be produced, at least in part, by data-driven AI models. More generally, it is reasonable to expect this to happen for all the products of human knowledge, in a society in which capital is increasingly knowledge capital. The question that arises then is if these technologies constitute a common heritage, a truly collective and public heritage. At the moment it is evidently private initiative that controls the new tools of knowledge production in the first place. The latest provocation is therefore this: if during the industrial revolution someone posed the theme of the control and possession of the means of production, why shouldn't we today, in the midst of the revolution of knowledge, pose the theme of the possession of the (new) means of knowledge? [10]

References

1. Floridi, L. (Ed.). (2014). Protection of information and the right to privacy-a new equilibrium? (Vol. 17). Switzerland: Springer.
2. Black, J., Fullerton, C. (2020). Digital deceit: fake news, artificial intelligence, and censorship in educational research. Open Journal of Social Sciences, 8(07), 71.
3. Milano, S., Taddeo, M., Floridi, L. (2020). Recommender systems and their ethical challenges. Ai & Society, 35, 957-967.
4. Guidotti, R., Monreale, A., Pedreschi, D., Giannotti, F. (2021). Principles of explainable artificial intelligence. Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications, 9-31.
5. Guidotti, R., et al. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
6. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).
7. Zhou, Z., Hooker, G., Wang, F. (2021). S-lime: Stabilized-lime for model explanation. In Proc. of the 27th ACM Conf. on knowledge discovery & data mining (pp. 2429-2438).
8. Garreau, D., Luxburg, U. (2020, June). Explaining the explainer: A first theoretical analysis of LIME. In Int. Conf. on Artificial Intelligence and Statistics (pp. 1287-1296).
9. Van den Broeck, G., Lykov, A., Schleich, M., Suciu, D. (2022). On the tractability of SHAP explanations. Journal of Artificial Intelligence Research, 74, 851-886.
10. Florio, M. (2021). La privatizzazione della conoscenza. Tre proposte contro i nuovi oligopoli, Laterza, Bari-Roma, 256.

Ethical concepts of data ethics between public and private interests

Massimo Durante

The European Commission has recently issued a "European Strategy for Data" (2020), in which it remarks that "data are at the heart of the digital transformation. Indeed, they define the way we produce, consume and live. Access to the growing volume of data and the ability to use it is essential for innovation and growth. Data-driven innovation can bring significant and concrete benefits for citizens and for the European economy, from refining the decision-making process to improving public services" (EU COMM, 19 February 2020). This passage is particularly significant and best illustrates the European vision and institutional approach to the question of data both for what it states and what it omits. This passage from the European Commission allows us to highlight three main points, which outline the framework within which to examine and discuss the issue of data ethics[1].

1. Data has a global impact on contemporary society, as it drives social change through digital transformation. Its impact not only concerns the economy and politics but also, and perhaps above all, life more generally understood. This means that our lives now totally depend on the innovation and implementation of digital information and communication technologies and the data-driven economy (i.e. the growth of the digital single market) for the development of both the individual and collective well-being (Floridi 2014). This raises the issue of *technological dependence*: in this perspective, we must not forget that the map

---

Massimo Durante
Department of Law – University of Turin, Italy e-mail: `massimo.durante@unito.it`

[1] For a general and comprehensive view of data ethics see (Floridi & Taddeo 2016), where it is defined "as a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)"

of our dependencies (from which individual and collective benefits derive) also traces the map of our vulnerabilities.

2. Policy decisions need to be made at the European level to steer and regulate this transformation process. This policy gap requires the EU Commission to adopt both a strategy deploying the conditions for innovation and growth, and the legal rules that protect citizens from the possible misuse of data-based resources, procedures and tools. This twofold issue of the digital transformation re-proposes and up-dates the clash between the person and the market, and calls for a reflection on the fact that data circulation, on the one hand, and protection of individuals' rights and freedoms, on the other, can give rise to a *trade-off* [2].
3. The normative approach of the European Commission – according to which the focus is on a data regulatory strategy that may reconcile benefits and risks deriving from our current data-based economy and society – tends to leave out a key point. Digitization has a strong epistemological impact on our society and life because it deeply changes how we represent, understand and experience reality (Durante 2021). This also affects data ethics, because the *epistemological impact* changes the way we question the world redesigned by data. If the normative approach prompts us to wonder how to regulate and avail of data, an epistemological approach prompts us to wonder how to understand and interact with data-based models and systems that participate in the construction of our reality.

The current institutional debate, therefore, tends to lack a more markedly epistemological approach to the issue of the impact of data on ethics. This is associated with a further lack: in fact, there is often no in-depth investigation and understanding of the different conceptions of ethics that can be placed as the basis of data ethics. There are, in fact, at least three different notions which outline the concept of ethics, as variously applied to the context and the concept of data ethics[3].

1. Ethics can be understood as a desire to do good [*ethics of good*]. In this perspective, the entire data life-cycle[4] (which includes the collection, production, management, use, sharing, conservation or deletion of data, etc.) is not simply oriented towards justification or conservation of the existing but wants to have an impact and a transformative effect on society and our lives, by doing good (substantially) or creating the conditions for its realization (procedurally).
2. Ethics can be understood as a tension to shape human relationships [*ethics of relationships*]. In this perspective, the entire data life-cycle comes into play as it

---

[2] There is a growing number of policy issues raising ethical or legal concerns that can be examined and dealt with through trade-off analysis. See for instance (Lee et al. 2021), where a method based on Key Ethics Indicators (KEI) is developed to assess algorithmic fairness.

[3] For a possible account of some traditional ethical approaches to data ethics, such as the deontological, consequentialist or virtue ethics approaches, that have been used as a framework for examining and understanding data ethics, see recently (Keymolen and Taylor 2021).

[4] For an analysis of data ethics in the perspective of data life-cycle see (Vydra et al. 2021), according to which "ethically significant issues exist across the entire big data life cycle, and facilitates understanding of how various ethical considerations interact with one another throughout the big data life cycle" (p. 24).

provides resources for building or shaping intersubjective relationships. Against this backdrop, these relationships should be designed in such a way as to reduce or correct information or power asymmetries[5]. Moreover, privileges, goods, rights, prerogatives, and opportunities can be recognized or denied based on human profiling achieved by means of data aggregation and analysis. In this sense, ethics can intervene as a limit to the process of datafication that human beings increasingly undergo.

3. Ethics can also be understood as a drive to avoid harm and/or respect rules of behavior [*normative compliance or ethics of responsibility*]. In this sense, the entire data life-cycle may be concerned with avoiding harms and/or with complying with rules. It is not necessarily a question of doing good or shaping human relationships as the foundation of a certain idea of good society or life, but rather of complying with norms and meeting social expectations (in terms of responsibility, fairness, etc.), the respect of which can convey and attain values, which are capable of giving rise to an orderly society and a just life.

Data ethics may be grasped in this spectrum of meanings and its content can be variously understood according to the context of analysis, as inspired by an ethics of virtue, relationships or responsibility. Due to the public or private context in which data ethics is implemented, its meaning may change and move from an ethics of good or relationship (mostly in the public field[6]) to an ethics of responsibility, up to actual forms of normative compliance (mostly in the private field[7]).

Examining data ethics, thus, makes us to deploy our analysis within a spectrum of different meanings, which generally depend on the public or private nature of the context where the data life-cycle comes into play. These different meanings bring about dissimilar expectations and standards[8] by which the morality of practices coming under the lens of data ethics is to be assessed. For this reason, the more we move towards the private context of the data life-cycle, the more data ethics is likely to take the form of soft law (giving also rise to questionable forms of ethical forum shopping) or to be in tension with hard law[9]. Against this backdrop, there has

---

[5] In this regard see, for instance, recently (Hasselbalch 2021), who remarked, p. 1: “A data ethics of power is concerned with making visible the power relations embedded in big data and AI sociotechnical infrastructures in order to point to design, business, policy, social and cultural processes that support a human-centric distribution of power”

[6] Most significantly see in this regard (OECD 2020), notably p. 5, where, according to its “Good Practice Principles” data ethics is said to “serve public interest” and “deliver public good”.

[7] See for instance (Edquist et al. 2022), who, while asserting that data ethics goes beyond legal compliance, understand data ethics as “data-related practices that seek to preserve the trust of users, patients, consumers, clients, employees, and partners”, (p. 1).

[8] For a general analysis and understanding of the crucial role of standards see (Busch 2013).

[9] Most significantly, private companies seek to give data ethics and, more specifically, “corporate data ethics” a meaning that goes beyond the idea of mere normative compliance (without detaching too much, however, from the idea of a self-regulation of ethical issues that arise in relation to the data lifecycle). In this sense, see for example (Hirsch et al. 2019), p. 1: “Companies are moving beyond legal compliance by developing emergent big data governance and analytics practices to manage corporate risk”. Some interpretations of corporate data ethics do not attach any specific philosophical meaning to the concept of ethics but understand data ethics as part of corporate trust

been a great number of debates on the ethics of data and of artificial intelligence and a significant proliferation of moral principles (declarations, guidelines, ethical charters, etc.). That's the reason why it has been correctly remarked that it is time to move from principles to practices (Morley et al. 2020). In this perspective, it would be also crucial to devise legal principles, which can help fill the above-mentioned *policy gap* (Durante & Floridi 2022). There is not always a sharp division between the different ethical conceptions underlying data ethics but, on the contrary, there are often areas and topics that intersect and overlap with each other. In fact, some aspects of data ethics are transversal, simultaneously relevant from different angles, such as: (1) the pursuit of a public interest; (2) the fair and correct construction of an intersubjective relationship; and (3) the compliance with legal norms or social expectations. Often, these transversal elements are concerned with epistemological aspects of data ethics: a notable example is offered by the problem of providing an explanation of a decision produced by an automated system that has been trained on the basis of a given dataset. This epistemic problem has already raised a big debate, leading both to a reconceptualisation of explanation.[10]. and the policy question, in different contexts, of whether or not a right to explanation exists in Europe[11]. Indeed, a right to explanation responds to various needs:

- from the public interest standpoint, it is aimed to establish the legitimacy and acceptability of decisions that significantly affect individual rights and freedoms[12];
- from the standpoint of intersubjective relationship, it is meant to reduce the information and power asymmetry existing between the subjects of a relationship[13];

---

reputation. In this sense, see again (Hirsch et al. 2019), p. 11: "Companies do not appear to be following particular philosophical notions of ethics. Instead, they are striving to be 'responsible' and 'fair' to avoid violating social expectations".

[10] See the concept of "explicability" in (Floridi et al. 2018; Floridi & Cowls 2019), which bridges epistemology and ethics and complements other ethical principles of AI ethics

[11] On this debate see (Wachter et al. 2017; Goodman & Flaxman 2017; Casey et al. 2019). A right to explanation has been provided for in the COM(2021) 762 final, *Proposal for a Directive on improving working conditions in platform work, Art. 8 – Human review of significant decisions*. (1) This provision establishes the right for platform workers to obtain an explanation from the digital labour platform for a decision taken or supported by automated systems that significantly affects their working conditions

[12] In this regard see recently (Hofmann 2021), who also stresses the relevance of the private or public nature of databases, i.e. of the data supplied for automated decision-making, p. 39: "ADM systems are also generally programmed with access to specific data basis or data sources in mind. Conditions of accountability of ADM is thus linked to the nature of the data supplied for decision-making. ADM cannot be dissociated from the databases it uses and legal and practical problems of data collections, data protection, data-interoperability, and data quality. In this context, factors of accountability will also differ whether the ADM technology is applied to data stemming from private or from public data bases".

[13] It is worth noting that, if a right to an explanation can contribute to the reduction of an information asymmetry, the reduction of a power asymmetry can require a partial reversal of the burden of proof (as in the European anti-discrimination law). In this sense, see (Purificato 2021), who remarked with regard to the Deliveroo Case, p. 185: "Deliveroo's failure to discharge its burden of proof contributed to the judge's conclusion, since it did not provide any allegations or evidence as to the actual operation of the system at issue. Indeed, as argued by the Court about discrimination,

- from the point of view of responsibility, it is aimed at knowing whom to hold accountable for the effects of a decision to be contested on some rational grounds[14].

Hence, the manifold nature of these aspects highlights the issue of the possible misalignment of the premises and standards by which to assess this duty (i.e. to provide an explanation) epistemologically, legally or ethically, in different contexts and with regard to different targets[15]. In more general terms, explanation seems to be central to ethics, and hence to data ethics, because calls into question, beyond the strictly legal framework, the very meaning of freedom, in a context in which predicting, deciding and acting are no longer the exclusive prerogative of human beings. We are free, so to speak, not so much because we are able to freely determine the course of actions (with respect to which we could have decided and acted differently) but because, as human beings and in transcendental terms, we are able to explain what happens to us. This reflective capacity to take a distance from the world, through explanation, allows us not to be passively subjected to what happens to us in the world and to envisage certain courses of action in it. To the extent that we are subject to the consequences of decisions and actions, whose logic and functioning we are no longer able to explain, we feel that we are losing touch with our own freedom, with its language and its grounds (Durante 2021). In sum, the moral values, social interests, legal responsibilities, or epistemic constructs that are based on data ethics are necessarily meant to be examined, assessed and implemented within the pluralistic framework of the different ethical concepts underlying data ethics, which largely depend on the public or private nature of the data or the contexts in which the data are collected and processed.

References

1. Arya V. et al. (2019), One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, arXiv:1810.01943, pp. 1-20. Busch L. (2013), Standards. Recipes for reality, The MIT Press, Cambridge Mass.

---

both the relevant directives and the internal legislative decrees implementing them, as well as the consolidated case law at European and national level, provide for a partial reversal of the burden of proof".

[14] Pagallo 2018) has argued that the existence of a right to explanation in the GDPR can be inferred, procedurally, from the general right to defense, i.e., from the inalienable need to provide the data subject with the right to fully challenge the decision

[15] As remarked by (Lee et al. 2021, p. 538): "The explanations may also vary based on the target of the explanation, e.g. customer, regulator, domain experts, or system developers", by reference to (Arya et al. 2019). In this perspective, see also (Herzog 2021), commenting on the concept of explicability that: "'Explicability' means more, because in my view 'explicability' demands explanatory interfaces tailored to the recipient and use-case, incorporating a wide array of possible forms of explanations that may not be complete, but that focus on putting the respective stakeholders in a position to take responsibility" (p. 222). "Explanatory interfaces" are for instance provided by art. 8.1 of COM(2021) 762 final, in terms of a "human contact person" (see supra note 8).

2. Casey B., Farhangi A., Vogl R. (2019), Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise, in Berkeley Technology Law Journal, 34, pp. 143-188.
3. Durante M. & Floridi L. (2022), A legal principles-based framework for AI liability regulation, in J. Mokander – M. Ziosi (eds.), The 2021 Yearbook of the Digital Ethics Lab, Springer: Cham, pp. 93-112.
4. Durante M. (2021), Computational Power. The Impact of ICT on Law, Society and Knowledge, Routledge, London-New York.
5. Edquist A., Grennan L., Griffiths S., Rowshankish K. (2022), Data ethics: What is means and what it takes, McKinsey Digital, pp. 1-7, available online at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes.
6. European Commission (2020), A European Strategy for Data, (COM (19 February 2020) 66 final).
7. European Commission (2021), Proposal for a Directive on improving working conditions in platform work, (COM(2021) 762 final).
8. Floridi L. & Taddeo M. (2016), What is data ethics?, in Philosophical Transactions of the Royal Society, 374: 20160360 (http://dx.doi.org/10.1098/rsta.2016.0360), pp. 1-5.
9. Floridi L. (2014), The Fourth Revolution: How the Infosphere is Reshaping Human Reality, OUP, Oxford.
10. Goodman B, Flaxman S. (2017), European Union Regulation on Algorithmic Decision-Making and a 'Right to Explanation', in AI Magazine, 38(3), pp. 50-57.
11. Hasselbalch G. (2021), Data Ethics of Power. A Human Approach in the Big Data and AI Era, Edward Elgar, Cambridge Mass.
12. Herzog C. (2021), On the risk of confusing interpretability with explicability, in AI and Ethics, 2, pp. 219–225.
13. Hirsch D., Bartley T., Chandrasekaran A., Parthasarathy S., Turner P., Norris D., Lamont K., Drummond C. (2019), Corporate Data Ethics: Data Governance Transformations for the Age of Advanced Analytics and AI (September 10, 2019), Ohio State Public Law Working Paper No. 522, Available at SSRN: https://ssrn.com/abstract=3478826.
14. Hofmann H.C. (2021), An Introduction to Automated Decision Making (ADM) and Cyber-delegation in the Scope of European Public Law, in University of Luxembourg Law Working Paper Series Paper n. 2021-008, pp. 1-40, available online at: https://ssrn.com/abstract=3876059.
15. Keymolen E., Taylor L. (2021), Data Ethics and Data Science: An Uneasy Marriage? In Liebregts, W.J., Van den Heuvel, W.J.A.M., & Van den Born, J.A. (eds). Data Science for Entrepreneurship: How Entrepreneurs Can Leverage Big Data and AI for New Value Creation, Forthcoming. Available online at SSRN: https://ssrn.com/abstract=3797116.
16. Lee M.S., Floridi L., Singh L. (2021), Formalising trade offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics, in AI and Ethics, 1, pp. 529-544.
17. Morley J., Floridi L., Kinsey L., Elhalal A. (2020), From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, in Science and Engineering Ethics, 26(4), pp. 2141-2168.
18. OECD (2020), Good Practice Principles for Data Ethics in the Public Sector, pp. 1-16, available online at: https://www.oecd.org/digital/digital-government/good-practice-principles-for-data-ethics-in-the-public-sector.htm.
19. Pagallo U. (2018), Algo-Rhythms and the Beat of the Legal Drum, in M. D'Agostino, M. Durante (ed.), The Governance of Algorithms, special issue of Philosophy & Technology, 31(4), pp. 507–524.
20. Purificato I. (2021), Behind the scenes of Deliveroo's algorithm: the discriminatory effect of Frank's blindness, in Italian Labour Law e-Journal, 1(14), pp. 169-194.
21. Vydra S., Poama A., Giest S., Ingrams A., Klievink B. (2021), Big Data Ethics: A Life Cycle Perspective, in Erasmus Law Revue, 1, pp. 24-44 (doi: 10.5553/ELR.000190).

Being a statistician in the big data era: A controversial role?

Giancarlo Manzi

Abstract This paper illustrates some aspects of the controversial relations between statistical competence and trustfulness in results, mainly from the point of view of the statistical profession with respect to new professions in the Big Data era. Who is a statistician? How can she/he considered trustful? What is the relationship between being a statistician and, say, a data scientist? These are among the questions we try to answer in this paper.

Key words: Statistician, deontology, data analysis ethics standards, chartered professional

1 Introduction

In his first address to the Royal Statistical Society as Honorary Secretary, William Guy [2] illustrated for the first time what the proper function of a Statistical Society should be, gave the statistician a formal status listing the tasks she/he should accomplish, also suggesting the means to obtain them, and gave a comprehensive view of which areas the Society should be concerned with, including:

> *education, crime, industry, health, wealth, manufacture, commerce, special branches of industry and production.*

One century and a half later, with the advent of data science, this list cannot be fully given, as data analysis pervades now almost every field of human knowledge.

The above quote matches very well with another famous one by John Tukey about being a statistician [3]:

> *The best thing about being a statistician is that you get to play in everyone's backyard.*

Department of Economics, Management and Quantitative Methods and Data Science Research Center, University of Milan - Via Conservatorio 7, 20122 Milan, Italy, e-mail: giancarlo.manzi@unimi.it

This of course does not mean that the statistician's task is to only provide back-office support for every other field, but to improve its progress with the help of data analysis, and - again thinking of what data science is nowadays - that statistics should be a *front yard* discipline.

However, can we say today that statistics can be really considered a truly front line discipline? As it is used in many fields it is also *contaminated* by many fields. ”Improvised” statisticians are all around us and they are not arrested like fake physicians, psychologists, etc. are. Why is that? This is the question we want to address in this paper, i.e. what is the professional stakes that a data expert should have and, once assessed this, what possibilities and limitations she/he must have in relation to data ethics and the new context of the big data era.

The rest of this paper is organised as follows. Section 2 illustrates the state of the art of working with data with four examples in countries with different approaches on to the professional aspects of being a statistician. Section 3 focuses on the educational case where we will deal with the problem of how to teach students about the treatment of human subjects' data. This case is important because of the ”statistical reasoning” to be a key part of a ”healthy” data ethics training. Section 4 lists some of the most important controversies a statistician normally faces when acquiring, developing and analysing data especially with respect to policies implemented in western countries. Section 5 contains some considerations about some possible actions that can be taken to let the statistician ”work in peace” and the relations with other related professionals (including the data scientist).

2 The chartered statistician: state of the art and current rules

In 2011 The UN Committee for the Coordination of Statistical Activities [5] defined the ideal profile of ”international statisticians” as

> *international official statisticians being a professional group with its own distinct profile and human resources management needs*

and listing some activities the international statistician should perform, from compiling and harmonising data collected from secondary providers, usually from official sources, to developing indicators to facilitate inter-country comparison of complex issues, from supporting national statistical capacity building to managing, maintaining and disseminating statistical databases for analytical and monitoring purposes.

At national level the profession of statistician is defined by national statistical agencies and/or by national statistical societies, but differences are striking. In the following we briefly illustrate four examples of how the statistical profession is regulated in countries representing different ”statistical traditions”. In Italy there are professions organised in registers and professions non-organised registers. The statistician is mainly a self-regulated profession and there is a national statistical association, offering the possibility to be accredited as statistician both at national and European level through the Federation of European Statistical Associations. There was an *Esame di stato in discipline statistiche*, which was a formal recognition of statistical competency from the state, but it was suppressed in 2018, due to lack of applications. The Royal Statistical Society in the UK offers two routes to become Chartered Statistician (CStat), a standard route for which applicants should meet Society's standards for a Graduate Statistician regarding statistical knowledge and at least one of the three academic criteria with respect to the degree of the applicant. Among these academic requirements one can present also the Society's Graduate Diploma which is a thorough examination aiming at testing the statistical knowledge level of the applicant. In France the professional regulation for statisticians is dominated by the search for professional independence. Professional independence has always been recognized as essential to the credibility of official statistics and statistical trustfulness. It was established by a new law reforming economic competition in France and a surveillance organism (*l'Autorité de la statistique publique*) has been *ad hoc* created. The Pstat is a professional statistician status given by the American Statistical Association in the US which is awarded based on criteria covering the education, experience, and demonstrated competence of applicants. Thus, it is a portfolio-based rather than an examination based designation.

3 The education case

With a view to spreading statistical culture among young people and in schools, many countries are far behind. If we talk about data ethics, the matter gets even worse because in some countries it is completely absent. Statistics educators should have a responsibility to educate their students about the ethical aspects related to the collection of sensitive data. Once this is achieved the attitude of treating the data in the right way with respect to privacy and security should come naturally.

In the US the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report has paved the way for instructors to present introductory statistics to students in a way that it is both approachable and engaging. Raman et al. [4] hints that this report can be exploited to boost the awareness of being a statistician especially ethically, but still lacks a bit of practical implementation. They propose an action of consultation with ethicists that would help to frame curricular choices in light of the challenges statisticians face. Ethicists can best help not by proclaiming abstract moral principles nor by suggesting how to elicit students' compliance, but by helping statisticians and their students better understand what they can, may, and must do when studying and engaging in statistics.

Therefore, together with the standard subjects (probability, mathematical statistics, etc.) statistics student should be taught philosophy, especially ethics.

4 Controversies

Every statistician/data scientist has had at least one controversy in her/his career about data retrieval/data acquisition. Web scraping, for example, is still considered illegal in some countries. In the EU the GDPR law about data privacy and security has been an important milestone for the protection of individual privacy rights. The problem is the implementation and interpretation of the law which in many cases are adopted by governmental and university bodies in a very restrictive way, while in reality the law - if you read it well - allows many different methods for the possibilities of processing data for informational and progress purposes.

A figure who, for reasons related to his education and culture, is entrusted with the confidentiality of data - like a doctor is required to keep the data she/he acquires from the patient confidential - would be welcome. It is true that data-driven information acquisition, if not correctly conducted, can distort reality to extreme cases where collective opinion can be guided in an extremely dangerous way, so that it has a very dangerous epistemological impact on societies [1], but it is also true that protection about these extremes should not conduct towards other extremes in the direction of censoring the progress of knowledge.

5 Discussion with some proposals

The famous Ronald Fisher’s quote:

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of*

may be thought of as a comparison between the medical and the statistical profession. A statistician should be able to know *everything* about the purpose of a study, *before* implementing it. A patient admitted to an emergency room usually says everything about his/her pain/disease to the emergency doctor whatever sensitive topic this would imply.

”The data doctor” is an online 2018 British comedy about a digital guru (”The Data Doctor”) of Indian origins called in by a ”patient” (’Bash the entertainer’) as a data expert to provide the perfect gift for his girlfriend’s (Cheyenne) birthday. The data expert decides that a two-step research should be conducted: (i) social media analysis of behavioural attitudes to know the best areas from which Cheyenne is ”secretly attracted” and (ii) a spectacular gift presentation to increase satisfaction. (i) is performed with social network analysis ending up with the result that Cheyenne is attracted by wolves and wrestlers and likes belts; (ii) is performed through collecting data from interviews to professional wrestler, with the result of hiring a black wrestler for a wrestler-style presentation. Therefore the gift will be a wrestler champion belt with a wolf represented in it. The gift will be given to Cheyenne by a black wrestler.

The parody is based on mocking the figure of the data doctor with some stereotypes (he is Indian and therefore good with data and is in striking contrast with the wrestler from the point of view of the physical appearance, the wrestler is a black wrestler, etc.) and on the usefulness of data analysis. If this might represent the public thought on the importance of data there is plenty of room to be worried.

Unfortunately decades of denigration and ignorance on the statistics role in the societies, even with the striking efforts of statistical societies, still don't give the statistician the credit she/he deserves. The ”good contamination” (for example from computer science) of the profession still is not perfect, and the ”bad contaminations” are still there. Good contaminations should be encouraged: the collaboration from people from different background, especially the ones forming the ”wonderful three” (i.e. statistics, computer science and subjects where data science is mainly used) should be increased starting from dissemination work since primary school. The ”lies, damned lies and statistics” mantra has not helped in the past, reflecting the attitude of politicians with respect to statistics. Still statistics and data science are thought of as some mysterious disciplines. On the contrary, artificial intelligence is not, maybe because it can be considered under the umbrella of hard science (engineer, physics, etc.). Therefore, also statistics should be put under the same umbrella, whereas in some cases like in Italy is something which is conceived only at disposal of humanities subjects like economics. A strong work of dissemination is needed and the qualified statisticians/data scientist should be awarded the position it deserves among other professions, forgetting the rhetoric of the ”no-need to be a professional in this field because everyone can achieve it”. More free access to data (even sensitive data) with very few exceptions should be granted to statisticians/data scientists. Privacy and security should be assured by the professional status of being the real ”data doctor”.

References

1. Durante M. (2021) *Computational Power. The Impact of ICT on Law, Society and Knowledge*. Routledge: London-New York.
2. Guy, W.A. (1865) On the original and acquired meaning of the term ”Statistics”, and on the proper functions of a Statistical Society: also on the question whether there be a Science of Statistics; and, if so, what are its nature and objects, and what is its relation to political economy and ”Social Science”. J. Statist. Soc. Lond., 28, 478–493.
3. Leonhardt, D. (2000) John Tukey, 85, Statistician; Coined the Word ’Software’. *The New York Times*.
4. Raman, R., Utts, J., Cohen, A.I., Hayat, M.J. (2023) Integrating Ethics into the Guidelines for Assessment and Instruction in Statistics Education (GAISE). *The American Statistician*, DOI: 10.1080/00031305.2022.2156612.
5. United Nationds (2011) Defining the profile of international statisticians and recruitment policies. *UN Committee for the Coordination of Statistical Activities - 18th Session - Luxembourg, 7-9 September 2011*.

Forecasting relative humidity using LoRaWAN indicators and autoregressive moving average approaches

Previsione dell'umidità relativa utilizzando indicatori LoRaWAN e modelli con struttura autoregressiva a media mobile

Renata Rojas Guerra, Anna Vizziello and Paolo Gamba

Abstract This paper analyzes the contribution of the signal strength in LoRaWAN networks in forecasting the relative humidity (RH). To this aim, the received signal strength indicator (RSSI) values from eight LoRaWAN transmitter nodes are considered as regressors in the autoregressive integrated moving average (ARIMA), beta autoregressive moving average (βARMA), and Kumaraswamy autoregressive moving average (KARMA) models. These three approaches are compared to identify the best predictive model for the RH and verify their ability to capture the influence of the RSSI measurements. The results show that the ARIMA class presented the best performance, and the βARMA appeared as a competitive alternative. Moreover, it has been proved that the inclusion of the RSSI signals improved the RH forecasts in all fitted approaches.

Key words: double-bounded variables, received signal strength indicator, unit ARMA models

1 Introduction

Wireless sensor networks operating outdoors are exposed to fluctuations in weather conditions, which may influence the reliability of wireless channel transmission and the connectivity between the nodes [2, 5]. This relationship has been examined by

---

Renata Rojas Guerra
Universidade Federal de Santa Maria, Address of Institute, e-mail: renata.r.guerra@ufsm.br

Anna Vizziello
Department of Electrical, Computer and Biomedical Engineering, University of Pavia & CNIT, Research Unit of the University of Pavia, Pavia, Italy e-mail: anna.vizziello@unipv.it

Paolo Gamba
CNIT, Research Unit of the University of Pavia, Pavia, Italy e-mail: paolo.gamba@unipv.it

some authors, especially to verify the impact of temperature and humidity to forecast signal strength indicators [2] or investigate their correlation [9].

However, it is also useful to conduct research on the use of signal strength for weather estimation and prediction, since it could be of support when there are limitations in collecting information about some weather parameters.

The aim of this paper is to examine if the received signal strength indicator (RSSI) in LoRaWAN networks can contribute to forecasting the air relative humidity (RH), which is collected at different time stamps. The effect of the RSSI is computed by including regressors in three autoregressive moving average (ARMA) approaches. We consider the classical autoregressive integrated moving average [3] (ARIMA) models, which have proven essential to forecast a wide variety of variables. Thus, we compare its prediction capability with the unit ARMA models based on the beta [7] (βARMA) and the Kumaraswamy [1] (KARMA) distributions.

The βARMA and KARMA models were introduced to deal with time series that assume values in the standard unit or double-bounded intervals. They are alternatives to the ARIMA's Gaussian assumption since the beta and Kumaraswamy distributions can naturally accommodate asymmetries and heteroscedasticity. Recently, the forecast accuracy of these approaches has been compared to obtain an adequate model to analyze mortality rates [6], the percentage of energy stored in hydroelectric plants [8], and the monthly RH in different Brazilian cities [1]. Hence, the present work contributes to this literature by bringing the novelty of using the RSSI for RH forecasting and analyzing this variable with hourly frequency.

The rest of this work is structured as follows: In Section 2, the theoretical background of the dynamical time series models is described. It also gives the data source and a descriptive analysis of the variables of interest. Section 3 carries out the RH estimation and analyzes the contribution of the RSSI in forecasting accuracy. Finally, Section 4 presents the concluding remarks.

2 Material and Methods

The RH forecast is computed using autoregressive moving average approaches with and without the presence of regressors. In this section, we discuss the theoretical background related to the dynamical models employed in the study and describe the analyzed data set.

2.1 The ARIMA model

The ARIMA model is a combination of autoregressive (AR) and moving average (MA) components with the integration filter (I), enabling the modeling of the series of the three components or a subset of them. Let $y_1, \ldots, y_n$ be a time series of interest. The ARIMA(p,d,q) takes the following form

$$\nabla^d y_t = (1-B)^d y_t, \tag{1}$$

where ∇^d refers to the difference operator of order d, B is the backward shift operator. Finally, y_t is given by

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j r_{t-j}, \tag{2}$$

where $\alpha \in \mathbb{R}$ is a constant, $\phi_1, \ldots, \phi_p$ are the coefficients of the autoregressive terms with order p, $\theta_1, \ldots, \theta_q$ are the moving average coefficients with order q, and r_t is the white noise series that assumes Gaussian distribution with zero-mean and constant variance.

The methodological steps of the modeling process can be found in [3]. In this work, the ARIMA models are fitted using the forecast package implemented in the R programming language. From this package, we can fit a regression model with ARIMA errors to consider the effect of regressors in the analysis.

2.2 *The unit ARMA models*

The unit ARMA are dynamical models introduced as alternatives to the classic ARIMA models when the object of study is restricted to a double-bounded interval. Let $y_1, \ldots, y_n$ be a time series such that each $y_t \sim \mathrm{D}(\mu_t, \sigma_t)$, where $y_t \in (0,1)$, $t = 1, \ldots, n$, and $\mathrm{D}(\mu_t, \sigma_t)$ denotes a two-parameter probability density function (pdf) in which μ_t and φ_t are location and precision (or dispersion) parameters, respectively. The unit ARMA models have the following specification for the location parameter, conditionally to the previous information set up to $t-1$,

$$\eta_t = g(\mu_t) = \alpha + x_t^\top \beta + \sum_{i=1}^{p} \phi_i \left\{ g(y_{t-i}) - x_{t-i}^\top \beta \right\} + \sum_{j=1}^{q} \theta_j r_{t-j}, \tag{3}$$

where η_t is the linear predictor, $g(\cdot)$ is a link function such that $g : \mathbb{R} \to (0,1)$ is strictly monotonic and twice differentiable, $\alpha \in \mathbb{R}$ is a constant, x_t is the k-dimensional vector containing the covariates at time t, $\beta = (\beta_1, \ldots, \beta_k)^\top$ is the k-dimensional vector of parameters related to the covariates, $r_t = g(y_t) - g(\mu_t)$ is the error term, while $\phi = (\phi_1, \ldots, \phi_p)^\top$ and $\theta = (\theta_1, \ldots, \theta_q)^\top$ are the autorregressive and moving average coefficients, respectively. Thus, the unit ARMA models will differ on the assumption of the random component's distribution.

The βARMA model [7] is obtained considering that y_t is conditionally beta distributed, where μ is the conditional mean and φ is a precision parameter. The KARMA model [1] has the random component following the Kumaraswamy distribution and assumes that μ is the conditional median with φ being a dispersion parameter. In this work, we use the logit link function and the maximum likelihood method to estimate for both βARMA and KARMA models. The models are fitted from the implementations in the R programming language[1].

2.3 Data Source and Description

The data used in this study are taken from the available open-access literature reported by [4]. We are interested in the relative humidity (RH) from a weather station model Davis Vantage Pro2 and its relation with the received signal strength indicator (RSSI) values collected by eight LoRaWAN transmitter nodes (Tinovi PM-IO-5-SM), namely RSSI_01, ..., RSSI_08. The combined hourly dataset contains 2029 records with average hourly values from November 10, 2020, 23:20:00 to March 5, 2021, 05:00:00 GMT. However, the RSSI_08 has lower data counts, and the RH has eleven missing data corresponding to the period from February 2, 2021, 12:00:00 GMT, to February 2, 2021, 22:00:00 GMT. Thus, we perform the analysis considering 1721 records corresponding to 75 days over November 22, 2020, 15:20:00 to February 2, 2021, 11:00:00 GMT. Five RSSI transmitter nodes presented one missing value value in this final data set. We replace them with the mean of the corresponding node to perform the time series analysis.

3 Results and Discussion

We are interested in verifying if the RSSI signals contribute to forecasting RH. For this purpose, we use the frameworks described in Section 2 with and without using the RSSI signals as regressors. We refer to KARMAX, βARMAX, and ARIMAX for the models with the proposed regressors. The data set is divided into 86% ($n = 1474$) for the train set and 14% ($n = 247$) for the test set to assess the forecasting performance. The training set is used exclusively for model development. Then the test sample is used to evaluate the established model in one-step-ahead forecasts.

The final ARIMA models are obtained from the auto.arima function from the forecast R package, which allows for 1determining the best order parameters using the Akaike information criterion (AIC). A similar procedure is performed for the βARMA and KARMA models. For each class, we tested all models with orders $p \in \{1,2,3\}$ and $q \in \{1,2,3\}$ and selected those with the smaller AIC. Subsequently, the forecasting analysis was performed for the test set.

Figure 1 compares the original values and the forecasts of the fitted models. Overall, all models yielded accurate forecasts for the RH series. The KARMA model yielded the worse result, especially for hours with a fast decay in the RH values. It is the case, for instance, of the values observed on January 31. We also provide

---

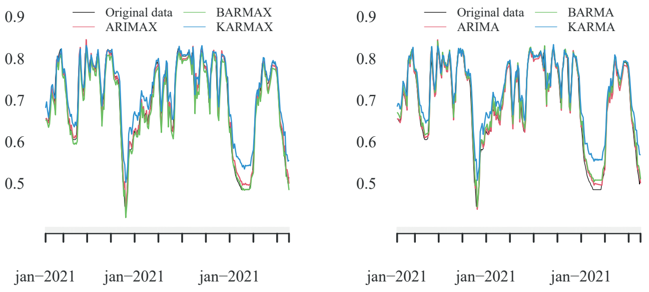
[1] See https://github.com/vscher/barm for βARMA implamentations, and https://github.com/fabiobayer/KARMA for those from the KARMA.

some accuracy measures to corroborate the insights provided by the visual inspection. The mean absolute percentage error (MAPE), mean absolute error (MAE), and root-mean-square error (RMSE) are the calculated measures. The smaller their values, the best is the forecasting performance. Table 1 presents the results for the selected models. We also calculate the percentage of increase in the measures concerning the ARIMAX and provide the order of each selected model.

Notice the inclusion of the RSSI signals did not change the order of the final model in each class. Also, the ARIMAX was superior in all measures, followed by the ARIMA without regressors. As observed in Figure 1, the KARMAX model did not perform well, presenting RMSE, MAE, and MAPE, respectively, 32,2%, 40,8%, and 45,6% higher than the ARIMA model. The βARMAX is more competitive, with an RMSE of 11.5% higher. The advantage of using the models based on the beta and Kumaraswamy distributions is that they are restricted to the range of possible values for the RH, i.e., the unit interval. However, the accuracy measurements did not stand out when compared with the ARIMA. They evince the usefulness of the RSSI signals in forecasting the RH. The inclusion of these indicators improved the forecasting results in all classes of models. We also conducted a preliminary experiment to verify the accuracy of the ARIMA and ARIMAX in replacing missing values. Both methods included RSSI observations at the time stamps of missing RH measurements. They performed better than just replacing the missing RH values with its mean for up to $k = 12$ missing observations, and the ARIMAX was superior for $k \in \{1,2\}$. Therefore, for the analyzed data set, the ARIMAX is the best option to predict RH values among the analyzed classes of dynamical models.

Table 1 Accuracy measures and the percentage difference with respect to the ARIMAX.

Model	Accuracy measures			Percentage difference		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
KARMAX$(2,3)$	0.033	0.026	4.154	37.200	40.800	45.600
BARMAX$(2,1)$	0.024	0.017	2.431	11.500	10.200	7.000
ARIMAX$(2,1,2)$	0.021	0.015	2.261	–	–	–
KARMA$(2,3)$	0.037	0.029	4.622	43.500	46.100	51.100
BARMA$(2,1)$	0.022	0.017	2.525	4.500	8.300	10.500
ARIMA$(2,1,2)$	0.021	0.016	2.277	0.200	0.800	0.700



(a) (b)

Fig. 1 Test and forecasted values for the RH. (a) Moddels using the RSSI values measurements as regressors. (b) Moddels without regressors.

4 Concluding remarks

We investigated the contributions of the RSSI in LoRaWAN networks to forecasting air RH. To this aim, we compare three different ARMA approaches that were fitted with and without the presence of regressors. They include the classical ARIMA model and unit ARMA models based on the beta and the Kumaraswamy distributions. Despite the theoretical suitability of the unit ARMA for double-bounded random variables, the ARIMA model presented better accuracy measurements, and the βARMA appeared as more competitive than the KARMA. Finally, in response to the question that motivated this study, all three approaches presented better results when including the RSSI values as regressors. Thus, we can conclude that the RSSI measurements can contribute to the RH estimation. These results can be helpful, for example, when there are limitations in collecting information about the RH or replacing missing observations in the observed time series.

References

1. Bayer, F.M., Bayer, D.M., Pumi, G.: Kumaraswamy autoregressive moving average models for double bounded environmental data. Journal of Hydrology **555**, 385–396 (2017)
2. Bhat, S.A., Huang, N.F., Hussain, I., Sajjad, U.: Correlating the ambient conditions and performance indicators of the LoRaWAN via surrogate Gaussian process based bidirectional LSTM stacked autoencoder showkat. IEEE Transactions on Network and Service Management pp. 1–1 (2023)
3. Box, G.E., Jenkins, G.M., Reinsel, G.C.: Time series analysis: forecasting and control. John Wiley & Sons (2011)
4. Goldoni, E., Savazzi, P., Favalli, L., Vizziello, A.: Correlation between weather and signal strength in LoRaWAN networks: An extensive dataset. Computer Networks **202**, 108,627 (2022)
5. Luomala, J., Hakala, I.: Effects of temperature and humidity on radio signal strength in outdoor wireless sensor networks. In: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1247–1255 (2015)
6. Melchior, C., Zanini, R.R., Guerra, R.R., Rockenbach, D.A.: Forecasting Brazilian mortality rates due to occupational accidents using autoregressive moving average approaches. International Journal of Forecasting **37**, 825–837 (2021)
7. Rocha, A.V., Cribari-Neto, F.: Beta autoregressive moving average models. TEST **18**(3), 529–545 (2009)
8. Scher, V.T., Cribari-Neto, F., Pumi, G., Bayer, F.M.: Goodness-of-fit tests for βARMA hydrological time series modeling. Environmetrics **31**, e2607 (2020)
9. Yi Lim, N.C., Yong, L., Su, H.T., Yu Hao Chai, A., Vithanawasam, C.K., Then, Y.L., Siang Tay, F.: Review of temperature and humidity impacts on RF signals. In: 2020 13th International UNIMAS Engineering Conference (EnCon), pp. 1–8 (2020)

Interpretability of Machine Learning algorithms: how these techniques can correctly guess the physical laws?

Marzio De Corato, Alfio Ferrara and Silvia Salini

Abstract The Machine learning algorithms (MLA) provide a formidable tool for making progress among different sciences [1]. Among them, remarkable results were obtained for physical sciences [2]; however, despite the high accuracy in predictions that can be obtained with these algorithms, using them for base scientific research also requires to have an interpretation of their machinery. Furthermore, it is worth mentioning that, apart from being a requirement for scientific purposes [2], interpretability is a requirement imposed on algorithms by the GDPR [3]. Moreover, as shown by Miller in [4], the interpretability of a MLA is strictly connected to finding the causal connection between the features analysed: therefore, if one is interested in going beyond the statistical correlation, he/she has to face how to make the MLA used interpretable [5]. While for some MLA, the interpretation is straightforward, for instance, in the case of linear regression, for others, like the neural networks and the support vector machines, such insight seems less evident. The interpretability issue was faced previously by a restricted set of authors ([3, 4, 6] and Ref. therein) with respect to the community that uses the MLA algorithm. In this study, we propose a systematic investigation of how a selected set of MLA algorithms can capture the generating laws for an input dataset. For this purpose, we started with datasets generated by a physical law or from real data (both taken from astronomy). While for the first case, the public datasets were considered, such as the NASA dataset of exoplanets [7] as well the hazardous asteroids [8], for the second case, the data were generated starting, for instance, from the gravitational law.

Marzio De Corato
Università degli Studi di Milano - Dipartimento di Informatica, Via Giovanni Celoria, 18, 20133 Milano MI, e-mail: `marzio.decorato@unimi.it`

Silvia Salini
Università degli Studi di Milano - Dipartimento di Economia, Management e Metodi Quantitativi, Via Conservatorio, 7 20122 MILANO (MI) e-mail: `silvia.salini@unimi.it`

Alfio Ferrara
Università degli Studi di Milano - Dipartimento di Informatica, Via Giovanni Celoria, 18, 20133 Milano MI, e-mail: `alfio.ferrara@unimi.it`

In this last case, other features were considered: in particular, these were generated with a different type of noise added to the correct input features. In the end, for these cases, we have datasets for which the underlying generating laws are known. Once prepared these datasets, an output variable was considered based on the known laws. After these steps, the following MLA algorithms were considered for the analysis: Neural networks (with different architectures), Support Vector Machines, Logistic Regression, Quadratic Discriminant Analysis, Random Forest [9], and graphical models [10]. After the mentioned algorithms were trained and tested, we considered the standard interpretation techniques [11] such as the Partial Dependence Plots, as implemented in the *iml* R package [12] to get an insight into the machinery of the algorithms considered. This outcome was compared with the prior knowledge about the generating law of the datasets. In this way, one obtains an assessment of the algorithms' accuracy and how well these approximate the underlying generating law. Given such validation on how the MLA correctly guess the physics of the input dataset, one can consider moving more safely on a real dataset in which the underlying laws are less known.

Key words: Algorithm intepretability, Machine Learning algorithms, ML algorithms for physics

References

1. Eric Mjolsness and Dennis DeCoste. Machine learning for science: state of the art and future prospects. *science*, 293(5537):2051–2055, 2001.
2. Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.
3. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
4. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
5. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
6. W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.
7. `https://exoplanetarchive.ipac.caltech.edu/applications/DocSet/index.html?doctree=/docs/docmenu.xml&startdoc=1`.
8. `https://cneos.jpl.nasa.gov/about/neo_groups.html`.
9. Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.
10. Jonas Haslbeck and Lourens J Waldorp. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.06871*, 2015.
11. Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
12. Christoph Molnar, Bernd Bischl, and Giuseppe Casalicchio. iml: An r package for interpretable machine learning. *JOSS*, 3(26):786, 2018.

The Role of BERT in Neural Network Sentiment Scoring for Time Series Forecast

Basili R., Croce D., Iezzi D.F., Monte R.

Abstract Sentiment scores measure the strength of customer sentiment when evaluating a product or service. This score is expressed as positive (and negative) for a numerical value between 0 and 100, where 100 is the most favourable possible result, and 0 is the least. This paper aims to combine a product's sales volume time series with the sentiment score time series of tweets generated by the BERT-NN within a state space model. We apply this model to the monthly sales volume of the Fiat L500 time series from August 2012 to Dec 2018.

Key words: BERT, ETS, Neural Network, Sentiment Scoring, state-space Model

1 Introduction

State-space models are structural models for fitting and predicting time series. Although naturally arising in multivariate contexts where the explanatory variables of a phenomenon are only partially or indirectly observable, state-space models also embody ARIMA models and can account for seasonal effect (see Durbin-Koopman

---

Basili Roberto
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: basili@info.uniroma2.it

Croce Danilo
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: croce@info.uniroma2.it

Iezzi Domenica Fioredistella
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: stella.iezzi@uniroma2.it

Monte Roberto
Department of Civil Engineering and Computer Science Engineering, Tor Vergata University, Rome, ITALY e-mail: roberto.monte@uniroma2.eu

(2012) [4] 1.2, 3.4, see also Harvey (1990) [5]). A simplified form of state-space models, the so-called ETS model, has been developed in the works of Holt (1957) [7], Winters (1960) [12], and Hyndman (2008) [8]. It is currently used in the most diverse industrial applications. The ETS model decomposes a time series into three components: a trend-cycle component, split in turn into a *level* component and a *slope* component, and a *seasonal* component. Furthermore, these components share a common innovation. The level, slope, and seasonal components are considered not observable. Therefore, the ETS model is essentially a state-space model with three non-observable state processes and the time series of interest as a single observation process.

Taking for granted the reasonably simple idea that consumer sentiment for a product influences the sales volume dynamics, the question arises of how to account for this sentiment. Referring to the ETS model, similar to Iezzi & Monte (2022) [9], we propose applying a state-space model in which we interpret one hidden component, the slope, as consumer sentiment for the product while observing a proxy of consumer sentiment, the Bert score. In contrast to the ETS approach, our goal is to use the information conveyed by the two signals to improve sales volume forecasting and, at the same time, obtain a consumer sentiment forecast. Therefore, we introduce a state-space model with three hidden state processes, one of which is consumer sentiment, and two observation processes, the time series of interest, in this paper, the Fiat 500L monthly sales volume.[1] and the Bert score proxy of consumer sentiment (see Yu et al. (2012) [13] for another approach). We use BERT in a neural network (Bert-NN) to build the proxy for consumer sentiment. Introduced by Devlin et al. in 2019 [3], BERT has rapidly become a highly regarded pre-trained neural model in the natural language processing community for its ability to tackle a wide range of language processing tasks. Its adoption by Google in 2020 further reinforced its status as a leading model in the field. BERT stands out for its bidirectional nature, which enables it to consider contextual information from both the previous and subsequent tokens in a given text. Combined with its unsupervised pre-training, BERT can effectively encode text data and generate high-quality representations that can be fine-tuned for various NLP tasks.

The remainder of the paper is organized as follows. Section 2 presents the predictive models. Section 3 discusses BERT and neural networks. In Section 4, we present the data and main results.

2 Our State Space Models

We use as a benchmark the ETS-AAA model introduced by Hyndman et al. (2008)[8]. This can be written as follows:

---

[1] See the website: https://www.carsitaly.net/fiat-car-sales_italy.htm

$$\begin{aligned} y_t &= \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \\ \ell_t &= \ell_{t-1} + b_{t-1} + \lambda\varepsilon_t, \\ b_t &= b_{t-1} + \beta\varepsilon_t, \\ s_t &= s_{t-m} + \gamma\varepsilon_t, \end{aligned}$$

where y_t is the value of the time series of interest at time t, the hidden variable ℓ_t [resp. b_t, resp. s_t] is the *level* [resp. *slope*, resp. *seasonality*] of the Holt-Winters decomposition of y_t, and the variable ε_t represents the innovation term at time t with variance σ^2. Parameter σ^2 is determined together with the parameters λ, β, γ, and the initial states of the model in the estimation procedure.
Our idea is to attribute to the slope variable b_t the role of consumer sentiment for the product (see also [9]). Of course, we cannot expect to perfectly observe such a variable, which is retained as a hidden variable, while we observe a proxy of it, which BERT-NN builds. As a consequence, in this paper, we study a state-space model where the state and observation equations take the following forms:

$$\text{state equations} \begin{cases} \ell_t = \beta_{\ell,\ell}\ell_{t-1} + \beta_{\ell,b}b_{t-1} + \sigma_{\ell,\ell}w_t^{(l)}, \\ b_t = \beta_{b,b}b_{t-1} + \sigma_{b.b}w_t^{(b)}, \\ s_t = s_{t-m} + \sigma_{s,s}w_t^{(s)}, \end{cases}$$

$$\text{observation equations} \begin{cases} y_t = \beta_{y,\ell}\ell_t + \beta_{y,b}b_t + \beta_{y,s}s_t + \sigma_{y,y}w_t^{(y)}, \\ z_t = \beta_{z,b}b_t + \sigma_{z,z}w_t^{(z)}. \end{cases}$$

Here, y_t is still the observed sales volume time series of the product, and z_t represents the observed BERT-NN score on the hidden sentiment variable b_t. To add more flexibility to our model compared to the ETS-AAA model, we introduce the additional parameters $\beta_{\ell,\ell}, \ldots, \beta_{z,b}$, Moreover, we introduce independent innovations $w_t^{(l)}, \ldots, w_t^{(z)}$, with variances $\sigma^2_{\ell,\ell} \ldots, \sigma^2_{z,z}$, respectively. All these parameters, together with the initial states of the models, are estimated in a recursive procedure using the functions of the *MARSS* [6] *R* library.

3 Transformed-based Sentiment Analysis

To capture the sentiment expressed in text for use in a state-space model, we treated the semantic processing problem as a classification task and utilized the BERT neural classifier to address it. This approach is based on the work of Devlin at al. (2019) [3], and Vaswani et al. (2017) [11]. BERT provides a sentence encoding model capable of producing contextualized lexical embeddings for individual words and an encoding vector for the whole sentence. This is achieved through a *pre-training* stage applied to millions of unlabeled texts, primarily based on acquiring an expressive and robust language and text model. By stacking a dedicated network, BERT can easily be adapted to various and diverse tasks through *fine-tuning*. Usually, a shallow multilayer perceptron is represented by a dense layer to optimize task-specific parameters.

In this work, we adopted a fine-tuning process for BERT dedicated to sentence classification, i.e., operating on a single text given as input. It enables customization of the final classifier to suit the specific problem and fine-tun all network parameters, including those of BERT, over just a few epochs. This prevents "catastrophic forgetting" of the linguistic information gained during pre-training. It is important to note that pre-training imposes no bias on the target language. The language model learned by BERT can be acquired regardless of the language used in the input texts used in the pre-training stage. This has led to the creation of multilingual language models, such as XLM-RoBERTa, as demonstrated in Conneau et al. (2020) [2]. Here, we fine-tuned such a multilingual architecture on the SENTIPOLC dataset in Basile et al. (2021) [1] that includes tweets in Italian annotated about subjectivity with polarity labels reflecting the writer's sentiment. In the tweet analysis, we generalize the simplified view that a tweet can be either positive or negative. Instead, we acknowledge that a text can express both positive and negative polarity. Hence, two classifiers are necessary to estimate the two independent probabilities $p(\texttt{pos}|t)$ and $p(\texttt{neg}|t)$ of a message expressing a positive and negative sentiment, respectively. This allows mapping the classification task into two binary classification tasks[2]. To fine-tune the targeted sentiment analysis task, we adopt the XLM-RoBERTa architecture to generate the final hidden vector $\vec{t} \in \mathbb{R}^h$ (with h the dimensionality of the embedding space) corresponding to the first input token (`[CLS]`) as the aggregate representation of each micro-post. Two different classifiers are applied, one for the `pos` class and one for the `neg` one. For each class, we stacked a classification layer with weights $\texttt{W}_{pos} \in \mathbb{R}^h$ and $\texttt{W}_{neg} \in \mathbb{R}^h$. The output probability $p(\texttt{pos}|t)$ is estimated by evaluating the sigmoid function, i.e., $sigmoid(\vec{t}\,\texttt{W}_{pos}^{\intercal})$ and $sigmoid(\vec{t}\,\texttt{W}_{neg}^{\intercal})$. The two binary cross-entropy losses $\mathcal{L}_{pos}$ and $\mathcal{L}_{neg}$ are evaluated against the annotated data. The final loss is combined in a multi-task fashion as in Liu (2019) [10], i.e., $\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$. After fine-tuning the architecture, each tweet is processed through a BERT-based neural network, which assigns both a probability that the text is positive and a probability that it is negative[3]. Time forecasting needs to account for subjectivity and polarity signals over time. A specific aggregation method was employed to translate the classification evidence from individual tweets into signals that reflect collective information for a given period. While other methods exist, the chosen strategy is simple and easily implementable. First, probability values are used to map tweets to classes through thresholds. All tweets having $p(\texttt{pos}|t) \geq \tau_{\texttt{pos}}$ are considered positive, while all tweets $p(\texttt{neg}|t) \geq \tau_{\texttt{neg}}$. To summarize the expression of sentiment in messages $t \in T^m$ that refer to a specific time (i.e., the month m), we define the probability $p(\texttt{pos}|T^m) \approx \frac{|\{t \in T^m | p(\texttt{pos}|t) \geq \tau_{\texttt{pos}}\}}{|T^m|}$. This corresponds to the

---

[2] A tweet is judged fully positive when associated with $p(\texttt{pos}|t) = 1$ and $p(\texttt{neg}|t) = 0$. A perfectly negative tweet is mapped to to $p(\texttt{pos}|t) = 0$ and $p(\texttt{neg}|t) = 1$. As a result, a neutral tweet corresponds to $p(\texttt{pos}|t) = 0$ and $p(\texttt{neg}|t) = 0$, while a contrastive tweet expresses both polarities with $p(\texttt{pos}|t) = 1$ and $p(\texttt{neg}|t) = 1$.

[3] For example, t = "*Fiat 500L, la "macchina" che fa anche il caffe' @anonymizedauthor!*" is considered by the classifier to be $p(\texttt{pos}|t) = 0.72$ and $p(\texttt{neg}|t) = 0.09$.

percentage of tweets observed during m classified as positive. The same holds for the negative polarity, i.e., $p(\mathtt{neg}|T^m) \approx \frac{|\{t \in T^m | p(\mathtt{neg}|t) \geq \tau_{\mathtt{neg}}\}}{|T^m|}$. The initial empirical evidence suggests a correlation between the target distribution and the negative polarity expressed in the messages. As a consequence, we fed the state space model with the distribution of $p(\mathtt{neg}|T^m)$ with a threshold value of $\tau_{neg} = 0.7$.

4 Main Results

To test our model, we consider the FIAT 500L monthly sales volume time series, from August 1, 2012, to December 31, 2018. We collected a corpus of 20,137 tweets for the same period and measured sentiment using BERT-NN. As benchmarks, we apply the ETS-AAA model to both the sales volume and the sales volume logarithm time series. The necessity of introducing the logarithmic time series is dictated by the different scales of the sales volume time series and the sentiment signal. Table 1 summarizes the results of our analysis. Fig. 1 [resp. 2] shows the last part (from Dec 2016 to December 2018) of the fitted and predicted results from log-scaling the ETS-AAA model [resp. from the BERT-NN model] for FIAT 500L sales volume [resp. sales volume logarithm].

Models	**logLik**	**AIC**	**BIC**	**AICc**	**RMSE**	**MAE**	**MAPE**	**SMAPE %**	**MASE**
ETS-AAA data	-606.879	1247.759	1285.738	1259.759	867.621	706.553	48.772	18.469	0.698
ETS-AAA log-data	-68.734	171.469	209.448	183.469	0.597	0.469	6.512	3.110	1.440
BERT-NN (y) log-data	7.442	9.115	35.924	11.611	0.287	0.207	2.797	1.376	0.635
BERT-NN (z) log-data	7.442	9.115	35.924	11.611	0.033	0.025	35.631	19.064	0.651
diffuse initial state									
BERT-NN (y) log-data	84.288	-144.575	-117.766	-142.079	0.423	0.318	4.298	2.145	0.977
BERT-NN (z) log-data	84.288	-144.575	-117.766	-142.079	0.037	0.028	35.922	21.680	0.718
random initial state									

Table 1 Validation measures - ETS - AAA and BERT-NN for Fiat 500L

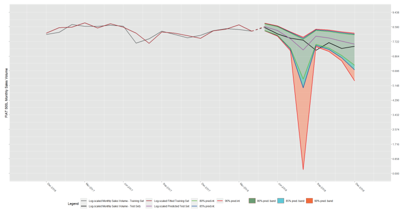


Fig. 1 Log-scaling of ETS-AAA data

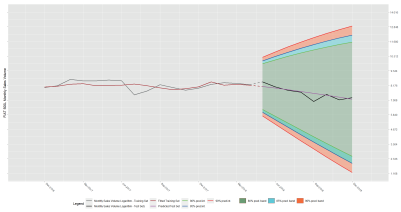


Fig. 2 BERT-NN (y) log-data

While the ETS-AAA for the original time series outperforms the ETS-AAA for the logarithmic time series, the BERT-NN with a diffuse initial state seems to outperform both models. However, the training set fitting of the latter model presents non-optimal residuals in terms of serial correlation because it seems unable to filter the seasonal component of the time series. On the contrary, despite the higher MASE, the BERT-NN model with a randomly determined initial state performs better overall. Moreover, evaluating the performance of the BERT-NN state-space model, we should consider that we obtain not only the fit and forecast of the time series of interest but also the fit and forecast of the consumer sentiment time series. The latter might be an interesting piece of information to advise the management for a better-tailored advertising campaign.

Given the above results, we believe the state-space model approach is promising and worthy of further studies.

References

1. V. Basile, N. Novielli, D. Croce, F. Barbieri, M. Nissim, and V. Patti. Sentiment polarity classification at EVALITA: lessons learned and open challenges. *IEEE Trans. Affect. Comput.*, 12(2):466–478, 2021.
2. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451, Online, July 2020.
3. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
4. J. Durbin and S.J. Koopman. *Time series analysis by state space methods, 2nd ed.* Oxford University Press, Oxford; New York, 2012.
5. A.C. Harvey. *Forecasting, structural time series models, and the Kalman filter*. Cambridge Univ. Press, Cambridge u.a., 1990.
6. E.E. Holmes, J. Eric, E.J. Ward, Scheuerell M.D., and K. Wills. *MARSS: Multivariate Autoregressive State-Space Modeling*, 2021. R package version 3.11.4.
7. C.C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.
8. R.J. Hyndman, A.B. Koehler, J.K. Ord, and R.D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach.* Springer Series in Statistics. Springer Berlin Heidelberg, 2008.
9. D.F. Iezzi and R. Monte. Sales forecast and electronic word of mouth: the power of feelings. In *Proceeding of the 16th International Conference on Statistical Analysis of Textual Data*, pages 489–494, Naples, Italy, July 2022. Valdistat press in coedizioni Erranti.
10. X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L . Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
12. P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.
13. X. Yu, Y. Liu, X. Huang, and A. An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734, 2012.

Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment

Andrea Sciandra, Matilde Trevisani and Arjuna Tuzzi

Abstract Diagnostics is a crucial component of any topic modelling application. However, available measures seldom offer indisputable and consistent solutions. We analyse the score distribution of a large set of intrinsic measures by varying two model inputs: text length and topic number. The first aim is to identify an ideal text length (or range of) by exploring per-length diagnostic distributions over the topic number. The second aim, once the optimal text length has been set, is to select the best model (or candidates) by comparing different specifications that include document metadata. We will also detect any conflict or ambivalence in the solutions produced by the different diagnostics.

Key words: diagnostic measures, topic modelling, structural topic modelling, model selection

1 Introduction

Diagnostic measures are a crucial component of any topic modelling application. This is because several decisions need to be made before estimating the final model for meaningful results to be obtained. However, identifying and justifying these choices is a challenging journey, and available measures seldom offer indisputable and consistent solutions. In general, selecting an appropriate topic model (TM) involves a variety of trade-offs and judgments by the human researcher. In this study,

---

Andrea Sciandra
Dipartimento FISPPA, Università di Padova, e-mail: andrea.sciandra@unipd.it

Matilde Trevisani
DEAMS “Bruno de Finetti”, Università di Trieste, e-mail: matilde.trevisani@deams.units.it

Arjuna Tuzzi
Dipartimento FISPPA, Università di Padova, e-mail: arjuna.tuzzi@unipd.it

we discuss the role of diagnostic measures towards selecting the most appropriate topic structure of a diachronic corpus.

Using a TM as an unsupervised tool involves focusing on how the learned topics align with human evaluations and help differentiate between aspects of a discourse. Until recently, the evaluation of such models has been ad hoc and application-specific, ranging from a fully automated intrinsic approach to a manually crafted extrinsic approach. Intrinsic evaluation, based on statistical measures, can be problematic because the measures do not account for domain relevance. Meanwhile, extrinsic evaluations are hand-constructed and often costly to perform for domain-specific topics. In any case, the real-world deployment of topic models requires time-consuming expert verification and model refinement to gain semantically meaningful topics within the domain of analysis.

Because of the ability of intrinsic measures to standardise, automate and scale the evaluation of TMs, the analyst generally picks one or more diagnostics of this type to be guided in the landscape of possibilities from which to choose the best model. Two broad classes can be envisaged: diagnostics that measure the predictive accuracy of the model (of which perplexity and marginal probability are the most well-known and widely applied) and diagnostics that assess the quality of topics (of which semantic coherence and Kullbach-Leibler (KL) topic divergence are among the most frequent instances, although the diversity of metrics is greater in this class). Moreover, any diagnostic can be variously implemented. Importantly, assessing a TM based on its predictive ability generally involves choices that are misleading, if not conflicting, in their judgements based on the quality of topics.

In this paper, we analyse the score distribution of a large set of intrinsic measures by varying two model inputs: text (or text partition) length and topic number. While topic number selection has been extensively studied, the impact of text length on a model's performance has been rarely addressed. In particular, some studies focus on the relationship between text length and topic number [8] or propose a document partition to improve model estimation [5]. In this study, we widen the research scope by first seeking to identify an ideal text length (or range) that optimises the selected diagnostics by exploring their distribution over the topic number. Once we have determined the optimal document chunking, we will analyse what the same diagnostics suggest for model selection by comparing alternative specifications that include document metadata. We will also detect any conflict or ambivalence in the solutions produced by the different diagnostics.

Within the panorama of topic modelling, we chose for our simulation study the structural topic model (STM, [7]) because it is a natural extension of the most famous and widely used TM, i.e. the Latent Dirichlet Allocation (LDA): it allows for correlation both between topics and between the topics and document-level covariates. At present, STM is very popular among topic modelling practitioners compared to other LDA generalisations. Moreover, it suits our application in which the effect of corpus metadata on topic determination is of interest.

Section 2 summarises the STM and introduces the data. Section 3 presents the simulation plan and selected diagnostics. Section 4 shows the results of a preliminary pilot study carried out on reduced dataset and simulation scope.

2 Model and data

STM is an unsupervised method for identifying the topical structure of a collection of texts. The method incorporates observable metadata information (i.e. covariates at the text level) to capture their effects on topics. STM can be conceptually divided into three components. The first is a topic prevalence model, which controls how words are allocated to topics as a function of covariates:

$$\gamma_\kappa \sim Normal_p(0, \sigma_k^2 I_p), \quad for\ k = 1 \ldots K-1, \tag{1}$$

$$\theta_d \sim LogisticNormal_{K-1}(\Gamma' x'_d, \Sigma), \tag{2}$$

where Γ is the matrix of coefficients for the topic prevalence model specified by Equations (1) and (2), d stands for document, k is the number of topics, X is the matrix for topic prevalence, and σ is a k-dimensional hyper-parameter vector; The second is a topical content model that controls the frequency of the terms in each topic as a function of covariates:

$$\beta_{d,k,v} = \frac{exp(m_\nu + k^{(t)}_{k,\nu} + k^{(c)}_{y_d,\nu} + k^{(i)}_{y_d,k,\nu})}{\sum_\nu exp(m_\nu + k^{(t)}_{k,\nu} + k^{(c)}_{y_d,\nu} + k^{(i)}_{y_d,k,\nu})} \quad for\ \nu = 1 \ldots V\ and\ k = 1 \ldots K, \tag{3}$$

where $k^{(t)}_{k,\nu} + k^{(c)}_{y_d,\nu} + k^{(i)}_{y_d,k,\nu}$ is a collection of coefficients for the topical content model, and m_ν is the marginal log-transformed rate of term ν; The third is a core observation model that combines models (1), (2), and (3)

$$\mathbf{z}_{d,n} \sim Multinomial_k(\theta_d), \quad for\ n = 1 \ldots N_d, \tag{4}$$

$$\mathbf{w}_{d,n} \sim Multinomial_V(\mathbf{B}\ \mathbf{z}_{d,n}), \quad for\ n = 1 \ldots N_d, \tag{5}$$

The core observation model allows for correlations in the topic proportions using the logistic normal distribution. The topic prevalence (which describes the association of a document with a topic) and the topical content (which describes how the words are used within a topic) components enable the expected text-topic proportion and, respectively, the topic-word probability to vary as a function of the observed text-level covariates X rather than arising from global parameters shared by all texts.

The diachronic corpus under scrutiny is a collection of all end-of-year addresses of the Italian Presidents of the Republic. Time (i.e. years of the speeches) and President (i.e. speakers) are the covariates employed to test their effect on the STM components. The corpus includes 73 addresses (1949-2021) delivered by 11 presidents. The corpus is available in both original and lemmatised versions and is continuously updated through the collation of digitalised text with audio-visual recordings. Word selection relies on part of speech (POS) information and frequency.

3 Simulation plan and selected diagnostics

We consider and discuss two main problems: 1) determining the length of the texts under scrutiny and the opportunity to work with equal-sized chunks; and 2) choosing the best model (which involves topic number selection) by comparing alternative specifications (i.e. metadata). Concerning the first, although speech length varies considerably, the presidential addresses generally represent medium-length texts (i.e. longer than a social media post, shorter than a novel). This raises the following questions: would length standardisation improve model performance? Is there a text length more appropriate to topic modelling? Given these questions, we compare per-length diagnostic score distributions over the topic number for the original setting (consisting of the whole documents) and for standardised settings (obtained by chunking the original speeches into equal text fragments). Given that we chose to first select content words (by POS tagging and frequency threshold) and then split the documents, we decided to make chunks constituted of 10 to 100 words that increment by a 10-word step. The selected content words cover roughly 50 percent of a sentence; a chunk of 10 words corresponds to a short-medium 20-word sentence, and a chunk of 100 corresponds to a 200-word text that is close to the shortest original documents. Lastly, related to the first aim, we discuss the assumption that each text is necessarily multi-topic, also in relation with the length of the text itself. As for the second objective, once we have applied the best chunking option (including no chunking), we will compare diagnostic score distributions over the topic number to select the best model from the different combinations of both covariates and model specifications (i.e. year and/or President included as prevalence and/or content model).

We chose the most widely used intrinsic measures for TM evaluation to compose the set of diagnostics under investigation. Within the class of diagnostics for addressing a predictive task, we calculate the following: (p1) held-out log-likelihood, (p2) residuals, (p3) perplexity and (p4) model posterior probability. The first two are provided by the `stm` R package and correspond to the (p1) log-likelihood of the held-out set of words, according to the document completion method [9] and the (p2) multinomial dispersion of model residuals (i.e. when the model is correctly specified, the multinomial likelihood implies a residual dispersion $\sigma^2 = 1$). Perplexity (p3) is the most well-known diagnostic in topic modelling and is defined as the inverse geometric mean of the per-word probability of a held-out set of words. (p4) is the model posterior probability under a Bayesian estimation approach [4]. The class of topic quality diagnostics includes countless examples. However, a ‘red thread’ that allows a synthetic interpretation can be borrowed from the literature of psychology (i.e. self-definition) and organisation theory (‘category’ definition), which explains that a well-defined topic requires the co-presence of distinctiveness, coherence and continuity. A non-exhaustive list of quality measures contains the following: (q1) semantic coherence, (q2) exclusiveness, (q3) consistency and differentiation (CD) scores, (q4) between-topic (cosine) similarity [6], (q5) symmetric Kullbach-Leibler divergence between the singular value distribution of the topic-word matrix and the row L_1 norm of document-topic matrix [1], (q6) the Jensen-

Shannon divergence between all pairs of topics [3] and (q7) distinctiveness whence saliency [2]. (q1) and (q2) are provided by the `stm` package, though only for those specifications without content covariates (to overcome the limitations of existing packages, we developed ad hoc R procedures for each measure). We propose using (q3) to synthetise the trade-off between (q1) and (q2) by calculating the L_2 norm of the two min-max normalised measures to pinpoint the top-right region of the plane generated by the two metrics.

In this work, we first focus on the p1–p2 and q1–q3 measures. The idea is to extend the empirical part to the other diagnostics presented above. We repeat the estimation of each configuration (length, topic number, model) 30 times to control the variability associated with model initialisation (i.e. different random seed values). The topic numbers range from three to 50.

4 A pilot study

For our pilot study, we chose to work with sole nouns that occur at least 10 times (612 lemmas). Figure 1 shows the first two predictive (p1–p2; top left and middle panels) and topic quality diagnostics (q1–q2 on the same plane; synthesis q3, the CD score; in bottom left and middle panels) required for text length selection. Both the p1–p2 and q3 measures clearly favoured longer chunks, so we chose a chunk size of 50 nouns. This represented the longest chunk possible and appeared to exalt interpretability while maintaining both the highest levels of held-out log-likelihood and the lowest levels of residuals. The CD score provided two more insights: (1) equal-sized chunks ensure better results than whole documents, and (2) each CD curve reaches its optimum level around k within 10 to 15 topics. This indicates that chunking does indeed produce a form of standardisation. A topic number that is slightly higher than the number of presidents may suggest a combination of factors, such as the distinctive influence of the presidents' personal traits and the evolution of socio-political historical facts. Once the ideal chunk length was fixed, the next step was to determine the optimal model specification. We experimented with different settings to test the effect of the Year (smoothed with splines) and/or President covariates on the topic prevalence (θ) and/or the topical content (β) components. These settings were as follows: (i) no covariates for neither θ nor β; (ii) Year for θ and no covariate for β; (iii) President for θ and no covariate for β; (iv) Year for θ and President for β; (v) no covariate for θ and President for β; (vi) President on θ and Year on β; (vii) no covariate for θ and Year for β. Models that included the President factor in the topical content were found to perform far better (Figure 1, top-right). Given a topic, a different lexicon characterised the Presidents. The ideal topic number was determined at slightly above 10 or 20 across all models. By picking the best or second-best topic number, most models (except for the best couple with President in β) performed similarly on the predictive side (bottom-right).

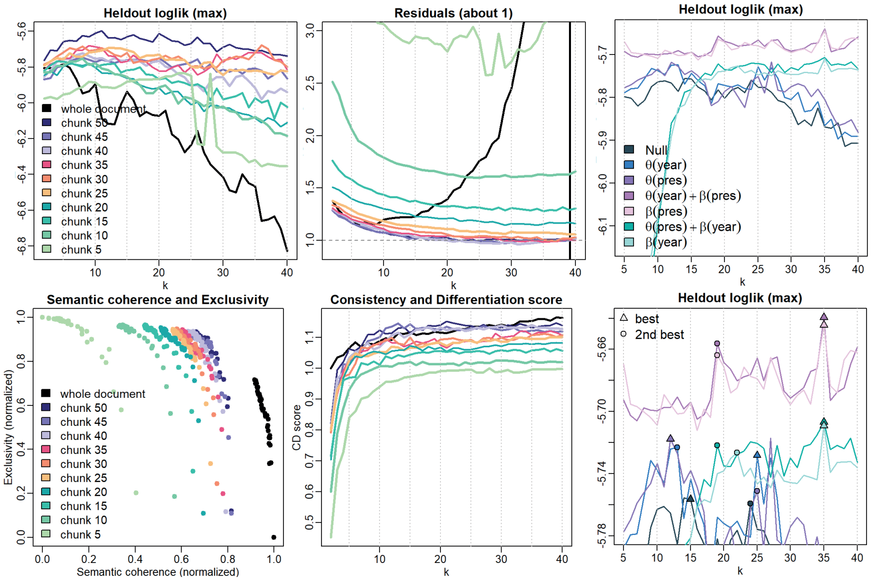


Fig. 1 (Left and middle panels) Per-length predictive (p1-p2; top) and topic quality (q1-q3; bottom) diagnostic distributions for choosing the best text length; (right panels) predictive distributions for choosing the best model structure (p1, top; p2, bottom)

References

1. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In Zaki M. J., Yu J. X., Ravindran B., Pudi V. (eds) Advances in knowledge discovery and data mining, 391-402 (2010)
2. Chuang, J., Manning, C., Heer, J.: Termite: visualization techniques for assessing textual topic models. Proceedings of International Working Conference on Advanced Visual Interfaces (AVI 2012), 74-77 (2012)
3. Deveaud, R., SanJuan, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. Document numerique 17(1), 61-84 (2014)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235 (2004)
5. Guo, C., Lu, M., Wei, W.: An improved LDA topic modeling method based on partition for medium and long texts. Annals of Data Science 8(2), 331-344 (2021)
6. Juan, C., Tian, X., Jintao, L., Yongdong, Z., Sheng, T.: A density-based method for adaptive lda model selection. Neurocomputing - 16th European Symposium on Artificial Neural Networks 2008 72(7-9), 1775-1781 (2009)
7. Roberts, M.E., Steward, B.M., Airoldi, E.M.: A Model of Text for Experimentation in the Social Sciences. J of the American Statistical Association 111(515), 988-1003 (2016)
8. Sbalchiero, S., Eder, M.: Topic modeling, long texts and the best number of topics. Some problems and solutions. Quality and Quantity 54(4), 1095-1108 (2020)
9. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimmo, D.: Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09), pp 1105-1112 (2009)

Mapping the thematic structure of Data Science literature with an embedding strategy

Antonio Irpino, Michelangelo Misuraca and Giuseppe Giordano

Abstract The term Data Science covers a family of methods aiming at extracting useful information from large amounts of complex data for decision-making purposes. However, there is a lack of a shared definition, and the diverse communities of scholars and practitioners involved in this research field influence the different conceptualisations. To overview the different methodological approaches and application domains of Data Science and try to depict its thematic structure, we propose a scoping review of the last 10-year reference literature. In the framework of science mapping analyses, we employed a topic detection strategy relying on word embedding to define and describe the how and the what of Data Science, contributing to the ongoing debate about the nature and the peculiarities of this research field.

Key words: Science Mapping, Topic detection, BERT model

1 Introduction

The evolutionary process of Data Analysis began a long time ago, paralleling the development of ICT and the growing availability of large data volumes as well as new types of data in any domain of interest. At the beginning of the 1970s, J.P. Benzécri – exposing the basic principles of Data Analysis from a statistical perspective – stated the importance of computers in the analysis of complex phenomena and preached the necessity of considering the greater number of dimensions to provide a suitable representation of these phenomena [1]. This vision, in some ways

---

Antonio Irpino
University of Campania L. Vanvitelli, Caserta (Italy), e-mail: antonio.irpino@unicampania.it

Michelangelo Misuraca
University of Calabria, Rende (Italy) e-mail: michelangelo.misuraca@unical.it

Giuseppe Giordano
University of Salerno, Fisciano (Italy) e-mail: giuseppe.giordano@unisa.it

ahead of his time, had its roots in the new conception of Statistics as an exploratory science originally proposed by J.W. Tukey [19], and anticipated highly topical issues such as the use of big data. Over the years, communities of scientists from diverse backgrounds presented their views of Data Analysis, somewhat with convergent positions. About thirty years ago, whereas C.R. Rao defined Statistics as the methodology for extracting information from data and expressing the amount of uncertainty in decisions people make [17], G. Piatetsky-Shapiro introduced the term *Knowledge Discovery in Databases* (KDD) [15] to denote the process of extracting implicit, previously unknown, and potentially useful information from data, giving a well-defined goal to Data Mining techniques. Between the 1990s and the 2000s, there was a close competition between statisticians and computer scientists, with the proposal of methodologies and practical solutions that often overlapped and differentiated just for the formalization and the terminology used by the two distinct scientific communities. The advent of the new century has seen the term *Data Science* increasingly employed as a replacement for Data Analysis. According to some authors, Data Science has to be seen as a new generation of Statistics [4]. In contrast, other authors argued that Data Science is a consolidation of several interdisciplinary fields [7, 2] or even a new body of knowledge [5].

Due to the difficulty of providing an unambiguous and shared definition of Data Science and identifying the conceptual, methodological and applicative characteristics of the domain, in this paper, we aim to contribute to the ongoing debate by performing a scoping review on the primary methodological approaches as well as the application domains. For this purpose, we analysed the abstracts of publications concerning Data Science appearing in scientific journals from 2013 to 2022 in order to map the thematic structure of the last 10-year reference literature. To accomplish this task, we implemented a topic modelling strategy based on *word embedding* [8], an algebraic representation of texts that captures the semantic structure and considers the context surrounding each word. The use of *BERTopic* [9] allows leveraging the embedding logic in the topic detection procedure, creating interpretable themes related to the different methods of Data Science and their practical uses and keeping the most important words employed in the themes' description.

The paper is organised as follows. Section 2 introduces the research methodology and the data selection method. Section 3 outlines the preliminary results, provides some remarks and leads the way for the future developments of the research.

2 Methodology and data selection

Documents written in natural language incorporate the information in a form that is difficult to analyse from a quantitative point of view because the textual content is composed of unstructured data. The most used algebraic model for representing textual content as structured data is the so-called *vector space model* (VSM) [18], in which a document or a text chunk is transformed in a p-dimensional vector (where p is the number of words belonging to the text, i.e. its vocabulary). The encoding scheme beneath this model – known as *bag of words* (BoW) – treats texts as multi-sets of the words that compose them, retaining their multiplicity (in terms of word occurrences) but disregarding grammatical and syntactic roles. The adoption of BoW simplifies the computational treatment of extensive collections, but at the same time, it limits the possibility of considering the context of the use of the different words. Resorting to words' co-occurrences allows dealing with this drawback since the context may be (partially) recovered by the relational structure of words in text chunks, with a level of detail that can go from the entire document to its single component clauses. An alternative to VSM is the so-called *word embedding* [8], whose primary goal is to map words to a vector of real numbers in a high-dimensional space. Differently from BoW, word embedding captures both the semantic and syntactic meaning of words.

It is possible to distinguish several types of word embedding, but they can generally be categorised into three groups:

- *prediction-based approaches*, which predict a word given its context or a context given a word, like *Word2Vec* [12];
- *hybrid-based approaches*, which predict a word taking into account its co-occurrences with the other words used in the collection, like the *Global Vectors for Word Representation* (GloVe) [13];
- *transformer-based approaches*, which use particular neural networks known as *transformers* [20] to generate contextualised embeddings for words, like the *Bidirectional Encoder Representations from Transformers* (BERT) [6].

Basically, the architecture of the different embedding models is based on neural networks, with only one or two hidden layers for prediction-based and hybrid-based approaches (i.e., with a *shallow learning* perspective) and several hidden layers for transformer-based approaches (i.e., with a *deep learning* perspective). The main advantage of the latter is that it is possible to produce a different embedding for each word depending on the given context, whereas other approaches produce the same embedding even with a different context. Another advantage of transformer-based approaches is that it is possible to employ pre-trained models tuned for a specific purpose. Starting from the original version of BERT, several models for specific kinds of texts (e.g., [10]) or for given languages (e.g., [16]) have been developed and applied to cope with several Text Mining tasks.

Aiming to perform a topic detection procedure on Data Science literature, we decided to employ the BERTopic model here. This approach relies on a three-fold strategy for determining which are the topics addressed in a collection of texts and describing them by means of the most relevant words used in the texts themselves. In the initial step, a pre-trained language model is used to obtain document-level information, capturing the contexts of the documents. Subsequently, a dimensionality reduction is performed on the embedding space before creating semantically similar document clusters for representing distinct topics. Finally, an adapted TF-IDF weighting scheme is used to select the most relevant words in each cluster-topic and obtain, in this way, a topic description. The separate steps used in BERTopic allow using different alternative solutions. In this early stage of our study, we decided to employ – as in the original configuration of the analysis – the *uniform manifold approximation and projection* (UMAP) [11] to reduce dimensionality and the *hierarchical density-based clustering* (HDBSCAN) [3] to find topics.

3 Preliminary results, remarks and future developments

The collection of scientific publications concerning Data Science has been obtained from the *WOS* (Web of Science) database, searching for the records containing the term Data Science as an author keyword. Up to March 2023 10th, we obtained 3,876 entries, but after applying some filters on publication type (→ Article, or Proceeding Paper, or Review Article, or Book Chapters), publication stage (→ final), publication language (→ English), and publication years (→ 2013–2023), we reduced the set to 3,680 entries. Removing all the records without an abstract, we obtained a final dataset of 3,460 entries. To bring out the thematic structure of Data Science literature, we employed the BERTopic model on the abstract field of the set.

In Fig. 1, we show the yearly production of papers according to the publication type. As we can see, there was an increasing production of contribution about Data Science, with a wider share of articles.

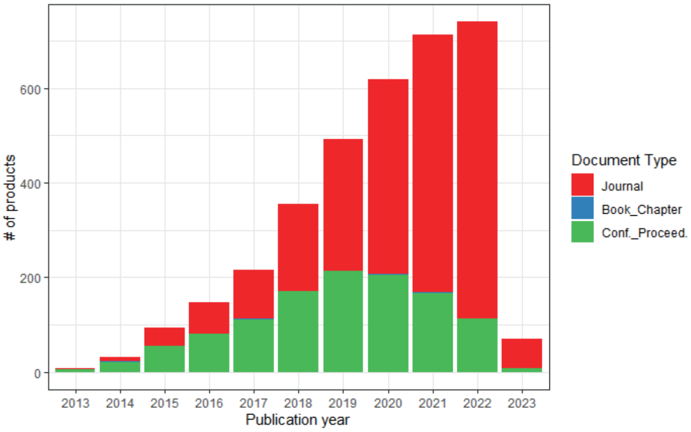


Fig. 1 Documents type production over the last 11 years. Note that 2023 is truncated up to March.

For the sake of interpretability, we set the model to produce 1 and 2-grams. After removing the stopwords, the model provided 57 topics. After performing the UMAP and the HDBSCAN step we obtained the topic map shown in Fig. 2

In the following, we reported the main interesting findings:

- topics related to the education field arise on the right part of the map and show a strong interest among scholars in how to educate new generations about data science;

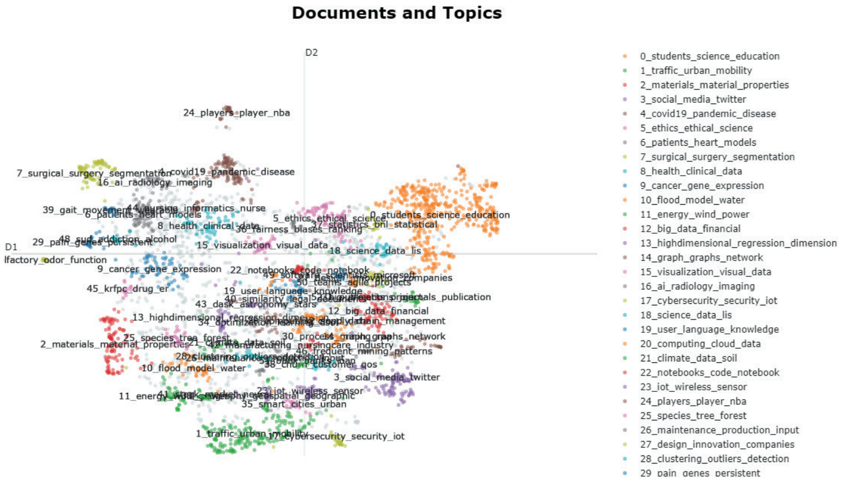


Fig. 2 Map of the documents and the topics generated by the BERTopic algorithm.

- topics related to the health-care of data science techniques (left and left top part of the map) shows a variety of interests;
- at the center of the map, topics related to the computer science and statistical aspects of data science arise;
- on the bottom topics related to the environmental and energy application arise;
- application of data science to sport (the little cluster at the top of the map) and cybersecurity issues (the little cluster at the bottom) shows some niche topics.

The beforementioned findings are just a very brief list of results. Other aspects will be shown during the presentation and are eligible for further interpretation.

From a methodology point of view, a deep investigation of the method calls for further reflections and potential improvements. Some issues are:

- an abstract of a document is classified as belonging to a single topic, and this can be too restrictive;
- topics in research may have a rapid evolution, an analysis using dynamic models could return some interesting patterns;
- papers are classified by the authors through the keyword. Furthermore, journals belong to some peculiar research subject categories. It could be interesting to study to what degree the topics of a paper are consistent with the keywords or the subject categories.

References

1. Benzécri, J.P.: L'analyse des données. Dunod, Paris (1973)
2. Blei, D.M., Smyth, P.: Science and data science. Proc. Natl. Acad. Sci. U.S.A. **114**, 8689–8692 (2017)
3. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining, pp.160-172. Springer, Berlin (2013)
4. Cleveland, S.: Data Science — An Action Plan for Expanding the Technical Areas of the Field of Statistics. ISI Rev. **69**, 21–26 (2001)
5. Concolato, C.E., Chen, L.M.: Data Science — A New Paradigm in the Age of Big-Data Science and Analytics. New Math. Nat. Comput. **13**, 119–143 (2017)
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT — Pre-training of deep bidirectional transformers for language under- standing. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186. ACL, Minneapolis (2019)
7. Hayashi, C.: What is Data Science? Fundamental Concepts and a Heuristic Example. In: Hayashi, C., Yajima, K., Bock, H.H., Ohsumi, N., Tanaka, Y., Baba, Y. (eds) Data Science, Classification and Related Methods, pp. 40-51. Springer, Tokyo (1998)
8. Hinton, G.E.: Learning distributed representations of concepts. In: Morris, R.G.M. (ed.), Parallel distributed processing — Implications for psychology and neurobiology, pp. 46-61. Clarendon Press, London (1989)
9. Grootendorst, M.: BERTopic — Neural topic modeling with a class-based TF-IDF procedure. arXiv (2022) doi: 10.48550/ARXIV.2203.05794
10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT — a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, **36**, 1234–1240 (2019)
11. McInnes, L., Healy, J., Melville, J.: UMAP — Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv (2018) doi: 10.48550/ARXIV.1802.03426
12. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting Similarities among Languages for Machine Translation. arXiv (2013) doi: 10.48550/ARXIV.1309.4168
13. Pennington, J., Socher, R., Manning, C. D.: Glove — Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543. ACL, Doha (2014)
14. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv (2018) doi: 10.48550/ARXIV.1802.05365
15. Piatetsky-Shapiro, G.: Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Mag. **11**, 68–70 (1991)
16. Polignano, M., Basile, V., Basile, P., de Gemmis, M., Semeraro, G.: AlBERTo — Modeling Italian Social Media Language with BERT. Ital. J. Comput. Linguist., **5**, 11–31 (2019)
17. Rao, C.R.: Statistics and Truth — Putting Chance to Work. International Co-operative Publishing House, Burtonsville (1989)
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
19. Tukey, J.W.: The Future of Data Analysis. Ann. Math. Statist. **33**, 1–67 (1962)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: von Luxburg, U., Guyon, I. (eds.), Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010. Curran Associates Inc., Red Hook (2017)
21. Weihs, C., Ickstadt, K.: Data Science — The impact of Satistics. Int. J. Data Sci. Anal. **6**, 189–194 (2018)

Critical Visual Explanations. On the Use of Example-Based Strategies for Explaining Artificial Intelligence to Laypersons.

Spiegazioni visive critiche. Uso di strategie basate su esempi per spiegare l'intelligenza artificiale ai non addetti ai lavori.

Beatrice Gobbo

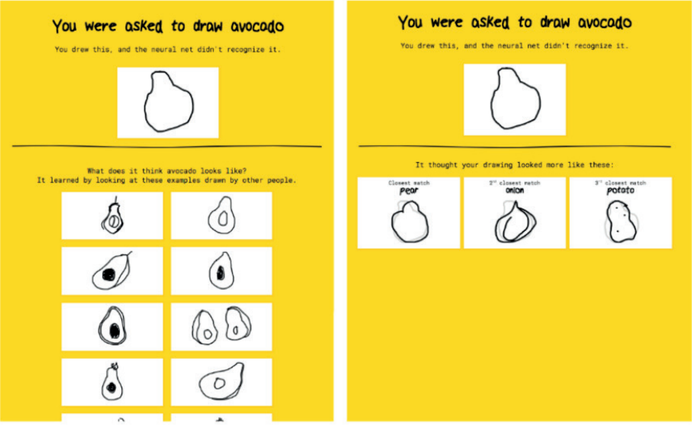
Long Abstract. Explainable Artificial Intelligence (XAI) has started to develop both as a field and as a collection of techniques for enabling humans to understand and interpret Artificial Intelligence (AI) mechanisms and decisions that permeate the lives of many human beings today. [16] While much effort was initially put into making AI systems and algorithms intelligible to their developers and creators, over time, interest expanded towards newcomers, domain experts and laypersons [18, 4], opening new challenges and research opportunities. For instance, promoting algorithmic fairness, accountability, trust, ethics, and awareness at all levels has become a significant challenge in terms of governance. [1] As a consequence, expanding audiences and scopes have made room for other disciplines. Indeed, surveys and State-of-the-Art Reports (STARs) on AI Explainability witness the proliferation of scientific contributions increasingly leaning towards multidisciplinary cooperation where efforts in organising and classifying AI explanations have been made considering a large variety of parameters including, for instance, *strategies*, *media* and *audience* [8, 11]. Specifically, *strategies* define how the line of explanation reasoning goes. *Media* and *audiences* define the visual, audio or text apparatus staged for supporting the explanation and its target public. [8] Among the numerous combination that could be found in the literature [11], this paper will discuss and critique from an information design perspective the use of example-based [16] visual explanations [19] for AI addressed to laypersons. While some oversimplification and misunderstanding risks have emerged when addressing with expert users [14], example-based explanations supported by visual means are considered efficient when targeted to laypersons [13, 7, 10, 15].

A study carried on by Cai and colleagues in 2019 aimed at evaluating the effectiveness of comparative and normative visual explanations for a sketch recognition system on a sample of laypersons [2] demonstrates that *normative explanations* are more effective than *comparative explanations*. In other words, it demonstrates that

---

Beatrice Gobbo
CIM — Centre for Interdisciplinary Methodologies, Warwick University, e-mail: `beatrice.gobbo@warwick.ac.uk`

the sample of lay users involved in the study, when trying to understand the reason why a sketch recognition system did not recognise their drawings (see the screenshot depicting two examples of Visual Explanation provided by the system in Figure 1), prefer to establish a norm for what drawings look like in the target class (1 , left) instead of understanding the system relying on the comparison between the user's drawing and similar drawings from alternative classes (1 , right).



You were asked to draw avocado

You drew this, and the neural net didn't recognize it.

What does it think avocado looks like?
It learned by looking at these examples drawn by other people.

You were asked to draw avocado

You drew this, and the neural net didn't recognize it.

It thought your drawing looked more like these:

Fig. 1 The image juxtaposes normative (left) and comparative (right) example-based visual explanations. Source: [2] using QuickDraw

Hence, although example-based visual explanations are effective when dealing with laypersons, possible limitations emerging from this case study concern the risk of relying on widespread sample datasets [3] and providing biased explanations of complex, unstable and situated systems such as Artificial Intelligence machines. In this regard, it is worth mentioning that Artificial Intelligence algorithms have been defined as socio-technical *assemblages* [9], and *megamachines* [5] since they are socially, historically and materially entangled with data. Thus, the contribution revolves around the discussion on how to engage laypersons in understanding such complex machines using example-based visual explanations and how the concept of *critical information visualisation* [6] could support their design process. The principles of critical information visualisation are meant to help designers and researchers to formulate questions during the design, use, and study of information visualisations. Similarly, this paper aims to provide researchers with — preliminary — conceptual tools to critically design and stage example-based visual explanations for laypersons. [1].

- *Enrich to counteract the norm* — In line with the idea of *plurality* proposed by Dörk et al. [6], exposing multiple perspectives could enhance the audience's

---

[1] Note that we do not criticise the explanation of the system *per sé* — if the system is explainable — but the way the examples are displayed while explaining it.

ability to identify with the complex system. For instance, the ’Anatomy of an AI System’ [17] provides a ramified visual explanation involving different actors, materials and technologies.

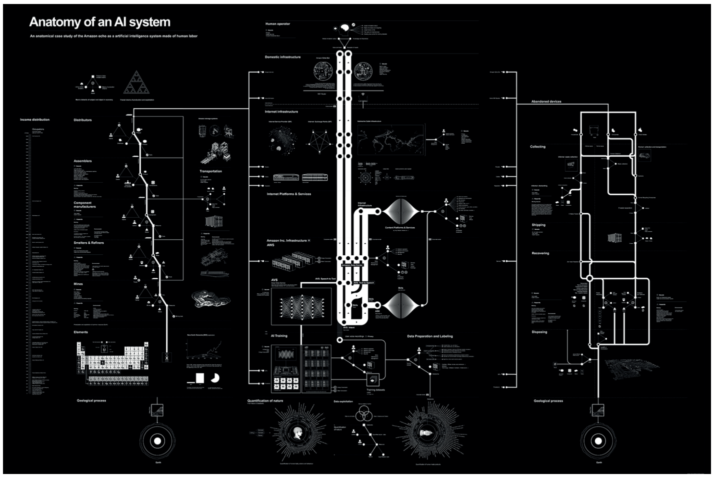

Anatomy of an AI system
An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor
Income distribution
Distributors

Fig. 2 The Anatomy of an AI System is a project by Share Lab, which explain how Amazon Echo by Alexa works using a multi-perspective range of examples. Source: [17]

- *Dissect to disclose the process* — Drawing on the *disclosure* principle proposed by Dörk et al., being critical in presenting example-based visual explanations implies being able to inform the audience about design decisions, such as the reason why particular examples have been presented instead of others. For instance, Mauri and colleagues present a collection of visual posters explaining a collection of statistical models (including some AI algorithms) to laypersons using examples. There, visual explanation fragments are usually accompanied by texts specifying the reason behind the choice of each example.
- *Simulate to empower the audience* — Simulation is a well-known procedure in computing and statistics. Indeed, many examples of visual explanations simulate the functioning of Artificial Intelligence systems enabling the users to interact with them. Moreover, the simulation of examples was revealed to be convincing when dealing with experts and newcomers [12]. Here, we are promoting simulation to support example-based strategies to empower the audience. Indeed, users could have the opportunity to play with their data or produce data with the algorithm in a real-time scenario (i.e., the ”Simplicial Depth Measure” and ”Page Rank” examples proposed by Mauri et al. [15]).

In the first part, the author presented example-based visual explanations as effective tools for explaining AI to laypersons. In the second part, the use of examples has been challenged. Finally, by drawing on the principle of *plurality*, *contingency*, *disclosure*, and *empowerment* proposed by Dörk et al., the paper outlined the need to promote a critical approach to the design of example-based visual explanation and denoted an increasing need for collaboration, where computer scientists, statisticians, social scientists and information designers cooperate in co-design settings.

References

1. Ananny, M., and Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society 20*, 3 (2018), 973–989. _eprint: https://doi.org/10.1177/1461444816676645.
2. Cai, C. J., Jongejan, J., and Holbrook, J. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray California, Mar. 2019), ACM, pp. 258–262.
3. Ciston, S. A critical field guide for working with machine learning datasets.
4. Correll, M. Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, Association for Computing Machinery, p. 1–13.
5. CRAWFORD, K. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.
6. Dörk, M., Feng, P., Collins, C., and Carpendale, S. Critical InfoVis: exploring the politics of visualization. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris France, Apr. 2013), ACM, pp. 2189–2198.
7. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. Expanding Explainability: Towards Social Transparency in AI systems. Publisher: arXiv Version Number: 1.
8. El-Assady, M., Jentner, W., Kehlbeck, R., Schlegel, U., Sevastjanova, R., Sperrle, F., Spinner, T., and Keim, D. Towards XAI: Structuring the Processes of Explanations. 13.
9. Gillespie, T. 167The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*. The MIT Press, 02 2014.
10. Gobbo, B. Embalming and dissecting artificial intelligence. visual explanations for the general public.
11. Gobbo, B., Elli, T., Hinrichs, U., and El-Assady, M. xai-primer.com — A Visual Ideation Space of Interactive Explainers. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (New Orleans LA USA, Apr. 2022), ACM, pp. 1–4.
12. IEEVIS. Workshop on visualization for ai explainability.
13. Jin, W., Fan, J., Gromala, D., Pasquier, P., and Hamarneh, G. Euca: A practical prototyping framework towards end-user-centered explainable artificial intelligence. *ArXiv abs/2102.02437* (2021).
14. Kim, B., Khanna, R., and Koyejo, O. O. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc.
15. Mauri, M., Vantini, S., Gobbo, B., Elli, T., Aversa, E., Benedetti, A., Briones Rojas, M. D. L. A., and Colombo, G. Making posters to understand statistics: towards a didactical approach in communication design. In *DRS2022: Bilbao* (2022), pp. 1–18.
16. Molnar, C. *Interpretable machine learning: a guide for making black box models explainable*, second edition ed. Christoph Molnar, Munich, Germany, 2022.
17. ShareLab. Anatomy of an ai system.
18. Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, May 2021), ACM, pp. 1–16.
19. Tufte, E. R. *Visual explanations: images and quantities, evidence and narrative*, twelfth printing, april 2019 ed. Graphics Press, Cheshire, 2019. OCLC: 1126611319.

Visualising unstructured social media data: a chart-based approach

Visualizzare i dati non strutturati dei social media: un approccio basato su grafici

Elena Aversa

Abstract This paper tackles the visualisation of unstructured data generated on social media platforms. Due to the inherently noisy, multimodal, and heterogeneous nature of these datasets, which offer a variety of information fields, a flexible approach has been chosen for both analysis and visualisation. Drawing from the Information Design field, the aim of this study is to show that a chart-based approach, where different visual models map different data variables, provides an effective means for presenting and analysing unstructured social media data.

Key words: information visualisations, visual models, unstructured data, social media platforms

1 Introduction

The past few years have been marked by a surge in the use of social media platforms [2]. This trend has been particularly evident during social issues [9] such as the COVID-19 pandemic, as shown by multiple studies [3] [6] [8]. The increasing reliance on social media as a primary source of information [19] has produced a plethora of data in various formats, including text, images, and videos. This study focuses on the visualisation of such unstructured data, drawing from the field of information design. What visual models can be used to visualise unstructured data generated on social media platforms? In this paper, ”visualise” not only means presenting data in ”graphical or pictorial form to make the information easy to understand” [5], but also and above all, the exploration of the dataset aimed at allowing reasoning and enabling diverse interpretations of the same phenomenon. Visualization tools are crucial for effectively navigating large datasets from the web and

---

Elena Aversa
Design Department, Politecnico di Milano, Milano. e-mail: elena.aversa@polimi.it

discovering valuable insights, as noted by Dahyot et al. [4]. The objective is to illustrate visual models that can effectively represent unstructured social media data based on a set of potential research questions. To address the aforementioned objective, the study examines an ”active controversy” [17], the earthquake that struck Turkey and Syria on February 6, 2023.

2 Related Work: Visual models in information design

In this paper, the term ”information visualization” refers to Isabel Meirelles’s [11] definition of ”visual displays in which graphical approaches play a central role in communicating information in a meaningful way”. In this paper, those ”graphical approaches” are called ”visual models”, namely different ways to present visual variables that encode the data dimensions. Among the various visual models available [7] [14], this study focused on the following:

1. Temporal models → how data points change over time;
2. Relational models → connections among data points;
3. Hierarchical models → hierarchical relationships as parts of wholes;
4. Visual analysis → visual representation of collections of images.

3 Methodology

This research follows a chart-based approach to data visualisation [10], focusing on how different visual models can be used to map different variables ranging from qualitative to quantitative and visual data.

3.1 Data collection

The dataset analysed in this paper consists of 10156 tweets collected using the 4CAT Capture and Analysis Toolkit [12][1] with the query ”earthquake AND turkey AND syria has:images -is:retweet”[2] for a single day, February 6th, 2023, the day of the earthquake. The structure of the dataset encompasses both qualitative and quantitative data, ranging from timestamp to engagement count, the body of the tweet, and images attached, among others. The focus was limited to those tweets with an

---

[1] 4CAT is a software suite created by OILab and the Digital Methods Initiative at the University of Amsterdam, designed to gather data from various online sources and analyse it using analytical processors.

[2] ”has:images” restricts the collection to tweets that have images, while the ”-is:retweet” exclusion ensures that tweets that are retweets of tweets already included in the dataset are not collected.

engagement [3] greater than 100. This was done to concentrate on tweets that reached a wider audience. As a result, the number of tweets subjected to analysis was 487.

3.2 Research questions and data analysis

Five diverse visual models were utilised to address the five research questions.

3.2.1 RQ.1 What type of content circulated on Twitter within the first 24 hours after the earthquake in Turkey and Syria?

The analysis of both the corpora and attached media allowed for the identification of three types of content: links to external websites, videos, and images. In this case, the visual model used to map all the variables considered was the Beeswarm plot. The Beeswarm plot is employed to depict the distribution of data points along a continuous axis. Each data point is represented as a dot, and the vertical dimension prevents the overlap between them, thereby revealing their distribution.

3.2.2 RQ.2 Who were the actors involved in the circulation of the analysed tweets? And what relationships are set up between them?

Networks are employed to visualise complex relationships among a substantial number of elements. These relationships are depicted through lines (edges) linking the different elements (or nodes). Specifically, the visualisation focused on the relationships between ”actors” (intended as users who shared the tweets) and the hashtags present in the corpus. Twitter hashtags feature significantly in discussions on the platform, to the extent of being considered a means of facilitating a distributed discussion among users who are not connected through their network of followers [13].

3.2.3 RQ.3 What is the main subject matter of the analysed tweets?

To identify the subject of the conversation, a hierarchical visual model, the word tree, was employed. Wordtrees facilitate the recognition of repeated words within the contextual framework of a phrase while exhibiting the structure of said context, allowing for comprehensive exploration [18]. The analysis of the corpus of tweets was carried out using three separate queries as the roots, namely, the keywords served for the data collection phase. In the visualisations, the dimension of

---

[3] The engagement for each tweet is determined by adding together the counts of ”likes,” ”quotes,” ”retweets,” and ”comments”.

each word also represents the frequency of occurrence - the larger the word, the more frequently it appeared within the specified context.

3.2.4 RQ.4 How is the event represented through images?

The last research question explores the role of images, as non-linguistic pictorial elements, in generating meanings and reinforcing narratives. According to Aiello and Parry [1], images are a crucial component of communication, identity, and connection in social media environments. The selected visual model is the grid, useful to cluster images by colour or similarity and thus have a comprehensive overview of the collection. The images displayed are 707, a larger number compared to the tweets because, in some tweets, more than one image was published. The sorting in the grid was done by colour.

3.3 Results

3.3.1 RQ.1 → Time-series: Beeswarm

Figure 1 displays the Beeswarm[4] of the content circulated. Each dot stands for a single tweet, and the size reflects the level of engagement it received. The tweets are organised on the Y-axis into three groups: tweets shared with images, tweets that include images and external links in their corpus, and tweets with videos and images. The first cluster resulted to be the densest and most engaging, as well as the most consistent over time. Meanwhile, tweets that included both links and images were fewer in number but maintained their stability over time. Tweets containing videos and images, on the other hand, emerged later and in smaller quantities.

3.3.2 RQ.2 → Relational models: Network

The network[5] shown in figure 2 clarifies not only the involved actors but also the elements that establish connections between them: the hashtags. The size of the elements depends on the type of connections; for example, if a hashtag is used by many actors, its size will be larger and vice versa. Despite the earthquake affecting mostly Turkish territory, the visualisation shows that actors are mentioning the Syria and Turkey hashtags equally. This raises questions about which actors are participating and why, and whether they are verified entities, news agencies or automated bots.

---

[4] The visualisation was realised using RAWGraphs[10] and polished and annotated on Adobe Illustrator.

[5] The graph was created using Gephi, an open-source tool for network exploration, and was spatialised using the Force Atlas 2 layout. Adobe Illustrator was then utilised to improve the visual appearance of the graph and add annotations.

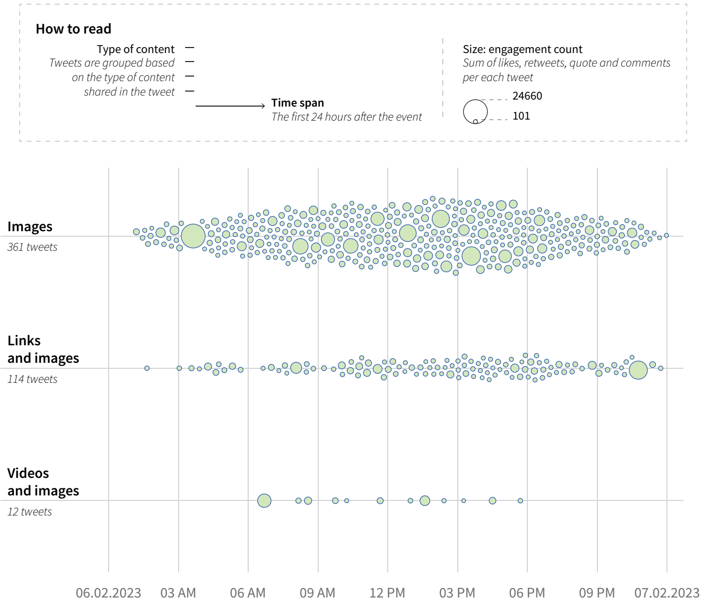


Fig. 1 Beeswarm of the content circulated on Twitter in the first 24 hours after the earthquake, with the keywords “Syria”, “Turkey” and “earthquake” in the corpora of the tweet. The reference time is CET.

3.3.3 RQ.3 → Hierarchical models: Wordtree

Wordtrees use a branching structure to illustrate the connection between a pre-selected word and other words in the phrase, adding context to their usage. Figure 3 displays the three wordtrees[6] created from a merged corpus of 35154 words. The visualizations show the most frequent words following ”earthquake”, ”Turkey”, and ”Syria,”. The word ”Turkey” is consistently followed by the word ”Syria,” while the opposite is not as prevalent. A similar trend is observed when the root word is ”earthquake.”

3.3.4 RQ.4 → Visual analysis: Grid of images

The final research question refers to that part of the data which is neither text nor numerical: images. To understand how the event was visually depicted, the 707 analyzed images were arranged in a grid[7], and sorted by colour. As shown in figure

---

[6] The three visualizations were created using the Wordtree feature on 4cat and refined with Adobe Illustrator

[7] The visualization was created with ImageSorter, an open-source software developed by the Visual Computing Group at the HTW Berlin.

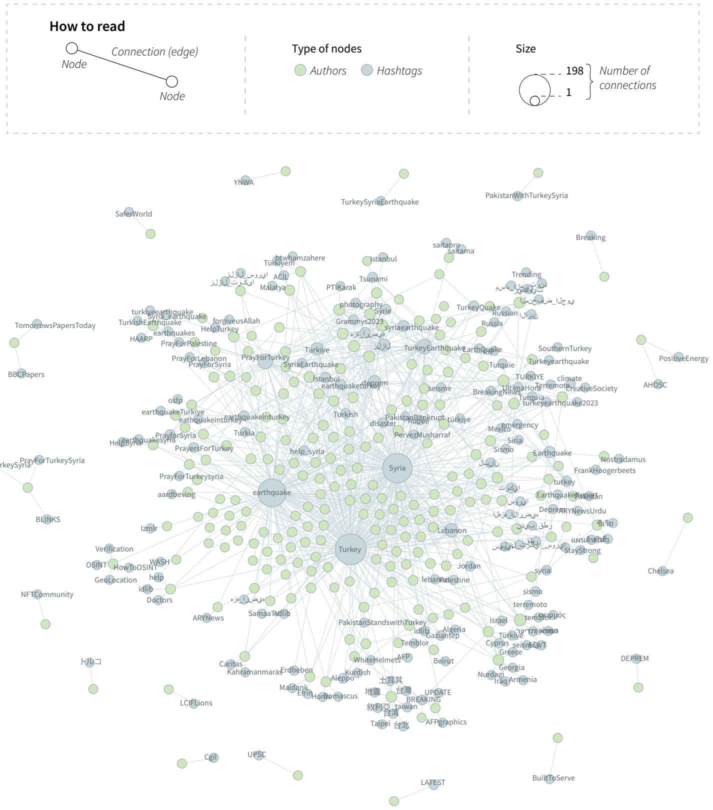


Fig. 2 Network of the actors and hashtags with labels. Actors' labels are not included to ensure privacy.

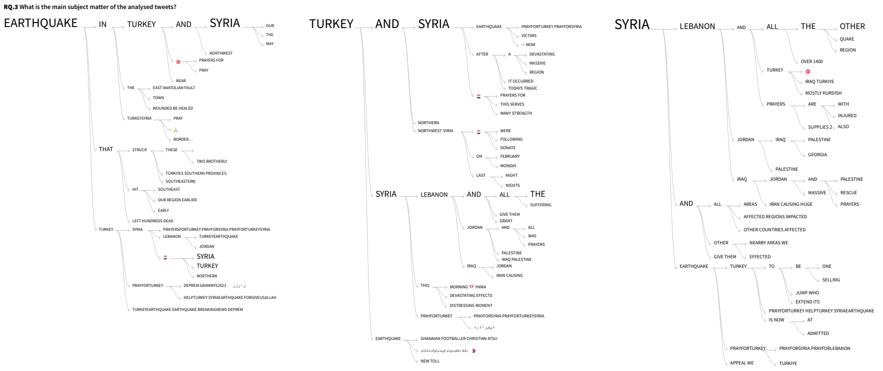


Fig. 3 Wordtrees of the tweets corpora with the roots "turkey" "Syria" "earthquake".

4a, four clusters emerged mainly: photographs, charts and maps, images with text, and graphic images. Many of these images are repeated in the grid: this indicates common usage in tweets. Figure 4b shows a picture that was shared across nine tweets. It’s noteworthy that the photograph in question was not actually taken in Syria or Turkey, but rather, it was a stock photo under the classification of ”rescue dog”.

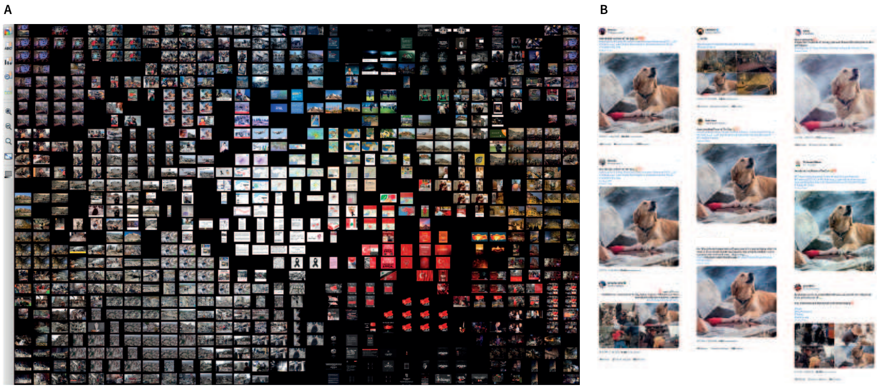


Fig. 4 A-Grid of images sorted by colour. B-Repeating images.

4 Conclusion

This paper presented a chart-based approach for visualising and exploring unstructured data generated on social media platforms. The purpose of this study is to emphasise the need for a flexible approach to visualising social media data, which is now ubiquitous. The study remains at a general level of data analysis. Still, it opens up avenues for reflection on visualising unstructured data, web-generated data, and the chart-based approach.

References

1. Aiello, G. (2019). Visual communication: Understanding images in media culture (1st edition). SAGE Publications.
2. boyd, danah. (2015). Social Media: A Phenomenon to be Analyzed. Social Media + Society, 1(1), 205630511558014. https://doi.org/10.1177/2056305115580148
3. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The COVID-19 social media infodemic. Scientific Reports, 10(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

4. Dahyot, R., Brady, C., Bourges, C., and Bulbul, A. (2015). Information visualisation for social media analytics. 2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), 1–5. https://doi.org/10.1109/IWCIM.2015.7347082
5. Kirk, A. (2019). Data visualisation: A handbook for data driven design (2nd edition). SAGE.
6. Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., and Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak. Journal of Travel Medicine, 27(3), taaa031. https://doi.org/10.1093/jtm/taaa031
7. Ferdio: The Dataviz Project. https://datavizproject.com/
8. Goel, A., and Gupta, L. (2020). Social Media in the Times of COVID-19. JCR: Journal of Clinical Rheumatology, 26(6), 220–223. https://doi.org/10.1097/RHU.0000000000001508
9. Marres, N. S. (2005) No issue, no public: democratic deficits after the displacement of politics. Thesis, fully internal, Universiteit van Amsterdam, pp. 90- 93. Ipskamp Printpartners.
10. Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAWGraphs: A Visualisation Platform to Create Open Outputs. Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter, 1–5. https://doi.org/10.1145/3125571.3125585
11. Isabel Meirelles. 2013. Design for information. Rockport Publishers, Beverly, MA. ISBN: 9781592538065
12. Peeters, S., and Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. Computational Communication Research, 4(2), 571–589. https://doi.org/10.5117/CCR2022.2.007.HAGE
13. Rambukkana, N. (A c. Di). (2015). Hashtag publics: The power and politics of discursive networks. Peter Lang.
14. Ribecca S: The Dataviz Catalogue. https://datavizcatalogue.com/
15. Rogers, R. (2013). Digital methods. Mit Press.
16. Röhle, B. R. T. (2012). Digital Methods: Five Challenges. In D. M. Berry (A c. Di), Understanding Digital Humanities (pp. 67–84). Palgrave Macmillan UK. https://doi.org/10.1057/97802303719344
17. Venturini, T. (2010). Diving in magma: How to explore controversies with actor-network theory. Public Understanding of Science, 19(3), 258–273. https://doi.org/10.1177/0963662509102694
18. Wattenberg, M., and Viegas, F. B. (2008). The Word Tree, an Interactive Visual Concordance. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1221–1228. https://doi.org/10.1109/TVCG.2008.172
19. Westerman, D., Spence, P. R., and Van Der Heide, B. (2014). Social Media as Information Source: Recency of Updates and Credibility of Information. Journal of Computer-Mediated Communication, 19(2), 171–183. https://doi.org/10.1111/jcc4.12041

From teaching Statistics to designers to teaching Statistics through design

Michele Mauri and Simone Vantini

Abstract This paper draws on our experience in introducing statistics to design students. Information design professionals today cannot avoid working with data: creating visualizations, or data driven narrative, designing dashboards, or using it as part of their creative process in the development of new services and products. However, approaching the teaching of statistics with frontal, theoretical lessons risks encountering disinterest or rejection from the students, who see such information as something too far from their needs and practice. To overcome this issue, we developed an approach in which the classical didactical approach in statistics is reversed: rather than starting from general principles and then moving to specific application, we proposed to the students to make a research on a particular statistical method and present it visually to their peers by means of paper posters. In detail, this approach brought to the development of 14 different posters in the course of the last two years that will be exposed at the end of the talk.

Key words: Algorithms, Information Visualization, Visual explanations, Design education, Posters

1 Introduction

This paper draws on our experience in introducing statistics to design students. Information design professionals today cannot avoid working with data: creating visualizations, or data driven narrative, designing dashboards, or using it as part of their creative process in the development of new services and products. It is therefore

---

Michele Mauri
DensityDesign Lab, Department of Design, Politecnico di Milano, e-mail: michele.mauri@polimi.it

Simone Vantini
MOX, Department of Mathematics, Politecnico di Milano, e-mail: simone.vantini@polimi.it

needed to provide, in design education, an overview of what it means to work with data [5, 3]. This knowledge is meant to highlight two aspects that influence the design process, from the micro to the macro level. On the micro level, designers need techniques and tools to work directly with data when they need it in their daily work. On the macro level, knowing the basics of statistics is useful to understand and have a critical stance on the process followed by others in the creation of the data that designers are using for the creation of graphics, interfaces or services.

Given this overall goal, difficulties emerge when outlining a didactical approach that can fit in a design curricula highly oriented on practical activities, based on studio-courses. Students are used to learn-through-doing activities [4], in which the needed knowledge is provided by teachers and assistants and directly used in the development of a real or mock-up project. Approaching the teaching of statistics with frontal, theoretical lessons risks encountering disinterest or rejection from the students, who see such information as something too far from their needs and practice. To overcome this issue, we developed an approach in which the classical didactical approach in statistics is reversed: rather than starting from general principles and then moving to specific application, we proposed to the students to make a research on a particular statistical method and present it visually to their peers.

2 Course Description and Topics

This approach was developed over several years in a studio-course at Politecnico di Milano called “Final Synthesis”. Studio-courses are based on laboratorial activities, in which students develop one or more projects over the semester, and they are evaluated on the outcome of their work. Students usually work in groups, allowing them to develop and test their teamwork abilities. The course “Final Synthesis” is the last studio-course followed by design students in the Communication Design Master at Politecnico di Milano. All the students are therefore at the end of their education and the goal of the course, as the name outlines, is to face them with design challenges in which they can show all the acquired knowledge. The duration of the course is four months and it carries a value of 18 ECTS (European Credit Transfer and Accumulation System), which equates to 180 hours of classroom instruction and a comparable amount of independent study.

The course is held in parallel sections with the same goals, but on different topics. One of these sections, coordinated by researchers affiliated to DensityDesign Lab, focuses on the topic of the visual representation and communication of complex phenomena, with an in depth focus on information visualization and data visualization. The faculty of the course is composed of design teachers, a semiotic teacher, and a statistics one. The course, described in a previous publication [2], is divided in three main modules: the first one focused on statistics, the second and third one more related to information design and public communication. By framing the first module as a warm-up project, students are asked to represent statistical algorithms with a static paper poster. The choice of the algorithms was related to the following issues:

- We favor algorithms provided with a simple intuitive explanation not necessarily requiring an explicit mathematical formalization. For this reason we focus on algorithms which rely on a geometrical exploration of data and not on a probabilistic modeling of data.
- We focus on algorithms which could be splitted in consecutive and possibly iterative steps which are more prone to a visual and diagrammatic representation.
- To trigger students' curiosity we finally focus on algorithms for which some popular and compelling real world applications can be identified.

Here comes a partial list of the algorithms selected by the students: hierarchical clustering, k-means clustering, edge-betweenness community detection, Cheng and Church biclustering, principal component analysis, kernel density estimation, depth measures, K-nearest neighbor classifier, classification trees, control charts, and bootstrap estimation.

The request to design a printed poster is specifically aimed at nudging the students to leverage on their visual abilities to overcome the constraint of the printed papers, identifying a clear visual structure, balancing texts, visualization and illustrations, and providing a coherent narrative. The union of the two (statistical algorithms and the design of poster) allows also to leverage on the field of "visual explanation", a rising approach for making technologies, and in particular digital ones, understandable to a larger public [1].

3 Course structure

In the course preparation, the statistics teacher identified ten possible topics to be represented by the students. At the beginning of the course, in the first lesson the overall goal was explained, and the identified topics were briefly presented with the exclusive use of naive pictures and examples of applications. Student groups made their choice, and started their research. The module lasted five weeks, and each group had a weekly 1-to-1 30-minute review of the project with the statistics professors and design one.

The work advancement was roughly divided as follows:

- Week 1: collection of source, understanding of the topic, first sketches on paper. The goal is to check that students identified sounding sources and properly understood the topic. Teacher elicit them to delve into details of the method: how it works, which other statistical knowledge is needed to understand and explain it.
- Week 2: first draft of the overall structure, identification of narrative strategy. Students have to identify the best strategy to explain their methods: if showing its working step by step, if by showing the result and explaining the process backward, or if showing it as a general schema

- Week 3: consolidation of choices, definition of the language, identification of the relevant knowledge to be shown. At this point of the module, students face with the dilemma of knowing more that they can represent, and they have to make choices and focus on the most salient and peculiar feature of the method
- Week 4: general consistency. While most of the choices have been made, the students have the opportunity to identify small and bigger inconsistencies, both in the language, in typography and on visual choices
- Week 5: fine tuning. At this point, if the previous steps were properly followed, students have time to focus on the fine tuning and testing of the result.

The above description illustrates the average process, which in the practice may vary according to students' ability to understand the topic and find a suitable visual solution and narrative strategy. However, during the weekly reviews the teacher helps the students in passing through those checkpoints both to keep them on track and also to test their understanding of the topic. The results are then presented by each group in front of the class and of the whole course faculty, both the two teachers that followed them in this module, and by the other ones.

4 Conclusions

This didactical approach brought to the development of 14 different posters in the course of two years. The results, while with different levels of completeness and correctness, proved to be a suitable tool to introduce statistical concepts to design students. Many of them proved also to be results almost ready for real dissemination of the illustrated statistical methods. A selection of the posters will be presented at the conference venue.

Furthermore the experience posed the basis for adopting, conversely, didactical approaches from the field of design to statistics. In detail, the didactical approach which is typical of a design course was tested also with students of the Master of Science in Mobility Engineering in a course dedicated to network analysis. The strong heterogeneity of the students of this master makes it too difficult to set a level of formalization that could suit most of the students. Also in this case, we build the course around the group projects dedicating only 50% of the lectures to theory and lab sessions with the remaining 50% of the course instead dedicated to project revisions and to class and faculty presentation. The students' feedback to this seminal experiment was extremely positive. The most relevant feedbacks were two: the students perceived the learning experience as specifically personalized for each group and adapted to the group specific strengths and weaknesses; the continuous evaluation was not perceived as a mere grading process but rather as a further opportunity of personalized learning.

References

1. Kolkman, D. The (in)credibility of algorithmic models to non-experts. *Information, Communication & Society 25*, 1 (Jan. 2022), 93–109.
2. Mauri, M., Colombo, G., Angeles, B., and Ciuccarelli, P. Teaching the Critical Role of Designers in the Data Society: The DensityDesign Approach. In *Proceedings of DRS Learn X Design 2019: Insider Knowledge* (Ankara, Turkey, 2019), N. A. Börekçi, D. Koçyıldırım, F. Korkut, and D. Jones, Eds., METU Department of Industrial Design, pp. 183–195.
3. Moere, A. V., and Purchase, H. On the role of design in information visualization. *Information Visualization 10*, 4 (Oct. 2011), 356–371.
4. Özkar, M. Learning by Doing in the Age of Design Computation. In *Computer-Aided Architectural Design Futures (CAADFutures) 2007*. Springer Netherlands, Dordrecht, 2007, pp. 99–112.
5. Swanson, G. Educating the Designer of 2025. *She Ji: The Journal of Design, Economics, and Innovation 6*, 1 (2020), 101–105.

Forecasting Spatio-Temporal Data with Bayesian Neural Networks

Previsione di Dati Spazio-Temporali Tramite l'Utilizzo di Reti Neurali Bayesiane

Federico Ravenda, Mirko Cesarini, Stefano Peluso & Antonietta Mira

Abstract Nowadays researchers have to deal with both geo-referenced and temporal data in a wide range of fields i.e., in medicine and in economics. We propose a Bayesian Spatio-Temporal Neural Network to achieve two goals: (1) create a model that can infer both spatial and temporal components; and (2) combine the flexibility of Neural Networks (which does not need to satisfy any statistical assumption) with the uncertainty quantification of a Bayesian model. Our architecture is compared with the established INLA on COVID-19 data, with clearcut better prediction performances of the proposed method.

Key words: Entity Embedding, INLA, Probabilistic Deep Learning, Node2Vec, Time Series

1 Motivation

The research question that guides this work is as follows: *Can a flexible neural network (NN) compete or even improve the performances of traditional statistical Spatio-Temporal models?*

Federico Ravenda
Università of Milano-Bicocca, Milano, Italy, e-mail: `federico.ravenda@unimib.it`

Mirko Cesarini
Università degli Studi Milano-Bicocca, Milano, Italy, e-mail: `mirko.cesarini@unimib.it`

Stefano Peluso
Università degli Studi Milano-Bicocca, Milano, Italy, e-mail: `stefano.peluso@unimib.it`

Antonietta Mira
Università della Svizzera italiana, Lugano, Svizzera & Università degli Studi dell'Insubria, Como, Italy, e-mail: `antonietta.mira@usi.ch`

The idea of this work is to transpose the most relevant aspects of a hierarchical Bayesian framework [1] capable of modeling both the spatial and temporal components [2], into a deep learning architecture. To achieve this, the Bayesian learning paradigm is applied to NNs. The main contribution of this work is the joint combination of:

- **Bayesian Neural Networks** for forecasting purposes and uncertainty quantification;
- **Probabilistic Layers** to model the outcome variable;
- **Embeddings** of different types to synthesize temporal and spatial components;
- **Recurrent**, **Convolutional**, and **Attention based** layers to model dynamical temporal data sequences.

The proposed Neural Network Architecture, applied on COVID-19 dataset, reaches satisfactory goodness-of-fit performances, delivers precise predictions of events (combined with uncertainty quantification), works over varying size time intervals, and outperforms in terms of predictions a traditional parametric spatio-temporal model, INLA [3], as discussed in Section 3.

2 Methods

2.1 Modelling Spatial and Temporal Components with Embeddings

Embeddings are numerical representations of categorical variables, commonly used in machine learning [4]. They capture the semantic meaning of the variable by mapping it to a dense vector of real numbers. Models built with embeddings can learn the relationships between the variables in a continuous space rather than as discrete values, and this can result in an improved performance. In the Neural Networks (NN) literature two methods are proposed to embed the spatial component. They can be seen as complementary methods that synthesize different spatial characteristics, and for this reason, we opt to jointly fed them as inputs to our Bayesian Spatio-Temporal Neural Network (BSTNN) architecture. The first one, **Node2vec** [5], represents the geographic information. It is used for learning low-dimensional representations of nodes in a graph: different locations are represented as vertices (nodes) and adjacent locations share a common edge (edges' weights represent the distance between locations). In contexts where the geographic location is important to explain the phenomenon of interest (nearby sites have similar behaviours), embeddings are expected to be very informative. Another way to embed spatial component is by synthesizing "*context*" information (locations with similar behavior have similar latent representation). Entity Embedding [6] serves this purpose: the idea is to map categorical variables into Euclidean spaces using a function approximation problem where categories are turned into *"Entity"* (a.k.a. *category*) Embeddings. It is expected that values with similar function outputs are close to each other in the embedding space.

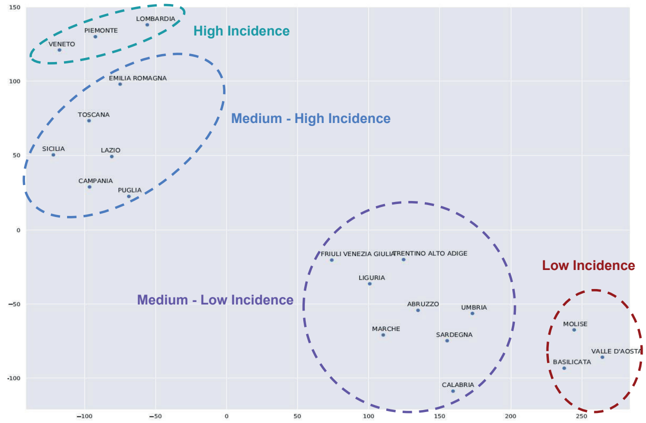


Fig. 1: A plot of the embeddings extracted using the Entity Embedding Approach. For visualization purposes t-SNE was applied to represent embeddings in 2 dimensions.

In the same way, temporal components e.g., month and year, can be modelled. To give an intuition on how embeddings work we can observe Fig. 1. In this case, data come from the COVID-19 Daily-Deaths time-series for each of the 20 Italian regions (data are available at the following link https://github.com/pcm-dpc/COVID-19) and the goal is to embed each region in a latent space using the Entity Embedding approach. It can be observed that regions that have had a similar incidence of deaths from COVID-19 lie close together in the dimensional-reducted space (after applying t-SNE [7] to the latent representation vectors for visualization purposes), creating natural clusters between regions which can be intuitively justified.

2.2 Model Architecture

We want to build a NN architecture that is able to learn patterns of space-time data. Furthermore, our NN will account for uncertainty in its forecasts, like a traditional Bayesian statistical model.
One of the characteristics of this architecture is that, even as the forecast interval increases, it is able to provide satisfactory predictions with low error scores compared with other traditional models, as will be discussed in Section 3. We now introduce the BSTNN architecture (Fig. 2), focusing on the most original aspects. Suppose we want to analyze the temporal trend of a particular phenomenon in N different sites. Formally, for each site $j = 1, \ldots, N$ a phenomenon is observed in T temporal instants and we want to make predictions for the following times, up to horizon $T + h$.
The architecture in Fig. 2 is structured as follows. In the first layer there are N inputs, as many as the considered sites. Each input goes through two convolutional layers (**Conv1D**) which create a temporal smoothing and find relevant patterns in

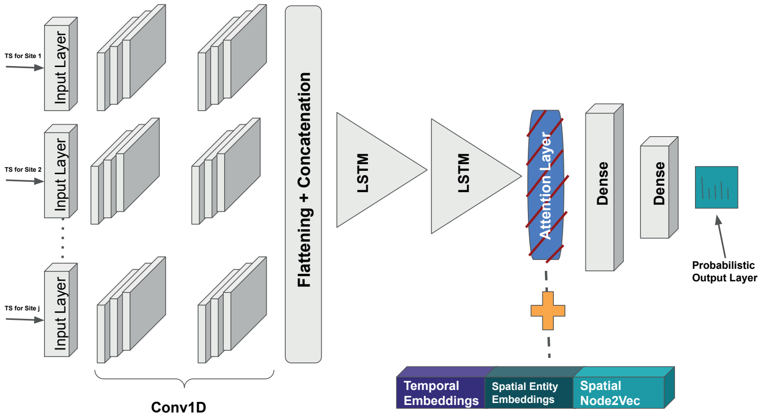


Fig. 2: Skeleton of the BSTNN architecture.

the time series. The output of each convolutional layer is flattened [8] [1] and reshaped to feed the next layers. At this point two LSTMs [9] are stacked, followed by an Attention layer [10] to extract sequential information from the data. The information is then concatenated to the temporal and spatial embeddings. At each time instant the same input data will feed the network N times, each time coupled with the spatial embedding related to the site whose prediction is to be generated.

The rationale for providing both data and spatial embeddings as network inputs is that the embeddings will supply useful information to compute the output for a specific site, while taking into account the input data collected from several sites. Once the embeddings are concatenated with Convolutional layers, Dense layers with decreasing number of neurons are executed, up to the Output layer (a dense layer or a Probabilistic layer). One key aspect of our NN is that it is Bayesian. To transpose the Bayesian framework into Deep Learning the main idea is the following. Consider a neuron weight w_i. In a standard (deterministic) NN, this weight is estimated with a single value $\hat{w}_i$, learnt via backpropagation. In a Bayesian NN weights are treated as random variables and thus their prior distribution is specified and then, using the Bayesian paradigm, their posterior distribution is estimated, given the data. This allows for uncertainty quantification on the output of the network, which is not just a single point estimate but a posterior distribution. This makes the network more robust to overfitting, as it can incorporate the uncertainty in its own parameters. In Bayesian statistics, regression problems are commonly solved with Markov Chain Monte Carlo (MCMC) [11]. MCMC methods work well for small problems e.g., 10 to 100 parameters, but not for larger networks such as Deep Learning models with millions of weights. There are two computationally more efficient alternative approaches to obtain an approximation of the posterior: the ***Variational Inference (VI)*** Bayes [12] and the ***Monte Carlo (MC) dropout*** [13]. The proposed BSTNN

---

[1] Flattening is a procedure required to prepare data for the subsequent layers

architecture has been implemented both with a Variational Inference approach and MC dropout. Results shown in Section 3 are obtained with MC Dropout.

3 Results

The data source is the github repository of the Italian Civil Protection. We focus on COVID-19 *New-Daily deaths* in each Italian region. This phenomenon is subject to high variability and strong shocks; for this reason, capturing the signal and purifying it from the noise requires the use of complex models. Performances are evaluated over a forecast range of **7**, **14** and **21** days. The test set is from `2022-09-21` to `2022-10-11`. The goodness of our BSTNN model is compared with that of a state-of-the-art statistical estimation methodology, INLA [3]. Performances are evaluated using *Mean Absolute Error* (**MAE**) at regional level (Table 1 & Fig. 3 (**a**)) and on each test horizon (Fig. 3 (**b**)).

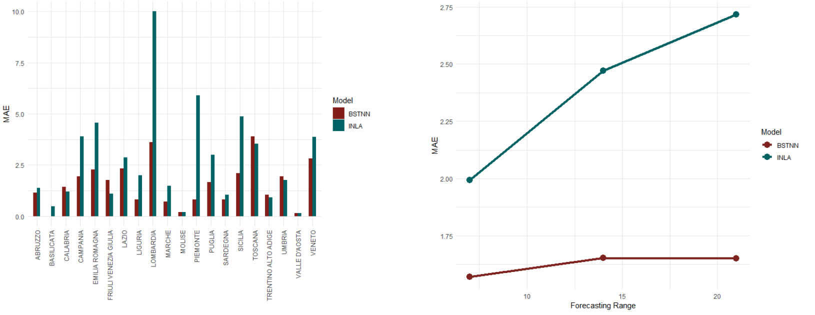
Table 1 shows that the BSTNN model performs, for most regions, better than INLA (best performer highlighted in bold). Another interesting result can be seen in Fig. 3 (**b**): as the forecast interval increases, the INLA (green line) forecasts significantly worsen in terms of MAE, while BSTNN (red line) predictions **21** days ahead, also deteriorates, but much less showing a more stable predictive power.

Region	BSTNN	INLA
ABRUZZO	**1.14**	1.38
BASILICATA	**0.00**	0.48
CALABRIA	1.43	**1.19**
CAMPANIA	**1.95**	3.90
EMILIA ROMAGNA	**2.38**	4.57
FRIULI VENEZIA GIULIA	1.86	**1.10**
LATIUM	**2.33**	2.86
LIGURIA	**0.81**	2.00
LOMBARDY	**4.00**	10.00
MARCHES	**0.71**	1.48
MOLISE	**0.19**	**0.19**
PIEDMONT	**0.81**	5.90
APULIA	**1.90**	3.00
SARDINIA	**0.81**	1.05
SICILY	**2.10**	4.86
TUSCANY	4.57	**3.52**
TRENTINO ALTO ADIGE	1.05	**0.90**
UMBRIA	2.05	**1.76**
AOSTA VALLEY	**0.14**	**0.14**
VENETO	**3.00**	3.86

Table 1: MAE scores for every model conditioned on each region. The lower the score the better the model. In **bold** the best model for each region.

References

1. Ravenda, F., Cesarini, M., Peluso, S. & Mira, A. Spatio-Temporal Forecasts with Bayesian Neural Networks. In Progress.
2. Wikle, C., Zammit-Mangion, A. & Cressie, N. Spatio-temporal Statistics with R. (Chapman,2019)



(a)

(b)

Fig. 3: Panel (a) shows the MAE scores for every model in each region: green (red) lines represent INLA (BSTNN) performances. Panel (b) shows how MAEs worsen as the forecast range increases (**7**, **14**, and **21** days) for INLA (green) and BSTNN (red).

3. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal Of The Royal Statistical Society: Series B (statistical Methodology). **71**, 319-392 (2009)
4. Hancock, J. & Khoshgoftaar, T. Survey on categorical data for neural networks. Journal Of Big Data. **7**, 1-41 (2020)
5. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining. pp. 855-864 (2016)
6. Guo, C. & Berkhahn, F. Entity embeddings of categorical variables. ArXiv Preprint ArXiv:1604.06737. (2016)
7. Maaten, L. & Hinton, G. Visualizing data using t-SNE.. Journal Of Machine Learning Research. **9** (2008)
8. Jin, J., Dundar, A. & Culurciello, E. Flattened convolutional neural networks for feedforward acceleration. ArXiv Preprint ArXiv:1412.5474. (2014)
9. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Computation. **9**, 1735-1780 (1997)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention is all you need. Advances In Neural Information Processing Systems. **30** (2017)
11. Dellaportas, P., Forster, J. & Ntzoufras, I. On Bayesian model and variable selection using MCMC. Statistics And Computing. **12**, 27-36 (2002)
12. Kingma, D. & Welling, M. Auto-encoding variational bayes. ArXiv Preprint ArXiv:1312.6114. (2013)
13. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. International Conference On Machine Learning. pp. 1050-1059 (2016)
14. Dürr, O., Sick, B. & Murina, E. Probabilistic deep learning: With python, keras and tensorflow probability. (Manning Publications,2020)

Oracle-LSTM: a neural network approach to mixed frequency time series prediction

Alessandro Bitetto, Paola Cerchiello

Abstract In the context of macro-economic indicators there are two main concerns regarding the frequency of the variables. The first is related to MIxed DAta Sampling (MIDAS), i.e. some indicators are reported annually, some quarterly, others monthly. The second deals with the need of forecasting predictions between reporting dates, e.g. before the end of the year, and it is known as ”nowcasting”. Existing methods rely on the alignment of high-frequency input data to low-frequency target variable by the means of lagged variables and their temporal-decaying weighting. We develop a two-steps algorithm that makes use of two Recurrent Neural Networks. The first, called Oracle, is a Deep Autoregressive network and predicts the target variable at high-frequency given past information. The second, called Predictor, is Long-Short Term Memory (LSTM) network and learns the relationship between Oracle’s predictions and high-frequency input data. The prediction error is a weighted average of three terms: the first two compare the observed low-frequency target with predictions of both Oracle and Predictor, respectively, the other compares the Predictor’s high-frequency predictions with the Oracle’s ones. Our model is tested on both simulated data, where we know the generated high-frequency data, and real macro-economic data. Our results show better performances compared to classical approaches.

Key words: Mixed frequency data; Artificial Neural Networks; LSTM; Nowcast

---

Alessandro Bitetto
University of Pavia, San Felice al Monastero, Pavia 5, Italy, e-mail: alessandro.bitetto@unipv.it

Paola Cerchiello
University of Pavia, San Felice al Monastero, Pavia 5, Italy, e-mail: paola.cerchiello@unipv.it

1 Introduction

This paper arises in the context of prediction with mixed-frequency time series. We want to cover the problems that usually emerge when dealing with forecasting by the means of macro-economic indicators. The first problem is related to different frequencies of both input and target time series. Generally, some indicators like GDP are reported annually and some official proxies can be found at most with quarterly frequency. Instead, other useful predictors, such as interest rates and stock prices, are available at daily or even hourly/minutes/seconds frequencies. This leads to asymmetry in the available information and usually summary aggregations are required in order to align all the time series and to apply any prediction algorithm. The second problem deals with the delays in reporting the official values of macro-economic indicators, that are available after the expected reference date, e.g. the GDP annual figures are usually published at the end of the first quarter of the following year. Therefore, there is the need to get the expected values right after the end of the year and to monitor the evolution of the indicator at higher frequencies, e.g. to have monthly estimates of GDP. This problem is usually referred to ”nowcast” [1]. Within the existing literature, several methods have been proposed, ranging from classical econometric models to more complex Artificial Neural Networks (ANN) architectures. The milestone in modelling mixed frequency time series was the mixed frequency data sampling model (MIDAS) [2], which extends the auto-regressive structure with a wise alignment of the different frequencies and their weighted lagged contribution subjected to parsimony constraints. It has the advantage of requiring few parameters to be estimated, thanks to the regularisation constraints, despite of the polynomial limitation in modelling the time series. Recent ANN architectures [3, 4, 5] push forward the same idea by taking advantage of their ability of non-parametric modelling the non-linear relationship in the data. They have the advantage of capturing non-linearities beyond the polynomial form but, due to their highly complex architecture, they have a big number of parameters to be estimated, thus requiring large amounts of data. It is important to notice that, although daily and monthly data can account for large numbers of data points, annual indicator are, at their best, limited to hundreds of observations. Therefore, ANN can be trained with at most hundreds of samples, a number that can hardly be sufficient for achieving good and reliable performance. Trying to combine the advantages of both previous approaches, we develop a two-steps algorithm that makes use of two Recurrent Neural Networks. The first, called Oracle, predicts the low-frequency target variable at the given high-frequency past information. The second, called Predictor, learns the relationship between Oracle’s high-frequency predictions and high-frequency input data. Our model is tested on both simulated data, where we know the generated high-frequency data, and real macro-economic data. Our preliminary results show better performances compared to classical approaches, motivating further investigation and improvements on the proposed methodology.

2 Methods

The classical approach of MIDAS [2] is formally different from auto-regressive models, also known as distributed lag models, because the predictors' lags have different frequencies. Supposing we have a single low-frequency target variable and k high-frequency predictors, the model can be expressed as:

$$y_t - \alpha_1 y_{t-1} - \cdots - \alpha_p y_{t-p} = \sum_{i=0}^{k} \sum_{j=0}^{l_i} \beta_j^{(i)} x_{tm_i - j}^{(i)} + \varepsilon_t, \tag{1}$$

where y_t is the low-frequency univariate process, p is the lag for the target variable, $x_\tau^{(i)}$, $i = 1, \ldots, k$ is the i-th high-frequency process with frequency m_i (e.g. $m_i = 30$ if $x^{(i)}$ is daily and y_t is monthly), so that $\tau = (t-1)m_i + j$, $j = 1, \ldots, m_i$, l_i is the number of high-frequency timesteps within each low-frequency interval and ε_t is the error. Eq. 1 can be written into a compact form:

$$\boldsymbol{\alpha}(B) y_t = \boldsymbol{\beta}(L)^\top \boldsymbol{x}_{t,0} + \varepsilon_t, \tag{2}$$

where B and L are the lag operators for y_t and $x^{(i)}$, respectively. In order to estimate the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the target and input variables must be aligned, meaning that all the high-frequency predictors $x^{(i)}$ must be lagged by the corresponding m_i. For example, given a quarterly y_t, the monthly predictor $x_t^{(1)} = x_t$ and the weekly predictor $x_t^{(2)} = z_t$, so that $m_1 = 3$ and $m_2 = 12$, we have:

$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} \alpha_1 + \begin{bmatrix} x_6 & \ldots & x_1 \\ \vdots & \vdots & \vdots \\ x_{3n} & \ldots & x_{3n-5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_5 \end{bmatrix} + \begin{bmatrix} z_{24} & \ldots & z_1 \\ \vdots & \vdots & \vdots \\ z_{12n} & \ldots & z_{12n-23} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \vdots \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{3}$$

where $\beta_t^{(1)} = \beta_t$ and $\beta_t^{(2)} = \gamma_t$. It is worth noticing that each row of Eq. 3 corresponds to a sample's observation and that the number of lags to be used is selected a-priori. At this point, any estimation algorithm can be used to estimate the parameters. This leads to the Unrestricted MIDAS (U-MIDAS). However, given the high number of parameters, it is possible to impose some sparsity constraints on the $\boldsymbol{\beta}$, so that the contribution of each fixed lag can be weighted with an appropriate decay function $\Phi(\boldsymbol{\delta_i})$, where $\boldsymbol{\delta_i}$ are the parameters of the function for each predictor i. The aim of this function is to push as many $\beta_t^{(i)}$ to zero. In order not to add further parameters to be fitted, the Φ function should not be too complex and low-order polynomials or exponentials are usually used. Recently, the MIDAS model has been coupled with Artificial Neural Networks. In [6], the authors propose the ANN-MIDAS where the high-frequency time series are firstly aligned and combined through the decaying weights and then fed into a fully-connected neural network to predict the low-frequency target. Both, number of lags and parameters of function Φ are estimated through back-propagation. Similarly, [7] adds Attention mechanism and Long-Short Term Memory (LSTM) block to the fully-connected part of the network. Attention automatically selects the appropriate number of lags for each series, and selects the most relevant features to be used in the LSTM recurrent block. As explained in Section 1, the MIDAS model has the limitation of modelling only a small set of non-linearities, whereas the ANN-MIDAS can suffer from the small amount of available target variables, which all high-frequency time series must be aligned to. Our suggested model combines two neural networks: one, defined as *Oracle*, aims to produce reliable high-frequency predictions of the low-frequency target variable, the other, defined as *Predictor*, learns the relationship between the predicted high-frequency target and the high-frequency input variables. In this way the Oracle can generate many target observations at once, that are then used as targets by the Predictor. Instead, in the setting of MIDAS/ANN-MIDAS only the observed low-frequency target samples can be used. For the sake of simplicity, let's suppose a monthly target variable y_t and a single daily predictor x_t, as shown in Fig. 1.

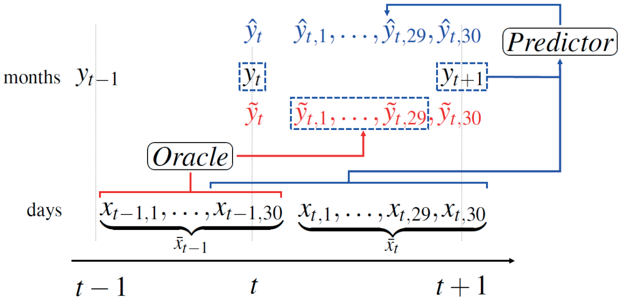

$\hat{y}_t$
$\hat{y}_{t,1},\dots,\hat{y}_{t,29},\hat{y}_{t,30}$
Predictor
months
y_{t-1}
y_t
y_{t+1}
$\tilde{y}_t$
$\tilde{y}_{t,1},\dots,\tilde{y}_{t,29},\tilde{y}_{t,30}$
Oracle
days
$x_{t-1,1},\dots,x_{t-1,30}$
$x_{t,1},\dots,x_{t,29},x_{t,30}$
$\bar{x}_{t-1}$
$\bar{x}_t$
$t-1$
t
$t+1$

Fig. 1 Architecture of Oracle-LSTM network.

The Oracle is a Deep Autoregressive network, consisting in a multilayer perceptron (MLP) $f : \mathcal{X}_1 \mapsto \mathbb{R}$, $\mathcal{X}_1 \subset \mathbb{R}^{l_1}$ a fully-connected ANN, that takes as input l_1 lagged observations of $x \in \mathcal{X}_1$ and returns the unknown daily prediction $\tilde{y} \in \mathbb{R}$ of y. The predicted $\tilde{y}$ are then used as target variable for the Predictor, namely a neural network with LSTM block $g : \mathcal{X}_2 \times \mathcal{Y} \mapsto \mathbb{R}$, $\mathcal{X}_2 \subset \mathbb{R}^{l_2}, \mathcal{Y} \subset \mathbb{R}^{l_1}$ that takes as input l_2 lagged observations of $x \in \mathcal{X}_2$ and returns the expected prediction $\hat{y} \in \mathbb{R}$ of $\tilde{y}$. During the training phase, the parameters of both networks are updated sequentially through back-propagation. During each epoch of the training, depending on the fixed lags l_1 and l_2, both f and g are repeatedly applied with a rolling window so as to predict all required $\tilde{y}$ and $\hat{y}$. Supposing being at prediction step t, where t corresponds to the last day of the month, we define $\bar{X}_{t-1} = \{x_{t-1,1},\dots,x_{t-1,30}\}$ the set of daily observations between $t-1$ and t, so that $(t-1,30) = t$. Then, we have:

$$f(x_f) = [\tilde{y}_{t,1},\dots,\tilde{y}_{t,29}] \tag{4}$$

$$g(x_g;[y_t,\tilde{y}_{t,1},\dots,\tilde{y}_{t,29},y_{t+1}]) = [\hat{y}_t,\hat{y}_{t,1},\dots,\hat{y}_{t,30}] \tag{5}$$

where $x_f \in \bar{X}_{t-1}, x_f \in \mathbb{R}^{l_1}$ and $x_g \in \{X_{t-1,30}\} \cup \bar{x}_t, x_g \in \mathbb{R}^{l_2}$. Once both models made the predictions, the loss for back-propagation is evaluated as the weighted average of three terms:

$$\mathscr{L} = w_1 \mathscr{L}_f([y_t, y_{t+1}], [\tilde{y}_t, \tilde{y}_{t,30}]) \tag{6}$$

$$+ w_2 \mathscr{L}_g([y_t, y_{t+1}], [\hat{y}_t, \hat{y}_{t,30}]) \tag{7}$$

$$+ w_3 \mathscr{L}_{f,g}([\tilde{y}_t, \tilde{y}_{t,1}, \dots, \tilde{y}_{t,30}], [\hat{y}_t, \hat{y}_{t,1}, \dots, \hat{y}_{t,30}]) \tag{8}$$

where $\mathscr{L}_f$ and $\mathscr{L}_g$ are the reconstruction loss of the observed low-frequency target and the prediction of the Oracle and Predictor, respectively, and $\mathscr{L}_{f,g}$ is the reconstruction loss of the Oracle's estimations $\tilde{y}$ and Predictor's predictions $\hat{y}$. The weights w_i are defined as hyperparameters and the Root Mean Squared Error (RMSE) is used for all three losses in order to normalize each contribution that is evaluated on a different number of observations. The parameters to be estimated are then the weights of the two models. Given the temporal nature of the data, we tune the parameters with a rolling-window cross-validation approach, where the window's size depends on the dataset's temporal length.

3 Data

Two sets of data have been tested: a simulated one and a real one. For the simulated one a mix of sinusoidal function and time-decay function have been combined so as to simulate seasonality with different periodicity and trends. Then, a target variable has been evaluated applying a different non-linear function to each input variable with the addition of noise. The input variables have been simulated with monthly pace and only values at the end of the year have been retained for the target variable. For the real dataset, annual GDP, monthly inflation, unemployment, balance of trades and current account for Italy from 1960 to 2021 have been used. To make the comparison fair, the same amount of data has been generated in the simulated dataset and target variables have been scaled into $[0,1]$ interval.

4 Empirical findings

Our preliminary results provide evidence that the proposed model performs better on both datasets compared to MIDAS and ANN-MIDAS. Table 1 reports the average RMSE for train and test set of the fitted models. Our model has a significantly lower error than MIDAS and improves the performance of ANN-MIDAS.

The results show that our model can be a good candidate in the set of tools used by central banks and governments to nowcast macroeconomic variables in a reliable way. Therefore, further investigation and improvements are required to assess robustness and interpretability of the predictions. Additional sensitivity test will be

Table 1 Average RMSE for train and test for MIDAS, ANN-MIDAS and Oracle-LSTM.

	MIDAS		ANN-MIDAS		Oracle-LSTM	
Dataset	train	test	train	test	train	test
Simulated	0.275	0.331	0.134	0.176	**0.105**	**0.122**
Real	0.318	0.379	0.197	0.212	**0.168**	**0.193**

performed and explainable methods must be applied to understand which input variable affect the most the model's outcomes. In particular, different architectures for both Oracle and Predictor should be tested, as well as mixed-frequency input data. Model-agnostic methods (such as Shapley values) should be used as benchmark and model-dependent (such as gradient-based explainer) should be applied for feature importance evaluation.

References

1. Bańbura, M., Giannone, D., Modugno, M., Reichlin, L. Now-casting and the real-time data flow. In Handbook of economic forecasting, vol. 2, 195–237 (2013).
2. Ghysels, E., Santa-Clara, P., and Valkanov, R. The MIDAS touch: Mixed data sampling regression models. UCLA: Finance. (2004)
3. Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In 8th International Conference on Learning Representations, ICLR 2020, (2020)
4. Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, Tomas Pfister, Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting, Volume 37, Issue 4, (2021)
5. Yitian Chen, Yanfei Kang, Yixiong Chen, Zizhuo Wang, Probabilistic forecasting with temporal convolutional neural network, Neurocomputing, Volume 399, (2020)
6. Qifa Xu, Xingxuan Zhuo, Cuixia Jiang, Yezheng Liu, An artificial neural network for mixed frequency data, Expert Systems with Applications, Volume 118, (2019)
7. Li, X., et al.: Attention-based novel neural network for mixed frequency data. CAAI Trans. Intell. Technol. 6(3), 301– 311 (2021)

Streamlined Variational Inference for Modeling Italian Educational Data

Gioia Di Credico, Claudia Di Caterina, Francesco Santelli

Abstract The streamlined version of the mean field variational Bayes (MFVB) algorithm for linear mixed models with crossed random effects allows simplifying calculations but may require one group's dimension to be moderate.Data collecting high school students' first term evaluations and INVALSI scores for Italian and Maths subjects perfectly comply with this setting: students are a vast random sample of those who enrolled at the university in 2019/20, while the number of tests is limited to 6. Three different MFVB product restrictions with incremental complexity are evaluated. All of them are convenient with respect to classic MCMC solutions from both a computational and a memory storage viewpoint. Results and interpretation of model coefficients are in line with the literature on educational data.

Key words: Crossed random effects, INVALSI, Mean field variational Bayes.

1 Introduction

Linear mixed-effects models are commonly used to analyze data with a continuous Gaussian outcome and multiple levels of variability arising from a grouped data structure. In order to account for the variability introduced by the nested or crossed structure of the observations, it may be convenient to include random effects treated as random variables in the model.

In the following, we focus on a crossed-data application where two levels of variability exist. Crossed designs imply that each combination of the group levels is represented in the data [1]. Our data refer to a random sample of students who enrolled in a Bachelor program at an Italian university in the academic year 2019/2020. Out-

---

Gioia Di Credico
Università degli studi di Trieste, Trieste, e-mail: `gioia.dicredico@deams.units.it`

Claudia Di Caterina
Università degli studi di Verona, Verona, e-mail: `claudia.dicaterina@univr.it`

Francesco Santelli
Università degli studi di Trieste, Trieste, e-mail: `fsantelli@units.it`

comes consist of students' evaluations during their 10th and 13th high school grades. Here, students and the type of tests define our two-group crossed-data design. As the groups size increases, model estimation gets slower and may even become unfeasible. Streamlined variational inference has recently been applied to overcome these estimation difficulties in random-effects models, e.g. by [3] for the nested group structure and by [4] for the crossed one. The key idea relies on the sparseness of the matrix to be inverted, which enables a quicker computation and less storage capacity. Nested data imply a so-called arrowhead block structure for this matrix [5], and non-zero sub-blocks can be easily identified to simplify the calculation of its inverse. In the crossed-data case studied here the matrix is less sparse, however, considering the most accurate restriction, the streamlined approach offers advantages when one group is moderate in size. Our motivating data exhibit such a feature: the number of students involved is very large (around 7000), while the tests whose mark is recorded on each student are limited to 6.

2 Methods

For each ith student, we assume the scores $\boldsymbol{y}_{ii'}$ on test i' follows a linear mixed model with two crossed random effects:

$$\begin{gathered}
\boldsymbol{y}_{ii'}|\boldsymbol{\beta},\boldsymbol{u}_i,\boldsymbol{u}'_{i'},\sigma^2 \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ii'}\boldsymbol{\beta}+\boldsymbol{Z}_{ii'}\boldsymbol{u}_i+\boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'},\sigma^2\boldsymbol{I}), \quad i=1,\dots,m, \\
\boldsymbol{u}_i|\boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}), \qquad \boldsymbol{u}'_{i'}|\boldsymbol{\Sigma}' \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}'), \quad i'=1,\dots,m',
\end{gathered} \tag{1}$$

where $\boldsymbol{X}_{ii'}$ is the $n_{ii'} \times p$ design matrix, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$, respectively of dimension $n_{ii'} \times q$ and $n_{ii} \times q'$, are the random effects matrices, $\boldsymbol{\beta}$ is the p-vector of fixed-effect coefficients, $\boldsymbol{u}_i$ and $\boldsymbol{u}'_{i'}$, respectively $q \times 1$ and $q' \times 1$, are the vectors of random effects, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are their $q \times q$ and $q' \times q'$ respective covariance matrices and σ^2 is the error variance.

The joint *a priori* density of the p fixed effects is $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta})$. For the error variance σ^2 and the random effects covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$, we consider the following family of marginally non-informative prior distributions [2]:

$$\begin{gathered}
\sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2},1/a_{\sigma^2}), \quad a_{\sigma^2} \sim \text{Inverse-}\chi^2(1,1/(\nu_{\sigma^2}s^2_{\sigma^2})), \\
\boldsymbol{\Sigma}|\boldsymbol{A_\Sigma} \sim \text{Inverse-G-Wishart}(G_{\text{full}},\nu_{\boldsymbol{\Sigma}}+2q-2,\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}}), \\
\boldsymbol{\Sigma}'|\boldsymbol{A_{\Sigma'}} \sim \text{Inverse-G-Wishart}(G_{\text{full}},\nu_{\boldsymbol{\Sigma}'}+2q'-2,\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}'}), \\
\boldsymbol{A_\Sigma} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A_\Sigma}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A_\Sigma}} = \{\nu_{\boldsymbol{\Sigma}}(s^2_{\boldsymbol{\Sigma},1},s^2_{\boldsymbol{\Sigma},2})\}^{-1}, \\
\boldsymbol{A_{\Sigma'}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A_{\Sigma'}}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A_{\Sigma'}}} = \left\{\nu_{\boldsymbol{\Sigma}'}(s^2_{\boldsymbol{\Sigma}',1},s^2_{\boldsymbol{\Sigma}',2})\right\}^{-1}.
\end{gathered} \tag{2}$$

In our application, the first group of random effects $\boldsymbol{u}_i$ ($i=1,\dots,m=7005$) corresponds to students enrolled at an Italian university in 2019/2020, and the second group $\boldsymbol{u}'_{i'}$ ($i'=1,\dots,m'=6$) corresponds to scores from assessments of Italian and

Math skills. Specifically, for each student, a written and oral score was recorded at the end of the first term and one written standardized INVALSI score (see Section 3) was recorded at the end of the final term. Each combination student/test, corresponding to the pair (i,i'), is observed $n_{ii'} = n = 2$ times, namely in the 10th and 13th grades of high school. The design matrix $\boldsymbol{X}_{ii'}$ has $p = 32$ columns, including the intercept. Moreover, $q = q' = 2$ because we consider both random intercepts and random slopes for the two groups, meaning

$$\boldsymbol{Z}_{ii'} = \boldsymbol{Z}'_{ii'} = \left[1\ x_{1,ii'j}\right]_{j=1,2},$$

where $x_{1,ii'j} = 1,2$ is the year indicator encoding the two high school grades. According to (1) and this set-up, the two scores of the ith student on the i'th test are modeled to be

$$y_{ii'j}|\boldsymbol{\beta},u_{0i},u_{1i},u'_{0i'},u'_{1i'},\sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0+u_{0i}+u_{0i'}+(\beta_1+u_{1i}+u_{1i'})x_{1,ii'j}+\sum_{k=1}^{31}\beta_k x_{k,ii'j},\sigma^2)$$

for $j=1,2$. The formula above shows that this modelling strategy allows for a different intercept and slope for every student/test combination. Heterogeneities among intercepts and slopes are defined by appropriate entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$.

We consider three product restrictions on the mean field approximation of the joint conditional density function of all parameters in (1) with covariance priors (2) ([4], Sect. 3):

$$q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}') = \begin{cases} q(\boldsymbol{\beta})q(\boldsymbol{u})q(\boldsymbol{u}')\,q(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{restriction I,} \\ q(\boldsymbol{\beta},\boldsymbol{u})q(\boldsymbol{u}')\,q(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{restriction II,} \\ q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')\,q(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{restriction III.} \end{cases} \quad (3)$$

Product restriction I has the simplest streamlined implementation and scales well to very large problems, but may produce small posterior variances as it sets all posterior correlations between $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{u}'$ to zero. Conversely, product restriction III allows for a full joint posterior covariance matrix of $(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')$, leading to higher inferential accuracy but challenging computing that can be streamlined for limited m'. A compromise is given by product restriction II, which includes posterior correlations between $\boldsymbol{\beta}$ and $\boldsymbol{u}$, for $\boldsymbol{u}$ larger than $\boldsymbol{u}$. For all the product restrictions, the prior distributions specification 2 leads to a fully factorization of the q-densities related to the covariance matrix and auxiliary variables [4].

The q-density parameters can be obtained using a coordinate ascent iterative algorithm. However, if applied naïvely, the potentially prohibitively large matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u})}$ requires storage and inversion. Product restrictions I, II and III lead to streamlined mean field variational Bayes (MFVB) algorithms with varying degrees of storage and computational overhead (see [4], Sect. 4).

3 Italian students' proficiency data

Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Università e Ricerca', and INVALSI. The dataset is known as MOBYSU. In the Italian School System, the scholastic assessment is in charge of the "Italian national institute for the evaluation of the school system" (INVALSI), which uses a set of standardized tests to evaluate the proficiency of students attending different schools at different years. Several domains are tested, and the main domains are Mathematical skills, English language, Italian language, and Science.

Our data regard the cohort of pupils that finished high school, achieved the Diploma in 2018/19, and then enrolled at university in 2019/20. Such students are more than 240000. To be included, students must have never failed a scholastic year, and must attend high school for the first time in the Italian school system. The response variable refers to students marks: four recorded at the end of the first term (Italian and Math, written and oral) and two throughout the INVALSI test (Italian and Math, written), during their 10th and 13th high school grades. Predictors involved in the analysis are listed in Tab.1. They include information on the socio-economic background, parental occupation and demographics.

Table 1 Model predictors, their description and reference categories.

Variable	Description	Reference
Gender	Male, Female	Female
Age	Reception (one year ahead), Regular	Regular
Nation	Foreigner, Italian	Italian
Student escs (EscsStud)	student socio-economic level	
School escs (EscsSch)	school socio-economic level	
School type (SchTy)	13 categories	Classical Lyceum
Work Mother (Work.M)	5 categories	Unemployed
Work Father (Work.F)	5 categories	Unemployed
Year	School year	
NUTS2 classification (NUTS2)	5 areas	Center
School	Private, Public	Public

4 Analysis and results

First term and INVALSI scores were centered to the national means and INVALSI were also scaled to standardize them to a common range and adapt to the prior distributions setting. Furthermore, we excluded students with missing information so that valid and complete data refer to 21228 students. The final model is fitted on a random sample of 33% of the units, corresponding to 7005 total pupils.

As hyperparameters, we set $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$, $\nu_{\sigma^2} = 1$, $\nu_{\boldsymbol{\Sigma}} = \nu_{\boldsymbol{\Sigma}'} = 2$, $s_{\sigma^2} = s^2_{\boldsymbol{\Sigma},1} = s^2_{\boldsymbol{\Sigma},2} = s^2_{\boldsymbol{\Sigma}',1} = s^2_{\boldsymbol{\Sigma}',2} = 10^5$. For each product restriction, we run the MFVB algorithm for 100 iterations. Computational times were about 8 minutes for the MFVB with product restriction I, 11 and 21 times longer for the MFVB with product restriction II and III. The model 1 with prior distributions as in 2 was also simulated through MCMC in Stan. In particular, 4 chains of 2000 iterations each (1000 warm-up; 1000 sampling) were simulated. In the following, MCMC inference is based on the sampling step draws. The running time for the MCMC setup was of 19 hours.

As expected, variational inference on the random effects and error variance components are relatively affected by the MFVB product restriction used, giving very similar results (Tab. 2). Differences between MFVB and MCMC on the tests random effects variability are likely due to a slow convergence of the MCMC chains, advised by a low effective sample size on the $\boldsymbol{\Sigma}'$ parameters. The product restriction impact is evident on the estimated variability of the fixed effects. While approximate posterior means of the fixed effects are the same across product restrictions, the least accurate (I) strongly underestimates the variability, while restrictions II and III lead to very similar results on all the coefficients, except for the year variable (see Fig. 1). On average, fixed marginal effects suggest that male students and those from islands and Southern Italy regions perform worse, while the Northeast is the area with best proficiency. The socio-economic status dimension has a significant positive effect, as expected, both at the individual and school levels. Lyceums record the best scores on average, with Scientific lyceum overperforming all the others. When parents are less involved in demanding jobs, students usually perform better. No clear effects are found for Italian nationality and students one year ahead. In both random intercept and slope, the variability carried by the student group is slightly larger than the item group one.

Table 2 Random effects standard deviation estimates (approximate posterior mean). Square root of diagonal entries of $\hat{\boldsymbol{\Sigma}}$ ($\hat{\boldsymbol{\Sigma}}'$) are denoted by $\hat{\sigma}_1$ ($\hat{\sigma}'_1$) and $\hat{\sigma}_2$ ($\hat{\sigma}'_2$). Correlation between random intercept and slope is denoted by $\hat{\rho}$ ($\hat{\rho}'$).

	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\rho}$	$\hat{\sigma}'_1$	$\hat{\sigma}'_2$	$\hat{\rho}'$	$\hat{\sigma}$
MFVB I	0.677	0.114	0.252	0.540	0.087	-0.578	0.931
MFVB II	0.678	0.114	0.251	0.540	0.087	-0.578	0.931
MFVB III	0.678	0.114	0.251	0.591	0.095	-0.523	0.931
MCMC	0.679	0.114	0.251	0.883	0.189	-0.201	0.931

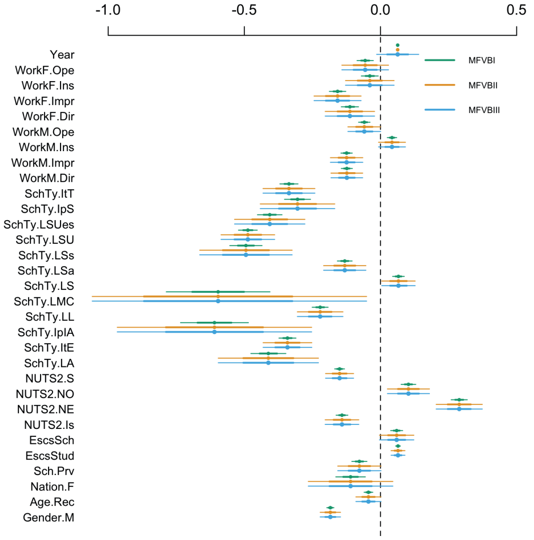


Fig. 1 Fixed effects: approximate posterior means (dots) and 95% credible intervals for the MFVB with the three product restrictions (I in green, II in orange and III in blue).

5 Conclusions

The work analysed Italian students' proficiency data using the streamlined MFVB algorithm based on three product restrictions. The code is not optimized and computational times are reported for comparative purposes only. Even so, the MFVB algorithms are much faster than standard MCMC solutions. Comparing the three MFVB product retrictions, the second appears to be an excellent compromise balancing estimation speed and results accuracy.

References

1. Baayen, R. H., Davidson, D. J., Bates, D. M.: Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. (2008) doi: 10.1016/j.jml.2007.12.005
2. Huang, A., Wand. M. P.: Simple marginally noninformative prior distributions for covariance matrices.Bayesian Anal. (2013) doi: 10.1214/13-BA815
3. Lee, C. Y. Y., Wand, M. P.: Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. Biometrical J. (2016) doi: 10.1002/bimj.201500007
4. Menictas, M., Di Credico, G., Wand, M. P.: Streamlined variational inference for linear mixed models with crossed random effects. J. Comput. Graph. Stat. (2022) doi: 10.1080/10618600.2022.2096622
5. Nolan, T. H., Wand, M. P.: Streamlined solutions to multilevel sparse matrix problems. ANZIAM J. (2020) doi: 10.1017/S1446181120000061

The use of magnetic resonance images for the detection and classification of brain cancers with D-CNN

Davide Mascolo, mascolo.2001991@studenti.uniroma1.it
Leonardo Plini, plini.2000543@studenti.uniroma1.it
Alessandro Pecchini, pecchini.1824164@studenti.uniroma1.it
Margaret Antonicelli, margaret.antonicelli@uniroma1.it

Sapienza University of Rome, Italy

Abstract Among the various oncological pathologies, brain cancer continues to be one of the most wide-spread, as well as lethal, diseases. Within this paper we used Keras and Tensorflow to implement state-of-the-art convolutional neural network (CNN) architectures, such as EfficientNetB0, Res-Net50 and VGG16, using Transfer Learning to detect and classify three types of brain tumors namely say – Glioma, meningioma, and pituitary. The dataset we used consisted of 3264 2-D MRI images and 4 classes. Due to the small number of images, various data augmentation techniques were used to increase the size of the dataset. Our proposed methodology consists not only of data augmentation, but also of various techniques of image denoising, skull strip-ping, cropping and bias correction. In our working proposal, the EfficientNetB0 architecture gave the best results providing a very high accuracy. The purpose of this document is to distinguish between normal and abnormal pixels and classify them more accurately.

Key words: Deep Learning, Convolutional Neural Network, Glioma, Meningioma, Pituitary, Transfer Learning

1 Introduction

A brain tumor is an abnormal growth of cells in brain tissue, which can be benign or malignant. The central nervous system is part of the brain and consists of two parts: the brain and the spinal cord. Together they control both voluntary functions, such as walking, talking, etc., and involuntary functions, such as breathing, digestion, and so on. The central nervous system is also the basis of sensory functions (sight, smell, touch, hearing and taste), of emotions and of all the so-called higher activities such as memory and learning.

Typically, the process of diagnosing a brain tumor begins after the patient has seen their general practitioner about the onset of symptoms. The doctor then evaluates the need for an in-depth study with the neurology specialist or prescribes instrumental diagnostic tests. In some cases, however, the onset of symptoms is sudden and requires urgent evaluation in the emergency department.

Magnetic resonance imaging (MRI) of the brain is the main test in case of suspicion of a brain tumor. Compared to CT, nuclear magnetic resonance (MRI) with and without the use of paramagnetic contrast medium (gadolinium), allows to identify lesions and nodules, provides 3D (three-dimensional) images, allows to identify the site, the dimensions, the extension of the disease and the relationships with the surrounding structures, the so-called ”eloquent” areas. Furthermore, the use of functional methods in MRI (diffusion and perfusion) can provide further information on the cellularity and vascularization of the analyzed regions.

Based on 3264 2-D MRI images, advanced analysis models will be evaluated, with the support of machine learning techniques, with the aim of distinguishing between normal and abnormal pixels and classifying them more accurately.

2 Method

Recently, deep learning methods have demonstrated excellent performance in analyzing medical images compared to traditional machine learning methods.[3] This is mainly due to the ability to extract features during the learning process and to the possibility of optimizing the network parameters in order to minimize the error committed.

For this study several networks based on the CNNs architecture were used: more in details, two convolutional networks were implemented from scratch which represent the benchmark with respect to compare the pretrained networks ResNet50, VGG16 and EfficientNetB0, used through the transfer learning technique[1].

The typical pipeline of an image-based cancer classification study using CNNs[2] is shown in Fig. 1.

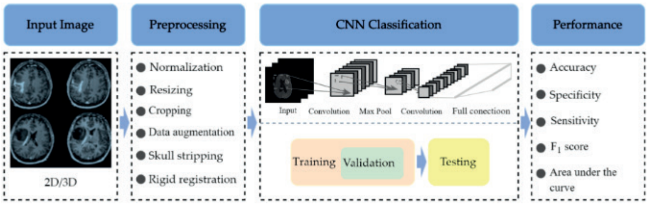


Fig. 1: Pipeline of CNNs-based Brain Cancer Classification

2.1 *Pre-processing*

As a first operation, the dataset was split into Training Set and Test Set. Subsequently the preprocessing operations were applied only on the Training data and during the training of the models, 10% of the Training data was used as Validation Set for parameters tuning. Fig. 2 shows the data split.

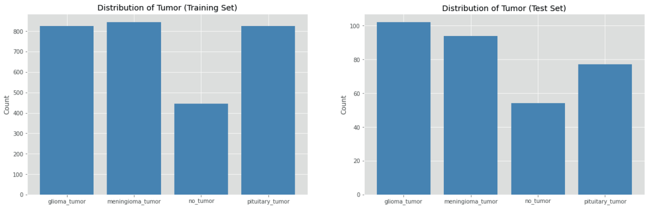


Fig. 2: Distribution of classes for Training Set and Test Set

2.1.1 Pre-processing of the image

1. Resize: The first pre-processing operation applied was to resize the image to a size of (150, 150).
2. Noise Reduction: To change the resolution of the original images, a Gaussian filter has been applied to reduce noise and blur the images.
   In Fig. 3 it's possible to see an example of an image filtered.

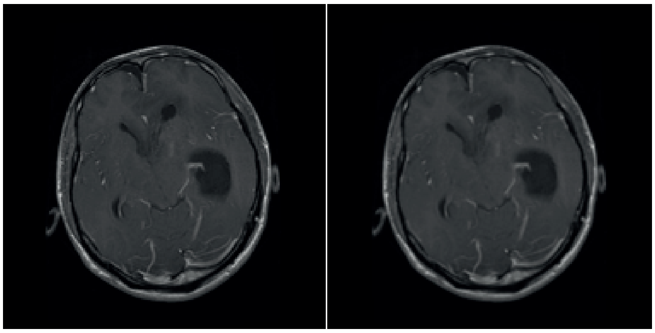


Fig. 3: Original Image and Filtered Image with Gaussian Method

3. Data Augmentation: To make the proposed methods more robust and reduce the overfitting effect, Data Augmentation was applied with various transformations, in particular:
   - rescale
   - rotation with an angle of 30 degrees
   - vertical and horizontal shift
   - image zoom
   - horizontal flipping

In Fig. 4 it is possible to observe the example of an original image and the respective transformations.

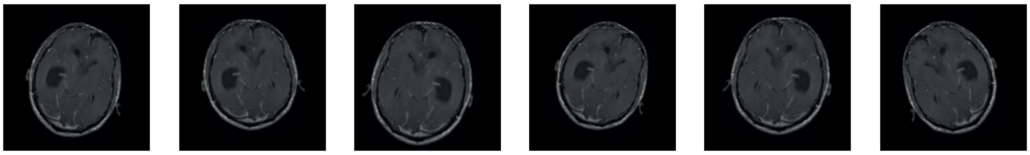


Fig. 4: Example of Original Image and Respective Transformations

4. One-Hot Encode: The last operation carried out was the transformation of the reference variable which was originally coded as numeric into a categorical variable.

2.2 Proposed Methodology

We chose to use a CNN architecture over any other model because CNN employs several convolutional filters to scan the whole feature matrix and perform dimensionality reduction, hence making them well suited for image processing and classification. Famous CNN architectures like VGG16, ResNet and EfficientNet-B0. All these architectures were implemented using Transfer Learning.

2.2.1 Configuration Model

To make it possible to compare the different classifiers, the same parameter configuration was used for the different models. To speed up the training a batch size of 32 was used and the models were trained for 15 epochs. The optimization algorithm chosen to train the models is Adam, because it is computationally efficient, easy to implement and works well for problems that involve large amounts of data or parameters.
The chosen loss function is the Categorical Loss-Entropy that is used to calculate and minimize the error of the model during the optimization process. This is a typical choice for a multi-class classification model.
The loss is defined by:

$$\sum_{i=1}^{outputsize} y_i log(\hat{y}_i) \tag{1}$$

To avoid overfitting and save the model with the best performance on the validation set, a model checkpoint procedure has been implemented. During the training epochs, the validation loss is monitored and if it does not decrease for two consecutive epochs, the Learning Rate is reduced. The parameters used were: factor = 0.3, patience = 2 and minimum value of delta = 0.001.
For the convolutional networks implemented from scratch, in addition to the above configuration, several techniques have been used to improve their performance. The first model contains 5 layers including input and output layers. Each layer consists of 2D convolution, Max Pooling 2D, Batch Normalization and Dropout operation. The final layer (Fully Connected Layer) is composed of 128 neurons connected to the 4 output neurons (one for each class to be provided) and managed by the softmax function. The activation function chosen is the ReLU function. For the dropout operation, the value of p is 0.2 and 0.3 for the fully connected layer.
In order to improve the performance of the first model, the second one contains 6 layers including input and output layers and the final layer is composed of 1024 neurons. In this case, the value of p is 0.25 and 0.4 for the fully connected layer. For performance evaluation, the following metrics were used: Precision, Recall, F1-Score.

3 Results and Discussion

In order to see the predictions of the models we employed a confusion matrix and drew conclusions from it. In this study we want to maximize the number of True Positive which means correctly classifying patients with the disease and minimizing False Negative, i.e. reducing the risk of not treating patients who are actually ill.

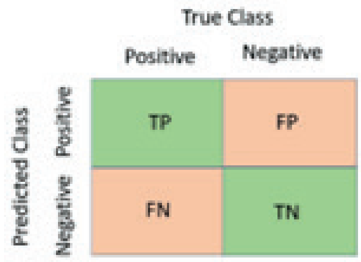


Fig. 5: Confusion Matrix

In Table 1 is possible to see the performances of the models,

Model	Precision	Recall	F1-Score
CNN (V.1)	0.79	0.78	0.78
CNN (V.2)	0.90	0.92	0.91
VGG 16	0.74	0.77	0.74
ResNet 50	0.95	0.97	0.96
Efficient Net - B0	0.97	0.95	0.97

Table 1: Comparison of performances

As we can observe from the table the best model is Efficient Net-B0, the result is not surprising considering the advanced architectu and the recent study of it. It is also interesting to note that the second convolutional network from scratch performs almost similarly the ResNet 50 network and even better than VGG 16. Despite several attempts and parameter configurations, the big loser is VGG 1 probably the poor performance is due to the few images contained in the dataset and to the network architecture.
In Figure 7 it is possible to observe the plot of the various metrics considered with respect to the number of epochs.
In Figure 8 we can see the Confusion matrix.

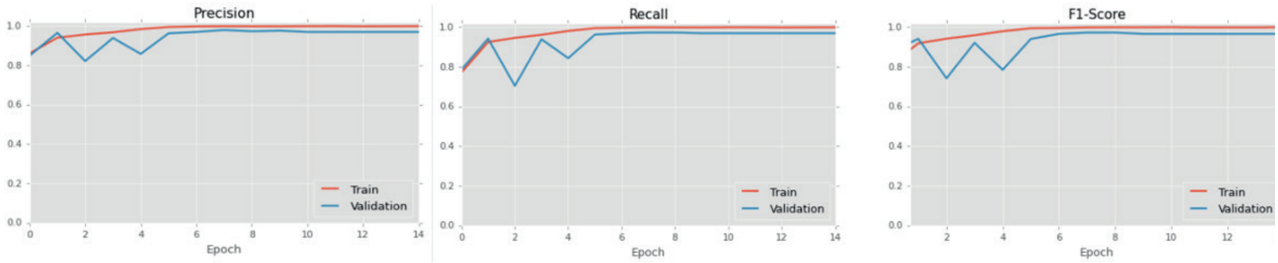


Fig. 6: Performance Metrics Vs. Epochs

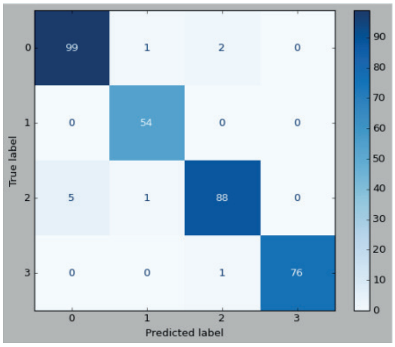


Fig. 7: Confusion Matrix Efficient Net - B0

For the interpretation of the confusion matrix, we remember that:

- 0: Glioma Cancer
- 1: No Cancer
- 2: Meningioma Cancer
- 3: Pituitary Cancer

4 Conclusions and Future Works

This study represents only one of the possible approaches to this type of problem. As we have seen, using a pretrained network does not always guarantee better performance than a network implemented from scratch and the results obtained strictly depend on the preprocessing operations carried out. In fact, for future works other preprocessing techniques could be applied such as image cropping, i.e. a technique that allows to modify the image by removing a portion that is not considered useful and focus attention on the part of the image needed for classification. Another improvement could be to perform object detection, i.e. detecting the actual location of the cancer with bounding boxes using segmentation and sophisticated detection algorithms such as YOLO (You Only Look Once) and SSD (Single-Shot Detector). In conclusion, future studies in this area may lead to better results, possibly using other pre-processing methods and further refining the model hyperparameters.

References

1. François Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: Proceed-ings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 1251–1258.
2. Nyoman Abiwinanda et al. “Brain tumor classification using convolutional neural network”. In: World congress on medical physics and biomedical engineering 2018. Springer. 2019, pp. 183–189.
3. Stefan Bauer et al. “A survey of MRI-based medical image analysis for brain tumor studies”. In: Physics in Medicine Biology 58.13 (2013), R97.

Modeling and clustering of traffic flows time series in a flood prone area

Modelli statistici e clustering per serie temporali dei flussi di traffico in un'area soggetta a inondazioni

Paola Zuccolotto, Giovanni De Luca, Rodolfo Metulini and Maurizio Carpita

Abstract Time series of traffic flows, recovered by mobile phone origin-destination signals, are used to monitor mobility and crowding in an area subject to flooding risk. We propose a time series model based on vector autoregressive with exogenous covariates combined to dynamic harmonic regression and a subsequent clustering procedure, aimed at obtaining groups of areas characterized by the common tendency to the occurrence of extreme events, that in this case study are extremely high incoming traffic flows.

Key words: Flooding risk, multivariate time series modelling, copula functions, tail dependence, time series clustering

1 Introduction

It is well known that extreme weather events often have huge social consequences for communities and individuals. Their immediate effects consist of loss of human life, devastation of crops, damage to goods, and deterioration of overall health and wealth conditions. Considered their social and economic impact, the statistical study of extreme weather phenomena can also be approached from a management perspective. In fact, natural Disaster Management ([7]) recommends the development

---

Maurizio Carpita, Paola Zuccolotto
Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 25122, Brescia, Italy e-mail: maurizio.carpita@unibs.it, paola.zuccolotto@unibs.it

Giovanni De Luca
Department of Management and Quantitative Sciences, University of Naples Parthenope, Via G. Parisi, 13, 80132 Naples, Italy e-mail: giovanni.deluca@uniparthenope.it

Rodolfo Metulini
Department of Economics, University of Bergamo, Via Caniana, 2, 24127, Bergamo, Italy e-mail: rodolfo.metulini@unibg.it

of a framework of exposure risk that can exploited in an early warning perspective. In this work we focus attention on floods. To draw flooding risk exposure maps is of fundamental importance, in order to face, in the best possible way, a flooding event. Such maps cannot ignore human presence and people mobility, but they traditionally assume a constant crowding over time. This assumption is far from reality, especially in metropolitan areas, so, a more detailed description of people presence and mobility is a critical issue to determine an accurate flooding risk. To do that, mobile phone network data have been used to obtain a dynamic monitoring of crowding in areas with hydrogeological criticality ([1]). Another approach consists of using mobile phone origin-destination signals, in order to recover information on traffic flows and then build statistical models, able to give accurate forecasts of people mobility. [6] proposed a model, based on combining vector autoregressive with exogenous covariates and dynamic harmonic regression. They applied the method to the case study of Mandolossa (an urbanized area subject to flooding, located on the western outskirts of Brescia) using hourly data from September 2020 to August 2021. The method worked quite well, but residuals exhibited a leptokurtic distribution with heavy tails determined by a number of extreme events (i.e., days with particularly high or low traffic flows). In this talk we propose to use the method of time series clustering based on copula functions, proposed by [3], in order to cluster the residuals time series with respect to their upper tail dependence. The aim of the analysis is to obtain clusters of areas for which extreme events (in terms of extremely high traffic flows) tend to occur together.

2 Data

Mobile Phone origin-destination (OD) data flows have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it) for the MoSoRe Project 2020-2022 and they refer to one year of hourly observations (from September 1st, 2020 to August 31st, 2021) of traffic among *Aree di CEnsimento* (ACEs) in the province of Brescia. OD data refer to the number of phone SIM cards connected to the TIM network that were retrieved during a 1-hour interval by the antenna in a given ACE i and, after five or more minutes, by the antenna in ACE j. [1] For each time interval t, and for selected ACEs (let say, i and j) in the province of Brescia, three types of flows are available: flows arriving in i (inflows), flows departing from i (outflows), and internal flows from i to i (internal).

---

[1] Two types of cards can be distinguished: human SIM (about 85% of the total SIM) and M2M technology machine SIM (about 15%). Since a user might have both a human SIM and some devices with an M2M machine SIM, we restricted our attention to human SIMs to avoid double counting of users.

3 Time series modeling

In this work we limit our attention to 4 ACEs inside the flood prone area of the Mandolossa and to 38 properly selected neighbour ACEs. Specifically, with the final aim of obtaining uncorrelated estimated residuals to whom perform the clustering, we estimate, for each single neighbour ACE j the following VAR model with eXogenous variables (**VARX**, [9]):

$$\mathbf{y}_{t,j} = \nu_j + \sum_{h=1}^{p} \mathbf{A}_{h,j}\mathbf{y}_{t-h,j} + \mathbf{B}_j\mathbf{x}_{t,j} + \varepsilon_{t,j}\,, j = 1,\ldots,38\,, \tag{1}$$

where $\mathbf{y}$ is a vector of length 3 made of inflows to i (where i is represented by the union of the 4 ACEs inside the Mandolossa), outflows from i and internal flows in i, and where $\mathbf{Bx}_t$ contains a two-way (i.e., daily and weekly periodicities) Dynamic Harmonic Regression (DHR) component (which is based on a combination of sine and cosine Fourier bases) and proper weekdays and month dummy variables.

The model recalls that used in [6], but lags of order smaller than 24 are here allowed. According to an AIC criterion, we model the DHR component by including 7 daily and 4 weekly Fourier bases. We then calibrate the model by choosing the autoregressive (AR) order based on the AIC, the Auto Correlation Function (ACF), the Partial ACF and the Ljung-Box test. After having tested different AR structures, we opted for a model with the first 25 lags (i.e., $p = 25$), that display a very limited autocorrelation with small values of ACF and with the Ljung-Box test almost always rejected (by varying the AR order). The final model has been used to obtain estimated residuals. Despite the analysis of all estimated residuals might be interesting, in this application we just use the ones related to the inflows, as they allow to cluster ACEs in terms of the dynamic of traffic to the area of the Mandolossa.

4 Time series clustering on upper tail dependence

In this Section we describe the clustering procedure we propose to define groups of time series for which extreme events (in this case, extremely high traffic flows) tend to occur together. To do that, we rely on the method originally proposed by [3], where time series clustering is performed on a dissimilarity matrix based on bivariate tail dependence coefficients, estimated by means of copula functions. A 2-dimensional copula ([8]) is a function denoted by

$$C : [0,1]^2 \to [0,1].$$

Given the random variables X_j, X_h, and their cumulative distribution functions $U_j = F_j(X_j), U_h = F_h(X_h)$, the 2-dimensional copula function applied to u_j, u_h, is equivalent to the joint distribution function,

$$C(u_j, u_h) = P(F_j(X_j) \le u_j, F_h(X_h) \le u_h)$$

that is

$$C(u_j, u_h) = F_X\left(F_j^{-1}(u_j), F_h^{-1}(u_h)\right).$$

Then

$$F_X(x_j, x_h) = C(F_j(x_j), F_h(x_h)).$$

Copula functions describe the joint distribution in a very flexible way, by combining the univariate marginal distributions of the variables and a copula function joining the margins. When a joint distribution is described by means of a copula function, some interesting features of the multivariate distribution can be easily recovered. Examples are the tail dependence coefficients (TDCs): given two variables X_j and X_j, the lower and upper TDCs are given, respectively, by

$$\lambda_{j|h}^{L} = \lim_{v \to 0^+} P(U_j \le v \mid U_h \le v)$$

and

$$\lambda_{j|h}^{U} = \lim_{v \to 1^-} P(U_j > v \mid U_h > v).$$

In case of tail independence, λ^L (λ^U) is null, while, when λ^L (λ^U) is in the range $(0,1]$ then the extremely low (high) values of the two variables are dependent, with stronger dependence as the coefficient value increases.

In this work we are interested to upper tail dependence, as events to be monitored are exceptionally high traffic flows. To cluster times series based on upper TDCs, the procedure proposed by [3] requires to obtain the Δ^S dissimilarity matrix Δ, containing the dissimilarities δ_{jh} between all the pairs of the N time series under study, with

$$\delta_{jh} = -\log(\lambda_{j|h}^{U}). \tag{2}$$

The dissimilarity matrix Δ_N is then used as a basis for the adopted clustering algorithm [3, 4, 2]. In this work we propose a clustering algorithm able to take into account, beyond dissimilarities, the spatial contiguity between areas. So, we introduce a new dissimilarity measure δ_{jh}^{θ} as a modification of (2),

$$\delta_{jh}^{\theta} = -\log(\lambda_{j|h}^{U}) + \theta c_{jh}, \tag{3}$$

where c_{jh} is a contiguity coefficient assuming value 0 when the jth and hth time series denote traffic flows coming from neighbouring areas, and 1 otherwise. The dissimilarity matrix obtained by (3) is denoted by Δ^{θ}.

The parameter $\theta > 0$ adjusts the impact of the contiguity coefficient in the dissimilarity between two time series, and has to be determined through an iterative procedure, as detailed in Algorithm 1.

Note that the quality of the clusterization is evaluated by the adopted internal clustering validation indices with reference to the dissimilarity matrix Δ. The rationale for this choice is that contiguity between areas is used to define a set of optimal clusterizations at given values of θ, but the final choice among them is done by

Algorithm 1 Upper tail dependence clustering with spatial structure

```
Require: Two dissimilarity matrices $\Delta$ and $\Delta^{\theta}$, obtained as in (2) and (3).
1: Define a sequence $\Theta$ of values, starting from 0, that could be plausible values for $\theta$ (e.g. $0.005, 0.01, 0.015, \ldots, 4$)
2: for $\theta$ assuming all the values in $\Theta$ do
3:     perform cluster analysis with a hierarchical agglomerative algorithm, using $\Delta^{\theta}$ as dissimilarity matrix
4:     identify the optimal number of clusters $k$, by cutting the dendrogram with the method proposed by [5]
5:     for the clusterization into $k$ groups, compute internal clustering validation indices (e.g. Average silhouette width, Dunn index, Calinski and Harabasz index, ...) on the dissimilarity matrix $\Delta$
6: end for
7: plot the graphics of the values of the internal clustering validation indices versus $\theta$, and decide its optimal value
```

selecting the one that ensures the best separation among clusters, only in terms of upper tail dependence. We carried out the procedure on the data described in Section 2. The estimated standardized residuals of model (1) applied to the 38 traffic flow time series have been used to derive the corresponding distribution functions $\hat{U}_{jt}$. For each of the $(38 \times 37)/2 = 703$ pairs $(\hat{U}_{jt}, \hat{U}_{ht})$, we estimated by Maximum Likelihood a set of elliptical and Archimedean copulas and selected the best one according to AIC. Once obtained the corresponding estimates of the upper tail dependence coefficients, we carried out the clustering procedure of Algorithm 1 with $\Theta = \{0.005, 0.01, 0.015, \ldots, 4\}$ and using a hierarchical agglomerative algorithm with complete linkage. As internal clustering validation indices we adopted the Average silhouette width, the Dunn index, the Calinski and Harabasz index, that all suggested an optimal value of θ around 0.04 (Figure 1).

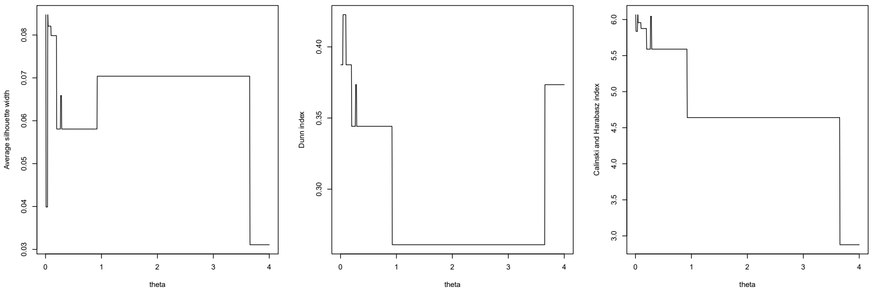


Fig. 1 Graphics of the values of the internal clustering validation indices versus θ (point 7 of Algorithm 1)

With $\theta = 0.04$ the areas turn out to be divided into 4 clusters, as displayed in Figure 2. We can observe that ACEs are grouped in a quite strong spatial neighbourhood structure, with extreme events occurring together in time in geographically contiguous areas. We found a group (coloured in blue) of ACEs located in the south outskirt of the Mandolossa. Those ACEs are characterized by a strong amount of streets going to the Mandolossa. Belongs to a second cluster (in purple) many of the ACEs that are not contiguous to the Mandolossa but with large streets connecting them with the Mandolossa itself. The other two clusters contain only a few ACEs, with Caino being a group by itself.

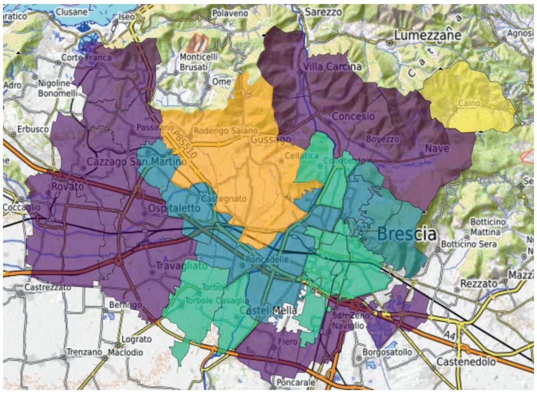


Fig. 2 Map of the 38 *Aree di Censimento* grouped with the upper tail dependence clustering with spatial structure applied to estimated residuals of flows to the Mandolossa (depicted in orange).

Acknowledgements his contribution has been developed for the Spoke 7 “CCAM, Connected Networks and Smart Infrastructure” of the Sustainable Mobility Research Center funded by the Italian NRRP (National Recovery and Resilience Plan) 2021-2026, which is part of the Next Generation EU (NGEU) Programme.

References

1. Balistrocchi M., Metulini R., Carpita M., Ranzi R.: Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, **20(12)**, 3485-3500 (2020)
2. D’Urso P., De Luca G., Vitale V., Zuccolotto P.: Tail dependence-based fuzzy *c*-medoids clustering of financial time series, *forthcoming*.
3. De Luca G., Zuccolotto P.: A tail dependence-based dissimilarity measure for financial time series clustering. Advances in Data Analysis and Classification, **5(4)**, 323-340 (2011)
4. De Luca G., Zuccolotto P.: Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach. International Journal of Approximate Reasoning **139**, 88–103 (2021)
5. De Luca G., Zuccolotto P.: Dynamic time series clustering with multivariate linkage and automatic dendrogram cutting using a recursive partitioning algorithm, *forthcoming*.
6. Metulini, R., Carpita M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research, 1-26 (2023)
7. Mishra D., Kumar S., Hassini E.: Current trends in disaster management simulation modelling research. Annals of Operations Research, **283(1)**, 1387-1411 (2019)
8. Sklar M.: Fonctions de repartition an dimensions et leurs marges. Publications de l’Institut de statistique de l’Université de Paris, **8**, 229–231 (1959)
9. Tsay, R.S.: Multivariate time series analysis: with R and financial applications. John Wiley & Sons (2013)

Global mobility trends from smartphone app data. The MobMeter dataset.

Indici di mobilità globali derivati dai dati di applicazioni smartphone. Il dataset MobMeter.

Francesco Finazzi

Abstract During the COVID-19 pandemic, mobility trends derived from smartphone data helped to understand the impact of lockdown and of the other measures of social distancing. The two main mobility products provided by Apple and Google are no longer available or updated, thus the need to replace them. This work describes the MobMeter dataset which provides time series of the daily average distance travelled by people and of the daily percentage of people at home for 17 global countries, uncertainty included.

Abstract *Durante la pandemia di COVID-19, i trend di mobilità derivati da dati smartphone hanno aiutato a comprendere l'impatto dei lockdown e delle altre misure di distanziamento sociale. I due principali prodotti di mobilità forniti da Apple e Google non sono più disponibili o aggiornati, da qui la necessità di sostituirli. Questo lavoro descrive il dataset MobMeter che fornisce serie temporali della distanza media giornaliera percorsa dalle persone e della percentuale giornaliera di persone a casa per 17 nazioni, incertezza inclusa.*

Key words: human mobility, COVID-19, lockdown.

1 Introduction

Apple's Mobility Trends Reports and Google's Community Mobility Reports (Cot et al., 2021) were global human mobility datasets made available to researchers during COVID-19 pandemic. Since the pandemic is no longer an emergency, both have been discontinued and they are no longer updated by the two Big Tech companies. Utility and importance of human mobility data, however, are not restricted to the COVID-19 pandemic (see for instance Tai et al. (2022) and Manawadu and Wijeratne (2022)) and the scientific community may benefit from new and similar

---

Francesco Finazzi
University of Bergamo, via dei Caniana 2, e-mail: francesco.finazzi@unibg.it

datasets. In this direction, MobMeter (Finazzi, 2022, 2023) offers updated human mobility trends for 17 global countries: Argentina (ARG), Chile (CHL), Colombia (COL), Costa Rica (CRI), Ecuador (ECU), Greece (GRC), Guatemala (GTM), Italy (ITA), Mexico (MEX), Nicaragua (NIC), Panama (PAN), Peru (PER), Philippines (PHL), Slovenia (SVN), Turkey (TUR), the United States (USA) and Venezuela (VEN). Specifically, MobMeter provides daily time series of the daily average distance travelled by people and of the daily percentage of people at home, uncertainty included. The data source at the base of MobMeter is the app of the Earthquake Network citizen science initiative (Finazzi and Fassò, 2017; Finazzi et al., 2020), which implements the first global smartphone-based earthquake early warning system.

The next sections discuss the peculiarities of smartphone app location data and present some results derived from the analysis of 3.8 million smartphone trajectories.

2 Smartphone app location data

Smartphone app location data are a sub-family of all location data that mobile devices can collect. First of all, they are not GPS data. On a smartphone, to avoid battery drainage, the GPS receiver is usually off unless the smartphone owner is using a tracking app or a navigator app. This means that the location acquired by the app (using the Wi-Fi or the 3G/4G/5G signals) is only approximated, with an accuracy that goes from 20 to 1000 or more meters. For the same reason, apps do not usually collect location data at high frequency. The maximum sampling rate is one location read every around 15/20 minutes. Additionally, apps can collect location data only if the smartphone is active and if the smartphone operating system is not (for any reason) preventing the app from reading the location. Smartphone app location data are thus characterised by a low sampling frequency, by a non negligible uncertainty and by extensive missing data.

Finally, smartphones with a given app installed are not necessarily a representative sample of the population and the app may have a high penetration in some regions and a low penetration in others. For instance, the app of the Earthquake Network initiative is more popular and more installed in seismic regions. Deriving mobility trends from smartphone app location data is a challenging tasks which calls for new statistical modelling approaches.

3 Location data analysis

Finazzi (2023) details how smartphone app location data are analysed to derive mobility trends at the country level and how their uncertainty is assessed. Specifically, the methodology developed in Finazzi (2023) aims at filtering out the smartphone “ghost” movements which are due to the uncertainty on the smartphone location rather than to a real movement across space of the smartphone. Also, the methodology aims at providing low-bias mobility trends by taking into account the population distribution across the country and the spatial distribution of the smartphones with the app installed. The non-parametric bootstrap technique is adopted to assess the uncertainty on the mobility trends.

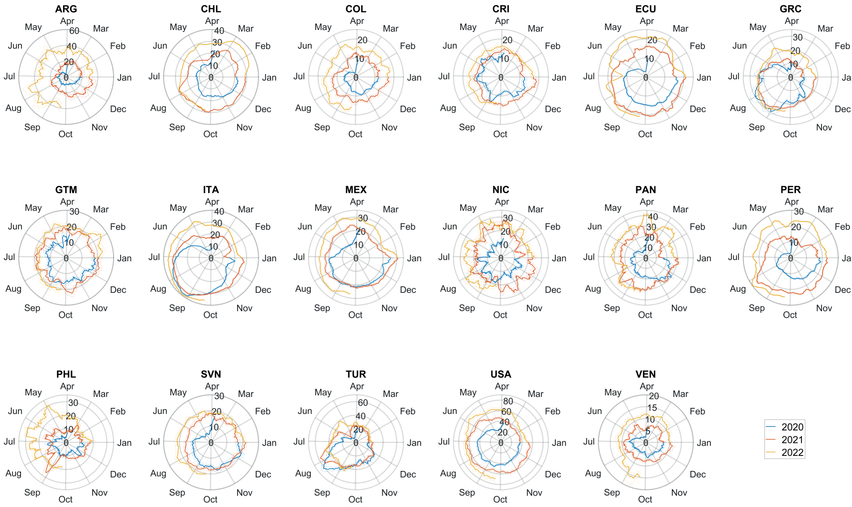


Fig. 1 Estimated daily average distance travelled by people (in kilometre) from March 24, 2020, to September 22, 2022. Dashed lines are 95% bootstrap confidence bands.

Figures 1 and 2 shows the estimated mobility trends for the above 17 countries from March 24, 2020, to September 22, 2022. Visual inspection allows to distinguish countries which are still recovering from the COVID-19 pandemic (characterised by "spiralling" plots) from countries which fully recovered and that only exhibit the natural seasonal variability. Bootstrap confidence bands are large or narrow depending on the number of smartphone users with the Earthquake Network app installed in each country.

On the other hand, Figure 3 shows how mobility trends are affected by social and natural events, like the social unrests in Ecuador and Panama of 2022 and like the Hurricane Agatha that struck the Oaxaca state of Mexico in May 2022.

4 Conclusions

Among location data, smartphone app data have unique peculiarities like sparseness and a non-negligible spatial uncertainty. On the other hand, if the app is popular, they tend to accurately describe human mobility trends. The MobMeter dataset provides

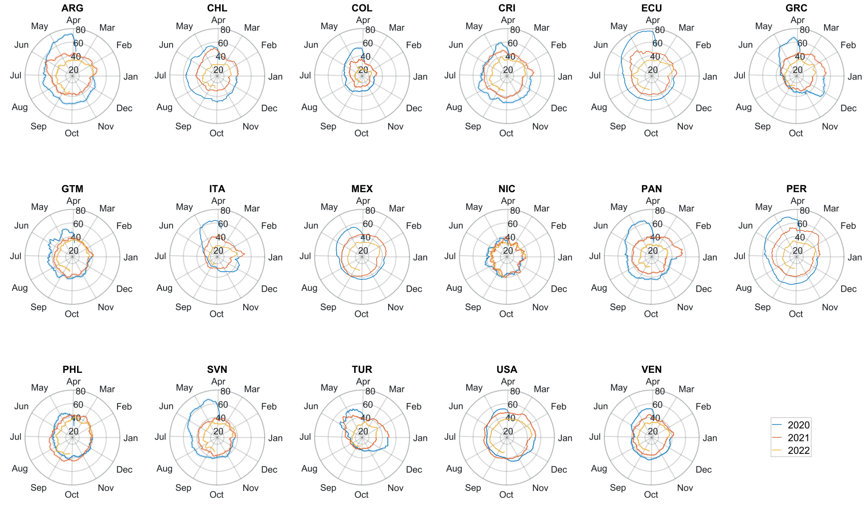


Fig. 2 Estimated daily percentage of people at home from March 24, 2020, to September 22, 2022. Dashed lines are 95% bootstrap confidence bands.

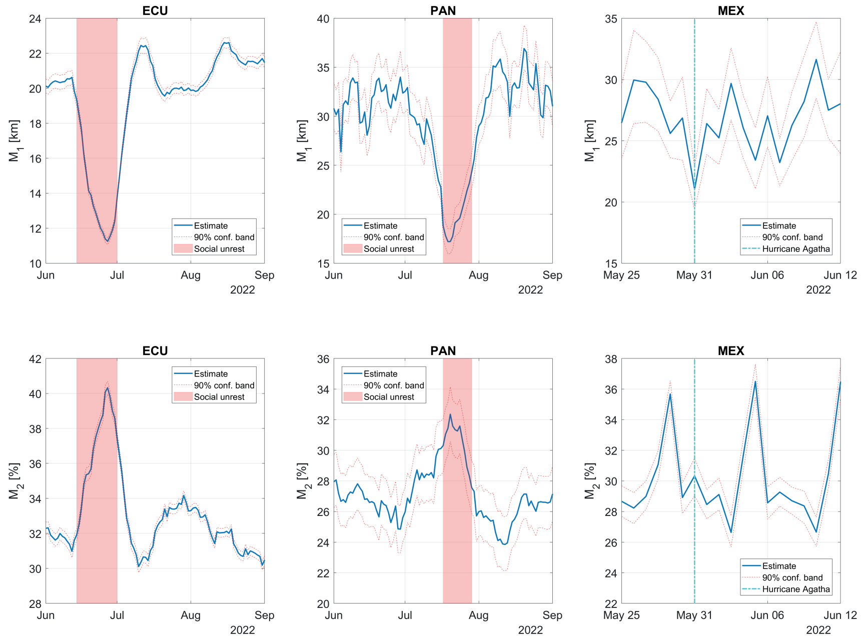


Fig. 3 Variations in mobility trends during 2022 social unrest in ECU and PAN and when Hurricane Agatha struck Oaxaca, MEX. M_1 refers to the daily average distance while M_2 to the percentage of people at home.

updated mobility trends for 17 global countries uncertainty included, and it partially replaces the discontinued mobility products of Apple and Google. In general, the body of statistical literature on smartphone app location data is very limited and many open problems require novel statistical modelling approaches.

References

Cot, C., Cacciapaglia, G., and Sannino, F. (2021). Mining google and apple mobility data: Temporal anatomy for covid-19 social distancing. *Scientific reports*, 11(1):4150.

Finazzi, F. (2022). Mobmeter: a global human mobility data set based on smartphone. https://zenodo.org/record/7635488.

Finazzi, F. (2023). Replacing discontinued big tech mobility reports: a penetration-based analysis. *Scientific Reports*, 13(1):935.

Finazzi, F. et al. (2020). Fulfilling the information need after an earthquake: statistical modelling of citizen science seismic reports for predicting earthquake parameters in near realtime. *Journal of the Royal Statistical Society*, 183:857–882.

Finazzi, F. and Fassò, A. (2017). A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stochastic environmental research and risk assessment*, 31:1649–1658.

Manawadu, L. and Wijeratne, V. (2022). Human mobility response to natural disasters and environmental change. In *Climate Change, Disaster and Adaptations: Contextualising Human Responses to Ecological Change*, pages 229–242. Springer.

Tai, X. H., Mehra, S., and Blumenstock, J. E. (2022). Mobile phone data reveal the effects of violence on internal displacement in afghanistan. *Nature human behaviour*, 6(5):624–634.

Spatio-temporal statistical analyses for risk evaluation using big data from mobile phone network

Analisi statistiche spazio-temporali per la valutazione del rischio utilizzando big data dalla rete di telefonia mobile

Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

Abstract In this short paper we summarize the ongoing work pertaining the use and combination of mobile phone data to characterize the spatio-temporal dynamic of the presence and the movements of people in the context of smart cities. We develop ad-hoc statistical approaches with the aim of developing small area indicators and forecasting traffic flows. The application of these strategies is related to the evaluation of flood risk in urban areas.

Key words: Vector autoregressive model; model-based functional cluster analysis; T-mode principal component analysis.

1 Introduction

Mobile phone data allow for a dynamic and fine-grained representation of human activities in urban systems. They are particularly suitable for the analysis and the mapping of the density of people's presences [1, 11] and movements [16, 12]. For this reason, they are increasingly adopted for the analysis of smart cities [2]. Some recent statistical applications can be found in [5, 4, 6, 9, 10].

Different sources of mobile phone data exist and might be used for different purposes. Here, we overview our ongoing work related to the development of statistical modeling and classification methods aimed at characterizing the spatiotemporal dynamic of people's presences and movements. The presented works concern the region of the Mandolossa (in the northwest outskirt of Brescia, Italy), which is an

---

DMS Statlab, Department of Economics and Management
University of Brescia, Contrada Santa Chiara, 50, Brescia. e-mail: selene.perazzini@unibs.it

Department of Economics
University of Bergamo, Via Caniana, 2, Bergamo. e-mail: rodolfo.metulini@unibg.it

DMS Statlab, Department of Economics and Management
University of Brescia, Contrada Santa Chiara, 50, Brescia. e-mail: maurizio.carpita@unibs.it

interesting case study because, being exposed to the risk of flooding, local authorities need several accurate information about people in the area. We restrict our attention to applications connected to three United Nations Sustainable Development Goals: 9 - Industry, innovation, and infrastructure; 11 - Sustainable cities and communities; 13 - Climate action. In particular, we focus on human exposure to flood risk, which is constituted by both the individuals staying in and those passing by the area. An accurate representation of human exposure should therefore account for the phenomena' static and dynamic characteristics. We show that the two aspects can be captured using different sources of mobile phone data. Their adoption allows for a multifaceted representation of human exposure at the small-area level, therefore providing a level of accuracy so high that it is rarely achieved by other types of data. We dedicate particular attention to the monitoring of traffic networks, which is a goal of Mission 3 of the Italian National Recovery and Resilience Plan (part of the Next Generation EU Programme). Furthermore, we show some examples of how different mobile phone data sources can be combined with each other's or with other types of data for the analysis of traffic.

2 Mobile phone data

We consider three types of mobile phone data:

- Crowding data (MPD), representing the Mobile Phone Density - i.e., the average number of mobile phone SIM cards in a squared cell of a pixel grid of dimension 150×150 meters in a 15-minute interval;
- Flows data (OD), representing the Origin-Destination flows - i.e., the number of SIM cards moving from one of the "Aree di CEnsimento" (ACE) of the Province of Brescia shown in the left map of Figure 1 to another in a 1-hour interval;
- Signals data (MDT), representing the data collected using the "Minimization of Drive Tests" technology that registers radio measurements of signals (i.e., phone calls, text messages, internet browsing, or technical operations on the network) transmitted over the 3G/4G mobile network from/to terminal devices with GPS enabled. The signals are collected in 15-minute intervals and geo-referenced on a grid of pixels measuring 10 meters per side. Though the MDT data only represents a sample of users, the accurate geolocation allows for small-area estimation. Since a mobile phone can produce multiple signals in 15 minutes, we refer to the number of grid cells from which MDT signals originated in an area.

The three types of data concern users subscribed to TIM and cover different periods and partially different (but overlapping) areas (see the left map of Figure 1). The MPD data was provided by the Municipality of Brescia in the context of a territorial monitoring project between 2014 and 2016. For this reason, they collect phone signals from April 1st, 2014, to August 11th, 2016. The OD and MDT data have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it) for the MoSoRe Project 2020-2022. The OD data cover the period between September 2020 and August 2021, while the MDT data refer to 5 days of November 2021 (namely Wednesday 10th, Friday 19th, Saturday 20th, Sunday 21st, and Monday 22nd). Indeed, MDT data require particular technologies to be activated and tested before data collection. For this reason, the data collection process is costly and takes time, and the produced MDT datasets typically cover short periods and small areas (see the right map of Figure 1). To overcome this issue, days for MDT data collection have been chosen in such a way as to represent a typical week.

It is worth noticing that, while the MPD and the OD data have already found applications in statistics, the MDT data have been very recently released and have been almost exclusively adopted in network engineering.

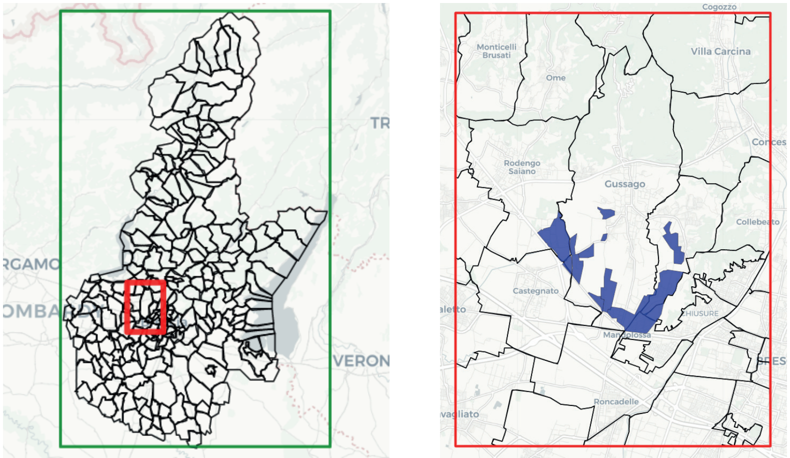


Fig. 1 Left: map of the ACEs of the Province of Brescia (black), and the areas captured by the MPD data (green) and by the MDT (red) datasets. Right: map of the ACEs (black) in the MDT dataset (red) and of the flood risk map with time to return equal to 20 years (blue).

3 An overview of statistical analyses

As a first step, we show that mobile phone data can be used to produce small-area indicators. In this regard, in [13] we use the MPD and the MDT data to represent respectively crowding and traffic intensity in the “Sezioni di CEnsimento” (SCE) (which are subdivisions of the ACEs) within the area captured by the MDT database. To this scope, since the MPD data are defined over a grid of quite large pixels, the number of users in each cell is distributed among the SCEs proportionally to the fraction of the area of the cell overlapping each of them:

$$MPD_{jt} = \sum_k MPD_{kt} \cdot \frac{Area(SCE_j \cap Cell_k)}{Area(Cell_k)} \tag{1}$$

where j indicates a SCE and t a time interval. As far as the traffic intensity is concerned, the MDT database is compared to a street map and is restricted to the phone signals originating from streets. Then, we count the number of cells of the MDT pixel grid corresponding to streets for each SCE j and each time interval t, and the obtained values are divided by the area of the street network in the j-th SCE:

$$MDT_{jt} = \frac{Number(\text{Streets Cells}_{jt})}{Area(\text{Streets}_j)}. \tag{2}$$

The two resulting sets of data (eq. 1 and 2) are analyzed and spatial patterns are investigated. At last, two indicators are defined using two T-mode principal component analyses [3] (see the left map of Figure 2 for an example). Given the strategic role of the road network in flood emergency management, three indicators capturing the main characteristics of the streets in the SCEs are also defined based on the available street maps. The joint analysis of the two mobile-phone-based and the three street-map-based indicators might be used to identify the areas with high concentrations of people or major connecting routes.

Mobile phone data can also be used for estimation and forecasting. In this respect, in [12] we use the OD data to analyze the traffic flows linked to the ACEs overlapping the flood risk map (see the right map of Figure 1). In that paper, a vector autoregressive model with dynamic harmonic components capturing complex seasonality [7] has been defined to forecast traffic flows:

$$\boldsymbol{Flow}_t = \boldsymbol{\nu} + \sum_{h=1}^{p} \boldsymbol{A}_h \boldsymbol{Flow}_{t-24\times h} + \boldsymbol{B}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \tag{3}$$

where $\boldsymbol{Flow}_t$ is a vector of length 3 containing the flows to, from, and within the ACEs exposed to floods. $\boldsymbol{\nu}$ is a constant vector of length 3, p is the autoregressive parameter, $\boldsymbol{A}_h$ is a 3×3 matrix of coefficients to be estimated, $\boldsymbol{\varepsilon}_t$ is the 3×1 vector of the error terms at time t, $\boldsymbol{x}_t$ is the vector of the l exogenous variables at time t, and $\boldsymbol{B}$ is the $3 \times l$ matrix of coefficients of the exogenous variables. The vector $\boldsymbol{B}\boldsymbol{x}_t$ is modeled using proper dynamic harmonic regression components. Despite the partial autocorrelation function of the estimated residuals showing significant first-order autocorrelation and a leptokurtic distribution with heavy tails, by means of a k-folds cross-validation we find that the model achieves satisfactory performance in forecasting both the number of people moving (see figure 2, that shows true versus fitted values for randomly chosen validation days) and the level of traffic intensity.

Then, in [14] we show that the estimation can be improved by combining the OD data with the MDT. We use the MDT data to estimate the proportion of traffic flows related to the flood-prone area in an ACE. The resulting ratios are then applied to the OD data as weights, in such a way as to obtain the traffic flows at risk. The combination of these two pieces of information allows us to produce statistical estimation and forecast of traffic flows for short time intervals at the small area level.

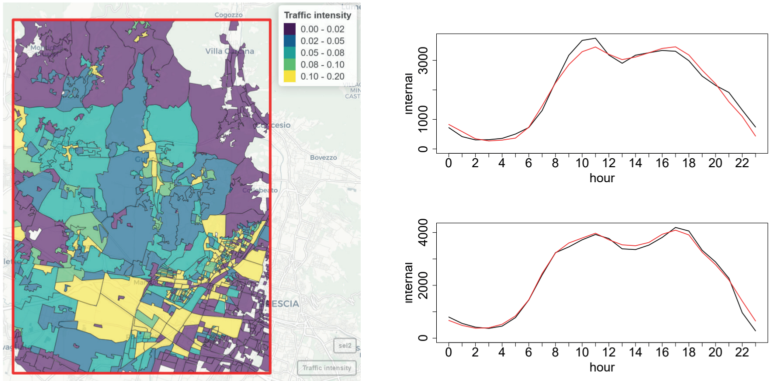


Fig. 2 Results of the analyses. Left: map of the traffic intensity indicator. Right: Observed (black) versus forecasted (colored) internal traffic flows. Validation days: Saturday February, 13th, 2021 (top), Tuesday July, 13th, 2021 (bottom).

This information can be used by local authorities to promptly activate traffic control actions aimed at preventing human losses and injuries.

4 Discussion and future developments

In this paper, three modern sources of mobile phone data have been presented, and their potential in the analysis of people's presence and movements has been shown through the discussion of some applications. We showed that the combination of different mobile phone databases can improve the accuracy of estimation. This aspect is still part of our current research. For example, we are working on a VARX model for forecasting traffic flows in risky areas, including traffic intensity and crowding indicators as regressors. Moreover, we presented some examples of how mobile phone data can be combined with other data sources to provide a multifaceted representation of a complex phenomenon. Nowadays, we are extending our research by introducing further data sources in the analysis and exploring the dynamics that link them. For example, in the context of the project "Data Science for Brescia", we are jointly analyzing mobile phone data and expenditure indices defined on Mastercard payment data to monitor the social and economic impacts of cultural events in the city of Brescia.

Acknowledgements This contribution has been developed for the Spoke 7 "CCAM, Connected Networks and Smart Infrastructure" of the Sustainable Mobility Research Center funded by the Italian NRRP (National Recovery and Resilience Plan) 2021-2026, which is part of the Next Generation EU (NGEU) Programme.

References

1. Balistrocchi, M., Metulini, R., Carpita, M., & Ranzi, R. Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, **20(12)**, 3485–3500 (2020)
2. Bibri, S.E., and Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable cities and society. **31**, 183–212 (2017)
3. Compagnucci, R. H., and Richman, M. B.: Can principal component analysis provide atmospheric circulation or teleconnection patterns?. International Journal of Climatology: A Journal of the Royal Meteorological Society, **28(6)**, 703-726 (2008)
4. Carpita, M., Manisera, M., and Zuccolotto, P.: Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia. In: Lombardo R., Camminatiello I., Simonacci V. eds. IES 2022: Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, Book of Short Papers of the 10th Scientific Conference of the SVQS, 575-581. PKE Press, Milano (2022)
5. Carpita, M., and Simonetto, A.: Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation. **5 (1)**, 31–41 (2014)
6. Curci, F., Kërçuku, A., Zanfi, F., Novak, C. et al.: Permanent and seasonal human presence in the coastal settlements of Lecce. An analysis using mobile phone tracking data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 57–71 (2022)
7. Hyndman, R.J., Athanasopoulos, G.. Forecasting: principles and practice. OTexts (2018)
8. Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis **71**, 92—106 (2014)
9. Manfredini, F., Lanza, G., Curci, F., et al.: Mobile phone traffic data for territorial research. Opportunities and challenges for urban sensing and territorial fragilities analysis. TeMA-Journal of Land Use, Mobility and Environment. **2**, 9–23 (2022)
10. Mariotti, I., Giavarini, V., Rossi, F., and Akhavan, M.: Exploring the “15-Minute City” and near working in Milan using mobile phone data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 39–56 (2022)
11. Metulini, R., and Carpita, M.: A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Social Indicators Research. **156 (2)**, 761–781 (2021)
12. Metulini, R., and Carpita, M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research. to be published. (2023)
13. Perazzini, S., Metulini, R., Carpita, M. Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Submitted to journal.
14. Perazzini, S., Metulini, R., Carpita, M. Integration of flows and signals data from mobile phone network for statistical analyses of traffic in a flooding risk area. Submitted to journal.
15. Pucci, P., Gargiulo, C., Manfredini, F., Carpentieri, G., et al.: Mobile phone data for exploring spatio-temporal transformations in contemporary territories. TeMA-Journal of Land Use, Mobility and Environment. **2**, 6–12 (2022)
16. Tettamanti, T., and Varga, I.: Mobile phone location area based traffic flow estimation in urban road traffic. Advances in Civil and Environmental Engineering. **1 (1)**, 1–15 (2014)

A Robust Approach to Profile Monitoring

Christian Capezza, Fabio Centofanti, Antonio Lepore, Biagio Palumbo

Abstract The main aim of statistical process monitoring (SPM) is to detect unusual operating conditions, which, by definition, result in an out-of-control (OC) state. On the contrary, when the process operates under standard conditions, it is said to be in control (IC). In modern industrial processes, data acquisition systems allow the collection of massive amounts of high-frequency data. Several examples may be found in the current Industry 4.0 framework, which is reshaping the variety of signals and measurements that can be gathered during manufacturing processes. In this context, the experimental measurements of the quality characteristics of interest are often characterized by complex and high dimensional formats that can be well represented as functional data, which are also referred to as profiles [10, 6].

The simplest approach for the SPM through one or multiple functional data is based on the extraction of one or more scalar features from them and applying classical techniques for multivariate data [8]. However, feature extraction is known to be problem-specific, arbitrary, and possibly masking useful information. This justifies the growing interest in profile monitoring [9], which is the monitoring of a process through quality characteristic observations in the form of one or multiple profiles.

Control charts are known as the main tools for SPM and are commonly implemented in two phases. The first is usually referred to as Phase I and is concerned with the identification of a clean data set to be assumed as representative of the IC state of the process, named Phase I sample or Phase I observations. This data set is then used to quantify the expected variation of a future observation to be used for the prospective process monitoring in the second phase, which is usually referred to as Phase II. However, the identification of Phase I observations in high-dimensional contexts is not an easy task due to the presence of several outliers in one or more components of the quality characteristic observation. Control charts are indeed very sensitive to the presence of outlying observations in the Phase I sample that can lead to inflated control limits and reduced power to detect process changes in Phase II.

Christian Capezza, Fabio Centofanti, Antonio Lepore, Biagio Palumbo
University of Naples Federico II, Department of Industrial Engineering, Piazzale Tecchio 80, Naples, e-mail: christian.capezza@unina.it

To deal with outliers, SPM methods use two common alternatives, namely the diagnostic and the robust approaches [5]. The diagnostic approach is based on standard estimates after the removal of sample units identified as outliers and may fail in detecting moderate outliers that are not always as easy to label correctly. The robust approach accepts all the data points and tries to find a robust estimator which reduces the impact of outliers on the final results [7].

Several robust approaches for the SPM of a multivariate scalar quality characteristic have been proposed in the literature [2, 4], while, to the best of authors' knowledge, a robust approach able to successfully capture the functional nature of a multivariate functional quality characteristic has not been devised so far. Moreover, in presence of many functional variables, the lack of robust approaches able to deal with outliers is exacerbated by the curse of dimensionality. In addition, multivariate robust estimators assume a casewise contamination model for the data only, where the majority of the cases is free of contamination. These traditional estimators are affected by the problem of propagation of outliers [3], when they consider outliers arising in each variable independently from the other ones, defined as cellwise outliers. Unfortunately, when the dimensionality of the data is high, under an independent contamination model, such as cellwise outliers, the fraction of perfectly observed cases can be rather small and the traditional robust estimators may fail. Moreover, as pointed out by Agostinelli et al. [1], both types of data contamination, casewise and cellwise, may occur together.

To perform the SPM of multivariate functional data in presence of casewise and cellwise outliers, we propose a new framework, referred to as robust multivariate functional control chart (RoMFCC). By means of a Monte Carlo simulation study, the ability of the RoMFCC in identifying mean shifts in the functional variables in presence of casewise and cellwise outliers is compared with other control charts already appeared in the literature. Moreover, we consider a real-case study on the SPM of a resistance spot welding (RSW) process in automotive body-in-white manufacturing that motivated this research.

Key words: Statistical Process Monitoring, Profile Monitoring, Casewise and Cellwise Outliers

References

1. Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H.: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test **24**(3), 441–461 (2015)
2. Alfaro, J., Ortega, J.F.: A comparison of robust alternatives to Hotelling's T^2 control chart. Journal of Applied Statistics **36**(12), 1385–1396 (2009)
3. Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H.: Propagation of outliers in multivariate data. The Annals of Statistics pp. 311–331 (2009)
4. Cabana, E., Lillo, R.E.: Robust multivariate control chart based on shrinkage for individual observations. Journal of Quality Technology pp. 1–26 (2021)

5. Hubert, M., Rousseeuw, P.J., Segaert, P.: Multivariate functional outlier detection. Statistical Methods & Applications **24**(2), 177–202 (2015)
6. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. Chapman and Hall/CRC (2017)
7. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust statistics: theory and methods (with R). John Wiley & Sons (2019)
8. Montgomery, D.C.: Introduction to Statistical Quality Control. Wiley (2012)
9. Noorossana, R., Saghaei, A., Amiri, A.: Statistical analysis of profile monitoring, vol. 865. John Wiley & Sons (2011)
10. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer (2005). DOI 10.1007/b98888. URL https://doi.org/10.1007/b98888

The FDA contribution to Health Data Science

Francesca Ieva

Abstract This contribution aims at presenting examples of Health Data Science where advanced methods based on Functional Data Analysis are used to bring value to clinical and biological problems.

Key words: Healthcare Research, Health Data Science, Health Analytics, Functional Data Analysis

1 General Background and motivations

Healthcare research generates a significant portion of big data from administrative routine, clinical practice, molecular sources, imaging investigations. This data is of paramount interest for defining a comprehensive fingerprint of patients' status to be used in primary and secondary prevention, therapy endpoints predictions and scoring [1]. This requires proper management and analysis in order to derive meaningful information from diversified data sources [2]. There are various challenges associated with each step of handling healthcare data, ranging from data access and integration, to development of advanced models for fingerprint extraction, to issues in late fusion of this information into suitable predictive models [3] [4]. That is why, to provide relevant solutions for improving public health, healthcare providers are required to be fully equipped with appropriate infrastructure to systematically generate and analyze data.

Among others, in this paper we focus on situations where the interest lies in dealing with time-varying processes, i.e., phenomena evolving over time. Examples are the dynamic monitoring of biological or vital signals, or models for longitudinal observations and covariates to be properly summarized and treated for plugging them

---

Francesca Ieva
MOX lab, Department of Mathematics, Politecnico di Milano, Milan , Italy & Health Data Science Center, Human Technopole, Milan, Italy e-mail: francesca.ieva@polimi.it

into Statistical and Machine Learning models and algorithms. In all these cases, Functional Data Analysis (FDA)[5] may be used as a proficient support to precision medicine, since it allows for developing powerful methods which account not only for baseline or cross sectional information, but also for the dynamic of the process at hand.

In particular, in Sect. 2 an overview of clinical applications where models exploiting FDA techniques are used will be presented, with the aim of highlighting FDA potential in supporting clinical practice and precision medicine approach. The first one (Sect. 2.1) concerns the extraction of dynamic information about patterns of care from Healthcare Utilization Databases. In the second case (Sect. 2.2), the dynamic monitoring of longitudinal biomarkers is presented. Finally, the third application (Sect. 2.3) is related to the assessment of genotype association with specific phenotypes of interest.

2 Case Studies

The case studies proposed in this Section are part of current researches carried out within the Health Analytics group at MOX lab (Department of Mathematics, Politecnico di Milano) and at the Health Data Science Center of Human Technopole. In particular, we exploit results coming from [6], [7] and [8].

2.1 Functional modeling of recurrent events on time-to-event processes

In clinical practice, it is often the case where the association between the occurrence of events and time-to-event outcomes is of interest; thus, it can bemodeled within the framework of recurrent events. The purpose of our study is to enrich the information available for modeling survival with relevant dynamic features, properly taking into account their possibly time-varying nature, as well as to provide a new setting for quantifying the association between time-varying processes and time-to-event outcomes.

In [6] and [7] we propose and discuss an innovative methodology to model information carried out by time-varying processes by means of functional data, modeling each time-varying variable as the compensator of marked point process the recurrent events are supposed to derive from. By means of Functional Principal Component Analysis, a suitable dimensional reduction of these objects is carried out in order to plug them into a Cox-type functional regression model for overall survival. We applied our methodology to data retrieved from the administrative databases of Lombardy Region (Italy), related to patients hospitalized for Heart Failure (HF) between 2000 and 2012. We focused on time-varying processes of HF hospitalizations and multiple drugs consumption and we studied how they influence patients' overall survival. This novel way to account for timevarying variables allowed to model self-exciting behaviors, for which the occurrence of events in the past increases the probability of a new event, and to quantify the effect of personal behaviors and therapeutic patterns on survival, giving new insights into the direction of personalized treatment.

2.2 A wavelet-mixed landmark survival model for the effect of short-term oscillations in longitudinal biomarker's profiles

In many chronic diseases, patient's disease progression and status can be monitored over time through easily measured biomarkers. Medical decisions regarding treatments are often made on the basis on such monitoring. Hence, it is important to have quantitative tools to exploit information given by such measurements. Since the final goal of medical decisions is the minimization of the risk of adverse events such as hospitalizations or death, survival models are very often the building blocks of such tools.

Recently two methods have been in widespread use for the modeling of longitudinal internal time-dependent covariates and survival: joint models and landmark models [9]. In joint models the longitudinal biomarker process is modeled through linear mixed effects models which allow to consider the subjects' specific trajectories of the biomarker through the inclusion of random effects into the model. These latent variables are used to model the effect of unobserved variables that are responsible of subjects' deviation from the overall mean trajectory specified through the fixed effects [10].

One of the main advantages of modelling the longitudinal process of the biomarker is that we can study the relationship between the rate of change of the biomarker and

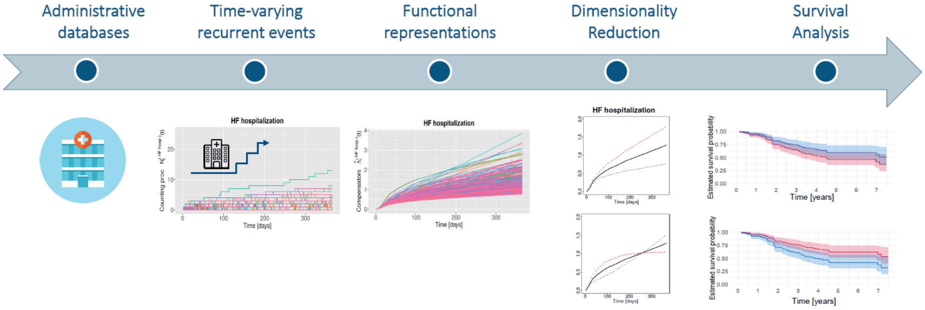


Fig. 1 Example of workflow for Healthcare Utilization Databases (HUD) exploitation. Functional data here represent the compensators of suitable stochastic processes describing re-hospitalizations over time and drug purchases for a given diseases of interest. They are analytically treated to be properly plugged into a predictive model for the main endpoint of the study (here long-term any cause survival).

the time-to-event process. However, a limitation of existing methods is that they don't take into account that the time scale of the survival process is very different from the time scale at which changes in the biomarkers happen in the human body. While in observational studies with the survival as the outcome of interest the follow-up period can be very long (e.g. years), intervals of measurements times are highly irregular because they depend of the clinical requirements. For example, during drugs up-tritation or acute disease periods more frequent measurements are required. As a consequence, such data is characterized by a very high observation period/time-between-measurements ratio. Sudden changes in biomarkers are very often important from a prognostic point of view. However, linear mixed effects model consider transient changes as measurement error when the length of the follow-up is much greater then their duration. Therefore, both joint models and landmark mixed models implicitly make a strong assumption, i.e., that short-term oscillations in biomarkers are either not present or they don't have effect on the risk of adverse events.

In [8] we propose a novel approach to study the association between a continuous timedependent longitudinal covariate and a time-to event outcome in order to overcome this limitation. Our method allows to identify and study the role of the short-term oscillations of the biomarker over time via a wavelet based functional approach and to set up a dynamic monitoring tool to support clinical decision making. The method is based on coupling a linear mixed effect model with a wavelet filter. The first allows to identify the effect of fixed covariates and the long-term effect of time. On the other hand, the wavelet filter is used to identify the subject-specific shortterm oscillations. The main idea is to combine a between-subjects model with a within-subject method such as a wavelet transform to obtain a functional and subject-specific representation of the biomarker trajectory over time [11]. Moreover, FDA offers methods to obtain functional objects from discrete and noisy longitudinal data and its application on time-dependent biomarker covariates offers a novel approach to extract biomarkers behaviors over time not observable with other methods.

2.3 Genomic trajectories

It is often the case in healthcare research that not the single measurement related to a quantity of interest is informative of the patient's status, but the evolution (a.k.a. behavioral pattern) of this quantity over time is. This is actually becoming particulary true in genomic studies, where the association between genomic traits and phenotypes are the goal of the analysis [12].

In such cases, the development and application of novel methodologies to study the genetic architecture of biomarkers' evolution over time, and their relationship with clinical outcomes of interest is the focus, and needs FDA application for both representing longitudinal trajectories and then modeling their effect on suitable endpoints as well as to assess their association with other clinical and biological traits.

Primary care and hospital administrative data can be exploited to derive longitudinal biomarker trajectories. Modelling Gene-biomarkers associations (considering single or multiple trajectories at a time) will result in the identification of representative biomarkers trajectory groups, and their association with individuals' genetic background. The considered biomarkers may include BMI, cholesterol, blood pressure, and others. Further clinical time varying information can be retrieved and included in modelling gene-biomarker relationships, to adjust for exogenous confounders influencing biomarkers' evolution, such as prescriptions and health conditions.

Once the trajectories are pointed out, further analyses will identify the effect of specific trajectory groups on the clinical endpoints (for the case of interest, ischemic stroke, coronary artery disease and diabetes). This enables novel insights on how biomarkers evolution over life time can increase the risk of adverse events and/or diagnoses. The application is carried out on UK Biobank data [13].

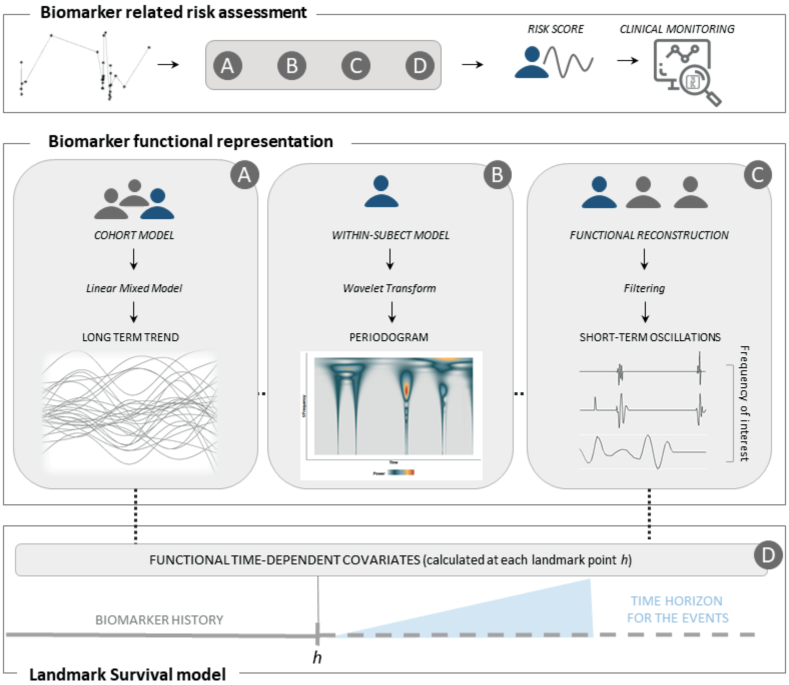


Fig. 2 Example of workflow for biomarkers monitoring in the context of Diskalaemya for Heart Failure patients.

3 Conclusions

In this paper we tried to emphatize the role FDA may have in different part of the health analytics process. The combination of advanced statistical methodologies with challenges proposed by recent clinical and biological problems may end up in a re-shaping of future directions of research in healthcare setting.

Acknowledgements This work was possible thanks to the substantial contribution of Dr. Caterina Gregorio, Michela Massi, Marta Spreafico and all the people dedicating their passion and capabilities to the healthcare research programs within the Health Analytics unit at Department of Mathematics of Politecnico di Milano and the Health Data Science Center of Human Technopole.

References

1. Schüssler-Fiorenza Rose SM et al: A longitudinal big data approach for precision health. Nat. Med. **25**, 792-–804 (2019)
2. Amal S, Safarnejad L, Omiye J A et al.: Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care. Frontiers in Cardiovascular Medicine, **9** (2022)
3. Stahlschmidt SR, Ulfenborg B, Synnergren J: Multimodal deep learning for biomedical data fusion: a review, Briefings in Bioinformatics, **23**(2): bbab569 (2022)
4. Acosta JN, Falcone GJ, Rajpurkar P et al. Multimodal biomedical AI. Nat Med 28, 1773–1784 (2022)
5. Ramsay JO, Silvermann BW: Functional Data Analysis. Springer-Verlag New York (2005)
6. Spreafico M, Ieva F: Functional modelling of recurrent events on time-to-event processes. Biom. J., **63**(5): 948-–967 (2021)
7. Spreafico M, Ieva F: Dynamic monitoring of the effects of adherence to medication on survival in Heart Failure patients: a joint modelling approach exploiting time-varying covariates. Biom. J., **63**(2) Special Issue: Novel Aspects in Biostatistics: 305–322 (2021)
8. Gregorio C, Barbati G, Ieva F: A wavelet-mixed effect landmark model for the effect of short-term oscillations in longitudinal biomarker's profiles on the risk of death: an application for the monitoring of potassium in Heart Failure. ArXiv preprint (2022) `https://doi.org/10.48550/arXiv.2204.05870`
9. Rizopoulos D, Molenberghs G, Lesaffre E M.E.H.: Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biom. J. **59**(6), 1261-–1276 (2017)
10. Tsiasis A, Davidian M: Joint modeling of longitudinal and time-to-event data: an overview on JSTOR. Stat. Sin. **14**(3), 809—834 (2014)
11. Unser M, Aldroubi A: A review of wavelets in biomedical applications. Proceedings of the IEEE **84**(4), 626—638 (1996)
12. Auton A et al: A global reference for human genetic variation. Nature **526**(7571), 68—74 (2015)
13. Sudlow C et al.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med **12**, e1001779 (2015)

A new topological weighted functional regression model to analyse wireless sensor data

Un nuovo modello di regressione per dati funzionali con dipendenza topologica per analizzare flussi di dati provenienti da sensori wireless

Andrea Diana and Elvira Romano and Antonio Irpino

Abstract Nowadays, modern sensor devices can generate a wide range of data types, including functional data, which can be challenging to analyse due to their complex dependencies. In order to effectively analyse these types of data, it is important to consider both the functional and topological dependencies, as well as any prior information that may be available. This may involve using advanced statistical techniques, such as functional data analysis or machine learning methods, to uncover patterns and insights from the data. In this paper, we focus on predicting the signal from such devices considering the topological structure induced by the connectivity sensor network. The approach we propose, is to pre-process the raw discrete functional data into smoothed functional data, which can help to reduce noise and make patterns in the data more clear. After that, a geographically weighted functional regression (GWFR) model is generalized to analyse functional data with topological dependencies induced by the connectivity of the network. Simulation studies with several types of connectivity were conducted to evaluate performances of the method. The method is motivated by an applicative study on the Intel indoor dataset (http://db.csail.mit.edu/labdata/labdata.html) to study the data prediction problem in a wireless sensor network.

Abstract *In questo articolo proponiamo di analizzare flussi di dati provenienti da sensori considerando la struttura topologica indotta dalla rete. L'approccio che introduciamo generalizza un modello di regressione geografico pesato per dati funzionali (GWFR) a dati con dipendenze topologiche. Il metodo, generalizzabile a*

Andrea Diana
Department of Mathematics and Physics, Universitá della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: andrea.diana@unicampania.it

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

Antonio Irpino
Department of Mathematics and Physics, Universitá della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: antonio.irpino@unicampania.it

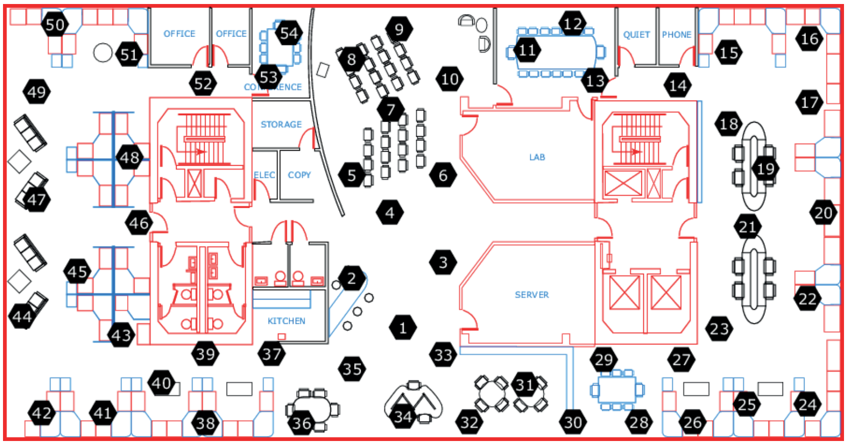


Fig. 1 Intel Berkeley Research Laboratory sensor network.

diversi contesti applicativi, nasce dal caso studio relativo al noto set di dati Intel indoor (http://db.csail.mit.edu/labdata/labdata.html).

Key words: Topological dependence, functional data, regression model, communication network

1 Introduction

This work is directly motivated by a real case study where signals are data consisting of 2.3 million readings collected by 54 sensors, including date, time, timestamp, node id, temperature, humidity, light, and voltage once every 31 seconds. Sensors data contain the coordinates of devices expressed in meters and relative to the upper right corner of the lab (Fig. 1) and aggregate connectivity information. Connectivity describes a communication network where a sender sends a message to a receiver with a certain probability of successful delivery. Unique IDs identify the sender and receiver, and the probability of successful delivery is specific to the sender-receiver pair, meaning that the probability of a message being successfully delivered from the receiver to the sender may differ. The additional information related to the probability of successful delivery may be helpful for making predictions about the relationships between signals in different spatial regions.

This type of communication network is useful in various real-world systems such as communication networks, transportation networks, medical images and social networks. In this sense, the proposed method can be useful in various applications.

The main aim of this work is to propose a topological weighted functional regression model (TWFR) for sensor data able to predict a continuous signal considering the communication network structure. It is a generalized spatial functional regression model for functional data spatially dependent. The spatial dimension is the grid of the communication network with the additional information related to the delivery of the sensors. TWFR can be used to identify patterns in the data that may indicate changes in the environment or the functioning of the sensors. TWFR can also be used to identify sensor failures or other issues that may affect the network's performance.

2 Topological Functional Data

Communication network data refers to data that describes the connections and interactions between individuals or devices in a network. This can include information on the number of connections between nodes, the strength of those connections, and the patterns of communication between nodes. Looking at the signal information over time as functional, communication networks data can be represented as graph or topological functional data.

We define topological functional data as undirected weighted attributed graph. Where the nodes of the graph represent the spatial data points and edges represent the relationships or connections between them, in general refers to the shortest path between two nodes or the probability of connection.

A **undirected weighted attributed graph** is a directed graph where some attributes are associated with the vertices and weights are associated with edges $G = (\mathscr{V}, \mathscr{E}, F_{\mathscr{V}}, F_{\mathscr{E}})$ where:

- $\mathscr{V}$ is a non-empty set of vertices (nodes);
- $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is a set of edges (links) which is symmetric;
- $F_{\mathscr{V}} : \mathscr{V} \to D_{\mathscr{V}}$ are the attributes associated with each nodes;
- $F_{\mathscr{E}} : \mathscr{E} \to \mathbb{R}^+$ are weights associated with each edge.

$D_{\mathscr{V}}$ is a description space containing numbers, categories, or functional data. For example, if $F_{\mathscr{V}} : \mathscr{V} \to \mathbb{N}$, $F_{\mathscr{V}}$ are called *types*.

In our case, a topological functional date is a undirected weighted attributed graph with spatial functional attributes, characterised by $D_{\mathscr{V}} = \mathscr{L}_2(T)^p$, so $F_{\mathscr{V}} : \mathscr{V} \to \mathscr{L}_2(T)^p$ is a vector of $p \in \mathbb{N}$ spatial functional data.

Let $\mathscr{V} = \{v_1, \ldots, v_i, \ldots, v_n\} \subseteq D \subseteq \mathbb{R}^d$ to be n nodes of undirected weighted attributed graph. Let $\{f_{v_1}(t), \ldots, f_{v_i}(t), \ldots, f_{v_n}(t)\}$ to be a set n of geostatistical functional data [2], [3]. $\mathscr{V}$ identify the n locations where the random functions $f_v(t)$ are located. Each function $f_v(t)$ is defined on $T = [a,b] \subseteq \mathbb{R}$ and is assumed to belong to a Hilbert space $\mathscr{L}_2(T) = \{f_{v_i} : T \to \mathbb{R},$ such that $\int_T f_{v_i}^2(t)dt < \infty\}$, with the inner product $\langle f_{v_i}, f_{v_j} \rangle = \int_T f_{v_i}(t) f_{v_j}(t)dt$. We assumed that the observed functions could be expressed according to the following model:

$$f_{v_i}(t) = \mu_{v_i}(t) + \varepsilon_{v_i}(t), \quad i = 1, \ldots, n \tag{1}$$

where $\varepsilon_{v_i}(t)$ are zero-mean residual functions and $\mu_{v_i}(t)$ is the mean one.

If we consider $F_{\mathscr{E}} : (v_i; v_j) \in \mathscr{E} \rightarrow h \in \mathbb{R}^+$, where h is defined as $h = \left\| v_i - v_j \right\|$ $\forall\, v_i, v_j \in \mathscr{V}$, we can see spatial functional data as also topological functional data. Topological functional data can be seen as an extension of spatial functional data, in which the spatial element is not associated with geographical location but rather with the degree of connection between the nodes of the network.

3 Topological weighted functional regression (TWFR) model

Topological weighted functional regression model (TWFR) we propose is a method that combines topological data analysis (TDA) ([1]) with spatial regression techniques to analyse topological functional data. The method uses topological features, such the probability in connectivity between two items like sensors, to analyse the underlying structure of the data and identify patterns that are not visible using traditional methods. TWFR applies a weighting scheme to the regression coefficients of a Geographically weighted regression model ([4]) , this allows the model to take into account the underlying structure of the data and the relationships between different features, which can improve the accuracy and interpretability of the model.

Consider i.i.d. regression data $(\chi_{v_1}(t), Y_{v_1}(\tau)), \ldots, (\chi_{v_n}(t), Y_{v_n}(\tau)) \sim P$ of the undirected weighted attributed graph $G = (\mathscr{V}, \mathscr{E}, F_{\mathscr{V}}, F_{\mathscr{E}})$, where each $(\chi_{v_i}(t), Y_{v_i}(\tau))$ is a multivariate topological functional stochastic process or a multivariate topological functional random field in $\mathscr{L}_2(T)^K \times \mathscr{L}_2(T_1)$, comprised of a response variable $Y_{v_i}(\tau)$ and a K$-$ dimensional vector of features (or predictors, or covariates) $\chi_{v_i}(t) = (X_{v_i,1}(t), \ldots, X_{v_i,K}(t))$. The topological weighted functional regression model (TWFR) specification follows the classical GWFR in [5] and is given by

$$Y_{v_j}(\tau) = \beta_0(\tau, v_i) + \sum_{k=1}^{K} \int_T X_{v_j,k}(t)\beta_k(t,\, \tau, v_i)dt + \varepsilon_{v_j}(\tau),\ j = 1, \ldots, n, \tag{2}$$

where $\beta_0(\tau, v_i)$ is the mean function at location v_i, $\beta_k(t, \tau, v_i)$ is the regression function for the $k-$th covariate at location v_j, and $\varepsilon_{v_j}(\tau)$ is a random error function at point v_j. Suppose, additionally, that the functional data can be approximated by a set of centred basis functions $\phi_k(t) = (\phi_{k,1}(t), \ldots, \phi_{k,m_1}(t))^T$ and $\psi(\tau) = (\psi_1(\tau), \ldots, \psi_{m_2}(\tau))^T$. The functional variables are expanded as $X_{D,k}(t) = C_k^T \phi_k(t)$ for each $k = 1, \ldots, K$, $Y_D(\tau) = A^T \psi(\tau)$ and $\beta_k(t, \tau, v_i) = \phi(t)^T B_{k,v_i} \psi(\tau)$ for each $k = 1, \ldots, K$ where C_k, A, and $B_{v_i,k}$ are matrices of dimensions $n \times m_1$, $n \times m_2$, $m_1 \times m_2$, respectively. The procedure estimation of $\beta_k(t, \tau, v_i)$ leads to the following equation:

$$(\mathbf{C}_k \boldsymbol{J}_\phi)^T \mathbf{W}_{v_i} (\sum_{k=1}^{K} \mathbf{C}_k J_\phi \mathbf{B}_{v_i,k}) J_\psi = (\mathbf{C}_k \boldsymbol{J}_\phi)^T \mathbf{W}_{v_i} A J_\psi \tag{3}$$

where $J_\phi = \int \phi^T(t)\phi(t)dt$, $J_\phi = \int \phi^T(t)\phi(t)dt$ and W_{v_i}, the weight matrix, is diagonal with following form

$$\mathbf{W}_{v_i} = [w_{v_i v_1}, w_{v_i v_2}, \dots, w_{v_i v_n}]^T \mathbf{I}_n. \tag{4}$$

The elements of $[w_{v_i v_1}, w_{v_i v_2}, \dots, w_{v_i v_n}]$ are the weights associated to the connectivity between the node v_i and all the other nodes $v_1, \dots, v_n$, i.e., they are related to $F_{\mathscr{E}}$, which are involved in the calibration of the model for the location v_i.

In the classic GWFR, the information from the geographic coordinates is transformed into a proximity matrix $\mathbf{W}_{v_i}$. If $\mathbf{W}_{v_i}$ is derived from a matrix of geographical distances, then we are examining spatially dependent functional data, but if $\mathbf{W}_{v_i}$ originates from a connectivity matrix, then we are working with topologically dependent functional data.

In the TWFR model, the matrix $\mathbf{W}_{v_i}$ is linked to $F_{\mathscr{E}}$, which weights the connectivity between nodes. This method is useful for examining phenomena where the influence between the attributes of nodes is dependent upon the edges of the graph. For example, in a study of Intel Berkeley Research Laboratory's sensor network (as seen in Fig. 1), two nearby sensors may be divided by a wall such that the influence between their attributes depends on the level of connection between them, rather than on their geographic location. In this instance, the connection level between two sensors is determined by the probability of one sensor of receiving the signal from the other.

4 Data prediction in wireless sensor networks

In this section, we apply the proposed methods and show the main results on the Intel indoor dataset (http://db.csail.mit.edu/labdata/labdata.html) to study the data prediction problem in a wireless sensor network. Functional data prediction can be used to improve data quality and reduce unnecessary data transmission. We focus on the study of micro clime changes in the Lab. We primarily evaluate the performances of the proposed conformal methods for studying the impact of the Temperature and the Light on the relative humidity. The data was collected by the Intel Berkeley Research Laboratory using Mica2Dot sensors between February 28th and April 5th, 2004, with the TinyDB in-network query processing system built on the TinyOS platform. The data consists of 2.3 million readings collected by 54 sensors, including date, time, timestamp, node id, temperature, humidity, light, and voltage once every 31 seconds. Sensors data contain the coordinates of devices expressed in meters and relative to the upper right corner of the lab (Fig. 1).

After a preprocessing steps, we use the functional form (Fourier basis expansion) for temperature ($TEMP$), humidity (HUM), and light (LUX) for estimating the relationship between the environmental variables using a TWFR model. The dataset contains the coordinates of sensors and the connection between them expressed by the average probability that the message sent from a sensor is correctly received by another one. To accurately simulate the dynamics of the phenomenon, we compared the GWFR, where W depends on the geographical distances between sensors, and TWFR model, where W depends on the connectivity between sensors induced by the probability of successfully delivery of a signal, using the classical indexes of goodness of fit for functional regression model like the Integrated Residual Sum of Squares (IRSS) and Squared Correlation Function $R^2(t)$. For the sake of brevity, we do not report the results of the analysis.

References

1. Chaza, F., Michel, B.: An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Front. Artif. Intell. 4:667963. doi: 10.3389/frai.2021.667963, (2021)
2. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric **21**, 224–239, (2010)
3. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)
4. Romano, E., Mateu, J., Butzbach, O. Heteroskedastic geographically weighted regression model for functional data, Spatial Statistics, **38**, (2020).
5. Yamanishi, Y. and Tanaka, Y.: Geographically weighted functional multiple regression analysis: A numerical investigation. Journal of Japanese Society of Computational Statistics **15**, 307–317, (2003).

Clustering for rotation-valued functional data

Clustering per dati funzionali a valori nel gruppo ortogonale speciale

Lise Bellanger and Aymeric Stamm

Abstract This methodological paper is motivated by the strong case study of monitoring gait for early detection of gait impairment in patients diagnosed with gait-affecting disorders. This leads to data with a great complexity that makes any kind of statistical analysis non-trivial. In effect, we collect functional data that evaluates on the Lie group of three-dimensional rotations. In this work, we develop sound statistical methods to enable joint clustering and alignment of such functional data by borrowing ideas from the existing k-means alignment approach.

Key words: functional data, k-means clustering, smooth manifold, tangent space, Lie group, Lie algebra, gait analysis.

1 Introduction

Functional data analysis pertains to the statistical analysis of samples of infinite-dimensional functions on which some regularity conditions can be assumed [6]. Such k-differentiable representations of discrete observations of such curves provides new insights into the temporal or spatial dynamics of the data at hand. An intrinsic property of functional data is the mixture of two sources of variability called amplitude and phase variability [10, 4]. The former captures variability in the co-domain of the curves while the latter captures variability in the domain of the curves. In some contexts, it can be desirable to separate these two sources of variability as they often come with enhanced interpretability. It is often the case for

---

Lise Bellanger
Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, France, e-mail: lise.bellanger@univ-nantes.fr

Aymeric Stamm
Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, France e-mail: aymeric.stamm@cnrs.fr

instance in clustering analysis where integrating alignment of curves in the clustering process yields more interpretable clusters. The k-means alignment algorithm has been proposed along these lines [8, 9] and showed a clear improvement in the resulting partition for a variety of concrete examples in various fields [1, 2, 5, 7]. However, it focuses only on $\mathbb{R}^p$-valued functional data.

In this paper, we present a natural extension of the k-means alignment clustering algorithm for rotation-valued functional data, that is functional data that evaluates on the special orthogonal group SO(3). The special orthogonal group is both a differentiable manifold and a Lie group with the property of being non-Abelian. Rotation-valued functional data are common in motion analysis when rotation motion of body parts is monitored in time. Joint clustering and alignment can be useful at two levels. In effect, it can be helpful (i) to resort to alignment with affine warping functions when computing individual gait patterns (IGP) to bring all gait cycles of the same individual at a given pace and (ii) to separate amplitude and phase variability to compare several IGPs expressed in percentage of the gait cycle, hence requiring boundary-preserving warping classes. We will therefore discuss extensions of the k-means alignment strategy for both cases.

2 Representation of rotation-valued functional data

There are a number of mathematical representations of a three-dimensional rotation in SO(3). One can use a 3×3 symmetric matrix R with $\det R = 1$, which is known as the *matrix representation*. One can use a 6-dimensional vector which stores the six values of the rotation matrix, which is known as the *vector representation*. One can use the Euler or Tait-Byan angle representation, which exploits the fact that every rotation in $\mathbb{R}^3$ can be described as the composition of three rotations around the axes of an orthogonal coordinate system. One can use the axis-angle representation which stores the axis of rotation as a three-dimensional unit vector $\mathbf{u}$ and the angle of rotation θ as a scalar value. Finally, one can use the *unit quaternion representation* which uses an hyper-complex number $\mathbf{q}$ of dimension 4 to encode the rotation as:

$$\mathbf{q} = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\left(u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k}\right), \tag{1}$$

where i, j and k are complex numbers such that $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$ and $\mathbf{ijk} = -1$. One can easily show that $\|\mathbf{q}\|_{\mathbb{H}} = \mathbf{q}^H = \sqrt{\mathbf{q}\mathbf{q}^\star} = \sqrt{\mathbf{q}^\star\mathbf{q}} = 1$, hence the name of unit quaternion. The space of unit quaternions equipped with the multiplication is known in algebraic topology as the compact symplectic group Sp(1) which is equivalent to the special unitary group SU(2) and topologically a 3-sphere $\mathbb{S}^3$.

The group Sp$(1) \cong$ SU(2) is exactly a double cover of the group SO(3) of 3-dimensional rotations since $\mathbf{q}$ and $-\mathbf{q}$ correspond to the same rotation. All these groups are Lie groups which are groups that are also smooth manifolds. For any Lie group G, one can define its corresponding Lie algebra $\mathfrak{g}$, which is its tangent space at the identity. The Lie algebra is a vector space equipped with an alternating bi-linear map called Lie bracket which satisfies the Jacobi identity and is anti-commutative. In particular, the vector space $\mathbb{R}^3$ can be seen as the Lie algebra of the Lie group of 3-dimensional rotations where the corresponding Lie bracket is the usual cross product $[x, y] = x \times y$.

In the following, we will use the unit quaternion representation of rotations. Moreover, we will restrict ourselves to functional data on uni-dimensional domains $T \subset \mathbb{R}$ as the primary target for statistical analysis are motion tracking. This leads us to introduce the following:

Definition 1 (rotation-valued functional datum). We will say that a random variable X is an $\mathrm{SU}(2)$-valued functional random variable when it evaluates in $L^2(\mathbb{R}, \mathrm{SU}(2))$, that is:

$$\begin{aligned} X : \Omega &\to L^2(\mathbb{R}, \mathrm{SU}(2)) \\ \omega &\mapsto \begin{aligned} X(\omega) : \mathbb{R} &\to \mathrm{SU}(2) \\ t &\mapsto X(\omega)(t) \end{aligned} \end{aligned}$$

3 Statistical analysis on the Lie algebra

It is non-trivial to directly extend statistical methods from the functional data analysis literature to rotation-valued functional data because of the nature of the co-domain of the realization of such random variables which requires computation of geodesics. However, since 3-dimensional rotations are a Lie group when equipped with the multiplication, one can think of projecting them onto the corresponding Lie algebra which is isomorphic to $\mathbb{R}^3$ and corresponds to the tangent space at the identity. In effect, statistical analyses on random functions evaluating in $L^2(T, \mathbb{R}^3)$ are more manageable. In particular, one can define the following:

Definition 2 (logarithm and exponential maps). Let $\mathbf{q} = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = w + \mathbf{v} \in \mathbb{H}$ be a quaternion. We define its *logarithm* as:

$$\log \mathbf{q} = \log \|\mathbf{q}\| + \frac{\mathbf{v}}{\|\mathbf{v}\|} \arccos \frac{w}{\|\mathbf{q}\|} \tag{2}$$

and its *exponential* as:

$$\exp \mathbf{q} = e^w \left(\cos \|\mathbf{v}\| + \frac{\mathbf{v}}{\|\mathbf{v}\|} \sin \|\mathbf{v}\| \right). \tag{3}$$

It is easy to show that the logarithm defined in eq.(2) maps the Lie group $\mathrm{SU}(2)$ to its corresponding Lie algebra $\mathbb{R}^3$ since for any $\mathbf{q} \in \mathrm{SU}(2)$, we have $\|\mathbf{q}\| = 1$ and thus $\log \|\mathbf{q}\| = 0$. Conversely, the exponential map defined in eq.(3) applied to a vector in $\mathbb{R}^3$, which can be viewed as a pure imaginary quaternion (*i.e.* with no real part $w = 0$), transforms it into a unit quaternion which represents the rotation of angle $\|\mathbf{v}\|$ around the axis $\mathbf{v}/\|\mathbf{v}\|$. We therefore have defined maps to bring unit quaternions in their tangent space $\mathbb{R}^3$ where statistical analysis can be carried out and to bring them back into $\mathrm{SU}(2)$ for interpretation.

Remark 1. While it is true that the logarithm maps the Lie group to its Lie algebra and the exponential maps the Lie algebra to the Lie group, we have that

$$\exp(\log q_1 + \log q_2) \neq q_1 q_2, \tag{4}$$

because the Lie group $\mathrm{SU}(2)$ is non-Abelian. Indeed, terms involving the Lie bracket of q_1 and q_2, which amount to zero in Abelian groups, must be taken into account. This is known as the Baker-Campbell-Hausdorff formula. It turns out that if rotations do not deviate too much from the identity (slow motion), the approximation $\exp(\log q_1 + \log q_2) \approx q_1 q_2$ can be appropriate to some extent, hence making statistical conclusions reached in the tangent space also valid in the original space.

4 The joint k-means and alignment method

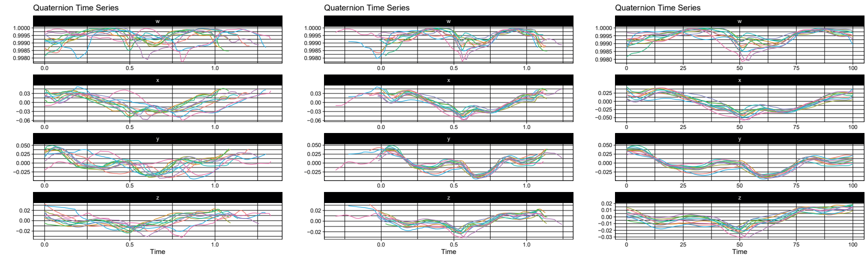
The principle behind the k-means alignment method is the foloowing. Let us assume that we have an n-sample $f_1, \dots, f_n$ of random curves on $T \subset \mathbb{R}$ that we want to cluster according to the k-means algorithm exclusively on the basis of their variability in amplitude, *i.e.* removing the phase variability. We define an appropriate class $\mathbb{W}$ of so-called warping functions of T and a compatible distance d between curves such that, if $h \in \mathbb{W}$, then $d(f_1 \circ h, f_2 \circ h) = d(f_1, f_2)$. In other words, the distance between any two functions f_1 and f_2 must be preserved if both functions are wrapped by the same warping functions and usually involves minimizing over the class $\mathbb{W}$ of warping functions the L^2-distance between one of the curve and the warped version of the second one. Using such a distance in the k-means algorithm effectively achieves a joint clustering (according to the amplitude variability) and within-cluster alignment of curves.

When the domain of the curves is the whole real line $\mathbb{R}$, it makes sense to use the affine class of warping functions for which a nice interpretation can be given when warping time. In effect, a non-null intercept indicates an initial delay or anticipation while a non-null slope indicate variation in velocity. The k-mean alignnment algorithm with affine warping functions for multivariate functional data has been described in [8] and we can apply it on rotation-valued functional data once mapped in $\mathbb{R}^3$ via the logarithm map.

When the domain of the curves is an interval of the form $T = [a, b]$, then we need a class of warping functions that is boundary-preserving. A natural choice pertains to resorting to the square-root slope function (SRSF) framework [9] which effectively aligns square-root slope functions of the original curves using the class of boundary-preserving warping functions. A version of the k-mean algorithm using the SRSF framework is currently available for univariate functional data only.

5 K-means for the analysis of individual gait patterns

We equipped a number of individuals with a motion sensor clipped on a belt at the level of the hip on their right side and registered the rotation of their hip over time on a time grid at 100 Hz frequency in the form of unit quaternions computed from data collected by 3 different 3-axis sensors: an accelerometer, a gyroscope and a magnetometer. As shown in Fig. 5, after a preprocessing step which consisted in correcting for small inconsistencies or artifacts in the collected data and segmenting the signal into gait cycles, we applied the proposed k-means alignment algorithm using affine warping functions and $k = 1$ on the gait cycles of the same individual in order to compute the so-called individual gait pattern which is effectively an average gait cycle representing the typical variation of hip rotation during one stride of the individual.



(a) (b) (c)

Fig. 1 Affine registration of individual gait cycles. Raw segmented gait cycles of the individual (a) are registered via the k-mean alignment algorithm using $k = 1$ and affine warping functions (b); the resulting registered curves are finally truncated to their common domain which is expressed in percentage of gait cycle (c) and the point-wise mean defines the individual gait pattern.

We used the above strategy to compute the IGP for a number of individuals. The goal is then to compare IGPs of patients diagnosed with gait-affecting disorders against healthy subjects. However, inter-subject gait variability comes from many different factors which we seek to isolate in order to focus as much as possible on the variability induced by the pathology. We believe that separating amplitude and phase variability will help us in this task. We extended the SRSF framework to accommodate rotation-valued functional data but it is currently limited to functional data on univariate domains only.

6 Discussion

We successfully extended existing k-means alignment strategies to the case of rotation-valued functional data. This proved to be helpful in defining the concept of individual gait pattern which turns out to a useful biomarker for studying gait impairment [3]. We plan to extend the SRSF version of the k-means alignment method to multivariate functional data to efficiently separate amplitude and phase variability in a sample of IGPs. Finally, we believe that the strategy of performing the statistical analysis on the Lie algebra of $\mathrm{SO}(3)$ is justified by the fact the hip rotation is a slow motion that does not deviate much from the identity, which should allow us to transfer the result of the statistical analysis back to the rotations themselves. However, in order to back up such a claim, we also plan to provide a k-means alignment method directly on the space of rotations using geodesics.

References

1. Mara Bernardi, Laura M. Sangalli, Piercesare Secchi, and Simone Vantini. Analysis of juggling data: An application of k-mean alignment. *Electronic Journal of Statistics*, 8(2):1817 – 1824, 2014.
2. Mara Bernardi, Laura M. Sangalli, Piercesare Secchi, and Simone Vantini. Analysis of proteomics data: Block k-mean alignment. *Electronic Journal of Statistics*, 8(2):1714 – 1723, 2014.
3. Pierre Drouin, Aymeric Stamm, Laurent Chevreuil, Vincent Graillot, Laetitia Barbin, Pierre-Antoine Gourraud, David-Axel Laplaud, and Lise Bellanger. Semi-supervised clustering of quaternion time series: Application to gait analysis in multiple sclerosis using motion sensor data. *Statistics in Medicine*, 42(4):433–456, 2023.
4. James Stephen Marron, James O Ramsay, Laura M Sangalli, and Anuj Srivastava. Functional data analysis of amplitude and phase variation. *Statistical Science*, pages 468–484, 2015.
5. Mirco Patriarca, Laura M. Sangalli, Piercesare Secchi, and Simone Vantini. Analysis of spike train data: An application of k-mean alignment. *Electronic Journal of Statistics*, 8(2):1769 – 1775, 2014.
6. J. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2006.
7. Laura M. Sangalli, Piercesare Secchi, and Simone Vantini. Analysis of AneuRisk65 data: k-mean alignment. *Electronic Journal of Statistics*, 8(2):1891 – 1904, 2014.
8. Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, 2010.
9. J Derek Tucker, Wei Wu, and Anuj Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013.
10. Simone Vantini. On the definition of phase and amplitude variability in functional data analysis. *TEST*, 21(4):676–696, 2012.

InstanceSHAP: An instance-based estimation approach for Shapley values

InstanceSHAP: un approccio di stima basato sulle istanze per valori di Shapley

Golnoosh Babaei[a] and Paolo Giudici[b]

[a]University of Pavia; `golnoosh.babaei01@universitadipavia.it`
[b]University of Pavia; Department of Economics and Management; `paolo.giudici@unipv.it`

Abstract

Explanations are necessary objects of a decision-making problem. However, complex Machine learning (ML) models that are usually used to provide decisions lack explanations. Model-agnostic explanation methods are the solutions for this problem and can find the contribution of each variable to the prediction of any ML model. Among these methods, SHapley Additive exPlanations (SHAP) is the most commonly used explanation approach which is based on game theory and requires a background dataset when interpreting an ML model. In this study we evaluate the effect of the background dataset on the explanations. In particular, we propose a variant of SHAP, InstanceSHAP, that use instance-based learning to produce a background dataset for the Shapley value framework. More precisely, we focus on Peer-to-Peer (P2P) lending credit risk assessment and design an instance-based explanation model, which uses a more similar background distribution. Experimental results reveal that the proposed model can effectively improve the ordinary shapley values and provide more robust explanations.

Abstract in italiano. La spiegabilità dei modelli di Machine Learning (ML) rappresenta un aspetto cruciale soprattutto nei processi decisionali. Tuttavia, i modelli di ML sono contraddistinti da un certo livello di complessità che può compromettere l'interpretabilità dei risultati che ne derivano. Al fine di risolvere questo problema, è possibile ricorrere a specifiche metodologie agnostiche, ossia indipendenti dal modello, che permettano di rilevare il contributo di ciascuna variabile alla previsione generata da qualsiasi modello di ML. Tra questi metodi, il metodo SHapley Additive exPlanations (SHAP) è l'approccio esplicativo più comunemente utilizzato. Partendo dalla teoria dei giochi, il metodo SHAP si basa su un set di dati di base in funzione del quale trarre l'interpretazione di un modello di ML. L'obiettivo di questo contributo consiste nel valutare l'effetto prodotto dal set di dati di base sulla spiegabilità. In particolare, proponiamo una variante di SHAP, chiamata InstanceSHAP, che utilizza l'apprendimento basato su istanze per generare un set di dati che funga da background per l'applicazione dei valori di Shapley. Più precisamente, il focus sarà rivolto alla valutazione del rischio di credito del prestito peer-to-peer (P2P) lending e sulla definizione di un modello di spiegazione basato sulle istanze, che utilizzi una distribuzione di background più simile. I risultati sperimentali rivelano che il modello proposto contribuisce effettivamente al miglioramento nella determinazione dei valori di Shapley fornendo, di conseguenza, una maggiore robustezza in termini di spiegabilità.

Keywords: Feature attribution; Shapley values; Machine Learning; Explainability.

A new paradigm for Artificial Intelligence based on Group Equivariant Non-Expansive Operators (GENEOs) applied to protein pocket detection

Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti,
Carmen Gratteri, Filippo Lunghini, Andrea Rosario Beccari, Carmine Talarico

Abstract Artificial Intelligence (AI) is now pervasive in everyday life, and it is quite often based on deep learning techniques. Deep learning continuously proves to be very effective in many applications, but its inherent opacity is also well known: deep learning experts cannot always explain AI decisions, even when coming from systems that were designed by themselves. Thus the need for eXplainable Artificial Intelligence (XAI).

Key words: XAI, GENEOs, pocket detection.

1 Introduction

Equivariant operators are proving to be increasingly important in deep learning, in order to make neural networks more transparent and interpretable [2, 7]. The

---

Giovanni Bocchi, Alessandra Micheletti
Department of Environmental Science and Policy, University of Milan, via Saldini 50, 20133 Milano, Italy, e-mail: giovanni.bocchi1@unimi.it, alessandra.micheletti@unimi.it

Patrizio Frosini
Department of Mathematics, University of Bologna, Piazza di Porta S.Donato 5, 40126 Bologna, Italy, e-mail: patrizio.frosini@unibo.it

Alessandro Pedretti
Department of Pharmaceutical Sciences, University of Milan, via Mangiagalli 25, 20133 Milano, Italy e-mail: alessandro.pedretti@unimi.it

Andrea R. Beccari, Filippo Lunghini, Carmine Talarico
Dompè Farmaceutici S.p.A., Via Tommaso de Amicis 95, 80123 Napoli, Italy e-mail: andrea.beccari@dompe.com, filippo.lunghini@dompe.com, carmine.talarico@dompe.com

Carmen Gratteri
Dept. of Health Sciences, Università degli Studi "Magna Græcia", Viale Europa, 88100, Catanzaro, Italy e-mail: carmen.gratteri.ext@exscalate.eu.

use of such operators corresponds to the rising interest in the so called "explainable artificial intelligence" [6, 14], which looks for methods and techniques whose functioning can be understood by humans. In accordance with this line of research, Group Equivariant Non-Expansive Operators (GENEOs) have been recently proposed as elementary components for building new kinds of networks [3, 4, 8]. Their use is grounded in Topological Data Analysis (TDA) and guarantees good mathematical properties to the involved spaces, such as compactness, convexity, and finite approximability, under suitable assumptions on the space of data and by choosing appropriate topologies.

A GENEO is a functional operator that transforms data into other data. By definition, it is assumed to commute with the action of given groups of transformations (equivariance) and to make the distance between data decrease (non-expansivity). The groups contain the transformations that preserve the "meaning" of our data, while the non-expansivity condition ensures that the operator simplifies the data metric structure. Both equivariance and non-expansivity are important: while equivariance reduces the computational complexity by exploiting symmetries of data, non-expansivity guarantees that the space of GENEOs can be finitely approximated.

In this paper we will introduce GENEOs and we will show promising results obtained in an industrial application, namely protein pocket detection.

2 Basic definitions and properties of GENEO spaces

Let us now formalize the concept of GENEO, as was introduced in [3]. We assume that a space Φ of functions from a set X to $\mathbb{R}^k$ is given, together with a group G of transformations of X, such that if $\varphi \in \Phi$ and $g \in G$ then $\varphi \circ g \in \Phi$. We call the couple (Φ, G) *perception pair*. We also assume that Φ is endowed with the topology induced by the L_∞-norm $D_\Phi(\varphi_1, \varphi_2) = ||\varphi_1 - \varphi_2||_\infty,\ \varphi_1, \varphi_2 \in \Phi$. Let us assume that another perception pair (Ψ, H) is given, with Ψ endowed with the topology induced by the analogous L_∞-norm distance D_Ψ, and let's fix a homomorphism $T : G \to H$.

Definition 1. A map $F : \Phi \to \Psi$ is called a *group equivariant non-expansive operator (GENEO)* if the following conditions hold:

1. $F(\varphi \circ g) = F(\varphi) \circ T(g)$ for every $\varphi \in \Phi$, $g \in G$ (equivariance);
2. $\|F(\varphi) - F(\varphi')\|_\infty \leq \|\varphi - \varphi'\|_\infty$ for every $\varphi, \varphi' \in \Phi$ (non-expansivity).

If we denote by F_{all} the space of all GENEOs between (Φ, G) and (Ψ, H) and we introduce the metric

$$D_{\text{GENEO}}(F_1, F_2) = \sup_{\varphi \in \Phi} ||F_1(\varphi) - F_2(\varphi)||_\infty, \qquad \forall F_1, F_2 \in F_{all}$$

the following main properties of F_{all} can be proven (see [3] for the proofs).

Theorem 1. *If Φ and Ψ are compact, then F_{all} is compact with respect to the topology induced by D_{GENEO}.*

Theorem 2. *If Ψ is convex, then F_{all} is convex.*

Theorem 1 guarantees that if the spaces of data are compact, then also the space of GENEOs is compact, thus it can be well approximated by a finite number of representatives, reducing thus the complexity of the problem. Theorem 2 implies that if the space of data is also convex, then any convex combination of GENEOs is still a GENEO. Thus when both properties hold we have an easy instrument to obtain new GENEOs starting from a finite number of them.

3 GENEOnet

We used GENEOs to build *GENEOnet* [5], a geometrical explainable machine learning method to detect pockets on the surface of proteins which are likely to host ligands (where ligands are usually drugs). Protein pockets detection is a key problem in the context of drug development, since being able to identify only a small number of good sites, allows a scientist to restrict the action of virtual screening procedures, saving thus both computational resources and time.

This problem is particularly suitable to be treated with GENEOs: on one side there is some important empirical chemical-physical knowledge that can not be directly embedded in the usual machine learning techniques, but can be injected in a GENEO architecture, and, on the other side, the problem shows a natural equivariance property, since if we rotate or translate a protein, its pockets will be coherently transformed in the same way. This suggests that pocket detection is equivariant with respect to the group of spatial isometries.

To apply GENEOs, input data have been discretized by surrounding each molecule with a bounded region divided into a 3D grid of voxels. In this way the data are modelled as bounded functions from the Euclidean space $\mathbb{R}^3$ to $\mathbb{R}^d$. We considered $d = 8$ distinct geometrical, chemical and physical potential fields, called *channels*, that were computed on each molecule [1].

The input data are fed to a layer of GENEOs chosen from a set of parametric families of operators, each one parametrized by one shape parameter $\sigma_i, i = 1, \dots, 8$. These families were designed in order to include the a priori knowledge of the experts of medicinal chemistry. We opted for convolutional operators with L^1 normalized kernels: if φ is one of the considered channel measured on a molecule, we compute

$$F_k(\varphi) = \int_{\mathbb{R}^3} \varphi(x) k(x-y)\, dy,$$

where k is a kernel. The behavior of such operators is determined by their kernels, thus by making the i-th kernel dependent only on one shape parameter σ_i, we have direct control on the action of each operator. We mainly used Gaussian kernels[2], or kernels having shapes of spheres, or of spherical crowns, assuming alternatively

---

[1] See [5] for further details on the specific channels.

[2] that is with general expression $k(x) = C \exp\left(-\frac{||x||^2}{2\sigma_i}\right)$

positive and negative values in different parts of the interior of the sphere or crown, and zero outside (see [5] for further details about the kernels). Nonetheless all the kernels are rotationally invariant functions, this fact, together with the properties of convolution, guarantees that the corresponding operators satisfy the key requirement to be equivariant with respect to the group of isometries of $\mathbb{R}^3$.

In the second step the d operators are combined through a convex combination, with weights $\alpha_1, \ldots, \alpha_d$, with $\alpha_i \in [0,1], \forall i$ and $\sum_{i=1}^{d} \alpha_i = 1$. The output of the convex combination is normalized to a function ψ from $\mathbb{R}^3$ to $[0,1]$. Here $\psi(x)$ can be read as the probability that a point $x \in \mathbb{R}^3$ belongs to a pocket. Finally, given a probability threshold $\theta \in [0,1]$, we get the different predicted pockets by taking the connected components of the superlevel set $\{\psi \geq \theta\} \subseteq \mathbb{R}^3$. The entire model pipeline is depicted in Figure 1.

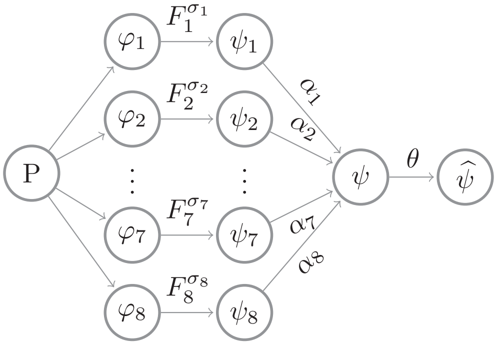


Fig. 1 Model workflow: input channels $\varphi_1, \ldots, \varphi_8$ are fed to the GENEOs $F_1, \ldots, F_8$ dependent on the shape parameters $\sigma_1, \ldots, \sigma_8$. The intermediate outputs $\psi_1, \ldots, \psi_8$ are combined through convex combination with weights $\alpha_1, \ldots, \alpha_8$ to get the final result ψ. To get predictions a thresholding operation with a parameter θ is applied obtaining the binary function $\widehat{\psi}$

The model that was described so far has a total of 17 parameters. The fact that the model only employs convolutional operators, and their linear combinations, allowed us to set up an optimization pipeline quite similar to a 3D Convolutional Neural Network (CNN), but with two fundamental differences. First of all GENEOnet has a really tiny set of parameters[3]. Additionally the convolutional kernels of the GENEOs are not learned entry by entry as in classical CNNs (in this way equivariance would not be preserved), instead the kernels are computed at each step from the shape parameters that are updated during the optimization.

4 Model training and comparison with other methods

In order to identify the unknown parameters, we chose to optimize a cost function that evaluates the goodness of our predictions, in terms of volume fraction of the

---

[3] For comparison DeepPocket [1], a recent approach that uses a 3D CNN, has 665 122 parameters.

cavity which contains the ligand that has been correctly detected. Eventually, after training, pockets are found as the connected components of the thresholded output of the model, resulting in a set of unranked pockets. Actually this representation is not much informative, since it is usually desirable to compute also the "druggability" of the identified cavities, that is a ranking score for the pockets, on the basis of their fitness to host a ligand. Thus we used the 'not thresholded' output ψ of the model to score the pockets, so that the final output consists in a list of pockets, ranked by their corresponding scores [5].

In order to identify the optimal model, we opted for a two-step optimization procedure: in the first step we generated $m = 200$ models $(\mathscr{M})_{k=1}^{m}$ optimized from $(T_k, IC_k)_{k=1}^{m}$, where T_k is a training set of size 200, subsampled from the whole dataset and IC_k are the randomly generated initial values of the parameters. In the second step each model was evaluated for its scoring capabilities, by computing H_1 (see (1) for the definition) on a validation set in order to select the one with highest H_1. This final model was evaluated on an independent (both from the training and the validation sets) test set to produce the results of next section.[4]

We compared the results of GENEOnet with other recent methods for protein pocket detection, based on ML techniques. We based our comparison on the scores assigned by the different methods to the pockets. In this way we compared the ability of the models to assign the highest scores to pockets that match the true ones. Given a dataset of proteins having only one ligand, and thus one "true pocket" each, we computed the following quantities

$$H_j = \frac{\#(\text{proteins whose true pocket is hit by the jth top ranked})}{\#(\text{proteins})} \tag{1}$$

and the corresponding cumulative quantities

$$T_j = \frac{\#(\text{proteins whose true pocket is hit within the jth top ranked})}{\#(\text{proteins})} = \sum_{i=1}^{j} H_j .$$

In this way different methods can be compared directly: if a model has a cumulative T_j that stands above all the others for all j, then that model is definitely better. The results reported in Table 1 show that GENEOnet has a better performance than all the other considered methods.

References

1. Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Priyakumar, U. D.: DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks.

---

[4] Dataset size: 12995, Training set size: 200, Validation set size 3073 (48 proteins in the intersection with the training set), Test set size: 9070 (totally disjoint from the other sets). Protein data retrieved from PDBbind v2020 dataset [11]

Table 1 T_j values for the methods in the comparison. In the last column the fraction of molecules whose correct pocket has been identified by at least one of the ranked cavities is reported. See [5] for the full analysis.

Method	T_1	T_2	T_3	T_4	$\sum_{j\geq 1} H_j$
GENEOnet [5]	**0.792**	**0.905**	**0.941**	**0.955**	0.975
P2Rank [10]	0.728	0.847	0.892	0.917	0.952
DeepPocket [1]	0.652	0.798	0.860	0.896	**0.978**
CAVIAR [13]	0.616	0.739	0.783	0.806	0.837
SiteMap [9]	0.424	0.502	0.529	0.542	0.558
Fpocket [12]	0.331	0.462	0.534	0.585	**0.978**
CavVis [15]	0.224	0.376	0.483	0.567	0.842

Journal of Chemical Information and Modeling (2021) doi: 10.1021/acs.jcim.1c00799

2. Anselmi, F., Evangelopoulos, G., Rosasco, L., and Poggio, T. : Symmetry-adapted representation learning. Pattern Recognition (2019) doi: 10.1016/j.patcog.2018.07.025
3. Bergomi, M. G., Frosini, P., Giorgi, D., and Quercioli, N.: Towards a topological–geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0087-3
4. Bocchi, G., Botteghi, S., Brasini, M., Frosini, P., and Quercioli, N.: On the finite representation of group equivariant operators via permutant measures. Annals of Mathematics and Artificial Intelligence (in press) (2023) doi: 10.1007/s10472-022-09830-1
5. Bocchi, G., Frosini, P., Micheletti, A., Pedretti, A. *et al.*: GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. (2022) preprint at arXiv:2202.00451.
6. Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.J. *et al.*: Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Scientific Reports (2021) doi: 10.1038/s41598-021-83922-6
7. Cohen, T. and Welling, M.: Group equivariant convolutional networks. In proceedings of the International Conference on Machine Learning (2016).
8. Conti, F., Frosini, P., and Quercioli, N.: On the Construction of Group Equivariant Non-Expansive Operators via Permutants and Symmetric Functions. Frontiers in Artificial Intelligence (2022) doi: 10.3389/frai.2022.786091
9. Halgren, T.: New method for fast and accurate binding-site identification and analysis. Chemical Biology & Drug Design (2007) doi: 10.1111/j.1747-0285.2007.00483.x
10. Krivak, R. and Hoksza, D.: P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics (2018) doi: 10.1186/s13321-018-0285-8
11. Liu, Z., Su, M., Han, L., Liu, J. *et al.*: Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Accounts of Chemical Research (2017) doi: 10.1021/acs.accounts.6b00491
12. Le Guilloux, V., Schmidtke, P., and Tuffery, P.: Fpocket: An open source platform for ligand pocket detection. BMC Bionformatics (2019) doi: 10.1186/1471-2105-10-168
13. Marchand, J.R., Pirard, B., Ertl, P., and Sirockin, F.: CAVIAR: a method for automatic cavity detection, description and decomposition into subcavities. Journal of Computer-Aided Molecular Design (2021) doi: 10.1007/s10822-021-00390-w
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0048-x
15. Simoes, T. M. C., and Gomes, A. J. P.: CavVis-A Field-of-View Geometric Algorithm for Protein Cavity Detection. Journal of Chemical Information and Modeling (2019) doi: 10.1021/acs.jcim.8b00572

Clustering Italian medical texts: a case study on referrals

Clustering di testi medici italiani: un caso studio sulle ricette dematerializzate

Vittorio Torri, Michele Ercolanoni, Francesco Bortolan, Olivia Leoni and
Francesca Ieva

Abstract In the medical domain, there is a large amount of valuable information that is stored in textual format. These unstructured data have long been ignored, due to the difficulties of introducing them in statistical models, but in the last years, the field of Natural Language Processing (NLP) has seen relevant improvements, with models capable of achieving relevant results in various tasks, including information extraction, classification and clustering. NLP models are typically language-specific and often domain-specific, but most of the work to date has been focused on the English language, especially in the medical domain. In this work, we propose a pipeline for clustering Italian medical texts, with a case study on clinical questions reported in referrals.

Key words: Natural Language Processing, Clustering, Administrative Databases, Medical Document

---

Vittorio Torri
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: vittorio.torri@polimi.it

Michele Ercolanoni
ARIA SpA (Azienda Regionale per L'Innovazione e gli Acquisti), Milan, Italy e-mail: michele.ercolanoni@ariaspa.it

Francesco Bortolan
UO Osservatorio Epidemiologico Regionale, Direzione Generale Welfare, Regione Lombardia, Milan, Italy e-mail: francesco_bortolan@regione.lombardia.it

Olivia Leoni
UO Osservatorio Epidemiologico Regionale, Direzione Generale Welfare, Regione Lombardia, Milan, Italy e-mail: olivia_leoni@regione.lombardia.it

Francesca Ieva
CHDS – Center for Health Data Science, Human Technopole, Milan, Italy
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: francesca.ieva@polimi.it

1 Introduction

In the last decade, Natural Language Processing (NLP) has seen significant advances and it has started to be applied also in the medical domain, where Electronic Health Records store nowadays a large number of textual documents, which often contain information that is not reported in the structured fields [5].

Medical documents present additional challenges with respect to those in the general domain, related to the specific lexicon, the high number of ambiguous abbreviations, the significant differences that exist in the way in which clinical documents are structured by different healthcare providers and the lack of annotated datasets.

Most of the work on NLP for the clinical domain is related to English documents, although in the last years, other languages are receiving increased attention. The extension of existing language models to different languages is not straightforward, due to the scarcity of (annotated) datasets and domain-specific resources [7]. In particular, there are only a few works which applied NLP to Italian medical documents, including both rule-based and machine-learning-based approaches [9, 6].

In this work, we present a pipeline for clustering Italian medical documents, applied to a specific case study: clinical questions that are reported in the referrals for specialized examinations.

In Italy, there is a standard format for referrals, containing various structured fields and two free-text fields: the *clinical question* and the *notes*. We focus on clinical questions since it is a field that must be compulsorily filled in with a description of the reason which determined the referral, i.e., the disease or the symptoms of the patient. For this purpose, we have been given access to a dataset of referrals filled-in in Lombardy Region in 2021, including their clinical questions.

Clinical questions are the only way to specify the reason for a referral since the Italian referral form does not provide any structured field for coding the disease or symptoms of the patients. Being able to automatically identify the disease code from the clinical questions would allow many subsequent analyses, both in the epidemiological field and on the appropriateness of prescriptions, two topics of paramount importance for the National Health System.

Referrals' clinical questions are not traditional clinical documents, in fact, they are administrative medical data. Administrative databases present the advantage of large coverage of the population, which is a relevant aspect for the above-mentioned types of analyses. Nevertheless, since they are meant neither for clinical nor for statistical use, they present some additional challenges with respect to traditional clinical documents: they often consist of one or two short sentences, missing punctuation and a proper syntactic structure, with multiple pieces of information not clearly separated (e.g., the proper clinical question, data on past examinations or past diseases, information on waiting times).

Due to the initial absence of an annotated dataset of clinical questions and since one of the most common problems in the medical domain is the absence of annotated datasets, we focused on the development of a clustering pipeline, which can be useful for different types of medical documents.

To the best of our knowledge, this is the first work on clustering of Italian referrals and more in general the first on clustering of Italian medical documents, with the exception of [1] which used proprietary software to cluster handovers of an Italian mental health institute.

The rest of the paper is organized as follows: in Section 2 we present the data and the clustering pipeline, in Section 3 we discuss preliminary results on the case study, together with limitations and possible future developments, while Section 4 contains conclusive remarks.

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 dataset used in this work consists of 5000 clinical questions from referrals filled in by physicians in Lombardy Region in 2021, related to specialized oncological examinations. These data have been manually labelled with respect to the most common types of cancer, identifying 26 clusters, plus a cluster of all the remaining referrals. The distribution among the different types of cancer is severely unbalanced: breast cancer has 1287 referrals, the *other* group has 1144 referrals, while nine other types of cancers are in the order of hundreds and the remaining ones in the order of tens of referrals.

2.2 Clustering pipeline

Figure 1 depicts the pipeline with its main steps: pre-processing, text representation, dimensionality reduction and clustering. In this section, we analyze each of them.

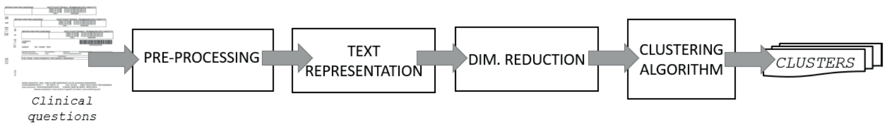


Fig. 1 Diagram of the clustering pipeline

2.2.1 Pre-processing

We considered the following steps: (i) lemmatization/stemmatization, (ii) abbreviations expansion, (iii) stop-words removal, (iv) typos correction, (v) low/high-frequency words removal. They are all common steps in NLP, with the exception of abbreviations expansion which is particularly relevant in the medical domain.

2.2.2 Text representation

In NLP there are two main types of representations that can be used: vocabulary-based representations and neural-network-based embeddings.

The most used vocabulary-based representation is the TF-IDF representation [2], where the entry for word w_j in document d_i of dataset D has the following value:

$$TF-IDF(w_j,d_i) = \frac{\text{\# occurrences of } w_j \text{ in } d_i}{\text{\# words in } d_i} \times log\left(\frac{|D|+1}{|d \in D : w_j \in d|}\right) \tag{1}$$

Vocabulary-based representations are sparse, high-dimensional and they do not take into account semantic similarity between words. Neural-network-based embeddings overcome this limitation since they produce fixed-length continuous vectors whose distance is related to the semantic similarity of the texts they represent. Here we focus on context-aware neural-network-based embeddings, i.e., those that are not fixed for a given word but depend also on the context in which it appears. BERT is a context-aware neural network model which has achieved state-of-the-art performances in many NLP tasks in the last years [4].

The original version of BERT has been developed for general-domain English documents, but many other versions of BERT have been subsequently developed to improve the performances on specific domains or different languages, using the same architecture but different training datasets. There are models for the medical domain [3], but they don't cover the Italian language, while there are models for the Italian language, but only for general-domain documents [8]. In our case study, we compare the results of Umberto, the most recent Italian version of BERT, with two other models that we have fine-tuned to cover the Italian medical domain:

1. **Umberto-medicina**: fine-tuned on documents extracted from the Medicine section of Italian Wikipedia (21.413 documents, 108 MB)
2. **Umberto-malattie**: fine-tuned on documents extracted from the Diseases section of Italian Wikipedia (4028 documents, 23 MB)

The datasets that we have used are much smaller than the original one (69GB), and for this reason, we adopted a transfer-learning approach, i.e. we fine-tuned Umberto instead of training a new BERT model from scratch.

2.2.3 Dimensionality Reduction

Vocabulary-based representations can easily reach hundreds of dimensions and even thousands. Neural-network-based embeddings have instead a fixed length, which is typically in the order of hundreds (768 for BERT). In both cases, there is a need to reduce their dimensionality before using them in a clustering algorithm. We applied PCA for this step, but other possibilities could be investigated in the future.

2.2.4 Clustering algorithm

The choice of the clustering algorithm depends on the type of representation, but also on the dataset. A relevant aspect, common to many clustering algorithms, is the distance metric to use. For vocabulary-based representations, cosine distance is a natural choice, while for neural-network-based embeddings both euclidean distance and cosine distance can be used.

3 Results

The results on the referrals dataset are measured in term of weighted F1-score, after having determined the best match between the clusters and the classes for which the data have been labelled. Table 1 shows a comparison between the TF-IDF representation and the different BERT-based embeddings. For TF-IDF we have used k-means as clustering algorithm, since it provided the best results, while for the BERT-based embeddings the best results have been achieved with HDBSCAN.

BERT-based language-models show large improvements with respect to the TF-IDF representation, and the fine-tuned versions of Umberto show a relevant improvement with respect to the standard Umberto. *Umberto-medicina* achieved the best results, probably because its fine-tuning dataset is larger than the one of *Umberto-malattie*.

Table 1 Results with different text representations

Text representation	W. Precision	W. Recall	W. F1
TF-IDF	0.167	0.223	0.184
Umberto	0.472	0.372	0.321
Umberto-medicina	**0.681**	**0.454**	**0.438**
Umberto-malattie	0.670	0.437	0.423

An analysis of the detailed results per type of cancer shows that that are significant differences between them. We have been able to identify clusters associated with 13 out of 26 cancer types. For most of the identified types of cancer, precision is high (> 0.90), while recall is low (< 0.50), with a few exceptions.

There are some types of cancers for which we identified multiple clusters that are associated with them. In particular, breast cancer, the largest class, has 5 clusters, while colon cancer has 3 clusters, in our best result. With TF-IDF the results are not only worse in terms of precision, recall and F1, but also in terms of number of clusters per cancer type, showing 9 different clusters for breast cancer. Some of these clusters referred to the same cancer type have a rationale for being separate, e.g., some of them contain indications of patients who have been under surgery, others are specifically referred to right/left breast, but this is not always the case. Further investigations are needed to understand how to avoid or merge these clusters.

4 Conclusions

In this work we have presented a pipeline for clustering Italian medical documents, addressing this problem for the first time. While many challenges remain open, relevant improvements have already been achieved with respect to baseline techniques, highlighting the benefits of BERT-based language models, even in this specific domain and without a large training dataset. Further investigations will be focused on the analysis of different dimensionality reduction techniques and on the problem of automatic merging different clusters related to the same topic.

References

1. Accordini, Monica, et al. Stories of change: The text analysis of handovers in an Italian psychiatric residential care home. Journal of Psychiatric and Mental Health Nursing 24.4, 232-242 (2017)
2. Aggarwal, Charu C. Mining text data. Data mining. Springer, Cham, (2015)
3. Alsentzer, Emily, et al. Publicly Available Clinical BERT Embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop (2019)
4. Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 4171-4186 (2019)
5. Iroju, Olaronke G., and Janet O. Olaleke. A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science 8, 44-50 (2015)
6. Lanera, Corrado, et al. Use of machine learning techniques for case-detection of varicella zoster using routinely collected textual ambulatory records: Pilot observational study. JMIR Medical Informatics 8.5, e14330 (2020)
7. Névéol, Aurélie, et al. Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics 9.1, 1-13 (2018)
8. Tamburini, Fabio. How ”BERTology” Changed the State-of-the-Art also for Italian NLP. CLiC-it (2020)
9. Viani, Natalia, et al. Information extraction from Italian medical reports: An ontology-driven approach. International journal of medical informatics 111, 140-148 (2018)

Classification of Recommender systems using Deep Learning based generative models

Contribution Title in Italian

Sanae FILALI-ZEGZOUTI, Oumayma BANOUAR and Mohamed BENSLIMANE

Abstract Recommendation systems have experienced great growth in the last two years. Many algorithms have been proposed to introduce new models or improve existing ones. It goes without saying that this is due to the explosive use of the internet for almost everything in our everyday life. The goal for companies using these recommendation systems is to increase profits by proposing products that the user had not even thought of at the beginning of his browsing, and that he seems to like and ends up buying, watching, or testing at least. Traditional recommender systems have been used for a long time. However, their combination with deep learning methods and generative models makes them now more efficient. This paper presents a comparative study of recent recommender systems using deep learning, based generative models.
Abstract *Abstract in Italian*

Key words: : Artificial intelligence, Machine Learning, Deep Learning, Generative models, Recommender systems, VAE, GAN, GPT

1 Introduction

There are several different types of Machine Learning models, including Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Reinforcement Learning, Generative Models, and Recommendation systems. Today, recommendation systems have evolved a lot. They have experienced spectacular growth and popularity in recent years, thanks to the advances and research made in this field. They are algorithms that use data about the user's behavior, preferences, and interactions to suggest items or actions that the user would be interested in. These systems are commonly used in e-commerce (like Amazon, Alibaba, and Netflix), media (like YouTube, and Spotify), and social media platforms (Facebook, Instagram, and LinkedIn) to personalize content and improve user engagement. They can be used also in other industries such as banking, healthcare [1], tourism [2, 3], education [4, 5], or travel. Recommendation systems may use different techniques such as Collaborative filtering [6, 7, 8, 9], or Deep Learning. In this paper, we present generative mod-

---

Sanae FILALI-ZEGZOUTI
Laboratory of Innovative Technologies, Sidi Mohamed Ben Abdellah University, Fez, Morocco e-mail: filali.sanae.stud@gmail.com

Oumayma BANOUAR
Laboratory of Computer Science Engineering and Systems, Cady Ayyad University, Marrakesh, Morocco e-mail: o.banouar@uca.ac.ma

Mohamed BENSLIMANE
Laboratory of Innovative Technologies, Sidi Mohamed Ben Abdellah University, Fez, Morocco

els based deep learning used for recommendation systems. We discuss their strengths and limits and compare them according to several criteria such as their components, complexity and training process.

2 Traditional types of Machine Learning models vs Generative models

Supervised Learning is where the model is trained on labeled data to predict outcomes for new, unseen data. There are two types of supervised Learning, Classification, and Regression. Examples of supervised learning algorithms include linear regression, logistic regression, and KNN [11, 10]. In unsupervised Learning, the model is trained on unlabeled data to find out patterns or structures in the data. Some types of unsupervised learning are Clustering, Dimensionality reduction, and Association rule learning. Examples of unsupervised learning algorithms include k-means [3, 7, 10], and PCA (Principal Component Analysis) [12, 13]. Semi-supervised Learning is a combination of supervised and unsupervised Learning. In this case, the algorithm is trained on a dataset that contains both labeled and unlabeled to improve the performance of a model. The model is generally trained on a small amount of labeled data and a large amount of unlabeled data. With reinforcement Learning [14], the model learns to make decisions by receiving rewards or penalties for certain actions. Some examples of reinforcement Learning models are Q-Learning, SARSA (State-Action-Reward-State-Action), and Monte Carlo (MC).

Generative models are a type of Machine Learning model that can generate new unseen data samples, that are similar to real data present in the training set. Generative models can be used in a variety of ways as data augmentation to increase the amount of training data, to generate synthetic data [17] when real data is limited or expensive to obtain, or as a pre-processing step to extract features that are used in the supervised or unsupervised model.

3 Recommendation systems

In Collaborative Filtering (CF) [6, 7], the past behavior of users is collected, analyzed and used to predict recommendations. There are two types of this approach: user-based and item-based.

- User based [6]: For example, if user 1 and user 2 have a similar behavioral history. Then, there is a high probability that they will be interested in the same products in the future. So later, when user 1 is interested in a product, that product will be recommended to user 2.
- Item based [6]: in this approach, a user will certainly like similar products to those he has already liked in the past.

Content-based (CB) [6] systems focus on the characteristics or attributes of the items, that a user had shown interest in them, in order to make recommendations for him. For example, if a user has already shown interest in a movie or a song, a content-based recommendation system could recommend videos of the same actor, genre, etc. Popularity-based Recommendation makes recommendations based on the popularity of items, which is simple but can be effective in some cases.

Hybrid Systems [8, 9] combine all the approaches above to make recommendations. In general, they are usually more scalable and effective, but also more expensive than Collaborative, and Content-Based Filtering.

To summarize, Collaborative filtering is good for handling a large number of users and items. While Content-based filtering is good for handling new users and items with no past interactions. Hybrid systems are good for taking advantage of the strengths of those different methods, overcoming the limitations of each method used individually. Nevertheless, they are more complex to implement and evaluate than the individual methods. They require more data and computer resources, and the results can be difficult to interpret in many cases. According to Hyeyoung Koet et al in [15], there has been a big interest in research on hybrid recommendation systems since 2014, which has resulted in an increase in the number of studies and research on this topic since this year.

4 Generative models for recommendation systems

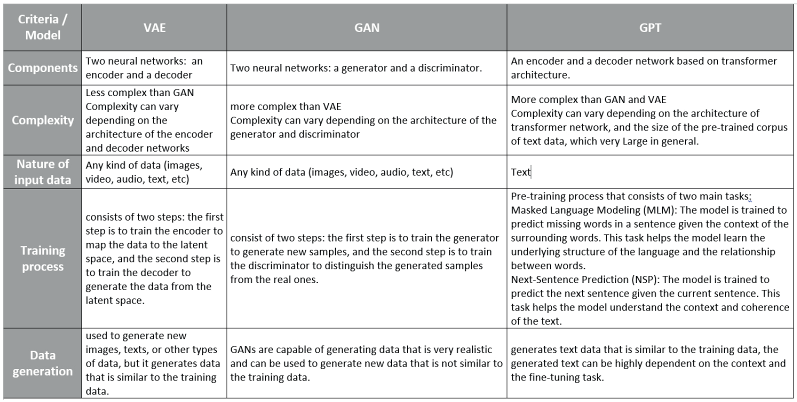
Deep learning is a subset of machine learning, called also Artificial Neural Network. It is inspired by the structure and function of the human brain. Deep learning is often used for complex tasks. It is based on the use of neural networks with multiple layers of interconnected nodes, called artificial neurons, which process and transmit information. The performance of deep learning algorithms increases with an increase in the amount of data and the layers of the neural network.
Deep Learning-based methods are good for handling large and complex datasets. They are more accurate than traditional methods, but they may require a large amount of data to train the model, and they can be computationally expensive and difficult to interpret. Deep Learning-based methods can be seen as an advancement or an alternative to traditional methods. We can combine all these different models to take advantage of the strengths of each, and it often improves performance in most cases. For example, using a combination of clustering and classification for customer segmentation. First, clustering is used to group similar customers together, and then a supervised learning model is trained to classify each cluster into specific segments. Another example consists of the combination of a Deep Learning model with recommendation systems by using generative models. In this case, Generative models can be used to generate new recommendations for users based on their past behavior by using complex Deep Learning algorithms. There are several types of generative models using deep learning. Table 1 listed recent works related to this approach. In the following section, we will focus on the models: VAE, GAN, and GPT. The other solutions are their variants. VAE is a type of generative model that learns to generate new data by learning a compact representation of the input data in the form of a continuous latent variable. The VAE consists of two main components: an encoder and a decoder. The encoder is a neural network that takes the input data and maps it to a lower-dimensional, continuous latent space. The encoder learns to compress the input data into a compact representation by maximizing the likelihood of the input data given the latent variables. The decoder is another neural network that maps the latent variables to the original input space. The decoder learns to generate new samples by sampling from the latent space and then mapping them back to the input space. VAEs are trained to minimize the

Table 1 Recent works using Generative models based deep learning

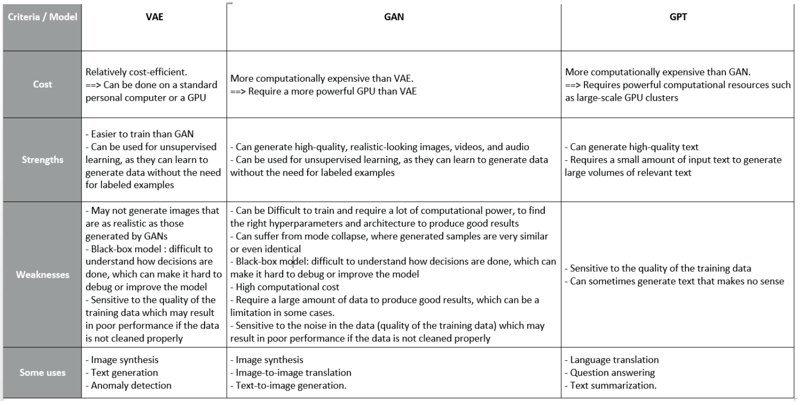
Generative models	Name of the Generative Model	Year
RBM	Restricted Boltzmann Machines	2006
VAE	Variational Autoencoder	2013
GAN	Generative Adversarial Network	2014
cGANs*	Conditional GANs	2014
AAE	Adversarial Autoencoder	2016
FBGM	Flow-based generative models	2016
DCGANs*	Deep Convolutional GANs	2016
InfoGANs*	InfoGANs	2016
Cycle GANs*	Cycle GANs	2017
WGAN*	Wasserstein GAN	2017
Transformer-based models	Transformer-based models	2017
GPT	Generative Pre-training Transformer	2018
StyleGAN*	Style-Based Generator Architecture	2018
BigGAN*	BigGAN	2019
GPT-3	GPT-3	2020
DALL-E	DALL-E	2021

*Variant of GAN

difference between the generated samples and the real samples. It can be used for tasks such as an image or speech generation, anomaly detection, and feature extraction. In [16], authors use a VAE model, upstream of other ML Algorithms (XGBoost, CatBoost and LightGBM), to recommend an optimal Nitrogen rate for rice growers. As data related to environmental and agricultural resources suffer usually from discrepancy and noise. Using VAE to reconstruct data, which means generate clean data, has allowed increasing of 4.32% yield compared to using normal data (without VAE). It is based on a dataset of 4884 records that contains characteristics of soil and climatic data, remote sensing indices, and farming practices. That information is collected from extended surfaces cultivated intensively with rice for the last 5 years prior to the article (2017–2021). The Neural Network is composed of 1024 nodes, and a 128-dimension hidden representation was obtained for the encoder. Here, some variants and extensions of VAE: Conditional VAE (CVAE), Adversarial Autoencoder (AAE), Wasserstein VAE, VQ-VAE, and Flow-based VAE model. GAN consists of two neural networks: a generator and a discriminator. The generator creates new data samples, and the discriminator attempts to distinguish between the generated samples and the real samples from the training dataset. The two networks are trained together in an adversarial way. GAN is used to generate realistic things that have never existed as images, video, text, and music, ... It can be used to generate synthetic data to train the recommendation model with more data, and then be more performant. One example in the medical field is about generating 3D synthetic lung [17]. It is used to solve the problem of lack of healthy lungs, to better train the model and so have good results, by using a 3D Progressive Growing GAN, named PGGAN, capable of generating unreal healthy lungs with good resolution. Another example is in the fashion field [18] where the authors proposed a multimodal transformer-based GAN with cross-modal attention. It explores simultaneously visual features and textual attributes. This model applied in fashion recommendation allows to generate a complete set of outfits and decides the number of items in the set. This model does not consider audio features, which could be a possible improvement to have better performance. There are many variants and extensions of GANs that have been proposed recently, including Conditional GANs, Deep Convolutional GANs, Wasserstein GANs, InfoGANs, Cycle GANs, and each of these algorithms has their strengths and weaknesses. Their use will depend on the problem and the types of input data. Generative Pre-trained Transformer (GPT) is a type of language model introduced by OpenAI in 2018. It uses a deep neural network with a transformer architecture to generate human-like text. GPT is pre-trained on a massive dataset of text from the internet and fine-tuned on specific tasks such as language translation, question answering, and text summarization. This allows it to generate high-quality, coherent text in a variety of languages and styles with remarkable accuracy. GPT-3 has been a real achievement in the field of language modeling and has generated a lot of enthusiasm in the research as well as the industry. It is the largest language model to date, with over 175 billion parameters. Today, GPT models are largely used in natural language processing and have become a major asset in the field. Below is a table comparing these three models based on the following criteria: architecture, complexity, nature of input data, training process, data generation, requirement, cost, strengths, weaknesses, and some uses.GPT is now the way toward conversational recommender systems (CRS) aiming to recommend high-quality items to users through interactive conversations[19, 20].



Criteria / Model	VAE	GAN	GPT
Components	Two neural networks: an encoder and a decoder	Two neural networks: a generator and a discriminator.	An encoder and a decoder network based on transformer architecture.
Complexity	Less complex than GAN Complexity can vary depending on the architecture of the encoder and decoder networks	more complex than VAE Complexity can vary depending on the architecture of the generator and discriminator	More complex than GAN and VAE Complexity can vary depending on the architecture of transformer network, and the size of the pre-trained corpus of text data, which very Large in general.
Nature of input data	Any kind of data (images, video, audio, text, etc)	Any kind of data (images, video, audio, text, etc)	Text
Training process	consists of two steps: the first step is to train the encoder to map the data to the latent space, and the second step is to train the decoder to generate the data from the latent space.	consist of two steps: the first step is to train the generator to generate new samples, and the second step is to train the discriminator to distinguish the generated samples from the real ones.	Pre-training process that consists of two main tasks: Masked Language Modeling (MLM): The model is trained to predict missing words in a sentence given the context of the surrounding words. This task helps the model learn the underlying structure of the language and the relationship between words. Next-Sentence Prediction (NSP): The model is trained to predict the next sentence given the current sentence. This task helps the model understand the context and coherence of the text.
Data generation	used to generate new images, texts, or other types of data, but it generates data that is similar to the training data.	GANs are capable of generating data that is very realistic and can be used to generate new data that is not similar to the training data.	generates text data that is similar to the training data, the generated text can be highly dependent on the context and the fine-tuning task.



Criteria / Model	VAE	GAN	GPT
Cost	Relatively cost-efficient. ==> Can be done on a standard personal computer or a GPU	More computationally expensive than VAE. ==> Require a more powerful GPU than VAE	More computationally expensive than GAN. ==> Requires powerful computational resources such as large-scale GPU clusters
Strengths	- Easier to train than GAN - Can be used for unsupervised learning, as they can learn to generate data without the need for labeled examples	- Can generate high-quality, realistic-looking images, videos, and audio - Can be used for unsupervised learning, as they can learn to generate data without the need for labeled examples	- Can generate high-quality text - Requires a small amount of input text to generate large volumes of relevant text
Weaknesses	- May not generate images that are as realistic as those generated by GANs - Black-box model : difficult to understand how decisions are done, which can make it hard to debug or improve the model - Sensitive to the quality of the training data which may result in poor performance if the data is not cleaned properly	- Can be Difficult to train and require a lot of computational power, to find the right hyperparameters and architecture to produce good results - Can suffer from mode collapse, where generated samples are very similar or even identical - Black-box model: difficult to understand how decisions are done, which can make it hard to debug or improve the model - High computational cost - Require a large amount of data to produce good results, which can be a limitation in some cases. - Sensitive to the noise in the data (quality of the training data) which may result in poor performance if the data is not cleaned properly	- Sensitive to the quality of the training data - Can sometimes generate text that makes no sense
Some uses	- Image synthesis - Text generation - Anomaly detection	- Image synthesis - Image-to-image translation - Text-to-image generation.	- Language translation - Question answering - Text summarization.

5 Conclusion

Generative models have been around since the early years of machine learning and artificial intelligence. But their growth has multiplied since the progress that deep learning models have achieved. Much recent research has combined deep learning and generative models to develop new architectures and techniques. These models have shown promising results in improving the performance of recommendation systems and providing more personalized recommendations.

References

1. Gaël Varoquaux, Veronika Cheplygina, Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine, Vol 5, Article number: 48, 2022.
2. Saman Forouzandeh, Mehrdad Rostamib, Kamal Berahmand, A Hybrid Method for Recommendation Systems based on Tourism with an Evolutionary Algorithm and Topsis Model, Pages 26-50, doi.org/10.1080/16168658.2021.2019430, 2022.

3. najmeh samani, Somayeh Aliyari, Mohammadreza Jelokhani, Developing a group urban tourism recommendation system based on the modied k-means algorithm and fuzzy best-worst method, researchsquare, doi: https://doi.org/10.21203/rs.3.rs-2500314/v1, 2023.
4. Sadia Ali, Yaser Hafeez, Mamoona Humayun, Nor Shahida Mohd Jamail , Muhammad Aqib , Asif Nawaz, Enabling recommendation system architecture in virtualized environment for e-learning, sciencedirect, Vol 23, Issue 1, Pages 33-45, doi.org/10.1016/j.eij.2021.05.003, 2022.
5. Karim Dahdouh , Ahmed Dakkak, , Lahcen Oughdir, and Abdelali Ibriz, Large-scale e-learning recommender system based on Spark and Hadoop, Journal of Big Data, vol 6, Article number: 2, doi.org/10.1186/s40537-019-0169-4, 2019.
6. Mohd Abdul Hameed et al., Collaborative Filtering Based Recommendation System: A survey, International Journal on Computer Science and Engineering, Vol 4, Issue 5, 2012.
7. HHui Jing, Application of Improved K-Means Algorithm in Collaborative Recommendation System, Journal of Applied Mathematics, Vol 2022, Article ID 2213173, doi.org/10.1155/2022/2213173, 2022.
8. Banouar Oumayma and Raghay Said, Enriching SPARQL Queries by User Preferences for Results Adaptation. International Journal of Software Engineering and Knowledge Engineering, vol. 28, issue 8, pp.1195–1221, 2018.
9. Ouedrhiri Oumayma, Banouar Oumayma, El hadaj Salah and Raghay Said, Intelligent recommender system based on quantum clustering and matrix completion, Concurrency and Computation Practice and Experience, vol.34, issue 3,DOI: 10.1002/cpe.6943, 2022.
10. R. Ahuja, A. Solanki and A. Nayyar, Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor, 9th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India, 2019, pp. 263-268, doi: 10.1109/CONFLUENCE.2019.8776969, 2019.
11. Airen, S., Agrawal, J. Movie Recommender System Using K-Nearest Neighbors Variants. National Academy Science Letters, 45, 75–82, doi.org/10.1007/s40009-021-01051-0, 2022.
12. Bandyopadhyay, S., Thakur, S.S. and Mandal, J.K. Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society. Innovations Syst Softw Eng 17, 45–52 , 2021.
13. Manolis G. Vozalis and Konstantinos G. Margaritis, A Recommender System using Principal Component Analysis, researchgate, 2008.
14. M. Mehdi Afsar, Trafford Crump, Behrouz Far, Reinforcement Learning based Recommender Systems: A Survey, arxiv.org, Version, v2, doi.org/10.48550/arXiv.2101.06286, 2022.
15. Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi, A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields, Electronics, vol.11, issue 1, doi:https://doi.org/10.3390/electronics11010141, 2022.
16. Miltiadis Iatrou, Christos Karydas, Xanthi Tseni, and Spiros Mourelatos, Representation Learning with a Variational Autoencoder for Predicting Nitrogen Requirement in Rice, Remote Sensing, vol.14, issue 23, doi: https://doi.org/10.3390/rs14235978, 2022.
17. Ferreira Artur, Pereira Tania, Silva Francisco, Vilares Ana T, Silva Miguel C , Cunha Antonio, Oliveira Helder P, Synthesizing 3D Lung CT scans with Generative Adversarial Networks, Proceedings of Annu Int Conf IEEE Eng Med Biol Soc, p:2033-2036. doi: 10.1109/EMBC48229.2022.9871481,Jul;2022.
18. Volokha Valery and Bochenina Klavdiya, Content-aware generative model for multi-item outfit Recommendation, Content-Aware Generative Model for Multi-item Outfit Recommendation. In: Groen, D., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds) Computational Science – ICCS 2022. ICCS 2022. Lecture Notes in Computer Science, vol 13350. Springer, Cham. https://doi.org/10.1007/978303108751612.
19. Zhang Yizhe, Sun Siqi, Galley Michel, Chen Yen-Chun, Brockett Chris, Gao Xiang, Gao Jianfeng, Liu Jingjing, and Dolan Bill, DIALOGPT: Large-Scale generative pre-training for conversational response generation. In Proceedings of System Demonstrations, ACL 2020. 270–278.
20. Aljbawi, Bushra, Health-aware Food Planner: A Personalized Recipe Generation Approach Based on GPT-2. Theses and Dissertations (Comprehensive). 2311, 2020.

Sparse Inference in Gaussian Graphical Models via Adaptive Non-Convex Penalty Function

Inferenza Sparsa nei Modelli Grafici Gaussiani Mediante Funzione di Penalizzazione Non-Convessa Adattiva

Daniele Cuntrera, Vito M.R. Muggeo, Luigi Augugliaro

Abstract In this paper, we propose a new penalized estimator for sparse inference in Gaussian Graphical Models, which is grounded on the adaptive non-convex penalty function initially proposed in [3]. The proposed estimator enhances several advantages with respect to the non-convex alternatives, such as SCAD and MCP, since we can control the amount of non-convexity of the objective function through a second tuning parameter and, consequently, we can remove the inferential problems related to the existence of multiple local minima. The performance of the proposed estimator is evaluated through a simulation study.

Abstract *In questo articolo, proponiamo un nuovo stimatore penalizzato per condurre inferenza sparsa nei modelli Gaussiani grafici, basato sulla funzione di penalità non convessa adattativa inizialmente proposta in [3]. Lo stimatore proposto presenta diversi vantaggi rispetto agli altri stimatori basati su funzioni di penalizzazione non convesse, come SCAD e MCP, in quanto possiamo controllare il grado di non convessità della funzione obiettivo attraverso un secondo parametro di tuning e, di conseguenza, possiamo rimuovere i problemi inferenziali legati all'esistenza di minimi locali multipli. La performance dello stimatore proposto è valutata attraverso uno studio di simulazione.*

Key words: Gaussian Graphical Models, high-dimensional data, non-convex penalty function, penalized inference, sparse inference

---

Daniele Cuntrera
University of Palermo, Department SEAS, e-mail: daniele.cuntrera@unipa.it

Vito M.R. Muggeo
University of Palermo, Department SEAS, e-mail: vito.muggeo@unipa.it

Luigi Augugliaro
University of Palermo, Department SEAS, e-mail: luigi.augugliaro@unipa.it

1 Introduction

Gaussian Graphical Models (GGM) are probabilistic graphical models based on the assumption that the random vector $X \sim N(\mu, \Sigma)$:

$$f(x;\mu,\Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-1/2(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\}, \tag{1}$$

The inverse of the covariance matrix, denoted by Θ, is called *precision matrix* and its off-diagonal elements, denoted as θ_{hk}, are the parametric tools by which density (1) factorizes according to the undirected graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$, where the edge-set $\mathscr{E}$ encodes the conditional dependence and independence structure among the p random variables. Formally, $(h,k) \notin \mathscr{E} \Leftrightarrow X_h \perp\!\!\!\perp X_k \mid X_{\mathscr{V}\setminus\{h,k\}} \Leftrightarrow \theta_{hk} = 0$ (see [5] for more details), therefore, our goal is to infer μ and Θ, as well as recover the conditional dependence structure encoded by $\mathscr{E}$. Suppose that a set of n i.i.d. observations are drawn from density (1). In principle, inference on $\mathscr{E}$ can be carried out by maximizing the profile log-likelihood function $\ell(\Theta) = \log\det\Theta - \mathrm{tr}(S\Theta)$, where S is the empirical covariance matrix. Then, the edge-set $\mathscr{E}$ the can be estimated by $\widehat{\mathscr{E}} = \{(h,k); \hat{\theta}_{hk} \neq 0\}$.

Although the procedure described above is theoretically well-founded, the application to real datasets is limited for two main reasons. Firstly, the number of measured variables is often larger than the sample size, implying the non-existence of the maximum likelihood estimator of the precision matrix. Secondly, the maximum likelihood estimator will exhibit very high variance even when the sample size is large enough. In terms of GGMs, this evidence translates into the assumption that Θ has a sparse structure; consequently, a number of authors have proposed a penalized approach to estimate Θ and $\mathscr{G}$ at the same time. We refer the interested reader to [1] for a review on sparse GGMs.

2 Sparse inference via adaptive non-convex penalty function

The proposed model. Following the approach presented in [7], in this paper, we propose to estimate Θ and $\mathscr{E}$ using a penalized approach grounded on the adaptive non-convex penalty function introduced in [3]. Formally, the proposed estimator is defined as follows:

$$\widehat{\Theta} = \arg\min_{\Theta \succ 0} -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P(|\theta_{hk}|/\nu_{hk}), \tag{2}$$

where $\ell(\cdot)$ is the log-likelihood and $P(|\theta_{hk}|/\nu_{hk}) = \nu_{hk}\int_0^{|\theta_{hk}|/\nu_{hk}} \exp\{-x^2/2\}\mathrm{d}x$. In the proposed estimator, $\rho > 0$ is the tuning parameter aimed to control the amount of sparsity in $\widehat{\Theta}$, i.e., if ρ is large enough, some $\hat{\theta}_{hk}$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{\mathscr{E}}$. The additional parameter ν_{hk} allows the penalty function to be flexible enough to mimic the widely popular penalty functions proposed in the literature, e.g. LASSO [6], SCAD [4] and MCP [8]. For example, noting that $\rho \lim_{\nu\to\infty} P'(|\theta_{hk}|/\nu_{hk}) = \rho$ we can conclude immediately that, for large enough ν-values, the results given by the proposed model are approximately equivalent to the one given by the graphical lasso (gLASSO) model [7]. Moreover, ν plays a central role in terms of stability since it acts on the degree of non-convexity of the proposed penalty function. More specifically, like SCAD model, the objective function in (2) can have multiple local minima when ν is small enough. Nevertheless, as discussed in [3], it is possibile to show that exists a unique lower bound, denoted by $\nu_{\min}$, such that the propose objective function have a unique global minimum, for any ρ-value.

Computational aspects: an efficient ADMM algorithm. We propose an efficient Alternating Direction Method of Multipliers (ADMM) to compute the proposed estimator. We begin defining the solution of our minimization problem (2) as the solution of the following equality-constrained minimization problem, with matrix variables Θ and Z:

$$\begin{aligned} \min_{\Theta, Z \succ 0} \quad & -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right), \\ \text{s.t.} \quad & \Theta - Z = 0. \end{aligned}$$

According to the standard ADMM theory, the augmented scaled Lagrangian function takes the form:

$$\mathscr{L}(\Theta, Z, U) = -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right) + \frac{\tau}{2}\|\Omega - Z + U\|_F^2 - \frac{\tau}{2}\|U\|_F^2, \tag{3}$$

where $\tau > 0$ is a penalty parameter, $U \succ 0$ is the scaled dual matrix and $\|\cdot\|_F$ denotes the Frobenius norm, respectively. Using (3), the solution of the problem (2) can be computed through the following procedure:

```
1: **repeat**
2:     $\Theta^{k+1} = \arg\min_{\Theta\succ 0} -\ell(\Theta) + \frac{\tau}{2}\|\Theta - Z^k + U^k\|_F^2$,
3:     $Z^{k+1} = \arg\min_{Z\succ 0} \frac{\tau}{2}\|\Theta^{k+1} - Z + U^k\|_F^2 + \rho\sum_{h,k}^{p} P(|Z_{hk}|/\nu_{hk})$,
4:     $U^{k+1} = U^k + \Theta^{k+1} - Z^{k+1}$
5: **until** convergence criterion is met.
```

As we shall show, the main advantage of the ADMM algorithm consists of the ability to split the initial complex problem into a series of simpler problems, each of which is easy to solve. In the remaining part of this section, we shall study in detail the minimization problems defined in Step 2 and 3.

Updating Θ. The problem in Step 2 has been studied in [2], where the authors show that Θ^{k+1} admits the solution in closed form. To get more insight into the updating formula, consider the first-order optimality condition of the problem in

Step 2, which can be rewritten in the following more convenient form:

$$\tau\Theta - \Theta^{-1} = \tau(Z^k - U^k) - S. \tag{4}$$

Let $Q\Lambda Q^\top$ be the spectral decomposition of $\tau(Z^k - U^k) - S$, then from the equation (4) we can immediately conclude that Θ^{k+1} can be written as $Q\widetilde{\Lambda}Q^\top$, where $\widetilde{\Lambda}$ is a diagonal matrix whose elements are solutions of the equation $\tau\widetilde{\Lambda} - \widetilde{\Lambda}^{-1} - \Lambda = 0$, i.e.,

$$\tilde{\lambda}_{ii} = \left(\lambda_{ii} + \sqrt{\lambda_{ii}^2 + 4\tau}\right)/(2\tau),$$

which are positive since $\tau > 0$.

Updating Z. Before delving into the technical details related to the update of the matrix Z, we note that the objective function in Step 3 takes the additive structure $\frac{\tau}{2}\sum_{h,k=1}^{p}\left\{(\theta_{hk}^{k+1} + U_{hk}^{k} - Z_{hk})^2 + \rho P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right)\right\}$, which implies that the minimization problem in Step 3 can be split into $p(p+1)/2$ univariate optimization problems that can be solved in parallel. Therefore, in the remaining part of this section, we focus on the sub-problem

$$Z_{hk}^{k+1} = \arg\min_{Z_{hk}} \frac{\tau}{2}(\theta_{hk}^{k+1} + U_{hk}^{k} - Z_{hk})^2 + \rho P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right).$$

Following [3], we solve the problem above using the local linear approximation (LLA) method [9], that is, Z_{hk}^{k+1} is computed as solution of a sequence of new minimization problems involving a new objective function obtained replacing the penalty function with a suitable local approximation. Formally, Z_{hk}^{k+1} is obtained by the following iterative procedure:

```
1: Let $\widetilde{Z}_{hk}^{k}$ be a starting value
2: repeat
3:     Let $w_{hk} = \exp\{-(\widetilde{Z}_{hk}^{k}/\nu_{hk})^2/2\}$
4:     $\widetilde{Z}_{hk}^{k+1} = \arg\min_{\widetilde{Z}_{hk}} \frac{1}{2}(\theta_{hk}^{k+1} + U_{hk}^{k} - \widetilde{Z}_{hk})^2 + \frac{\rho}{\tau} w_{hk}|\widetilde{Z}_{hk}|$
5: until convergence criterion is met
6: Return $Z_{hk}^{k+1} = \widetilde{Z}_{hk}^{k+1}$
```

In Step 4, we immediately recognize a weighted lasso problem; therefore, we have that the updating step of $\widetilde{Z}_{hk}^{k+1}$ admits solution in closed form, i.e.,

$$\widetilde{Z}_{hk}^{k+1} = S(\theta_{hk}^{k+1} + U_{hk}^{k}; \rho w_{hk}/\tau),$$

where $S(x;\lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator.

3 Simulation Study

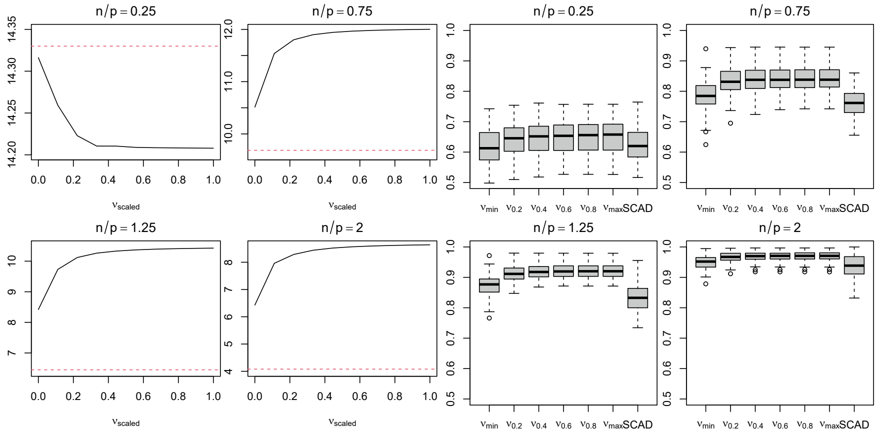
We run some simulation experiments to compare the behaviour of the estimator with the one obtained using the SCAD penalty function. As discussed above, the gLASSO model is a limiting case of our proposal, i.e., as the tuning parameters ν go to infinity, the estimates given by the estimator (2) are asymptotic equivalent to the ones given by gLASSO model. Therefore, in this study, we are aimed to evaluate the effects of ν on the entire path of $\widehat{\Theta}$. More specifically, we used an evenly spaced sequence of six ν-values, from ν_{min} to ν_{max} = 3, where the largest ν-value was chosen after a preliminary study to ensure that our estimates are approximately equal to gLASSO ones. To appreciate the effects of ν for different values of the ratio n/p, we set $p = 50$ and considered four different sample-size, i.e., $n = (13, 38, 63, 100)$. We defined the edge set as a collection of J disjoint edge sets encoding star structures to simulate a sparse precision matrix. Formally, $\mathscr{E} = \cup_{j=1}^{J} \mathscr{E}_j$, where $\mathscr{E}_j = \{(j,k) : k = (j+1), \dots (j+k_j)\}$. In our setting, $k_j = 24$ and, consequently, $J = 2$. The corresponding off-diagonal entries of Θ are simulated using a uniform distribution on the interval $(0.65, 1)$. In contrast, diagonal entries are computed to make the resulting precision matrix positively definite. For each possible combination of ν and n, we run 100 simulations.

The performance of the considered models is evaluated in terms of graph structure recovery and Frobenius distance between estimates and true precision matrix. Concerning the first aspect, we summarize the coefficient paths using the AUC based on the ROC curves in each simulation run. Concerning the Frobenius distances, we take the minimum value of the entire path of each estimator; then, we average the minima across replicates: in this way, we are looking at the best performance that all estimators can reach.

Figure 1 (a) displays the curves of the Frobenius distances varying ν and the value obtained using the SCAD penalty. When the number of parameters is much larger than the number of observations, SCAD has the maximum average Frobenius distance. At the same time, it is observed that using our proposal, the distance decreases as ν increases (although, in absolute terms, the decrease is not large). As the ratio n/p increases, it is observed that the lowest value of the distance to the true matrix Θ is obtained using the smallest value of ν, i.e. ν_{min}. The advantage concerning the opposite limit case (i.e., gLASSO) increases as the ratio n/p increases. Finally, by considering the capacity of the correct selection of non-zero coefficients (and thus the boxplots of AUCs), it is observed that our proposal always gives better results than SCAD. For different values of ν, the AUC tends to improve as the value of ν increases, but the differences decrease as the ratio n/p increases.

4 Conclusions

We have presented a new approach to estimate penalized graphical models, using the penalty function given for the first time on [3]. The penalty function is a trade-



(a) Frobenius distances (b) AUC

Fig. 1 Frobenius distances (left panel) and AUC (right panel) for SCAD (red dotted line in the left panel) and our proposal varying ν.

off between the LASSO and the L0 penalties. From the evaluations based on the simulations study, it can be seen that it may be crucial to propose a method for selecting the ν parameter, which certainly depends on the n/p ratio: for lower values, the ν optimum will tend to take high values (and thus, the solution is very close to the gLASSO ones); conversely, when the n/p ratio is large, the ν optimum will take values close to ν_{min}.

Acknowledgements. Luigi Augugliaro and Vito M.R. Muggeo gratefully acknowledge financial support from the University of Palermo (FFR2021-22).

References

1. Augugliaro, L., Mineo, A. M., and Wit, E. C. *ℓ_1-Penalized Methods in High-Dimensional Gaussian Markov Random Fields*. John Wiley & Sons, Ltd, 2016, ch. 8, pp. 201–265.
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*, 1 (2011), 1–122.
3. Cuntrera, D., Muggeo, V. M. R., and Augugliaro, L. Variable selection with unbiased estimation: the CDF penalty. In *51th Scientific Meeting of the Italian Statistical Society: Book of Short Papers* (2022), pp. 1835–1840.
4. Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*, 456 (2001), 1348–1360.
5. Lauritzen, S. L. *Graphical Models*. Oxford University Press, Oxford, 1996.
6. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

7. YUAN, M., AND LIN, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, 1 (2007), 19–35.
8. ZHANG, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*, 2 (2010), 894–942.
9. ZOU, H., AND LI, R. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*, 4 (2008), 1509–1533.

Bayesian causal inference from discrete networks

Federico Castelletti and Guido Consonni

Abstract We consider a collection of categorical variables whose joint distribution encodes a set of conditional independencies that can be represented through a Directed Acylic Graph (DAG). Focusing on one variable in the system, we are interested in evaluating the causal effect following an hypothetical intervention on another variable. The latter crucially depends on the underlying DAG structure which is typically unknown and accordingly must be inferred from the available data. We propose a Bayesian methodology which combines structure learning of DAGs and causal effect estimation.

Key words: Bayesian model selection, categorical data, causal inference, directed acyclic graph

1 Introduction

Graphical models based on Directed Acyclic Graphs (DAGs) are widely employed for representing dependence relationships between variables in multivariate settings [7]. Typically the underlying DAG structure is unknown and accordingly must be inferred from the available data, a process known as *structure learning*. To this end, several methodologies, both frequentist and Bayesian, have been proposed; see for instance [4, 5] and [2, 3] respectively. Under causal assumptions on the data generating mechanism underlying the graph, DAGs can be adopted for *causal inference*

Federico Castelletti
Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milan,
e-mail: federico.castelletti@unicatt.it

Guido Consonni
Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milan,
e-mail: guido.consonni@unicatt.it

purposes, namely to evaluate the effect on a variable following an intervention which is applied to another variable of the system [8].

When all variables are discrete, the allied multivariate joint distribution follows a categorical DAG model, whose parameter corresponds to a collection of conditional probabilities representing parent-child relations between nodes in the DAG. It can be shown that any causal effect between variables can be expressed as a function of the DAG-parameter. The latter is however dependent on the underlying DAG which is typically unknown. As a consequence, a unified approach which combines structure learning, parameter inference and causal effect estimation is required.

We propose a Bayesian methodology which computes the posterior distribution of any causal effect of interest based on a DAG model for categorical data. Our method can be extended to settings with uncertainty on the underlying DAG structure using Bayesian Model Averaging (BMA) techniques.

2 Methodology

2.1 *Data distribution and likelihood*

Consider a DAG $\mathscr{D} = (V, E)$, where $V = \{1, \dots, q\}$ is a set of nodes each associated with a categorical variable in the collection $\{X_1, \dots, X_q\}$, and $E \subset V \times V$ is a set of directed edges. Let $\mathscr{X}_j$ be the set of levels of X_j, $j \in V$, $\mathscr{X} := \times_{j \in V} \mathscr{X}_j$ the product space generated by the levels of the q variables and $x = (x_1, \dots, x_q) \in \mathscr{X}$ a generic element of $\mathscr{X}$. Additionally, for each node j, let $\mathrm{pa}(j)$ be its set of parents in the DAG, namely the set of all nodes $u \in V$ for which $u \to j$ is in $\mathscr{D}$.

For $S \subseteq V$, we let $X_S = (X_j, j \in S)$ and $\mathscr{X}_S := \times_{j \in S} \mathscr{X}_j$. We also let $\theta^{j \,|\, \mathrm{pa}(j)}_{x_j \,|\, x_{\mathrm{pa}(j)}} = \Pr(X_j = x_j \,|\, X_{\mathrm{pa}(j)} = x_{\mathrm{pa}(j)}, \boldsymbol{\theta})$ be the conditional probability for the event $\{X_j = x_j\}$, given the parent configuration $\{X_{\mathrm{pa}(j)} = x_{\mathrm{pa}(j)}\}$ with $x_{\mathrm{pa}(j)} \in \mathscr{X}_{\mathrm{pa}(j)}$, where $\boldsymbol{\theta}$ is the vector of all conditional probabilities. Under DAG $\mathscr{D}$, the joint probability factorizes as

$$p(X = x \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{q} p\big(X_j = x_j \,|\, X_{\mathrm{pa}(j)} = x_{\mathrm{pa}(j)}, \boldsymbol{\theta}\big), \tag{1}$$

where $X = (X_1, \dots, X_q)^\top$ and similarly $x = (x_1, \dots, x_q)^\top$. We refer the reader to [7] for further notation. Under a random sample of size n from (1) collected in the (n, q) data matrix $\boldsymbol{X}$, whose i-th row is $(x_{i1}, \dots, x_{iq})$, the likelihood function can be written as

$$p(\boldsymbol{X} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{q} \left\{ \prod_{k \in \mathscr{X}_{\mathrm{pa}(j)}} \left\{ \prod_{m \in \mathscr{X}_j} \left\{ \theta^{j \,|\, \mathrm{pa}(j)}_{m \,|\, k} \right\}^{n^{\mathrm{fa}(j)}_{(m,k)}} \right\} \right\}, \tag{2}$$

where $\mathrm{fa}_{\mathscr{D}}(j) = j \cup \mathrm{pa}_{\mathscr{D}}(j)$ is the family of node j in $\mathscr{D}$ and $n_{(m,k)}^{\mathrm{fa}(j)}$ the number of observations corresponding to configuration (m,k) of variables in $\mathrm{fa}(j)$, with $\sum_{j \in V} \sum_{m \in \mathscr{X}_j} \sum_{k \in \mathscr{X}_{\mathrm{pa}(j)}} n_{(m,k)}^{\mathrm{fa}(j)} = n$.

2.2 Prior parameter distribution

We now proceed by assigning a prior to the DAG-dependent parameter $\boldsymbol{\theta}$. Specifically, consider for each $j \in V$ and $x_{\mathrm{pa}(j)} \in \mathscr{X}_{\mathrm{pa}(j)}$ the vector-parameter

$$\left(\theta_{x_j \,|\, x_{\mathrm{pa}(j)}}^{j \,|\, \mathrm{pa}(j)}, x_j \in \mathscr{X}_j\right) := \boldsymbol{\theta}_{x_{\mathrm{pa}(j)}}^{j \,|\, \mathrm{pa}(j)}. \tag{3}$$

We introduce the following assumptions on the collection of parameters above.

- *global parameter independence*, i.e. independence across $j \in V$;
- *local parameter independence*, i.e independence across $x_{\mathrm{pa}(j)} \in \mathscr{X}_{\mathrm{pa}(j)}$.

Therefore, assuming (independent) Dirichlet distributions with hyper-parameters $\{\boldsymbol{a}_k^{j \,|\, \mathrm{pa}(j)}, j \in V, k \in \mathscr{X}_{\mathrm{pa}(j)}\}$ on each vector-parameter in (3), the joint prior on $\boldsymbol{\theta} = (\boldsymbol{\theta}_k^{j \,|\, \mathrm{pa}(j)}, j \in V, k \in \mathscr{X}_{\mathrm{pa}(j)})$ can be written as

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{q} \left\{ \prod_{k \in \mathscr{X}_{\mathrm{pa}(j)}} \mathrm{pDir}\left(\boldsymbol{\theta}_k^{j \,|\, \mathrm{pa}(j)} \,\middle|\, \boldsymbol{a}_k^{j \,|\, \mathrm{pa}(j)}\right) \right\}; \tag{4}$$

see also [6] for details and considerations on hyperparameter choices.

2.3 Posterior distribution

Both the likelihood function in (2) and the prior on $\boldsymbol{\theta}$ in (4) admit the same node-by-node factorization. Moreover, because of conjugacy of Dirichlet priors with categorical-data models, it is easy to show that the posterior of $\boldsymbol{\theta}$ is such that

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \boldsymbol{X}) &\propto p(\boldsymbol{X} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\ &= \prod_{j=1}^{q} \left\{ \prod_{k \in \mathscr{X}_{\mathrm{pa}(j)}} \mathrm{pDir}\left(\boldsymbol{\theta}_k^{j \,|\, \mathrm{pa}(j)} \,\middle|\, \boldsymbol{a}_k^{j \,|\, \mathrm{pa}(j)} + \boldsymbol{N}_{\mathrm{fa}(j)}^{k}\right) \right\}, \end{aligned} \tag{5}$$

where $\boldsymbol{N}_{\mathrm{fa}(j)}^{k}$ denotes the collection of counts for variables in $\mathrm{fa}(j)$, obtained by including only those observations corresponding to configuration $k \in \mathscr{X}_{\mathrm{pa}(j)}$. The posterior in (5) thus corresponds to a product of independent (posterior) Dirichlet distributions, so that direct sampling from $p(\boldsymbol{\theta} \,|\, \boldsymbol{X})$ is straightforward.

2.4 Causal effect estimation

Consider a DAG $\mathscr{D}$, implying the factorization in (1), and let $Y \in \{X_1,\dots,X_q\}$ be a variable of interest. We focus on the causal effect on Y following an intervention on $X_v \neq Y$, consisting in the action of setting $X_v = \tilde{x}$. Assuming a deterministic intervention [8] denoted by the do-operator $\mathrm{do}(X_j = \tilde{x})$, the *post intervention* distribution can be written as

$$p\big(X = x \,|\, \mathrm{do}(X_v = \tilde{x}), \boldsymbol{\theta}\big) = \begin{cases} \prod_{j \neq v} p\big(X_j = x_j \,|\, X_{\mathrm{pa}(j)} = x_{\mathrm{pa}(j)}, \boldsymbol{\theta}\big) & \text{if } x_v = \tilde{x} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

When X_v is binary so that $X_v \in \{0,1\}$, the causal effect on Y following an intervention on X_v is defined as

$$\gamma_v(\boldsymbol{\theta}) = \mathbb{E}\big(Y \,|\, \mathrm{do}(X_v = 1), \boldsymbol{\theta}\big) - \mathbb{E}\big(Y \,|\, \mathrm{do}(X_v = 0), \boldsymbol{\theta}\big). \tag{7}$$

Moreover, it can be shown [8] that

$$\begin{aligned} \gamma_v(\boldsymbol{\theta}) = & \sum_{k \in \mathscr{X}_{\mathrm{pa}(v)}} \mathbb{E}\big(Y \,|\, X_v = 1, \boldsymbol{X}_{\mathrm{pa}(v)} = k, \boldsymbol{\theta}\big) \,\mathrm{Pr}\,\big(\boldsymbol{X}_{\mathrm{pa}(v)} = k, \boldsymbol{\theta}\big) \\ & - \sum_{k \in \mathscr{X}_{\mathrm{pa}(v)}} \mathbb{E}\big(Y \,|\, X_v = 0, \boldsymbol{X}_{\mathrm{pa}(v)} = k, \boldsymbol{\theta}\big) \,\mathrm{Pr}\,\big(\boldsymbol{X}_{\mathrm{pa}(v)} = k, \boldsymbol{\theta}\big). \end{aligned} \tag{8}$$

It is straightforward to write the causal effect in terms of the DAG-parameter $\boldsymbol{\theta}$ as

$$\gamma_v(\boldsymbol{\theta}) = \sum_{k \in \mathscr{X}_{\mathrm{pa}(v)}} \left\{ \left(\theta^{Y \,|\, \mathrm{fa}(v)}_{1 \,|\, (1,k)} - \theta^{Y \,|\, \mathrm{fa}(v)}_{1 \,|\, (0,k)} \right) \theta^{\mathrm{pa}(v)}_{k} \right\} \equiv \gamma_v. \tag{9}$$

Importantly, since γ_v is a function of $\boldsymbol{\theta}$, inference on γ_v can be retrieved from the posterior distribution (5).

3 Learning causal effects under model uncertainty

Although omitted from the notation, the causal effect in (9) depends, through $\boldsymbol{\theta}$, on the underlying DAG $\mathscr{D}$, which is typically unknown. When the available data are purely observational, the latter cannot be estimated because DAGs are not identifiable. Accordingly the output of any structure learning procedure is typically a (potentially large) Markov equivalence class of DAGs [4] containing all DAGs which share the same conditional independencies. This represents an issue from the perspective of causal inference because the same intervention experiment can lead to distinct causal effects even when DAGs are Markov equivalent. Moreover, this actually happens whenever the parent set $\mathrm{pa}(v)$ in (8) varies across DAGs belonging to the same equivalence class.

To overcome this difficulty, a *conditional* strategy consists in estimating first an equivalence class and then, for each DAG within the class, obtaining an approximate posterior distribution of each causal effect. Specifically, let $[\mathscr{D}]$ be a Markov equivalence class of DAGs and focus on the causal effect γ_v in (8) for a given response variable Y and intervened node $v \in V$. For each DAG $\mathscr{D}$ we can first sample values of $\boldsymbol{\theta}$ from the posterior distribution $p(\boldsymbol{\theta}\,|\,\boldsymbol{X},\mathscr{D})$ where we now emphasize the dependence on DAG $\mathscr{D}$, and then compute corresponding posterior draws for the causal-effect coefficient γ_v by applying Equation (9), leading to the approximate posterior $\hat{p}(\gamma_v\,|\,,\boldsymbol{X},\mathscr{D})$. The previous procedure can be iterated across all DAGs in $[\mathscr{D}]$.

Finally, having assigned a prior $p(\mathscr{D})$, an approximate weighted (BMA) posterior distribution for γ_v can be computed as

$$\hat{p}(\gamma_v\,|\,\boldsymbol{X}) = \sum_{\mathscr{D}\in[\mathscr{D}]} \hat{p}(\gamma_v\,|\,\mathscr{D},\boldsymbol{X})p(\mathscr{D}\,|\,\boldsymbol{X}). \tag{10}$$

The previous strategy is predicated on a given (estimated) equivalence class of DAGs, so that (10) should be written as $\hat{p}(\gamma_v\,|\,\boldsymbol{X},[\mathscr{D}])$, where $[\mathscr{D}]$ is a Markov equivalence class. A more general procedure would also account for the uncertainty on $[\mathscr{D}]$ through a posterior distribution (see for instance [3]), and then proceed with a further BMA step relative to $p([\mathscr{D}]\,|\,\boldsymbol{X})$; see [1] for a similar strategy implemented in the Gaussian framework.

References

1. Castelletti, F., Consonni, G.: Bayesian inference of causal effects from observational data in Gaussian graphical models. Biometrics **77**, 136–149 (2020)
2. Castelletti, F., Consonni, G., Della Vedova, M., Peluso, S.: Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. Bayesian Analysis **13**, 1231–1256 (2018)
3. Castelletti, F., Peluso, S.: Equivalence class selection of categorical graphical models. Computational Statistics & Data Analysis **164**, 107304 (2022)
4. Chickering, D. M.: Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research **2**(3), 445–498 (2002)
5. Kalisch, M. and Bühlmann, P.: Learning Bayesian Networks: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8**, 613–36 (2007)
6. Heckerman, D., Geiger, D., Chickering, D. M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning **20**(3), 197–243 (1995)
7. Lauritzen, S. L.: Graphical Models. Oxford University Press, Oxford (1996)
8. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)

Sign-Flip tests for Spatial Regression with PDE regularization

Test con Sign-Flip per Regressione Spaziale con regolarizzazione alle equazioni alle derivate parziali

Michele Cavazzutti, Eleonora Arnone, Federico Ferraccioli,
Livio Finos, Laura M. Sangalli

Abstract We develop an innovative inference tool for Spatial Regression with Partial Differential Equation regularization. These regression models have a semiparametric structure, which combines a standard regression on space-varying covariates with a regression on a nonparamteric component. The problem of making inference in this class of models is challenging and little explored. We propose a modification of a resampling procedure, developed in [2]. The resulting test has higher power with respect to the original one. Based on the proposed nonparametric inference tools, we also develop nonparametric confidence intervals. The efficacy of the novel procedure is validated through empirical arguments. The proposed inference methods are employed to study chlorophyll-a concentration in the Mediterranean sea.

Key words: Semiparametric regression, smoothing, nonparametric inference

---

Michele Cavazzutti
MOX - Dipartimento di Matematica, Politecnico di Milano, Milano,
e-mail: michele.cavazzutti@polimi.it

Eleonora Arnone
Dipartimento di Management, Univesità degli studi di Torino, Torino,
e-mail: eleonora.arnone@unito.it

Federico Ferraccioli
Dipartimento di Scienze Statistiche, Univesità degli studi di Padova, Padova,
e-mail: Federico.ferraccioli@unipd.it

Livio Finos
Dipartimento di Scienze Statistiche, Univesità degli studi di Padova, Padova,
e-mail: Livio.Finos@unipd.it

Laura Maria Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Milano,
e-mail: Laura.Sangalli@polimi.it

1 Introduction

In this work we consider Spatial Regression with Partial Differential Equation regularization (SR-PDE), reviewed in [6]. Defining efficient and powerful uncertainty quantification tools for these semiparametric models has been shown to be nontrivial [2]. In [2], some first inference tools for the linear part of these models have been proposed. Specifically both parametric and nonparametric approaches are presented. We show that the resulting tests, although asymptotically exact, may suffer the effect of the bias induced by the smoothing penalization of the SR-PDE model, especially in small sample scenarios. Here we propose an empirically validated procedure, called Enhanced Eigen-Sign-Flip (Enhanced ESF), that allows to reduce the effect of the bias of the Eigen-Sign-Flip (ESF) test in [2]. The resulting test is shown to enjoy a higher power. The procedure is validated empirically, showing that its performance is always better or equal to the ESF test in [2]. Based on this proposal, We also provide an algorithm to produce nonparametric confidence intervals. Finally we apply the proposed methods to the study of chlorophyl-a concentrations in the Mediterranean sea.

2 Enhanced ESF

We consider a SR-PDE model. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. Let $\{\mathrm{p}_i\}_{i=1,\dots,n}$ be a set of n locations distributed over Ω at which we observe the value z_i of a variable of interest, and a vector of covariates $\mathbf{w}_i \in \mathbb{R}^q$. We assume there exist a vector $\boldsymbol{\beta} \in \mathbb{R}^q$ and a function $f : \Omega \longrightarrow \mathbb{R}$ s.t.

$$z_i = \mathrm{w}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i) + \varepsilon_i \tag{1}$$

for $i = 1,\dots,n$, where $\{\varepsilon_i\}_i$ is a set of i.i.d. errors with 0 mean and finite variance σ^2. The unknowns $\boldsymbol{\beta}$ and f are estimated by minimization of the functional

$$J(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \Big(\mathrm{z}_i - \mathbf{w}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i)\Big)^2 + \lambda \int_\Omega (Lf)^2, \tag{2}$$

where $Lf = 0$ is a PDE encapsulating available prior information on the problem under analysis. In SR-PDE the problem (2) is discretized using a finite element basis over the domain Ω. Let $\mathbf{f}$ be the discretization of the nonparametric term f usign such finite element basis, Ψ the matrix of the evaluation of the finite element basis functions at the observed spatial locations, P the discretization of the penalization term in (2), and $Q := I - \Psi(\Psi^\top \Psi)^{-1}\Psi^\top$. Then [6] show that the estimators for $\boldsymbol{\beta}$ and $\mathbf{f}$ are

$$\hat{\mathbf{f}} \; = (\Psi^\top Q \Psi + P)^{-1} \Psi Q \mathbf{z} \tag{3}$$

$$\hat{\boldsymbol{\beta}} = (W^\top W)^{-1} W^\top (\mathbf{z} - \Psi \hat{\mathbf{f}}). \tag{4}$$

To exemplify the application of the model, consider the data illustrated in Figure 1, concerning Chlorophyll-a concentration across the Mediterranean sea. These data are provided by GIOVANNI Modis Aqua satellite, together with the sea surface temperature and particulate organic and inorganic concentrations in the same locations. In this example, the Mediterranean sea is the domain of interest, Ω, discretized with the mesh reported in Figure 2. We consider a SR-PDE model where the observed chlorophyll-a concentration z is explained as the sum of a nonparametric component f and the effect of three covariates of interest: particulate inorganic carbon concentration w_1, particulate organic carbon concentration w_2 and the sea surface temperature w_3, as expressed in (1). In this context we want to asses whether the effect of the covariates is significant. Therefore we are interested in defining tests and confidence intervals for the vector of regression coefficients $\boldsymbol{\beta}$.
Different strategies are available to make inference in the context of semiparametric regression models. The interested reader can find a review of the most common ones in [4, 5, 7], and references therein. In the SR-PDE context, [2] tackles the problem by proposing two parametric approaches, based on pivotal quantities, and a nonparametric one, the Eigen-Sign-Flip, based on the resampling procedure proposed in [3]. The main drawback in these inference approaches is the fact that all the statistics taken into account are only asymptotically exact, while they are biased in finite sample scenarios, due to the presence of the penalization in the regularization. In particular, in the ESF test, the presence of the bias induces a greater variance than expected in the distribution of the sign-flipped statistic. However, we prove that the sign-flipping resampling procedure remains valid even if we selectively chose only a subset of components to be flipped. Thus we proposed a new test, called Enhanced ESF, that identifies the highly biased components, restricting the sign-flipping procedure to the unbiased ones, leading to an inference procedure with higher power.

3 Application to chlorophyll-a concentration in the Mediterranean sea

Chlorophyll-a concentration is relevant to study the ocean carbon cycle and the health of ocean environment. High chlorophyll concentrations may indeed highlight poor water quality due to pollution and water stagnation.

The development of micro-organisms responsible for the production of chlorophyll is influenced by multiple environmental factors, among which there are sun exposure, temperature and pollutants. These factors have a complex spatial distribution that may be addressed in a nonparametric way.

In this application, we aim to model the chlorophyll-a concentration via the SR-PDE model presented in Section 2. We model the concentration as the sum of a field f plus the effect of three covariates of interest, particulate inorganic and organic carbon concentration and sea surface temperature. We perform enhanced ESF test and we compare it with the existing parametric and nonparametric tests. In this case particulate inorganic concentration and particulate organic concentration co-

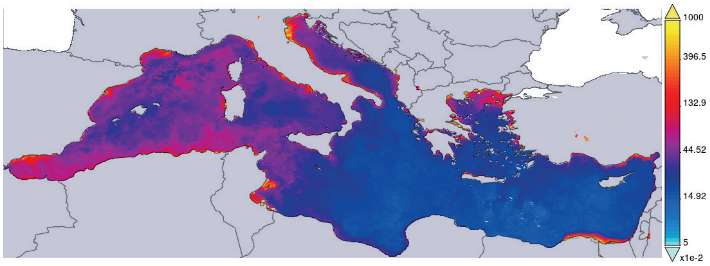


Fig. 1 Chloropyll-a concentration $[mg \cdot m^{-3}]$ represented with logarithmic color scale in Mediterranean sea in February 2019, reproduced using NASA application.

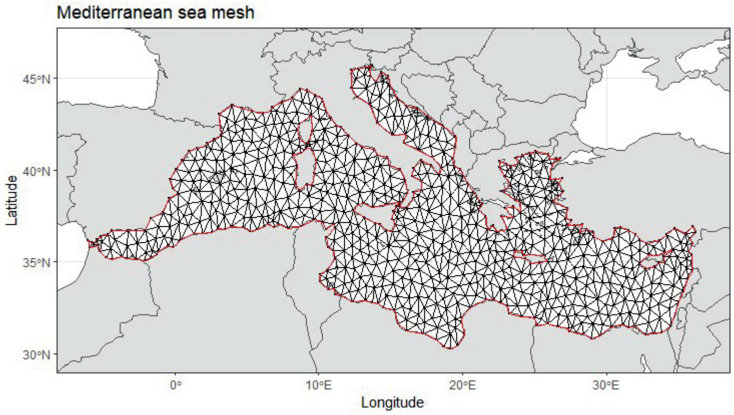


Fig. 2 Mediterranean sea mesh with 809 nodes.

variates are strongly significant, with similar estimates $\hat{\beta}_1 = 0.0474$ and $\hat{\beta}_1 = 0.138$ and p values $p_1 < 10^{-16}$ and $p_2 < 10^{-16}$. Notice that the model estimates a positive effect of the two covariates and the chlorophyll concentration, in agreement to what is known in literature. Indeed the inorganic particulate is the precursor of the Chlorophyll-a production process, while the organic particulate is clearly its byproduct. On the converse for sea surface temperature we obtain a weaker result, with estimate $\hat{\beta}_3 = -0.039$ and p value $p_3 = 0.0456$. Nontheless the estimate is coherent with what is expected from biological considerations: higher sea surface temperature induces higher thermal gradient in sea water, lifting the nutrients and microorganisms living on the seafloor. We deem that the weakness of the effect is due to the month chosen for the analysis, February, when this thermal gradient is in general reduced.

References

1. Demmler A., Reinsch C.: Oscillations Matrices with Spline Smoothing. Numerische Mathematik, **24**: 375-382 (1975)
2. Ferraccioli F., Sangalli L. M., Finos L.: Some first inferential tools for spatial regression with differential regularization. Journal of Multivariate Analysis, **189**: 104866 (2022)
3. Hemerik J., Goeman J., Finos, L.: Robust testing in generalized linear models by sign-flipping score contributions. Journal of the Royal Statistical Society Series B, **82**: 841-864 (2020)
4. Ruppert D., Wand M., Carroll R.: Semiparametric Regression. Cambridge University Press, Cambridge (2003).
5. Harezlak, D., Ruppert D., Wand M.: Semiparametric Regression with R. Cambridge University Press, Cambridge (2018).
6. Sangalli, L. M.: Spatial Regression With Partial Differential Equation Regularisation. International Statistical Review, **89**: 505–531 (2021)
7. Wood S. N.: Generalized Additive Models. Chapman and Hall/CRC, Boca Raton (2017). @articleboffi2010, author = Boffi, Daniele, title = Finite element approximation of eigenvalue problems, journal = Acta Numerica, year = 2010, volume = 1, pages = 1–120, numpages = 120

A novel sequential testing procedure for selecting the number of changepoints in segmented regression models

Una nuova procedura sequenziale per selezionare il numero dei punti di svolta in modelli di regressione segmentata

Andrea Priulla and Nicoletta D'Angelo

Abstract In this work, we address the problem of selecting the number of changepoints in segmented regression models. We propose a novel stepwise procedure and assess its performance through simulation studies. We demonstrate that our proposal behaves well with the Gaussian and Binomial responses.

Abstract *In questo lavoro, affrontiamo il problema della selezione del numero di punti di svolta nei modelli di regressione segmentata. Proponiamo una nuova procedura sequenziale e ne valutiamo la performance attraverso simulazioni. Dimostriamo che la nostra proposta funziona bene con risposte Gaussiane e Binomiali.*

Key words: Hypothesis testing, Segmented regression, Changepoints

Introduction

Segmented or broken-line models are regression models where the relationships between the response and one or more explanatory variables are piecewise linear, namely represented by two or more straight lines connected at unknown values i.e. *changepoints*. This paper deals with the problem of estimating the number of changepoints in segmented regression models. The aim is to propose a stepwise procedure and assess its performance via simulations. This procedure will be based on either the Davies or the pseudo-score test. The structure of the paper is as follows. Section 1 introduces the segmented regression model, and Section 2 reviews the sequential hypothesis testing procedure for the selection of the number of changepoints by Kim et al. (2000). Section 3 illustrates our proposal. Section 4 presents a simulation study to investigate the performance of our procedure. The paper ends with conclusions in Section 5.

Andrea Priulla · Nicoletta D'Angelo
Department of Economics, Business and Statistics, University of Palermo
e-mail: `andrea.priulla@unipa.it;nicoletta.dangelo@unipa.it`

1 Segmented regression models

The segmented linear regression is expressed as

$$g(E[Y|x_i,z_i]) = \alpha + z_i^T\theta + \beta x_i + \sum_{k=1}^{K_0} \delta_k (x_i - \psi_k)_+ \tag{1}$$

where g is the link function, x_i is a broken-line covariate and z_i is a covariate whose relationship with the response variable is not broken-line. We denote by K_0 the true number of changepoints and by ψ_k the locations of the changes in the relationship, which we shall call from now on *changepoints*. These are selected among all the possible values of x. The term $(x_i - \psi_k)_+$ is defined as $I(x_j > \psi_k)$, that is $(x_i - \psi_k)I(x_i > \psi_k)$. The coefficient θ represents the non broken-line effect of z_i, β represents the effect for $x_i < \psi_1$, that is the effect of x_i before the first estimated changepoint. Finally, δ_k is the vector of the differences in the effects. Throughout the paper, we only consider models with Gaussian iid errors.

The fundamental problem we deal with is identifying the number of changepoints K_0. The estimation of their locations ψ_k, and the broken-line effects β and δ may also be of interest. However, these are not addressed here. Much of the literature is concerned with the problem of determining the 'best' subset of independent variables. The two main approaches for variable selection are information criteria and hypothesis testing. In this paper, we focus on the hypothesis testing procedure.

2 Selection of K_0 through sequential hypothesis testing

A common approach to selecting the number of changepoints relies on a sequential hypothesis testing procedure (Kim et al., 2000). Typically, this consists in performing different hypothesis tests starting from $\mathscr{H}_0 : K_0 = 0$ vs. $\mathscr{H}_1 : K_0 = K_{max}$, where K_{max} is fixed a priori. Depending on the rejection or not of the null hypothesis, the procedure can test for the next hypothesis system by increasing the number of changepoints specified in $\mathscr{H}_0$ or decreasing the one postulated under $\mathscr{H}_1$. Pseudo-score and Davies' tests can be used to identify the correct number of changepoints. Both tests are originally proposed to test for the existence of a changepoint, that is:

$$\begin{cases} \mathscr{H}_0 : & \delta_k = 0 \\ \mathscr{H}_1 : & \delta_k \neq 0. \end{cases}$$

The Davies' Test (Davies, 1977) is an asymptotic test helpful in dealing with hypothesis testing when a nuisance parameter is present only under $\mathscr{H}_1$. Assuming fixed and known changepoints, the procedure i) computes K 'naive' test statistics for the difference-in-slope δ_1, ii) seeks the lowest value and the corresponding naive p-value (according to $\mathscr{H}_1$), iii) and then corrects the selected (minimum) p-value by means of the K values of the test statistic.

Considering the case of multiple changepoints $\psi_1 < \psi_2 < \cdots < \psi_k$ and relevant K test statistics, Davies defined an upper bound for the p-value given by

$$\text{p-value} \approx \Phi(-M) + V \exp\left(-M^2/2\right)(8\pi)^{-1/2}$$

where $\Phi(\cdot) \sim \mathscr{N}(0,1)$, $M = \max\left[S(\Psi_k)\right]_k$ is the maximum of the K test statistics, and $V = \Sigma_k(|S(\Psi_k) - S(\Psi_{k-1})|)$ is the total variation of $\{S(\psi_k)\}_k$. It is crucial to notice that, despite being helpful in testing for the existence of a changepoint, Davies' test is not the best tool to select the number of changepoints. This is because $\mathscr{H}_1$ states the existence of at least one additional changepoint, that is $K_0 > k$ when $\delta_k \neq 0$. In this respect, based on the rejection of the last test, the number of changepoints selected by Davies' test is not $K_0 = K_{max}$ but could be larger.

Alternatively, the pseudo-score test proposed by Muggeo (2016) is based on an adjustment of the score statistic. This approach requires quantities only from the null fit, and thus, it has the advantage that the estimation of the nuisance parameter under the alternative is unnecessary. The pseudo-score statistic is expressed as

$$s_0 = \frac{\bar{\varphi}^T(I_n - A)y}{\sigma\{\bar{\varphi}^T(I_n - A)y\}^{\frac{1}{2}}},$$

where $(I_n - A)y$ is the residual vector under $\mathscr{H}_0$, with I_n the identity matrix, A the hat matrix and y the observed response vector, and $\bar{\varphi} = \{\bar{\varphi}_1, \ldots, \bar{\varphi}_n\}^T$ is the vector of the means of the nuisance parameter ψ_k averaged over the range $\{\mathscr{L}, \mathscr{U}\}$, i.e. $\bar{\varphi} = K^{-1}\sum_{k=1}^{K}\varphi(x_i, \psi_k), i = 1, \ldots, n$, which does not depend on ψ_k, so the pseudo-score can be computed even under $\mathscr{H}_0 : \delta_k = 0$ when ψ_k is not defined.

3 Proposed sequential hypothesis testing

Based on permutation tests, the procedure of Kim et al. (2000) makes testing for more than two additional changepoints unfeasible with the pseudo-score test. Our proposal overcomes this problem by making it possible to test for any additional changepoints using a sequential procedure. Contrary to the procedure of Kim et al. (2000), our proposal has the advantage of not being limited to testing for a maximum number of additional changepoints K_{max}, fixed a priori.

For simplicity, we outline the algorithm when the maximum number of changepoints is $K_{max} = 3$, restricting the analyses to a contained number of changepoints.

The procedure is as follows:

1. Fit a segmented model to the data, with $\hat{K} = 1$ and test

$$\begin{cases} \mathscr{H}_0 : & \delta_1 = 0 \quad (K_0 = 0) \\ \mathscr{H}_1 : & \delta_1 \neq 0 \quad (K_0 \geq 1) \end{cases}$$

via the pseudo-score or Davies' test. If $\mathscr{H}_0$ is not rejected, then the procedure stops estimating $\hat{K} = 0$. Otherwise, we proceed with the algorithm;

2. Fit a segmented model with $\hat{K} = 2$ and test

$$\begin{cases} \mathscr{H}_0 : & \delta_2 = 0 \quad (K_0 = 1) \\ \mathscr{H}_1 : & \delta_2 \neq 0 \quad (K_0 \geq 2) \end{cases}$$

If $\mathscr{H}_0$ is not rejected, then the procedure stops estimating $\hat{K} = 1$. Otherwise, we proceed to fit the following model;

3. Fit a segmented model with $\hat{K} = 3$ and test

$$\begin{cases} \mathscr{H}_0 : & \delta_3 = 0 \quad (K_0 = 2) \\ \mathscr{H}_1 : & \delta_3 \neq 0 \quad (K_0 \geq 3) \end{cases}$$

If $\mathscr{H}_0$ is not rejected then $\hat{K} = 2$. Otherwise, $\hat{K} \geq 3$.

The p-value for each hypothesis is obtained via the Davies and the pseudo-score tests. Furthermore, we control for over-rejection of the null hypotheses at the overall level α using the Bonferroni correction, comparing each p-value with α/K_{max}. Of course, setting the Bonferroni correction to α/K_{max} means putting ourselves in the most conservative setting.

4 Simulation studies

We simulate four different scenarios, fitting models with different true values of the number of changepoints, namely $K_0 = 0, 1, 2, 3$. We consider three different sample sizes $n = 100, 250, 500$, including only one segmented covariate taking equispaced values between 0 and 1. The simulated segmented models are:

$$\begin{aligned} y_i &= 2 + 15x_i + \varepsilon_i \\ y_i &= 2 + 15x_i - 8(x_i - 0.2)_+ + \varepsilon_i \\ y_i &= 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \varepsilon_i \\ y_i &= 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \varepsilon_i, \end{aligned}$$

considering iid Gaussian errors with standard deviation equal to $\sigma = 0.3$. We fix $\alpha = 0.05$. For each K_0, we fit a set of four models with $\hat{K} = 0, 1, 2, 3$. We choose the best model by applying the procedure proposed in Section 3.

We refer to the `segmented` package (Muggeo, 2008) in R (R Core Team, 2019) for fitting the segmented regression models.

Table 1 reports the results in terms of the percentages of the correctly selected number of changepoints based on both the Davies' and the pseudo-score tests over 500 simulations. Conditional frequencies are reported in the rows, and therefore, a criterion that perfectly selects the right number of changepoints should report values equal to 1 in the main diagonal of the table and zeros in all the other entries.

Table 1: Percentages of the correctly selected number of changepoints based on 500 simulations and three different sample sizes $n = \{100, 250, 500\}$ - Gaussian response variable.

	$n = 100$				$n = 250$				$n = 500$			
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.982**	0.014	0.004	0.000	**0.994**	0.006	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.994**	0.004	0.002	0.000	**0.986**	0.006	0.008	0.000	**0.994**	0.002	0.004
2	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000
3	0.000	0.000	**0.682**	0.318	0.000	0.000	0.112	**0.888**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.986**	0.014	0.000	0.000	**0.976**	0.024	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.998**	0.002	0.000
2	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.999**	0.001
3	0.012	0.048	0.286	**0.654**	0.000	0.000	0.016	**0.984**	0.000	0.000	0.000	**1.000**

The Davies test underestimates the number of changepoints on average compared to the pseudo-score test. Other simulations studies, omitted for brevity, show that a sample size larger than $n = 500$ leads to the same results.

We perform other simulations considering the following logit models

$$
\begin{aligned}
y_i &= 2 + 4z_i + 15x_i + \varepsilon_i \\
y_i &= 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ + \varepsilon_i \\
y_i &= 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \varepsilon_i \\
y_i &= 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \varepsilon_i
\end{aligned}
$$

As for the binomial case, the overall performance of the considered tests is worse compared to the Gaussian case. For this reason, we chose to increase the sample size to $n = 2500, 5000, 10000$ for the simulations. Results are shown in Table 2.

5 Conclusions

In this paper, we addressed the problem of the selection of the number of changepoints in segmented regression models. A well-established procedure by Kim et al. (2000) required a procedure based on sequential testing for the existence of a changepoint. With this, testing for any additional changepoints is unfeasible. We have proposed a stepwise procedure that overcomes this problem, and we have as-

Table 2: Percentages of the correctly selected number of changepoints based on 500 simulations and three different sample sizes $n = \{2500, 5000, 10000\}$ - Binomial response variable.

	$n = 2500$				$n = 5000$				$n = 10000$			
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.996**	0.004	0.000
2	0.000	0.000	**0.866**	0.134	0.000	0.000	**0.872**	0.128	0.000	0.000	**0.874**	0.126
3	0.000	0.000	0.460	**0.540**	0.000	0.000	0.060	**0.940**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.988**	0.012	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.988**	0.012	0.000	0.000	**0.992**	0.008	0.000
2	0.000	0.000	**0.862**	0.138	0.000	0.000	**0.862**	0.138	0.000	0.000	**0.870**	0.130
3	0.000	0.000	0.352	**0.648**	0.000	0.000	0.064	**0.936**	0.000	0.000	0.000	**1.000**

sessed its performance via simulations. The results have shown that the proposed procedure based on sequential hypothesis testing behaves well both with the Gaussian and Binomial responses. These results have some limitations. We run the simulation studies fixing a small number of changepoints, namely $K_{max} = 3$. Of course, our method can be implemented to any fixed K_{max}. However, a small K_{max} is often reasonable in real-life applications, as shown in Priulla et al. (2021) which deals with higher education data.

References

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254.

Kim, H.-J., Fay, M. P., Feuer, E. J., and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in medicine*, 19(3):335–351.

Muggeo, V. (2008). segmented: An r package to fit regression models with broken-line relationships. *R NEWS*, 8/1:20–25.

Muggeo, V. M. (2016). Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. *Journal of Statistical Computation and Simulation*, 86(15):3059–3067.

Priulla, A., D'Angelo, N., and Attanasio, M. (2021). An analysis of italian university students' performance through segmented regression models: gender differences in stem courses. *Genus*, 77(1):1–20.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

On the numerical stability of the efficient frontier

Sulla stabilitá numerica della frontiera efficiente

Claudia Fassino and Pierpaolo Uberti

Abstract We show that the numerical instability of Markowitz model is due to the linear restrictions of the optimization problem and that it can be exacerbated by the choice of one single parameter, the expected return of the portfolio. The model instability depends on the fact that the two linear restrictions are badly scaled and almost collinear. On the basis of geometric arguments, using the notion of numerical rank of a matrix, we propose to opportunely restrict the choice of the expected return in order to reduce the instability. The effectiveness of our proposal is supported by a bunch of applications that are performed on real financial data.

Key words: Numerical Stability, Markowitz Model, Efficient Frontier

1 Introduction

Numerical instability is a well known issue of the efficient frontier, the solution of Markowitz optimal allocation problem, see [10]. For numerical instability of the model we intend that for small perturbations of the input data, the covariance matrix and vector of expected returns, the optimal portfolio identified by the model can show extreme variations. As a consequence, two portfolios that are very close on the efficient frontier can have huge differences in their composition. For these reasons the mean-variance model is very difficult to apply in practice. The instability generates poor out-of-sample results, see [3], when compared to performance of naive allocation rules.

In the financial literature the reasons of the numerical instability is discussed. First, the parameters of the model need to be estimated since they are unknown.

University of Genova
Genova, Via Balbi 5, 16126 e-mail: fassino@dima.unige.it

University of Milan-Bicocca
Milano, Piazza dell'Ateneo Nuovo 1, 20126, e-mail: pierpaolo.uberti@unimib.it

Then the estimation of the parameters is affected by uncertainty causing the in-sample frontier to be a biased estimator of the real efficient frontier, see [6]. The estimation error, identified by some authors as the main cause of instability in the model, see among the others [1] and [7], becomes worse when the size of the portfolio increases. In fact, while the number of assets grow linearly, the parameters in the covariance matrix grow quadratically, see [11]. Moreover, the solution of the optimization problem is a function of the inverse of the covariance matrix, that is potentially ill-conditioned and almost singular matrix when the returns of the assets in the portfolio are numerically close to be linear dependent. The literature proposed many alternative solutions to handle ill-conditioning: in [5] a Bayesian approach is discussed, [9] describe a procedure based on the shrinkage estimation, in [8] and [12] robust optimization techniques are introduced, and Lasso techniques are studied in [2]. It is clear from the above enumeration that the literature mainly focused on the covariance matrix that is used to define the objective function of the optimization problem.

Recently, in [4], it has been shown how a structural component of instability is hidden in the linear restrictions of the optimization problem. Our approach starts from the previous observation with the objective to fill the gap between the theoretical formalization of the model and its application in practice, using numerical mathematics techniques. In particular, we propose to restrict the interval of the values for the expected return parameter in order to preserve the numerical ranks of the matrices involved in the calculations. The impact of the choice of the parameter is showed, first, through some toy examples. Then we apply our proposal on real financial data. In particular, since the number of assets in the portfolio strongly affects the results, we apply our methodology to different databases where the number of assets ranges from 10 for the sectors of the S&P500 to 500 when we consider all of its constituents.

References

1. Best MJ, Grauer RR. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. Rev Financ Stud 1991;4(2):315–42.
2. Brodie J, Daubechies I, De Mol C, Giannone D, Loris I. Sparse and stable Markowitz portfolios. Proc Natl Acad Sci USA 2009;106(30):12267–72.
3. DeMiguel V, Garlappi L, Uppal R. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? Rev Financ Stud 2009;22:1915–53.
4. Fassino C, Torrente ML, Uberti, P. A singular value decomposition based approach to handle ill-conditioning in optimization problems with applications to portfolio theory. Chaos, Solitons & Fractals 2022;165, 112746.
5. Frost PA, Savarino JE. An empirical Bayes approach to efficient portfolio selection. J Financial Quant Anal 1986;21(3):293–305.
6. Kan R, Smith DR. The distribution of the sample minimum-variance frontier. Manage Sci 2008;54(7):1364–80.
7. Kan R, Zhou G. Optimal portfolio choice with parameter uncertainty. J Financial Quant Anal 2007;42(3):621–56.

8. Kim JH, Kim WC, Fabozzi FJ. Recent developments in robust portfolios with a worst-case approach. J Optim Theory Appl 2014;161(1):103–21.
9. Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 2003;10(5):603–21.
10. Markowitz H. Portfolio selection. J Finance 1952;7(1):77–91.
11. Pflug GCh, Pichler A, Wozabal D. The 1/N investment strategy is optimal under high model ambiguity. J Bank Financ 2012;36:410–7.
12. Rosadi D, Setiawan EP, Templ M, Filzmoser P. Robust covariance estimators for mean–variance portfolio optimization with transaction lots. Oper Res Perspect 2020;7.

Spatial regression with differential regularization over linear networks

Regressione spaziale con regolarizzazione differenziale su reti lineari

Aldo Clemente, Eleonora Arnone, Jorge Mateu, and Laura M. Sangalli

Abstract This work proposes a spatial regression method with differential regularization for data observed over linear networks. The method combines a maximum likelihood approach with a regularization penalty involving the laplacian operator, that has the purpose of controlling the roughness of the estimates. We apply the proposed regression estimator to a benchmark dataset containing the prices of houses sold in London during 2001.

Key words: spatial regression, linear networks, finite elements

1 Introduction

Nowadays an increasing amount of data is recorded over linear networks, such as road or rail networks or river networks. Typical examples in this respect are data concerning shared mobility or accidents and crimes in a city, referring to the road networks, or levels of pollutants along a river, for what concerns the river networks. For instance, Fig. 1 shows the London road network and the locations of houses

---

Aldo Clemente
MOX - Dip. di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy, e-mail: aldo.clemente@polimi.it

Eleonora Arnone
Dipartimento di Management, Università degli Studi di Torino, Corso Unione Sovietica 218 bis, Torino 10134, Italy, e-mail: eleonora.arnone@unito.it

Jorge Mateu
Department of Mathematics, Universitat Jaume I, Avinguda de Vicent Sos Baynat s/n, Castellón 12006, Spain, e-mail: mateu@uji.es

Laura M. Sangalli
MOX - Dip. di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy, e-mail: laura.sangalli@polimi.it

sold during 2001, with the color of the point marker referring to the price of the houses, in thousands of sterling. The modelling of data scattered over linear networks has recently attracted a strong interest in the literature, mostly concerning the problem of density/intensity estimation over linear networks, see, e.g., the review in [1]. The literature on spatial regression over these complicated spatial domains is instead scarcer. For instance, an application in the field of Geographically Weighted Regression over linear networks is discussed by Lu et al. [3]. Instead, Ladle et al. [2] and Ver Hoef [6] consider kriging methods for data observed over linear networks. This work develops a spatial regression method for data observed over linear network domains. The proposed method combines a maximum likelihood approach with a regularization penalty involving the Laplacian operator, that has the purpose of controlling the roughness of the estimates. In such respect, the proposed regression method can be seen as an extension to linear network domains of the literature on spatial regression with differential regularization (SR-PDE) reviewed in Sangalli [4]. In particular, consider n locations $\mathbf{p}_1, \dots \mathbf{p}_n$ over the linear network G. At location $\mathbf{p}_i$ we observe a variable of interest $y_i \in \mathbb{R}$ and a set of covariates $\mathbf{x}_i \in \mathbb{R}^q$. We consider the following model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i) + \varepsilon_i \qquad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is the unknown vector of regression coefficients, $f : G \to \mathbb{R}$ is the unknown field and $\varepsilon_1, \dots, \varepsilon_n$ are uncorrelated errors with zero mean and finite variance. SR-PDE estimates the unknown $\boldsymbol{\beta}$ and f by minimizing the following functional:

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i))^2 + \lambda \int_G (\Delta f)^2, \tag{2}$$

where λ is a positive smoothing parameter and Δ denotes the Laplacian operator.

In the estimation procedure of SR-PDE we have to deal with the presence of a function defined over the network domain. Hence, suitable functional spaces over linear network domains have to be introduced. In particular, we locally define spaces on each edge of the network and then we set continuity conditions between connected edges. Once the opportune functional setting is defined, we can prove some important theoretical properties for the proposed model, suitably extending the result obtained for spatial regression with differential regularization over simpler two-dimensional domains (see, e.g, [5]) The discretized version of the estimation problem (2) is obtained through Finite Element Method. Section 2 applies the proposed SR-PDE estimator to a benchmark dataset containing the prices of the houses sold in London during 2001.

2 London Houses Prices

In this section we consider a benchmark dataset containing the prices of houses sold in London (UK) during 2001 (see, e.g, [3]). Fig. 1 shows the London road network and the locations of houses, with the color of the point marker referring to the price of the houses, in thousands of sterling. The network is composed of 94498 nodes and 117062 edges, and has a very complicated geometry. The data consists of 1601 properties sold during the 2001 calendar year. The data and the London road network are available with the R package `shp2graph`. We treat as dependent variable y the overall price of the house and as explanatory variables the characteristics of the property such as the size of the property, the presence of at least two bathrooms and the percentage of the workflow in professional or managerial occupations in the census enumeration district in which the house is located.

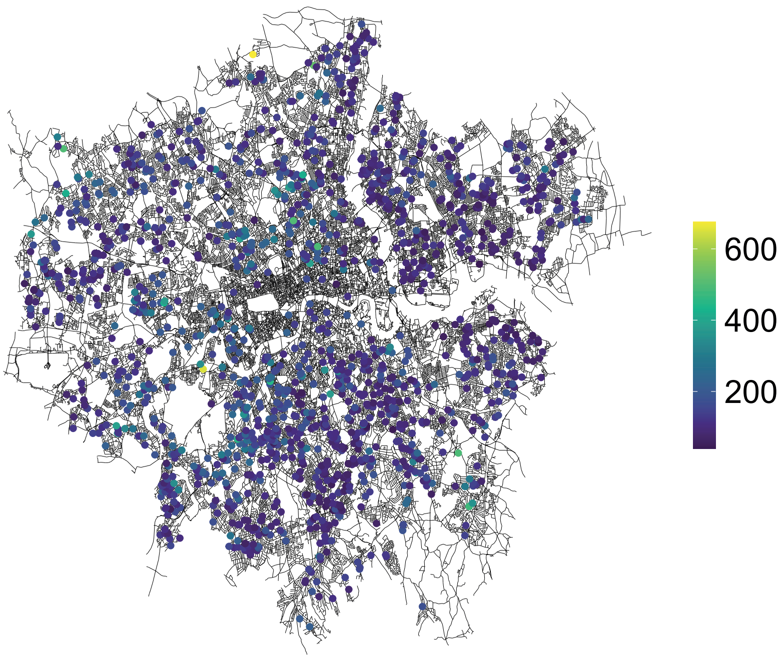


Fig. 1: Prices of houses sold in London (UK) in thousands of sterling.

We fit model (1)-(2) selecting the smoothing parameter through generalized cross validation. The three considered covariates have significant positive effect on the price. In particular, the most relevant is the one related to the presence of at least two bathrooms in the property.

Fig. 2 shows the SR-PDE estimate of the spatial component f. As one might expect, higher values of the estimate correspond to locations in the city center of London.

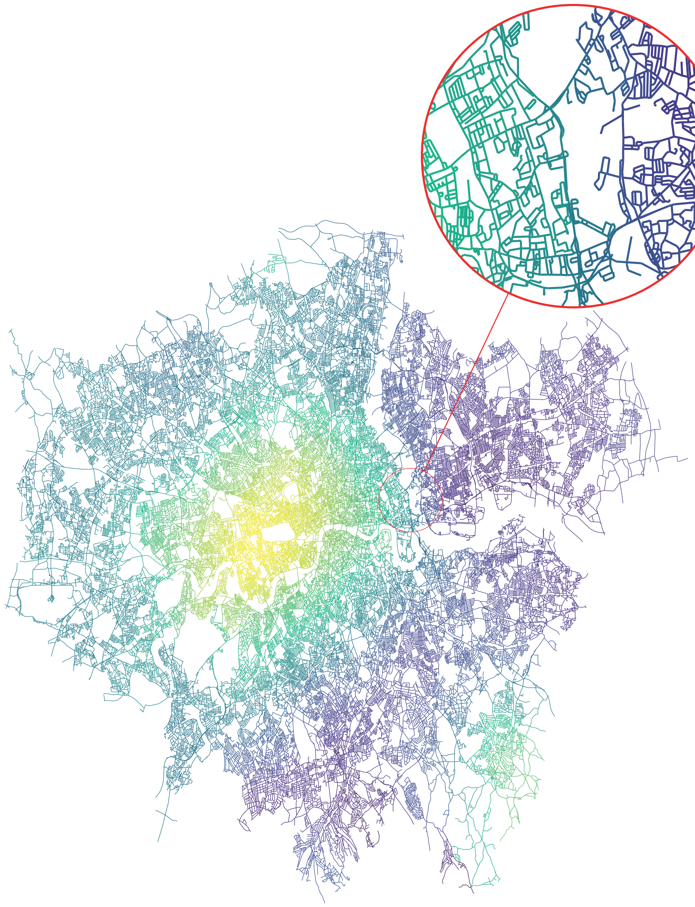


Fig. 2: Estimate of the spatial component of the selling price, for properties in London (UK). It highlights the effect of the spatial locations of the houses on the selling price.

The top right part of Fig. 2 shows a zoom in the city center. This plot highlights that the estimate complies with the complicated linear network domain. Indeed, roads very close in Euclidean distance, such as those on the opposite banks of the river Thames, may display different price values.

References

1. A. Baddeley, G. Nair, S. Rakshit, G. McSwiggan, and T. M. Davies. Analysing point patterns on networks—a review. *Spatial Statistics*, 42:100435, 2021.
2. A. Ladle, T. Avgar, M. Wheatley, and M. S. Boyce. Predictive modelling of ecological patterns along linear-feature networks. *Methods in Ecology and Evolution*, 8(3):329–338, 2017.
3. B. Lu, M. Charlton, P. Harris, and A. S. Fotheringham. Geographically weighted regression with a non-euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, 28(4):660–681, 2014.
4. L. M. Sangalli. Spatial regression with partial differential equation regularisation. *International Statistical Review*, 89:505–531, 2021.
5. L. M. Sangalli, J. O. Ramsay, and T. O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.
6. J. M. Ver Hoef. Kriging models for linear networks and non-euclidean distances: Cautions and solutions. *Methods in Ecology and Evolution*, 9(6):1600–1613, 2018.

An Estimation Tool for Spatio-Temporal Events over Curved Surfaces

Uno Strumento di Stima per Eventi Spazio-Temporali su Superfici Curve

Simone Panzeri, Blerta Begu, Eleonora Arnone, Laura M. Sangalli

Abstract Data coming from spatio-temporal Poisson point processes are attracting an increasing interest in many scientific fields. In this work we develop an innovative method for Spatio-Temporal Density Estimation with Partial Differential Equation regularization, capable to handle point patterns observed over complicated curved surfaces. We combine a non-parametric likelihood approach with a separable regularization in space and time. We define the functional space where to set the estimation problem, and we prove some important theoretical properties of the estimator. Leveraging on advanced numerical procedures, we develop an efficient and flexible method. Through simulation studies and real applications, we validate our method, assess its performances and highlight advantages over state-of-the-art techniques.

Key words: space-time density estimation, Poisson point process, functional data analysis, differential regularization, Riemannian manifolds.

Simone Panzeri
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza L. Da Vinci 32, Milano 20133, Italy, e-mail: simone.panzeri@polimi.it

Blerta Begu
School of Mathematics & Statistics, University College Dublin, Belfield Dublin 4, Ireland, e-mail: blerta.begu@ucdconnect.ie

Eleonora Arnone
Dipartimento di Management, Università degli Studi di Torino, Corso Unione Sovietica 218 bis, Torino 10134, Italy, e-mail: eleonora.arnone@unito.it

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza L. Da Vinci 32, Milano 20133, Italy, e-mail: laura.sangalli@polimi.it

1 Introduction

Nowadays, massive amounts of data are available thanks to recent technologies. In particular, point pattern data related to spatio-temporal events are attracting an increasing interest in many scientific fields, such as epidemiology, seismology, and many others. For instance, typical real applications may comprise disease surveillance and occurrence analysis of extreme natural phenomena. As a concrete example, consider the situation represented in Figure 1. Points correspond to the locations of powerful earthquakes occurred worldwide over different time windows, in the two-year period between 2020 and 2022. Clearly, points are mainly placed on elongated narrow regions corresponding to the faults in the Earth's crust originated by the action of plate tectonic forces. The interest here is in the identification of major critical areas, jointly combining both spatial and temporal analyses.

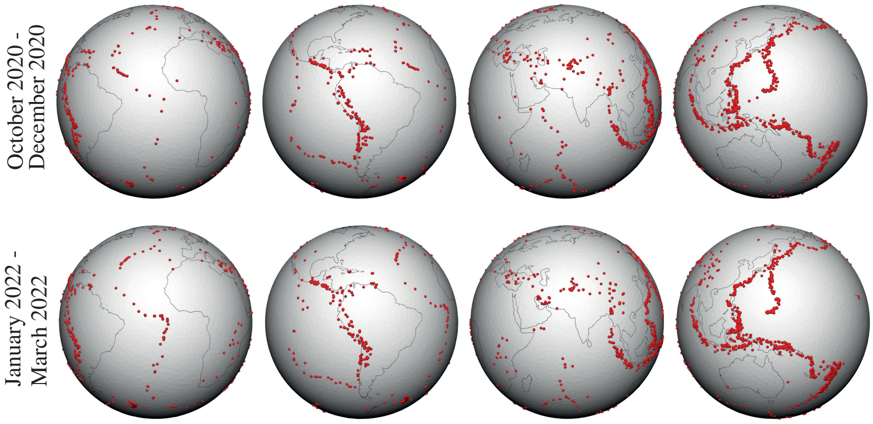


Fig. 1 Different views of locations of earthquakes of magnitude greater than 4.5, occurred worldwide during two different three-month time periods. Points are mainly placed on elongated narrow regions and therefore the signal turns out to be very skewed and anisotropic.

The main research goal usually consists in determining the characteristics of the underlying spatio-temporal point processes, that generate the observed data. In other words, this is equivalent to estimate the spatio-temporal probability density function associated with the distribution from which data are drawn. However, studying the mutual interplay of the temporal evolution and the spatial variability of the events is definitely not an easy task in a general framework.

Despite the broad literature produced in the fields of spatial (refer e.g. to [5, 6] and [12]) and spatio-temporal (e.g. [7] and [10]) Poisson point process analysis, state-of-the-art methodologies usually operate under some peculiar restrictive assumptions, either theoretically or computationally. In fact, most of the works are limited to deal with discrete-time data only, requiring an appropriate binning over T and therefore narrowing down the resulting analysis to local phenomena. For instance, this holds for the Spatio-Temporal Kernel Density Estimation (STKDE) model presented in [15] and for the implementation of the spatio-temporal Log-Gaussian Cox Processes (LGCP) described in [8]. Moreover, statistical methodologies to analyze spatio-temporal point patterns mainly act on very simple planar spatial domains or at most d-dimensional spheres, even though real case studies are set over regions with non-trivial curved geometries (e.g., complicated boundaries, sharp concavities, interior holes, etc.). On this point, STKDE operates solely with latitude-longitude data, whereas LGCP is mostly studied for two-dimensional spherical surfaces, as in [4]. Another research direction focuses on separable first-order spatio-temporal point processes, significantly simplifying the underlying problem. However, even in very basic settings, to the best of our knowledge ready-to-use implementations of models to solve density estimation problems over generic manifolds are not easily available.

In order to overcome such limitations, we propose an innovative methodology for Spatio-Temporal Density Estimation with a Partial Differential Equation Regularization (STDE-PDE) over two-dimensional curved domains. Due to the underlying mathematical model, STDE-PDE can be seen as a new addition to the class of non-parametric penalized methods, reviewed by [13], to analyze spatial and functional data over complex domains in various contexts. However, this novel approach has only been used to tackle density estimation problems solely in the spatial setting. In this regard, [9] and [1] present dedicated methods for planar and curved regions. We rather incorporate the time dimension into the analysis, inspired by the methods designed to solve spatio-temporal regression problems (refer to [3] and references therein).

2 Model

Let $\mathscr{M}_T := \mathscr{M} \times T$ be the spatio-temporal domain of interest, where $\mathscr{M}$ is a two-dimensional Riemannian manifold and T a bounded interval. Now let the pairs $\left\{(\mathbf{p}_i, t_i)\right\}_{i=1}^n \in \mathscr{M}_T$ be n independent and identically distributed observations drawn from a distribution F with density function $f : \mathscr{M}_T \to \mathbb{R}^+$ with respect to the Lebesgue measure. The goal of this work is to elaborate a procedure to estimate f over $\mathscr{M}_T$, starting from the observations $\left\{(\mathbf{p}_i, t_i)\right\}_{i=1}^n$. To this end, we propose to combine a non-parametric maximum likelihood approach with a separable regularization in both space and time based on differential operators. Thus, we obtain the estimate $\hat{g} = \log \hat{f}$ as the (unconstrained) minimizer of the objective functional $L(g)$, which comprises the likelihood and the regularization term $R(g)$. The contribution of the latter is controlled by the smoothing parameter $\lambda > 0$. Hence, we are able to balance the trade-off between the adherence of the model to data and the smoothness of the estimate eventually produced.

The content of this work has required significant efforts towards various directions. From a theoretical viewpoint, we appropriately generalize the results achieved in classical non-parametric literature for multivariate density estimation over simple domains (see [14] and [11]). In particular, the introduction of a regularization in time requires the model to be theoretically validated. In this regard, we first formally define a well-suited spatio-temporal functional space where to set the considered estimation problem. Then, we rigorously derive the well-posedness of such a problem under standard regularity assumptions. Afterward, we face methodological challenges to ensure the consistency of the proposed estimator. Additionally, we formally derive the equivalence of the considered problem with the associated spatio-temporal inhomogeneous Poisson intensity estimation version.

We leverage on an appropriate discretization procedure, based on finite elements in space and cubic B-spline basis functions in time, ensuring flexibility to handle spatial domains with non-trivial geometries. Moreover, concerning the implementation of the estimation procedure of STDE-PDE and the optimization aspects involved, we resort to advanced numerical techniques. This makes the proposed method efficient and computationally sustainable, even in the case of a large sample size or fine spatio-temporal discretizations.

Our contribution extends the functionalities provided by the `R/C++` `fdaPDE` library ([2]), providing the end user with a ready-to-use tool to solve effectively the considered problem.

3 Simulation study and application to seismic data

STDE-PDE has some important comparative advantages over state-of-the-art methods, whose details are here omitted for the sake of space. This holds in terms of both computational costs and accuracy of the estimate eventually produced.

We highlight those aspects extensively throughout several simulation studies with synthetic data, an example of which is graphically provided by Figure 2. We conduct the test over a spherical surface with the purpose of allowing comparisons with competing methods. We sample data from a mixture of time-dependent distributions via acceptance-rejection method.

In Figure 3 we instead illustrate some qualitative results that the proposed method achieves on data coming from a real spatio-temporal phenomenon, at two different time instants. We study the distribution of approximately twenty thousand locations of earthquakes of magnitude larger than 4.5, occurred worldwide in 2020-2022. Data are taken from the earthquake catalog of the United States Geological Survey. As expected, the pattern is highly localized over the faults in the Earth's crust, originated by the action of plate tectonic forces.

STDE-PDE proves to be able to capture several highly involved multi-modal signals in the described continuous-time framework, working successfully in all the scenarios considered in this work. Actually, as last remark, we stress that our method shows flexibility and efficiency over manifolds of various type and not only over simple spherical surfaces.

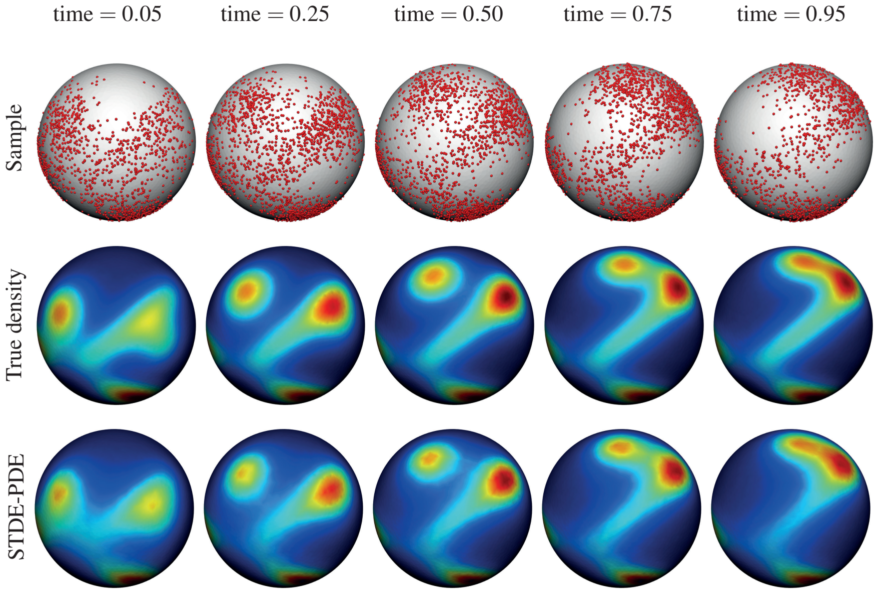


Fig. 2 Simulation case study with data generated via rejection sampling. Front view of the sample (first row), true density (second row) and STDE-PDE estimates (third row), captured at five different time instants in $T = [0, 1]$.

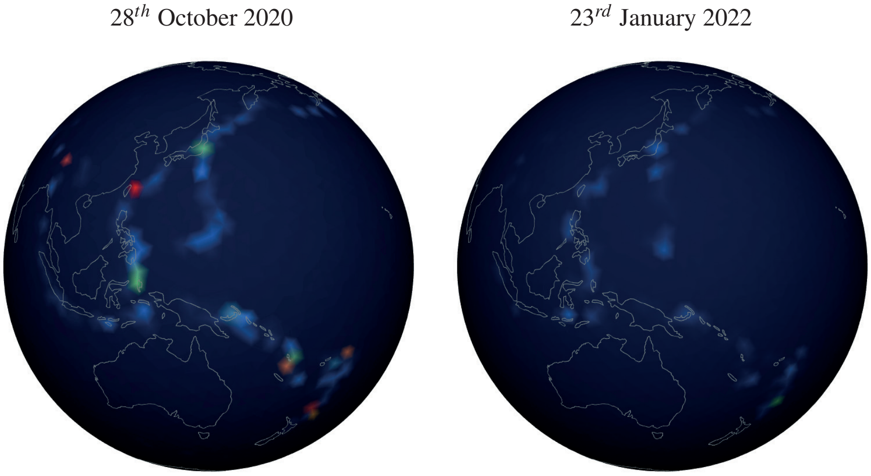


Fig. 3 Application case study. Main view over the globe (Australian, Eurasian and Pacific plates) of the STDE-PDE density estimates of earthquakes of magnitude greater than 4.5, captured at two different time instants, in the period from October 2020 to December 2020 and from January 2021 to March 2021 respectively. The signal turns out to be very skewed and anisotropic.

References

1. E. Arnone, F. Ferraccioli, C. Pigolotti, and L. M. Sangalli. A roughness penalty approach to estimate densities over two-dimensional manifolds. *Computational Statistics and Data Analysis*, 2022.
2. E. Arnone, L. M. Sangalli, E. Lila, J. Ramsay, and L. Formaggia. fdapde: Statistical analysis of functional and spatial data, based on regression with pde regularization. *R package version 1.1-11*, 2022.
3. E. Arnone, L. M. Sangalli, and A. Vicini. Smoothing spatio-temporal data with complex missing data patterns. *Statistical Modelling*, page 1471082X211057959, 2021.
4. F. Cuevas-Pacheco and J. Møller. Log gaussian cox processes on the sphere. *Spatial Statistics*, 26:69–82, 2018.
5. D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes, Vol.1. 2ed.* Springer, 2003.
6. D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes, Vol.2. 2ed.* Springer, 2008.
7. P. J. Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.
8. P. J. Diggle, P. Moraga, B. S. Rowlingson, and B. M. Taylor. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.
9. F. Ferraccioli, E. Arnone, L. Finos, J. O. Ramsay, and L. M. Sangalli. Nonparametric density estimation over complicated domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):346–368, 2021.
10. J. A. González, F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu. Spatio-temporal point process statistics: a review. *Spatial Statistics*, 18:505–544, 2016.
11. C. Gu and C. Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.
12. J. Møller and R. P. Waagepetersen. Some recent developments in statistics for spatial point patterns. *Annual Review of Statistics and Its Application*, 4:317–342, 2017.
13. L. M. Sangalli. Spatial regression with partial differential equation regularisation. *International Statistical Review*, 89(3):505–531, 2021.
14. B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.
15. Z. Zhang, D. Chen, W. Liu, J. S. Racine, S. Ong, Y. Chen, G. Zhao, and Q. Jiang. Nonparametric evaluation of dynamic disease risk: a spatio-temporal kernel approach. *PloS one*, 6(3):e17381, 2011.

Gromov-Wasserstein barycenters for optimal portfolio allocation

Baricentri Gromov-Wasserstein per allocazione ottima di portafolio

Alessandro Spelta, Nicolo Pecora, Mario Maggi

Abstract This paper develops a technique for portfolio optimal allocation which accounts for different market phases according to the behavior of a sentiment index. The algorithm we design to interpolate among risk-on, risk-off, and risk-neutral input portfolios, is based on the notion of Entropic Gromov-Wasserstein discrepancy. The resulting optimal portfolio is obtained as the barycenter of the input portfolios with barycentric coordinates that vary depending on the different market conditions. The comparison against suitable alternative portfolios reveals that our proposed technique outperforms benchmarks in terms of both return and risk.
Abstract *Questo documento sviluppa una tecnica per l'allocazione ottimale del portafoglio che tiene conto delle diverse fasi di mercato in base al comportamento di un indice di sentiment. L'algoritmo che progettiamo per interpolare tra portafogli risk-on, risk-off e risk-neutral si basa sulla nozione di discrepanza entropica di Gromov-Wasserstein. Il portafoglio ottimo è ottenuto come baricentro dei portafogli di input con coordinate baricentriche che variano a seconda delle diverse condizioni di mercato. Il confronto con portafogli alternativi rivela che la nostra tecnica proposta supera i benchmark in termini sia di rendimento che di rischio.*

Key words: Optimal Portfolio,Gromov-Wasserstein Barycenter, . . .

Alessandro Spelta
University of Pavia, e-mail: alessandro.spelta@unipv.it

Nicolo Pecora
Catholic University of Piacenza e-mail: nicolo.pecora@unicatt.it

Mario Maggi
University of Pavia e-mail: mario.maggi@unipv.it

1 Introduction

We propose an approach for optimal portfolio choice which dynamically adapts to market conditions. In order to deal with risk in a more flexible manner, we provide a generic nonlinear portfolio optimization model that incorporates both normal and crash risks. To measure market conditions we opt for a “SentimentBased” measure. We apply a set of technical indicators to the Fear & Greed series, which gauges the mood of the market, to determine whether investor behavior is pushing towards a boom or bust phase, or whether a neutral phase is dominating. To provide guidance on investment decision-making in the event that investors believe the market is moving toward or away from a risk-on/risk-off environment, we first propose to create risk-on and risk-off portfolios, which we call *input portfolios*, and then we design an algorithm to interpolate between them according to market conditions, by exploiting optimal transport metrics (see Villani, 2009, 2021). Within this paper, optimal transport is instrumental for computing distances among input portfolios, so to create a barycentric portfolio accounting for different market conditions. In other words, our optimal portfolio is a Wasserstein barycenter corresponding to optimal solutions of transportation problems for several input portfolios. We expand on that measuring distances among input portfolios in a metric-measure space (see Mémoli, 2007, 2011) and we design a methodology to compare portfolios by exploiting the notion of Entropic Gromov-Wasserstein (GW) discrepancy (see Peyré et al., 2016). This definition of discrepancy extends the Wasserstein distance between measure spaces to arbitrary metric-measure spaces, where a generic loss function allows us to work with matrices not necessarily positive and that does not necessarily satisfy the triangle inequality. The resulting technique defines the GW-portfolio as the barycenter of the input portfolios with barycentric coordinates that vary depending on the different market conditions. We make use of three sets of assets, namely ETF Bond, Stocks and ETF Short, to build risk-neutral, risk-on and risk-off input portfolios, while the Fear & Greed sentiment index is employed as a reference for market conditions. Results show that the GW-portfolio outperforms simple Markowitz allocation in terms of both Profits & Losses and Sharpe Ratio.

2 Methodology

2.1 Gromov-Wasserstein portfolio distance

We consider a portfolio composed by N assets as an entity described by a similarity matrix $S \in \mathbb{R}^{N \times N}$, being function of the correlation among its assets returns, and by a vector of weights w generated by an optimal asset allocation problem, such that $w_i \geq 0$, $i = 1, \ldots, N$ and $\sum_i w_i = 1$.

Following Mantegna and Stanley (1999), we define the similarity matrix S as:

$$S = 2 - \sqrt{2(1-C)} \tag{1}$$

where C represents the correlation matrix among asset returns. The similarity matrices S^p of the different p portfolios, $p = 1,...,P$, do not necessarily have the same size, and portfolios are not generally composed by the same set of assets. Additionally, the portfolios weights do not need to be computed by applying the same optimization methodology. Indeed, we compare portfolios using a soft-assign criterion, which measures the minimum distortion induced by a probabilistic map from the rows of one similarity matrix to the rows of another one. This criterion amounts to a regularized version of the Gromov-Wasserstein (GW) discrepancy between metric-measure spaces (Peyré et al., 2016).

We define the GW discrepancy between two portfolios $(S^1, w^1) \in \mathbb{R}^{N^1 \times N^1} \times \mathbb{R}^{N^1}$ and $(S^2, w^2) \in \mathbb{R}^{N^2 \times N^2} \times \mathbb{R}^{N^2}$ as:

$$\mathrm{GW}(S^1, S^2, w^1, w^2) \overset{\text{def}}{=} \min_{T \in \mathscr{C}_{w^1, w^2}} \left(\mathscr{E}_{S^1, S^2}(T) - \varepsilon H(T) \right), \tag{2}$$

where

$$\mathscr{E}_{S^1, S^2}(T) \overset{\text{def}}{=} \sum_{i,j,k,\ell} L\left(S^1_{i,k}, S^2_{j,\ell}\right) T_{i,j} T_{k,\ell}, \tag{3}$$

$$H(T) \overset{\text{def}}{=} - \sum_{i,j=1}^{N} T_{i,j} \left(\log\left(T_{i,j}\right) - 1\right), \tag{4}$$

where superscripts refer to portfolios while subscripts denote assets. The matrix T is a coupling between the two spaces on which the similarity matrices S^1 and S^2 are defined.

Function L in equation (3) represents a quadratic loss function that accounts for the misfit between the similarity matrices, namely:

$$L(a, b) \overset{\text{def}}{=} \frac{1}{2} |a - b|^2. \tag{5}$$

By introducing the 4-way tensor notation, the loss function can be expressed as follows:

$$\mathscr{L}(S^1, S^2) \overset{\text{def}}{=} \left(L\left(S^1_{i,k}, S^2_{j,\ell}\right) \right)_{i,j,k,\ell},$$

and thus, the objective function $\mathscr{E}_{S^1, S^2}(T)$ in equation (2) can be written as

$$\mathscr{E}_{S^1, S^2}(T) = \langle \mathscr{L}(S^1, S^2) \otimes T, T \rangle, \tag{6}$$

where $\otimes$ denotes the tensor-matrix multiplication.

The non-convex optimization problem in equation (2) can be solved by applying a projected gradient descent method (see Villani, 2009).

2.2 Portfolio Barycenters

To create the GW-portfolio, which mediates among input portfolios, we introduce the notion of *barycenter*. The design of a barycenter among similarity matrices is grounded on two primary factors: i) a measure of discrepancy between similarity matrices and ii) a definition of a mean. In our setting, the GW distance defines the discrepancy among similarity matrices, while the Fréchet mean is used to define the barycenter.

In formulae, the Fréchet mean is expressed as:

$$\min_{w \in \mathbb{R}^N} \sum_p \lambda^p \, \mathrm{GW}\left(S, S^p, w, w^p\right), \tag{7}$$

meaning that, given P input portfolios and a set of asset.

Aas in Cuturi (2013), the minimization over w can be solved with the Sinkhorn algorithm as:

$$w \overset{\text{def.}}{=} \prod_p \left(K^\top a_p\right)^{\lambda^p} \quad \text{and} \quad \begin{cases} b^p \leftarrow \frac{w}{K^\top a^p} \\ a^p \leftarrow \frac{w^p}{K b^p} \end{cases}$$

which finally provides the optimal weights of the GW-portfolio given the market condition expressed via the vector λ.

2.3 Portfolio Construction

In this Subection we exploit the GW barycenter to design a dynamic portfolio able to adjust to various market conditions. To distinguish among market phases, we apply the Kaufman's adaptive moving average (KMA) to the Fear & Greed sentiment index provided by the CNN. The Fear & Greed Index is the result of the combination of seven separate indicators, each of which tracks a different facet of stock market activity. They are the following: safe haven demand, market volatility, junk bond demand, stock price strength, stock price breadth, put and call options. The index measures the deviation between these individual indicators' averages and the amount by which they typically diverge. The index gives each indicator equal weighting when determining a score from 0 to 100, with 100 signifying the highest level of greed and 0 reflecting the highest level of fear. In particular, we compute KMA considering two different time horizons. The short-KMA has been parametrized with the values $[5, 2, 10]$ while for the long-KMA we have opted for $[10, 2, 25]$. The first input represents the number of days for the Efficiency Ratio while the second and the third are the number of days for the fastest and slowest exponential moving averages, respectively. We define a bull market (bl) phase when the Fear & Greed Index lies above both the short- and long-KMA while a bear market (br) is defined by the Fear & Greed below the two KMAs. Otherwise a neutral market (n) phase is called. Next we build three input portfolios each of which is composed by assets of different classes, by solving a Markowitz optimization problem. These portfolios are the stock portfolio with weights defined by the vectors w_{st}, a bond portfolio with weights w_b and a short portfolio with weights w_{sh}. We define combinations of the aforementioned input portfolios considering different weighting schemes $\alpha = [\alpha_{st}, \alpha_b, \alpha_{sh}]$ depending on the market conditions, and we call the resulting portfolio the Fear & Greed portfolio (FG-portfolio). In other words, the FG-portfolio weights (w_{fg}) are computed as a weighted sum of the input portfolio weights, that is:

$$w_{fg} = \alpha_{st} w_{st} + \alpha_b w_b + \alpha_{sh} w_{sh} \tag{8}$$

We consider several experiments which involve different asset proportions depending on the market conditions. For the bull market phase, we set the fraction of stocks α_{st} as the 90, 80 or 70% of the overall portfolio assets, the remaining 10, 20 or 30% is constituted by the fraction α_b while α_{sh} is zero. In neutral markets, instead, we opt for the following weighting scheme: α_{st} is fixed as 70, 60 or 50 %, α_b is set as 30, 40 or 50% and short assets are not present. Lastly, in bear markets, $\alpha_{st} = 0$, the majority of assets is constituted by bonds, fixing α_b equal to 70, 80 or 90%, and the remaining proportion is assigned to α_{sh}. Thus, during a bull market the FG-portfolio is mainly composed by stocks, with a residual fraction of ETF bonds. During neutral phases the same asset classes are involved but with a higher fraction of bonds. In a bearish market, instead, the FG-portfolio is mainly composed by bonds with a small fraction of short ETF assets.

Finally, the GW-portfolio weights are determined by using: i) the three different input portfolio weights w_{st}, w_b, w_{sh}, ii) their similarity matrices together with, iii) the similarity matrix of the overall assets, and iv) the FG-portfolio weights as initial barycenter value. Moreover, the λ vector, which defines the relative importance of each input portfolio, is set equal to the vector α. Portfolio Results In this Section we present the results of the application that has been described above. We begin by showing the cumulative Profits & Losses (P&L) of the GW-portfolio against suitable alternative investment strategies used as benchmarks, i.e. a portfolio composed by only stocks, a portfolio composed by ETF bonds and the FG-portfolio whose weights are given by Eq. (8). Moreover, we add a further portfolio named as the GW-portfolio Best Parameter in which we explicitly select the optimal Entropy parameter ε and the optimal initial risk profile value α.

To generate P&L we opt for multiple parameter configurations, namely we use i) different lengths of the in-sample windows from 150 to 225 days, with a step of 25 days, ii) different lengths of the rebalancing windows of the input portfolios, from 10 to 25 days with a step of 5 days iii) different risk profiles combinations iv) different values of the Entropy parameter $\varepsilon = \{.02, .04, .1\}$. All these parameter configurations give rise to 1296 different P&L. Moreover, we consider one basis point as transaction cost at each portfolio rebalancing.

Figure 1 reports the average P&L of the different portfolios and their respective standard deviation. We first notice that GW-portfolios, on average, outperform all the other alternatives. Moreover, the GW-portfolio with the best parameter configuration produces superior P&L with respect to the stock-based portfolio, even in the worst case scenario, as represented by the lower bound of the standard deviation.

Finally, we also notice that, during the Covid-19 market crash, the stock-based portfolio suffered the largest draw-down, compared with the GW-portfolios.

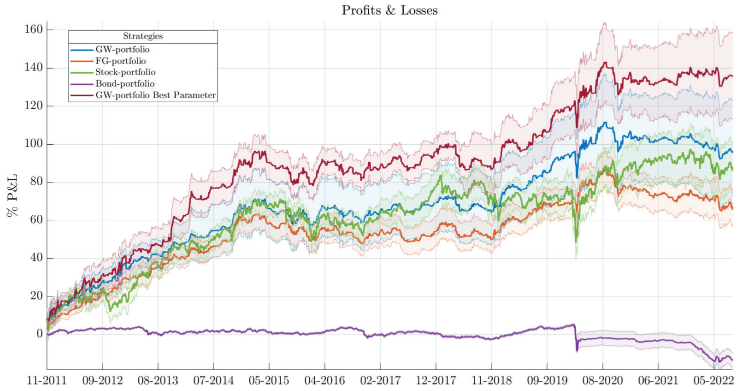


Fig. 1 Cumulative Profits and Losses. The figure reports the behavior of cumulative profits and losses associated with different portfolios.

3 Conclusion

We present a general nonlinear portfolio optimization which deals with risk in a more flexible way. Our methodology starts by defining risk-on, risk-off and risk-neutral input portfolios and interpolates between them using a Gromov-Wasserstein discrepancy measure. The resulting portfolio, that we named the GW-portfolio, is obtained as a Gromov-Wasserstein barycenter of the input portfolios, with barycentric coordinates defined by the dynamics of a market sentiment index. We demonstrate the usefulness of this methodology by comparing the GW-portfolio against suitable benchmark portfolios through different performance measures, which account for both return and risk.

References

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Mantegna, R. N. and Stanley, H. E. (1999). *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press.

Mémoli, F. (2007). On the use of gromov-hausdorff distances for shape comparison.

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487.

Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Online Job Advertisements: toward the quality assessment of classification algorithms for the occupation and the activity sector

Annunci di lavoro online: verso una valutazione della qualità degli algoritmi di classificazione per l'occupazione e l'attività economica

Elena Catanese, Francesca Inglese, Annalisa Lucarelli, Alessandra Righi, Giuseppina Ruocco

Abstract Online advertisements in job portals and company sites (OJAs) have great potential to get detailed and timely insights into labor market trends. To use this information is important to transpose the information expressed in the ads in NL textual into Standard international statistical classifications (eg., NACE, ESCO). We describe the strategy we followed for assessing the quality of the procedure used to classify these variables available in the OJAs Cedefop database. The evaluation of the efficiency of the supervised classifier algorithms focuses on the experimentation of procedures that try to combine human and machine learning intelligence in order to improve the accuracy of the classifiers. This work is part of the activities carried out within the ESSnet WIN project and aimed at defining a quality framework for OJA data.

Abstract *Gli annunci online nei portali di lavoro e nei siti aziendali (OJA) hanno un grande potenziale per ottenere informazioni dettagliate e tempestive sulle tendenze del mercato del lavoro. Per utilizzare queste informazioni è importante trasporre le informazioni espresse negli annunci in linguaggio naturale in classificazioni statistiche internazionali standard (es. NACE, ESCO). Nel lavoro descriviamo la strategia che abbiamo seguito per valutare la qualità della procedura utilizzata per classificare alcune variabili del database OJA di Cedefop. La valutazione dell'efficienza degli algoritmi di classificazione supervisionati si concentra sulla sperimentazione*

Elena Catanese
Istat, Rome, e-mail: catanese@istat.it

Francesca Inglese
Istat, Rome, e-mail: inglese@istat.it

Annalisa Lucarelli
Istat, Rome, e-mail: anlucare@istat.it

Alessandra Righi
Istat, Rome, e-mail: righi@istat.it

Giuseppina Ruocco
Istat, Rome, e-mail: giruocco@istat.it

di procedure che cercano di combinare l'intelligenza umana e l'apprendimento automatico al fine di migliorare l'accuratezza dei classificatori. Questo lavoro fa parte delle attività svolte nell'ambito del progetto ESSnet WIN e mira a definire un framework di qualità per i dati OJA.

Key words: Online job advertisements, web data, validation, classifiers evaluation

1 Introduction

The increasing amount of online advertisements in job portals and company sites (OJAs) has great potential to get detailed and timely insights into labor market trends. OJAs do not replace traditional data sources but can complement them to produce additional indicators, thus enriching the current official statistical production. The information acquired from OJAs, thanks to advances in web crawling technologies, machine learning, and big data techniques, is a relevant data sources for the analysis of online job vacancies.

The European Center for the Development of Vocational Training (Cedefop) has implemented a pan-European system for collecting and analyzing OJAs content [1]. Cedefop actively collaborates with Eurostat's Big Data Task Force and the European Statistical Systems Network (ESSnet), to explore the use of online job advertisements as a source of data for producing official statistics, as well as the challenges related to quality assurance. The Italian National Institute of Statistics (ISTAT) involved in the ESSnet Web Intelligence Network project (WIN). One of the main tasks of the project is the development of the use of OJAs and the improvement of the quality of the information produced. At the European level, OJAs derived from many online sources (job portals, company sites, social networks, employment websites, employment agencies, job search engines, online newspapers, public employment services, and employers organizations), are centrally collected, and stored in a Data Lab. The updated databases are disseminated on a quarterly basis with data available at a daily level. Databases are available starting from the third quarter of 2018 to the third quarter of 2022.

OJA data are made available in the Data Lab following all phases of collecting, processing, cleaning, standardizing, and classifying. The validation rules currently applied to the OJA dataset are both consistency and plausibility rules. Consistency with Eurostat's standard classifications (e.g. Multilingual Classification of Skills, Competences, Qualifications and Occupations ESCO, Statistical Classification of Economic Activities in the European Community NACE Rev. 2,) is required and the distribution of ads within categories of a classification is required to be reasonably stable over time and across data releases. Besides these rules, operating at the level of the data record or of a variable distribution, there are some others structural validation rules, which include the checking of correct naming of datasets and variables, absence of empty fields, etc.

This paper describes the strategy followed for assessing the quality of the procedure used to classify some variables such as occupation and economic activity, which are of particular interest for the labor market analysis. We consider some methods in order to both evaluate the quality of the classifiers (machine learning procedures) and contribute to a possible their improvement. This activity concerns the evaluation phase and follows the data validation process that depends on the efficiency of the classifier algorithm used and allows to check the internal consistency of the data but not its accuracy. The evaluation phase aims to increase the accuracy and efficiency of automatic classifiers that extract statistical variables from the text in natural language found in the job description and in other structured fields from the job portals. The activity of evaluation of the classifiers concerns the experimentation of procedures that try to combine human and machine intelligence in order to maximize the accuracy of the classifiers, assist human activities with machine learning, and increase the efficiency of classifiers. This work is still in progress and is part of the activities carried out within the WIN project and aimed at defining a quality framework for OJA data. In the next sections, we describe the classification algorithms within the Cedefop system; then we summarize the methods applied for the quality assessment of the algorithms. Finally, we present some preliminary results and future steps needed.

2 Classifier Evaluation

2.1 Classification algorithms

OJAs refer to advertisements published on job portals revealing an employer's interest in recruiting workers with certain characteristics for performing certain work. These advertisements normally include a lot of on the characteristics of the job (e.g. occupation, location, type of contract, working time and salary), characteristics of the employer (e.g. economic activity sector) and job requirements (e.g. education, skill and experience) and also on the advertisement itself (e. g. job portal and publishing and the expiring date of the ads). Part of this information is available only as natural language textual data. Therefore, to deal with this type of big data requires specific methodologies in terms of processing, classifying and analysing.

The classification algorithms, for extracting information from OJAs, are based on ontologies (keyword lists) and a machine learning model. Ontologies create a framework for processing and analysis of online vacancies and are defined for each language and for each categorical variable. The Cedefop system uses both standard and custom ontologies. Indeed, the keywords list is based on existing standard classifications and is enriched by experts [1].

Ontology based models try to classify OJAs according to the terms contained in the corresponding ontology. The process first seeks to rank the vacancy using text match and/or similarity to the relevant ontology. Practically, the raw data of each posting (job description and other structured fields, such as job title) is searched, and when a match is found, the related category is assigned as a result of the classification for the job posting. If the ontology models provided no results, then a machine learning algorithm previously trained is applied to classify the OJAs. To train the classifiers, Cedefop refers to a family of algorithms based on the same principle; the process must be applied separately for each variable considered and for each language. This approach is adopted because each model must be trained to best fit the characteristics of the domain and language [1].

Experts regularly validate the results of the machine classification process. The proposed adjustments are used to improve the accuracy of classification. The semi-automated augmentation process adds new terms and synonyms into ontologies through machine learning algorithms. Ontologies are also manually updated to incorporate new information. The results of corrections of misclassified job vacancies can also be incorporated into the training set and used to test the accuracy of the machine learning model [1].

2.2 Methods

Data annotation is the process of labelling the data and refers to the human classification of raw data. The set of data annotated with the same specification (annotated corpus) is a crucial input in the OJAs data classification process. OJAs data revision constitutes a way to improve the annotated corpus that can be used to train, validate and test machine-learning algorithms or to estimate the precision of the classifiers. Accuracy and consistency are the main measures used to evaluate the quality of annotated data: accuracy measures how close the labelling is to ground truth, while consistency refers to the degree of accuracy across the overall dataset [2].

In the WIN project, to evaluate the classification algorithms for the occupation variable, a stratified sample of OJAs containing the raw data and the classification results (ESCO) is selected for some European countries to accomplish the data annotation process. Data labelling was performed by reviewers who checked the selected OJAs, enhancing the correspondence with the classification results. Reviewers looked at ontologies associated with the occupation classification category and proposed changes to solve the identified problems. The data labelling tool used for data annotation is Doccano [3], an open-source text annotation tool for humans, providing several features to execute different tasks.

In addition to this, a descriptive analysis has been performed to verify the internal coherence between major occupational groups and economic activity sections. The aim was to measure the percentage share of occupations related to each economic section. To evaluate the classification algorithms for the economic activity variable, a different approach is being implemented. This method is based on the analysis of a testing sample, and consists in linking the company name deriving from the OJA with the information included in the Business Register. Then, the accuracy of the classifier is assessed, by comparing the attributed NACE code with the information from the statistical archive.

3 Results

The data annotation exercise on a sub-sample of 400 OJAs selected for Italy showed that in a rather high percentage (about 45 % of the cases) the two classifications of the OJAs textual data on the occupation applied (one based on the classification algorithm and the other on the manual annotation process) provided the same results. In 49% of the cases, the classification made by the algorithm did not correspond to that resulting from the manual annotation of the data. In a few cases, the manual annotation was not able to define a single occupation label for the advertisement (less than 1% of cases). In a very few cases (around 0.2%), either the advertisements contained no reference to the type of occupation required or it was impossible to classify the information at the 4-digit level of the ISCO-08 (International Standard Classification of Occupations). Furthermore, in less than 1%, the manual annotation was not able to define a single occupation label, but several ones were assigned to the advertisement; while in a non-negligible percentage, the sub-sample included announcements that were not really OJAs (about 4%).

The descriptive approach used to assess internal consistency between the economic activity and the occupation classification has been carried out at a macro level, by analysing the major, sub-major, and minor groups of ISCO-08 (International standard classification of occupations) in each economic activity section of the NACE classification (Sections B to S of the NACE Rev. 2 classification). In most economic activity sections, about 75% of the total amount of OJAs of the section has been classified into a type of occupation consistent with the main activity of the section.

In order to assess the quality of the NACE economic activity classifier, sub-samples of 500 OJAs have been extracted for some countries, including Italy. After a normalization and cleaning procedures of the company names the matching between the OJAs sub-samples and the Business register has been carried out. The results show a significant percentage of matching (around 60% of the sub-sample) but, on average, the accuracy of NACE classification of the company activity is around half of the matched cases.

4 Conclusion

Evaluating classifiers to improve OJA classifications and ontologies is an evolving phase of the OJA data mining process. A cyclical evaluation procedure based on the audit sample will be put in place, leading to iterative cycles of feedback and improvement. We will consider new approaches for selecting an audit sample of OJAs, for example, oversampling difficult categories. The evaluation dataset released peri-

odically will lead to 1) evaluation metrics for the classification algorithms (e.g., the accuracy rate); 2) suggestions for improvement of the sets of keywords, i.e. ontologies; 3) a set of human-coded data growing over time for training machine-learning models.
The preliminary activities described in this work constitute a first step towards the gold standard strategy with the aim to ensure high quality of annotated data. Defining gold standards for OJA variables of particular interest for the labour market analysis is the ESSnet project main approach for evaluating and improving the algorithm quality.

References

1. Cedefop, “Online job vacancies and skills analysis: a Cedefop pan-european approach”. 2019
2. J. Branka, V. Kvetan, J. Napierala “From the online job advertisements to official statistics – the aspects of quality assurance”. Q2022, Vilnius. 2022.
3. Doccano, https://doccano.github.io/doccano/

Linear Programming for Wasserstein Barycenters

Gennaro Auricchio, Federico Bassetti, Stefano Gualandi, Marco Veneroni

Abstract This paper presents a family of generative Linear Programming models for the computation of the Wasserstein Barycenter of a large set of two-dimensional images. Wasserstein Barycenters were recently introduced to mathematically generalize the concept of averaging a set of points, to the concept of averaging a set of clouds of points, such as, for instance, two-dimensional images. In Machine Learning terms, the Wasserstein Barycenter problem is a generative constrained optimization problem, since the values of the decision variables of the optimal solution give a new image that represents the average of the input images. Our family of Linear Programming models rely on different types of Kantorovich-Wasserstein distances used to compute a barycenter, and they are efficiently solved with a modern commercial Linear Programming solver. We numerically show the strength of the proposed models by computing and plotting the barycenters of all digits included in the classical MNIST dataset.

1 Discrete Kantorovich-Wasserstein Distances

In several Machine Learning problems, a fundamental step is the computation of a similarity measure between a pair of objects. These objects very often correspond to uncertain measure quantities, which are represented as probability density func-

---

Gennaro Auricchio
School of Electronics and Computer Science, The University of Southampton, e-mail: gennaro.auricchio@ecs.soton.ac.uk

Federico Bassetti
Dipartimento di Matematica, Politecnico di Milano, e-mail: federico.bassetti@polimi.it

Marco Veneroni
Dipartimento di Matematica, Università degli Studi di Pavia, e-mail: marco.veneroni@unipv.it

Stefano Gualandi
Dipartimento di Matematica, Università degli Studi di Pavia, e-mail: stefano.gualandi@unipv.it

tions or discrete N-dimensional histograms. The Kantorovich-Wasserstein distance is a mathematical metric which permits to compute the distance between probability density functions or N-dimensional histograms by solving a constrained optimization problem [15, 10]. Intuitively, the distance between two (discrete) measures is equal to the total cost of "transporting" all the mass of the first (discrete) distribution into the second. The cost for transporting a unit of mass from a location of the first distribution to a location of the second distribution is called the *ground distance*. In case of N-dimensional histograms, the locations are associated to the centers of the bins of the histograms, and the ground distance can be, for example, any of the standard norms: ℓ_1, ℓ_2, or ℓ_∞.

Given a discrete subset of n points in $\mathbb{R}^d$: $X \subseteq \mathbb{R}^d$, $X = \{p_1, \dots, p_n\}$, a discrete probability measure on X can be represented as $\tilde{\mu} = \sum_{i=1}^n \mu_i \delta_{p_i}$, where δ_{p_i} is the Dirac delta concentrated on $p_i \in X$ and $\mu = (\mu_1, \dots, \mu_n)$ is a vector in the unitary simplex $\mathscr{S}^n$, that is, $\mu \in \mathbb{R}^n_+$ and $\sum_{i=1}^n \mu_i = 1$. For sake of simplicity, we drop the Dirac delta from the notation and we identify $\tilde{\mu} = \mu \in \mathscr{S}^n$.

If we assume that the cost of transporting a unit of mass from position p_i to position p_j is given by the ground distance $d_{ij} = \|p_i - p_j\|_\alpha$, for a suitable norm $\|\cdot\|_\alpha$, (e.g., $\alpha = 1, 2, \infty$), the Kantorovich-Wasserstein distance $W(\mu, \nu)$ between the two discrete measures μ and ν on $X \subset \mathbb{R}^N$ can be defined as the solution of the following Linear Program:

$$\begin{aligned}
W(\mu,\nu) = \min \quad & \sum_{i=1}^n \sum_{j=1}^n d_{ij} x_{ij} && && (1)\\
\text{s.t.} \quad & \sum_{j=1}^n x_{ij} = \mu_i && i = 1, \dots, n && (2)\\
& \sum_{i=1}^n x_{ij} = \nu_j && j = 1, \dots, n && (3)\\
& x_{ij} \geq 0, && i = 1, \dots, n, j = 1, \dots, n. && (4)
\end{aligned}$$

Note that this problem is clearly a special case of the Transportation Problem [9, 11] and, hence, it can be formulated and solved as an uncapacitated minimum cost flow problem in a bipartite graph with $2n$ nodes and n^2 arcs [2].

In this paper, we are interested in the Kantorovich-Wasserstein distance between discrete probability measures when the distance d_{ij} in $\mathbb{R}^N$ is measured with the ℓ_1, ℓ_2 and ℓ_∞ norms.

2 Wasserstein Barycenters

A very interesting problem, which uses the Kantorovich-Wasserstein distance as a building block, is the **Wasserstein Barycenter problem** introduced in [1]: Given a set of m discrete measures γ_k, with $k = 1, \dots, m$, defined on a space $X \subseteq \mathbb{R}^N$, we have to find a discrete measure $\nu^* \in \mathscr{S}^n$ that has the minimal overall Kantorovich-

Wasserstein *distance* to all γ_k measures:

$$\mathscr{B}(\gamma) := \min_{\nu \in \mathscr{S}^n} \sum_{k=1}^{m} W(\gamma^k, \nu). \tag{5}$$

The new discrete measure ν^* is called the Wasserstein Barycenter, since it generalizes the idea of averaging of usual Euclidean barycenters. If we denote by γ_i^k the i-th element of vector γ^k, the discrete Wasserstein Barycenter problem is equivalent to the following Linear Program [3]:

$$\begin{aligned}
\mathscr{B}(\gamma) = \min_{y} \quad & \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} x_{ijk} && && (6)\\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ijk} = \gamma_i^k && i = 1,\ldots,n,\ k = 1,\ldots,m && (7)\\
& \sum_{i=1}^{n} x_{ijk} = \nu_j && j = 1,\ldots,n,\ k = 1,\ldots,m && (8)\\
& \sum_{j=1}^{n} \nu_j = 1 && && (9)\\
& x_{ijk} \geq 0, && i,j = 1,\ldots,n,\ k = 1,\ldots,m && (10)\\
& \nu_j \geq 0, && j = 1,\ldots,n. && (11)
\end{aligned}$$

The size of problem (6)–(11) depends on the size of each discrete measure γ^k, that is equal to n, and on the number m of given discrete measures: there are n^2m variables and $2nm$ constraints. Indeed, the size of this Linear Program in practice can become very large. Note that we have m bipartite graphs with n^2 arcs each, and hence, its solution raises very interesting computational challenges.

3 Related work

The state-of-the-art approach to compute Wasserstein Barycenters relies on entropic regularized formulations of the constrained optimization problem [8], which are solved with derivations of the Sinkhorn's algorithm [12, 7, 13]. The main advantages of this class of algorithms are two: (i) they are very easy to understand and to implement, and (ii) they can be implemented to run in parallel on multiple Graphics Processing Units (GPU). However, in the regularized formulation of Optimal Transport there is a crucial parameter that has to be tuned manually. As shown in [4], these methods can be numerically unstable and can provide solutions which are very far from the optimal value. Other approaches to compute the Wasserstein Barycenters include stochastic algorithms [6] and gradient descent based algorithms [14]. Unfortunately, Linear Programming approaches to compute Wasserstein Barycenters [3] are repeatedly considered to be extremely inefficient on larger instances [6, 14].

4 Our contribution

In this paper, we propose a new class of generative Linear Programming models to solve the Wasserstein Barycenter problem. Building on the results recently presented in [5], and depending on the norm used as ground distance in the pairwise Kantorovich-Wasserstein distances, we exploit the geometric structure of the problem to reduce the size of the linear programs. For the ℓ_1 and ℓ_∞ ground distances, we provide exact formulations for the computation of the barycenter of up to 3 200 images of size 28×28. For the ℓ_2 ground distance, with our approximation scheme we are able to approximate the optimal barycenter within a guaranteed percentage error.

References

1. Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis **43**(2), 904–924 (2011)
2. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows: Theory, Algorithms, and Applications. Cambridge, Mass.: Alfred P. Sloan School of Management, Massachusetts Institute of Technology (1988)
3. Anderes, E., Borgwardt, S., Miller, J.: Discrete Wasserstein barycenters: Optimal Transport for discrete data. Mathematical Methods of Operations Research **84**(2), 389–409 (2016)
4. Auricchio, G., Bassetti, F., Gualandi, S., Veneroni, M.: Computing Kantorovich-Wasserstein distances on d-dimensional histograms using (d+1)-partite graphs. Advances in Neural Information Processing Systems (2018)
5. Bassetti, F., Gualandi, S., Veneroni, M.: On the computation of Kantorovich-Wasserstein distances between 2D-histograms by uncapacitated minimum cost flows. SIAM Journal on Optimization **30**(3), 2441–2469 (2020)
6. Claici, S., Chien, E., Solomon, J.: Stochastic Wasserstein barycenters. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:999-1008 (2018)
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of Optimal Transport. In: Advances in Neural Information Processing Systems. pp. 2292–2300 (2013)
8. Cuturi, M., Doucet, A.: Fast computation of Wasserstein barycenters. In: International Conference on Machine Learning. pp. 685–693 (2014)
9. Flood, M.M.: On the Hitchcock distribution problem. Pacific Journal of Mathematics **3**(2), 369–386 (1953)
10. Santambrogio, F.: Optimal Transport for applied mathematicians. Birkäuser, NY pp. 99–102 (2015)
11. Schrijver, A.: On the history of the transportation and maximum flow problems. Mathematical Programming **91**(3), 437–445 (2002)
12. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics **21**(2), 343–348 (1967)
13. Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., Guibas, L.: Convolutional Wasserstein distances: Efficient Optimal Transportation on geometric domains. ACM Transactions on Graphics (TOG) **34**(4), 66 (2015)
14. Staib, M., Claici, S., Solomon, J.M., Jegelka, S.: Parallel streaming Wasserstein barycenters. In: Advances in Neural Information Processing Systems. pp. 2647–2658 (2017)
15. Villani, C.: Optimal Transport: old and new, vol. 338. Springer Science & Business Media (2008)

A multi-channel convolution approach for forecast reconciliation

Andrea Marcocchia and Serena Arima and Pierpaolo Brutti

Abstract Forecast reconciliation for hierarchies of time series, whose idea is that observed responses at each level have to add up to those observed at higher levels, is a theme of growing interest in the scientific community. The challenge is to exploit the high signal-to-noise ratio that characterizes the most aggregated data to boost the forecast on the more granular data. In this work, a new family of techniques is developed to solve the problem in a multi-hierarchical scenario, embracing both temporal and classical hierarchies. The idea is to leverage the predictive power of Convolutional Neural Networks (CNN) by designing a multi-channel matrix with the goal of making all the hierarchies simultaneously consistent. This approach has been tested on real and simulated datasets, and the new architectures achieve promising results.

Key words: forecasting, time series, forecast reconciliation, hierarchical time series, convolution, deep learning

1 Introduction

The goal of forecast reconciliation is that forecasted values at a level of the hierarchy add up to the predicted demands at higher levels. If forecasting at the different levels is done independently, we have forecast incoherence, meaning that the bottom level forecasts do not add up. The various components of the hierarchy can interact in a variety of complex ways: a change in one series can have an impact on other series at the same level, as well as on series at higher or lower levels. Reconciliation

---

Andrea Marcocchia
Sapienza University of Rome, e-mail: andrea.marcocchia@uniroma1.it

Serena Arima
University of Salento, e-mail: serena.arima@unisalento.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

is the process that fix incoherent forecasts.
In this work a new approach is proposed that exploits convolution techniques to perform reconciliation in the case of multiple hierarchies at the same time, so that all the hierarchies are exploited simultaneously to produce the best possible reconciled values. Most of the techniques in the literature are built to work with a single hierarchy. However, there are numerous real-world cases in which multiple hierarchies are valid simultaneously, and the information contained in such data structures and the gain they can bring must be exploited in the best possible way. The idea is to introduce an architecture that allows to work with multiple hierarchies in a Deep Learning framework, such that the output values are coherent for multiple hierarchies at once. The hierarchies can be of different types:

- multiple classical hierarchies: all hierarchies are administrative, spatial or groups;
- classical and temporal: one hierarchy is "classical", while the other is temporal;
- a mix of the two previous cases: the hierarchies can be more than two, so it possible to fall into a case that is a mix of the previous two.

The innovative and most important aspect of the presented technique is that the values obtained will be simultaneously consistent for all hierarchies considered, while exploiting information from all of them. This process is embedded in an end-to-end architecture, so that all information is also leveraged during the forecasting phase, and the two steps are not unnecessarily separated, returning in ouput coherent values.

2 Architecture

The proposed method makes use of Deep Learning, and it is specifically proposed as an end-to-end architecture. This means that a single loss function is optimized that keeps together both the forecasting and the reconciliation steps, instead of choosing a different loss and hyper-parametrer setting for the forecast and reconciliation stage. In the proposed architecture, the task of making the information available in the different levels of the hierarchy lies within the convolution layer. For this reason, it is essential to implement this method correctly, so that information circulates in the best possible way. In the basic scenario of a single hierarchy there are n bottom level series $b_1, b_2, \ldots, b_n$, and for each of these series the appropriate aggregate values can be created according to the hierarchical structure. Using a matrix formalization, there are n matrices such as the following one:

$$\left[\begin{array}{ccccccc} z_{1,1} & z_{1,2} & z_{1,3} & z_{1,4} & z_{1,5} & \cdots & z_{1,t} \\ z_{2,1} & z_{2,2} & z_{2,3} & z_{2,4} & z_{2,5} & \cdots & z_{2,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{k_1,1} & z_{k_1,2} & z_{k_1,3} & z_{k_1,4} & z_{k_1,5} & \cdots & z_{k_1,t} \\ b_1 & b_2 & b_3 & b_4 & b_5 & \ldots & b_t \end{array}\right] \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} H_1 \\ \left.\begin{array}{c} \\ \end{array}\right\} B \end{array}$$

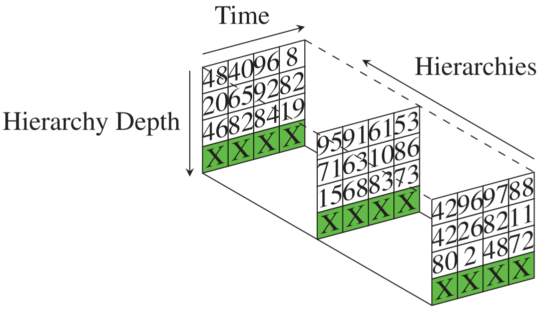
As the matrix shows, there is a bottom level series as the first row of the matrix, and all the aggregate values created according to the hierarchy. The values $z_{k_1,1}, z_{k_1,2}, \dots, z_{k_1,t}$ are the values for the first level of aggregation, and t is the time window observed in the dataset. The $z_{1,1}, z_{1,2}, \dots, z_{1,t}$ are the t values for the fully aggregated values according to the first hierarchy[1].

Following this approach, in case of a second hierarchical structure, that shares the bottom level series with the first hierarchy, it is possible to represent the data as follows:

$$\begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & y_{1,4} & y_{1,5} & \cdots & y_{1,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{k_2,1} & y_{k_2,2} & y_{k_2,3} & y_{k_2,4} & y_{k_2,5} & \cdots & y_{k_2,t} \\ z_{1,1} & z_{1,2} & z_{1,3} & z_{1,4} & z_{1,5} & \cdots & z_{1,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{k_1,1} & z_{k_1,2} & z_{k_1,3} & z_{k_1,4} & z_{k_1,5} & \cdots & z_{k_1,t} \\ b_1 & b_2 & b_3 & b_4 & b_5 & \dots & b_t \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix} y \\ \vdots \\ y \end{matrix}}\right\} H_2 \\ \left.\vphantom{\begin{matrix} z \\ \vdots \\ z \end{matrix}}\right\} H_1 \\ \left.\vphantom{b}\right\} B \end{array}$$

This representation is the more intuitive one following the previous approach. In fact, also in this case there are n matrices in output. The use of this data representation has the advantage of enclosing all the information related to a bottom level time-series in a single matrix, thus being able to exploit a convolution step exactly as done in the case of a single hierarchy. At the end of the convolution step, the values returned in the hidden state of the neural network will condense information from both hierarchies. Following exactly the same approach, it is possible to incorporate a temporal hierarchy within the same data structure, as long as the bottom-level series is in common. However, this way of representing the hierarchical data has also many limitations. In fact, the size of the resulting matrix can grow a lot, making the training step more complex as the number of hierarchies or the depth of the hierarchies increases. The biggest problem, however, concerns the way the information circulates within the matrix: the values of the last added hierarchies (H_2 in the previous example) are far removed from the bottom level series, as there is all the data from the other hierarchies between. In this way it can be very complicated through the use of a filter to be able to condense the information of the H_2 hierarchy with that of the bottom level series, as one would like. In this way, the matrix representation that will be obtained at the end of the training will not synthesize the information in the best possible way, because the series at the bottom level will not be able to take advantage of the information provided by the hierarchies included in the highest part of the input matrix. Only filters applied to the final part of the matrix will incorporate the series at the bottom level, with the problem of seeing only the data from the hierarchy immediately above.

The proposed solution is to structure the final matrix in a different way so that all the hierarchical matrices are simultaneously considered, also maintaining a degree of specificity so that different information are not mixed together. This can be done by using a multi-channel matrix, following an approach similar to that used for the RGB images in Computer Vision. In order to exploit in a better way the information contained into all the levels of the different hierarchies, these latter become different channels in the input matrix, with the bottom-level series always inside, instead of being stacked all together. The following image represents what the (multi-channel) input matrix might look like in the case of multiple hierarchies. In this case a new limitation is introduced: $k_1, k_2, \dots, k_w$ (the depths of the w considered hierarchies) must be identical, in order to have the different channels with the same shape.



where the X are the bottom level series, that is the same for each training sample. As it is possible to see from the above image, the advantage of this architecture is that it allows to apply a filter that includes all the information simultaneously. The filter to be used in the convolution step must also be multidimensional, specifically having a number of channels equal to the number of hierarchies. In this way the information is all considered simultaneously, but at the same time a specificity for each hierarchy is maintained.

The complete proposed method exploits an architecture similar to CNN + LSTM, in which in a first step there are layers that allow the information contained in the hierarchy to be shared through a convolution step, and then there is a forecast step exploiting techniques such as LSTM. The loss function can be evaluated in two different ways: focusing on the bottom level series (and rebuilding the whole hierarchy using a Bottom-up approach) or focusing on all the levels of the hierarchy. Considering that the general purposes when forecast reconciliation techniques are used is to improve the quality of the bottom level series, the first approach is used in the experiments.

3 Results

The proposed architecture has been tested over multiple dataset: the M5 dataset [4] of the Makridakis competitions and a simulated dataset. In the M5 dataset, the time-series represent the hierarchical unit sales of the world's largest retail company by revenue, Walmart, and the data comprise 3049 individual products from 3 categories and 7 departments, sold in 10 stores in 3 different states: the two hierarchy are built looking at the administrative organization and at the product classification.

The simulated dataset has a hierarchical structure that mirrors the M5 structure, but the time-series inside have an easier pattern: all the information are generated using an ARMA process. Both the dataset have multiple hierarchies defined. The common goal in analyzing the performances across datasets is to verify if the use of reconciliation techniques is useful in improving the quality of the original forecasts. In particular, the focus has been on the bottom level series, as these are the most complex and at the same time the most interesting forecasts. The results are evaluated according to the Mean Absolute Error (MAE). The proposed architecture is compared with multiple methods proposed in the literature, such as Optimal Reconciliation methods with OLS Estimator or WLS Estimator [2], Mint [3], Top-down and Bottom-up [5]. The Top-down approach is tested with three different approaches: Top-down with forecast proportions (TD-FP), average of the historical proportions (TD-GSA) and proportions based on the historical averages (TD-GSF). To properly perform the experiments, all the dataset are divided into training, test and validation. Specifically, 30% of the dataset is isolated to evaluate the test set, 20% as validation and the remainder as training. The splitting is performed on the bottom level series.

Table 1 MAE performances on test set

Model	Simulated	M5
Matrix Conv (multiple hierarchies)	0.0670	0.0921
MINT	0.0578	0.0971
WLS	0.0598	0.0974
OLS	0.0609	0.0981
BU	0.0571	0.0976
TD-FP	0.0810	0.1507
TD-GSF	0.0950	0.1099
TD-GSA	0.0890	0.1119

In both the datasets, the performances are evaluated by focusing on the series at the bottom level, since these are the ones of greatest interest. All the methods that require a base forecast algorithm to perform the reconciliation step (all except the end-to-end architectures that uses Deep Learning) use an Arima model to perform the base forecast. For the methods that work with only one hierarchy, one of the defined hierarchical structure is chosen to perform the reconciliation step.

For what concern the simulated dataset, it is possible to observe that the best result is obtained with the Bottom-Up method: this is quite intuitive since with this approach the focus is all on the series at the bottom level, and any errors that need to be corrected in the reconciliation stage affect the series at the higher levels of the hierarchy. The MINT algorithm also performs well for the reconciliation, but all the Optimal-reconciliation methods seem to be good. It is important to keep in mind that for the simulated dataset the generation process and the base forecast algorithm is the same, so it justifies the better performances of the classical methods with respect to the end-to-end architectures. Analyzing the M5 dataset, as can be seen from the previous table, the method that makes use of neural networks outperforms all the others. The result is due to the fact that in this case the data are much more complicated in their structure, so a simple Arima model is not enough to provide good base forecast. On the other hand, the end-to-end architecture provides a more flexible and power system to perform the forecast and the reconciliation.

4 Conclusions

Regarding the new reconciliation techniques that exploit convolution over multiple hierarchies, it has been observed that the results are in line, and sometimes better, with other methods in the literature. It should be emphasized, however, how computationally complex is to train these models in terms of the hardware resources required. Experiments on the M5 dataset are carried out by sampling the data in order to obtain a large number of time series to train the model robustly and generically. It is necessary because the proposed model have a large number of parameters, and the convolution steps are really computationally expansive: the use of the whole dataset is not feasible with the hardware resources at our disposal.

In addition, it is noted that the matrices data structure is very rigid, while the problem to be modeled requires more flexibility. In fact, although with the proposed method it is possible to avoid neighboring series in the hierarchy having an excessive distance, the matrix representation is still forced: for example the bottom level series is repeated for all the channels, producing some redundancy. A possible idea, that will be subject to further research is to use more flexible data structures, such as graphs, to perform the convolutional step.

References

1. Theodosiou, F. and Kourentzes N.: Forecasting with Deep Temporal Hierarchies. (2021) Available via http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1576994&dswid=3158
2. Athanasopoulos, G. and Hyndman R. et al. : Optimal combination forecasts for hierarchical. Computational Statistics & Data Analysis. **55**, 2579-2589 (2011)
3. Wickramasuriya, S. and Athanasopoulos G. and Hyndman R. : Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. Journal of the American Statistical Association. **114**, 1-45 (2018)
4. Makridakis, S. and Spiliotis E. and Assimakopoulos V. : M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting. (2022)
5. Orcutt G. et al. : Data Aggregation and Information Loss. The American Economic Review. **4**, 773-787 (1968)

Hedging global currency risk with factorial machine learning models

Copertura del rischio di cambio tramite apprendimento automatico dai fattori principali

Paolo Giudici, Paolo Pagnottoni, and Alessandro Spelta

Department of Economics and Management, University of Pavia.
`Corresponding authors and presenter: Paolo Pagnottoni,`
`paolo.pagnottoni@unipv.it, via San Felice 5, 27100 Pavia, Italy .`

Abstract

We propose a dynamic method to hedge foreign exchange risk of international equity portfolios. The method is based on the currency returns predictions obtained from a set of alternative machine learning models, built on the main factorial components of the time series currency returns. The analysis of several model performance indicators allows to conclude that accurate predictions of global factor returns, such as those obtained with non linear machine learning models, can improve currency risk hedging.

Proponiamo un metodo dinamico per coprire il rischio di cambio di portafogli azionari internazionali. Il metodo si basa sul prevedere i movimenti dei rendimenti valutari mediante le previsioni dei rendimenti dei principali fattori di rischio. A tal fine, costruiamo diversi modelli predittivi per la componente dei rendimenti globali dei fattori e ne studiamo la capacità di generare portafogli azionari con copertura dinamica. L'analisi di diversi indicatori di performance dei modelli consente di concludere che previsioni accurate dei fattori possono migliorare il grado di copertura del rischio di tasso di cambio.

Keywords: Currency returns, Currency risk factors, Mean-variance optimization, Time series machine learning.

1. Introduction

Global investment positions are nowadays substantial and have grown quickly in recent decades. When redenominated in the investors' home currency, the currency exposure resulting from international investments can significantly alter the underlying assets' overall risk-return profile, highlighting the key question of how investors should manage their foreign exchange exposure.

The main strategy for controlling foreign exchange (FX) exposure is mean-variance optimization, which establishes the best currency hedging positions. The strategy is theoretically interesting and includes requirements for both speculative hedging and risk management naturally. However, due to the well-known difficulties in forecasting exchange rates, this strategy, when used out-of-sample, suffers from severe estimating error in currency returns - see (3) -, leading to poor overall currency hedging performance - see (1; 2).

To solve the problem, in the paper we propose a novel methodology to dynamically determine currency hedge positions, that we denominate dynamic currency factor (DCF) hedging. The approach exploits the predictability of currency returns desumed from the predictability of the corresponding risk factors. Recent breakthroughs in international macro-finance by Lustig et al. (2011) and Verdelhan (2018) have found that the cross-section of currency returns can be explained as compensation for risk in a linear factor model that includes two global currency risk factors: dollar and carry. The dollar factor corresponds to the average return of a basket of currencies against the US dollar, while the carry factor reflects the returns from currency carry trade. These factors also account for a large proportion of the time-series exchange rate behavior in contemporaneous regressions, and thus, if their returns are predictable, currency returns are also partially predictable. We show that exploiting a predictable component in both global factors helps to mitigate the estimation error that typically hinders traditional mean-variance currency hedging, thereby delivering important investment gains.

Specifically, we build time series machine learning models to accurately forecast factor returns and improve the predictability of single currency returns.

To illustrate our proposed methodology from an empirical viewpoint, in the empirical part of the paper we take the perspective of a US investor who invests in a portfolio of G10 developed economies. The investor is assumed to have a predetermined long position in either foreign equities or bonds and desires to manage the Foreign Exchange exposure by forming optimal hedge positions within a mean-variance optimisation framework. To construct optimal hedge positions, monthly estimates of currency returns are initially formed by estimating currenciesâ exposures to dollar and carry factors (that is, factor betas). Factor returns are then forecasted using variables that have been theoretically motivated to drive either one or both factor returns, including Foreign exchange volatility (Merton, 1973; Menkhoff et al., 2012a; Cenedese et al., 2014), the average forward discount (Lustig et al., 2014), the TED spread (Brunnermeier and Pedersen, 2009; Brunnermeier et al.,2009), and commodity returns (Ready et al., 2017). These predicted global factor returns are combined with the estimated factor betas to form out-of-sample (OOS) forecasts of currency returns. The predicted currency returns are finally incorporated within the mean-variance optimizer to produce optimal, currency-specific, hedge positions.

Our empirical results show that the estimation error which frequently affects mean-variance currency hedging can be reduced and that significant economic investment gains can be achieved by taking advantage of a non-linear forecasting of components in both global factors.

2. Methodology

2.1 Framework for deriving optimal currency hedge positions

The main goal is to dynamically create the best currency hedge positions, from the viewpoint of a US investor. FX forward contracts are added to an existing portfolio of reference international equity indices in order to hedge FX exposure. Currency hedges are chosen to optimize a mean-variance investor's utility, whose objective function is used to determine hedge positions, and is specified as

$$\mu_{p,t} - \frac{\gamma}{2}\sigma^2_{p,t} \tag{1}$$

where $\mu_{p,t} = w'_t\mu_t$ represents the expected portfolio return for the following period, $\sigma^2_{p,t} = w'_t\Sigma_t w_t$ represents the portfolio risk, μ_t is the vector of expected excess returns in US dollars with variance-covariance matrix Σ_t. The vector w_t encompasses portfolio weights, and γ is the investor's level of risk aversion.

Portfolio weights in an unconstrained setting are given by $w^*_t = \frac{1}{\gamma}\Sigma^{-1}_t\mu_t$. In our setup, it is adequate to assume the underlying asset weights are predetermined by the portfolio manager, therefore determining the weights given to FX forward contracts is the ultimate goal (i.e., the optimal currency hedge positions). By first partitioning the expected return vector and related covariance matrix across the underlying equity indices and FX future contracts, the optimization issue can be rephrased as

$$\mu_t = \begin{pmatrix} \mu_{x,t} \\ \mu_{f,t} \end{pmatrix}, \quad \Sigma_t = \begin{pmatrix} \Sigma_{xx,t} & \Sigma_{xf,t} \\ \Sigma_{fx,t} & \Sigma_{ff,t} \end{pmatrix} \tag{2}$$

where the underlying indices and FX forwards are represented by x and f, respectively. Hence, the optimal weights in foreign exchange forwards are then given by

$$w^*_{f,t}(f \mid x) = \frac{1}{\gamma}\left(\Sigma^{-1}_{ff,t}\mu_{f,t}\right) - \delta_t w_{x,t} \tag{3}$$

where $w_{x,t}$ is the vector of weights and δ_t is the regression coefficient obtained from regressing underlying indices returns on long FX forward contracts returns, namely $\delta_t = \Sigma^{-1}_{ff,t}\Sigma_{fx,t}$.

Each element of $w^*_{f,t}$ is constrained to be between $-w_{x,t}$ (fully hedged) and zero (unhedged). This restriction mirrors the practice in currency overlay management that prohibits managers from taking speculative FX forward holdings above the position in the underlying security - i.e., a position cannot be overhedged or leveraged by FX forwards. We estimate δ_t and Σ_t each month using a 5-year rolling window and set the underlying portfolio weights to be equally weighted across indices.

2.2 Building currency factors

We firstly define the currency excess return. We use the symbols S and f to represent the log of the spot exchange rate and the forward exchange rate, respectively, both expressed in terms of foreign currencies per US dollars. The value of the domestic currency rises when the dollar value does. Simply put, the net log currency excess return for an investor who goes long in foreign currency i is

$$r^l_{i,t+1} = f_{i,t} - s_{i,t+1} \tag{4}$$

The investor purchases foreign currency or, equivalently, sells the dollar forward at the price (f_t) at time t and purchases dollars at the price (s_{t+1}) in the spot market in $t+1$. Similarly, for an investor who is short the foreign currency - and therefore long in the dollar - the net log currency excess return is given by:

$$r^s_{i,t+1} = -f_{i,t} + s_{i,t+1} \tag{5}$$

We divide the sample of currencies into four portfolios at the end of each period t based on the forward discounts

$$d_{i,t} = f_{i,t} - s_{i,t} \tag{6}$$

that observed at that time. Portfolio 1 comprises the currencies with the lowest interest rates or smallest forward discounts, and portfolio 4 contains the currencies with the highest interest rates or largest future discounts. They are ranked from low to high interest rates. By averaging the log currency excess returns for each portfolio j, we can get the log currency excess return $r^j_{i,t+1}$ for portfolio j. We make the assumption that investors will short all of the foreign currencies in the first portfolio in order to calculate returns.

According to predictions made by linear factor models, the average returns on a variety of assets can be attributed to risk premia related to their exposure to a limited number of risk factors. Our currency portfolios' principal component analyses show that two factors account for more than 65% of the difference in returns across these four portfolios. Since all portfolios load equally on the first principal component, which we labeled λ^{dol}, it can be said that it acts as a level factor and accounts for more than 65% of the common variation in portfolio returns. Given that portfolio loadings rise monotonically across portfolios, the second primary component, denoted λ^{car}, which accounts for about 22% of common variance, can be seen as a slope factor. The systematic component of the dollar factor appears to correlate with low frequency global business cycle conditions and can therefore be interpreted as capturing the level of global macroeconomic risk - see (4). The carry factor is a zero-cost portfolio, constructed by investing in high-yielding currencies while funding the position in low yielding currencies.

2.3 Expected currency returns and common risk factors

The expected currency return vector $\mu_{f,t}$ is a key input in eq. (3), given that the return at time $t+1$, to a US investor who enters a long forward contract on foreign currency i at time t is defined by eq. (4).

We model the cross-section of expected currency returns as function of two common risk factors, named the dollar and carry factors:

$$Er_{i,t} = \zeta(\lambda_t^{dol,car}) \tag{7}$$

Factor return predictability is then exploited for forecasting expected returns. In particular, we run predictive models of factor returns on a set of time $t-1$ predictor variables X_{t-1} described by:

$$\lambda_t^j = \psi_j(X_{t-1}); \quad j = \{dol, car\} \tag{8}$$

to derive multi-step-ahead conditional expected factor returns $E_t\lambda_{t+h}^j = \hat{\psi}_j X_t$ where $h = t+1, ..., t+H$ where H is the forecast horizon.

Given the estimated factor betas from eq. (7) and the expected factor returns (premiums) from eq. (8), conditional expected currency return over the following H periods for each currency pair i can be found as:

$$E_t R_{i,t+h} = \hat{\zeta}(E_t\lambda_{t+h}^j) \tag{9}$$

In our setting, we propose to estimate the relationships between Er_i and $\lambda^{dol,car}$, which we label as ζ, and those among $\lambda^{dol,car}$ and X (ψ_j) by means of time series machine learning models. Many of the machine learning algorithms are well known for delivering good forecasting performances and can be exploited to accurately forecast the dollar and carry factors and the conditional expected currency returns. In particular, we test different families of machine learning forecasting models for time series, namely Discriminant Analysis for time series (DA), Regression Trees (RT) and Support Vector Machine (SVM).

2.4 Currency factor predictors

As currency factor predictors we first exploit the forward currency discount $d_{i,t}$ represented in eq. (6). Secondly, since the carry factor returns exhibit a strong negative relationship with FX volatility we make use of exchange rates volatility measured as the daily squared returns currency pairs i against the US dollar. Namely, we exploit the change in FX volatility:

$$\Delta\sigma_t^i = \log\left(\frac{\sigma_t^i}{\sigma_{t-1}^i}\right)$$

where $\sigma_t^i = r_{i,t}^2$. Thirdly, given that tighter funding liquidity, as proxied by increases in the spread among the LIBOR and the Tbills, can also forecast carry returns, we employ changes in the LIBOR ($LIBOR_t$) and in the Tbills ($Tbill_t$) as exogenous predictors:

$$\Delta LIBOR_t = \log\left(\frac{LIBOR_t}{LIBOR_{t-1}}\right)$$

$$\Delta Tbill_t = \log\left(\frac{Tbill_t}{Tbill_{t-1}}\right)$$

Fourthly, it is established that higher commodity prices predict higher carry trade returns. Therefore, we employ two commodity price indices as currency factor predictors. The first one is the Invesco DB Commodity Index Tracking (CI^1) which tracks changes in the level of the DBIQ Optimum Yield Diversified Commodity Index Excess Return plus the interest income from the Fund's holdings of primarily US Treasury securities and money market income. The second one is the iShares S&P GSCI Commodity-Indexed Trust (CI^2) which tracks a set of futures contracts on an index composed of a diversified group of commodities futures. Also in this case, we use changes in the commodity price indices as currency factor predictors, namely

$$\Delta CI_t^1 = \log\left(\frac{CI_t^1}{CI_{t-1}^1}\right)$$

$$\Delta CI_t^2 = \log\left(\frac{CI_t^2}{CI_{t-1}^2}\right)$$

3. Data

For the purpose of illustrating our method, collect and analyze three main types of time series data: a) currency spot and future returns; b) international reference equity market indices; c) a set of currency factor predictors. Firstly, we consider the following countries and their respective foreign exchange rates: Australia (AUD), Canada (CAD), Switzerland (CHF), Europe (EUR), Great Britain (GBP), Hong Kong (HKD), Japan (JPY), Norway (NOC), New Zealand (NZD), Poland (PLN), Sweden (SEK), and South Africa (ZAR). For each currency, we collect spot and future closing prices. Secondly, we retrieve the equity prices of the reference stock market of each country[1]. Thirdly, we retreive daily time series observations related to the exogenous predictors discussed in Subsection 2.4. The analyzed time period ranges from 27 October 2008 to 30 December 2022.

4. Empirical results

In this Section we present the results of our methodology. The model is back-tested using a dynamic approach. For the in-sample estimation, we choose a 5-year rolling window, each one shifted by 22 working days, i.e. a trading month. For each of them, we produce out-of-sample predictions for the following month. Predictions are obtained by the selecting, for each rolling window, the best performing model across DA, RT and SVM in terms of cross-validation error. To ease computational burden, we use Bayesian optimization for training the set of different models and tuning their hyperparameters. Bayesian optimization finds an optimal set of hyperparameters for a given model by minimizing the objective function of the model, strategically selecting new hyperparameters for each iteration and tipically outperforming parameter grid search.

Figure 1 (upper panel) shows the cumulative payoff to investing in global equity portfolios, and specifically the cumulative payoff to a USD 1 invested in equally weighted global equity portfolios under different currency hedging frameworks from the perspective of a US investor. The payoffs to the linear DCF hedged portfolio, our proposed DCF hedged portfolio and the unhedged portfolio are highlighted. The profit and loss metric is complemented with two performance measures as highlighted in the central and lower panels of Figure 1, which illustrate the Sharpe and the Maximum Drawdown, respectively. Both measures are obtained with a rolling window of two years. Finally, we also report the empirical distribution of returns for the unhedged and hedged portfolio, along with the Kolmogorov-Smirnov test statistics and associated p-value.

We notice that hedged portfolios, on average, outperform the unhedged alternative. This suggests that hedging for global currency risk factors is determinant to the profitability of international equity porfolios. This is further confirmed by the sistematically superior values of the Sharpe ratios and Maximum Drawdown. Moreover, our proposed ML-hedging produces superior performances with respect to the linear hedged portfolio. The Kolmogorov-Smirnov test provides evidence against the null hypothesis that the returns of the two portfolios come from the same continuous probability distribution at all conventional significance levels ($p = 0.00073$).

References

[1] Gardner, G.W., Stone, D., 1995. Estimating currency hedge ratios for international portfolios. Financial Analysts Journal 51, 58–64.

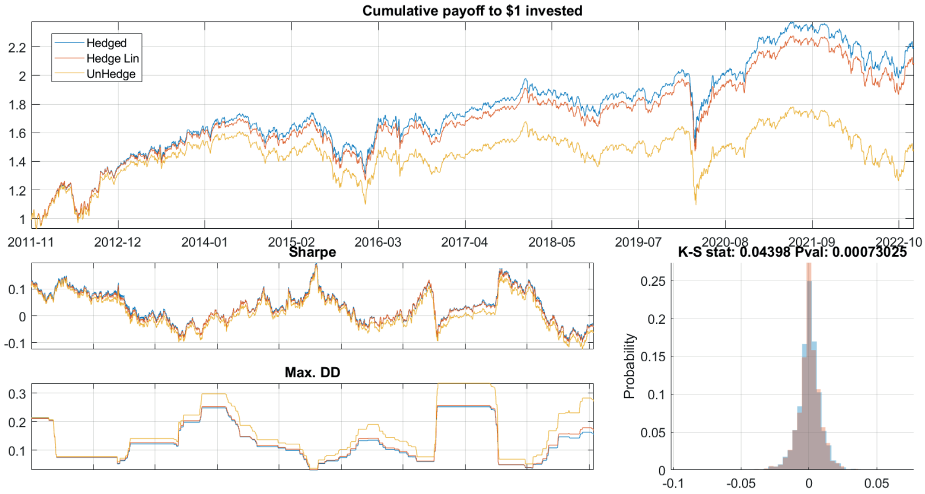
[2] Larsen, Jr, G.A., Resnick, B.G., 2000. The optimal construction of internationally diversified equity portfolios hedged against exchange rate uncertainty. European Financial Management 6, 479–514.

[3] Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of international economics 14, 3–24.

[4] Verdelhan, A., 2018. The share of systematic variation in bilateral exchange rates. The Journal of Finance 73, 375–418.

---

[1] For the EURO area, we select the DAX as reference stock market index.


Cumulative payoff to $1 invested

Sharpe

Max. DD

K-S stat: 0.04398 Pval: 0.00073025

Figure 1: **Performance of ML-hedged, linear hedge and unhedged portfolios**. The upper panel shows the cumulative payoff to a USD 1 invested in equally weighted global equity portfolios under different currency hedging frameworks (ML-hedged, linear hedge and unhedged) from the perspective of a US investor. The central and lower panels of show the Sharpe and the Maximum Drawdown, both measures are obtained with a two-year rolling window. The right lower panel reports the empirical distribution of returns for the unhedged and hedged portfolio, along with the Kolmogorov-Smirnov test statistics and associated p-value.

Predicting musical genres from Spotify data by statistical machine learning

Federica Biazzo and Matteo Farnè

University of Bologna, Department of Statistical Sciences, I-40126, Bologna, Italy
`matteo.farne@unibo.it`,
WWW home page: `https://www.unibo.it/sitoweb/matteo.farne`

Abstract. This paper presents a statistical machine learning application to musical genre classification. Specifically, we draw a sample of songs from Spotify pertaining to six musical genres, i.e. rock, rap, electronic dance music, rhythm and blues, latin, and pop, and we apply a decision tree, a random forest and a boosting algorithm to predict genre labels. Empirical results show that the latter two methods clearly outperform the first one, that speechiness is the audio feature with the largest classification power, and that rock and rap songs are the easiest to be classified.

Keywords: Spotify, musical genre, statistical machine learning

1 Introduction

Artificial intelligence (AI), defined as the ability of a system to learn through experience and solve problems in a continuously changing environment, is currently one of the most active research fields. AI systems, given their processing speed and the precision with which they make decisions, are widely used by companies in order to study the enormous amount of information they collect everyday about their customers, to understand their behaviour and to suggest products and services in line with their customers' interests and preferences [1].

One of the companies that put AI at the business core is Spotify. Founded in 2006, Spotify has quickly become the leader in music streaming, introducing an entirely personal way of listening to music [2]. By collecting information about its users, such as gender, listening history, and interactions with the user interface, Spotify is able to study its customer through Artificial Intelligence, make associations between different artists, tracks, podcasts and playlists, and propose intelligent and increasingly targeted suggestions by machine learning methods [3, 4].

In particular, playlists are the quintessential element of personalisation: users can create playlists that collect the tracks they want, they can share it with other users and they can create collaborative ones. Spotify by default suggests different types of playlist to different listeners: by artist, by music genre, by event. For each song, several audio features are automatically encoded and used to build a playlist according to a specific musical trait.

In this paper, our aim is to derive the most accurate prediction possible of the musical genre of a song starting from the knowledge of its audio features. We consider a large sample of songs from Spotify, pertaining to six musical genres: rock, rap, electronic dance music (EDM), rhythm and blues (R&B), latin and pop. For this purpose, we apply classification trees, random forests and gradient boosting, and we compare their accuracy. We then derive the best predictors of each musical genre among encoded audio features [5, 6].

The paper proceeds as follows. Section 2 presents the sample used and some descriptive analyses. Section 3 reports statistical analyses in detail, with a focus on performance comparison among the methods used and results interpretation. Section 4 finally summarizes the main findings.

2 Application

2.1 Data exploration

The first part of the work focuses on data collection by R package *spotifyr* [7]. Using Spotify's Web API, 33,042 tracks were randomly selected from the Spotify catalogue and the relevant catalogue information, like track name, popularity, genre, ... Their audio features were extracted for each of them. The ones we consider are the following [8]:

- *danceability* describes how suitable a track is for dancing;
- *energy* represents a perceptual measure of intensity and activity;
- *loudness* is the quality of a sound that is the primary psychological correlate of physical strength;
- *speechiness* detects the presence of spoken words in a track;
- *acousticness* is a confidence measure of whether the track is acoustic;
- *instrumentalness* predicts whether a track contains no vocals;
- *liveness* detects the presence of an audience in the recording;
- *valence* describes the musical positiveness conveyed by a track.
- *duration_ms* is the duration of the track in milliseconds.

2.2 Descriptive analysis

In the second part of the work, a descriptive analysis of the dataset was carried out with respect to musical genre. Our aim is to predict the six genres (rock, rap, EDM, R&B, latin and pop) from the audio features of musical tracks. Prior to this, we analyse audio features with respect to musical genres to display the variables that provide greater separation among genres (see Figure 1). A first screening of Figure 1 shows that in general the most discriminating variables are valence, danceability, energy and tempo. Instead, features such as instrumentalness and key do not contribute much to the classification by musical genre.

In more detail, we can note in Figure 1 that EDM tracks show very low values of acousticness, high energy and tempo, but among the genres, it is the one with the lowest values of valence. EDM tracks are found to sound more

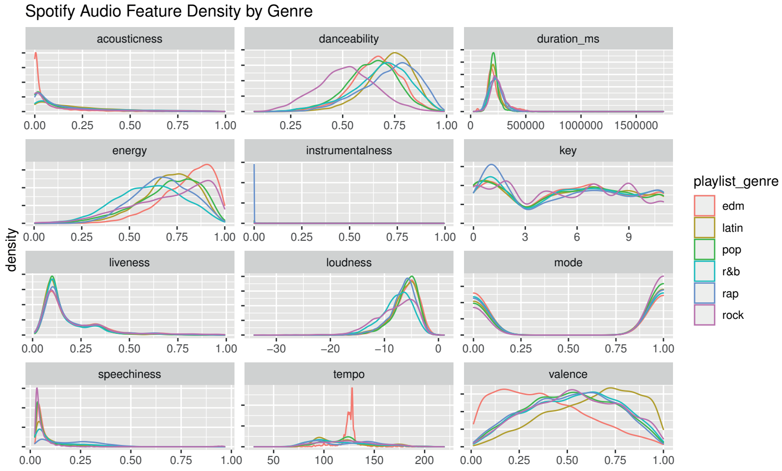


Fig. 1. Empirical densities of Spotify audio features by musical genre.

negative to the listener, transmitting sadness, anger or depression. Latin tracks, on the other hand, have the highest valence values and are generally suitable tracks for dancing, along with rap tracks. The rock genre differs from the others in that it has lower values of danceability, but still achieves fairly high values of energy. The R&B genre maintains values similar to pop genre except for energy, for which it reaches lower values compared to the other categories. Pop, latin and EDM present on average shorter tracks than the remaining genres.

We exclude 79 tracks with a too long duration, thus restricting the sample to 32,963 songs. The relationships among audio features are then studied. Correlation analsyis shows levels close to 0 except for a few feature pairs: the highest correlation is recorded between energy and loudness (0.66). This result is not surprising, since among the perceptual characteristics that contribute to determine the level of energy in a track, in addition to dynamic range, timbre, onset rate and general entropy, there is perceived loudness [8]. Given the high level of correlation, it was decided to remove loudness: the choice was guided by the fact that the density plot in Figure 1 indicates that energy appears more important in discriminating genres.

3 Genre classification

Once analysed the dataset, we try to successfully classify music genres using audio features. A training set consisting of 80% of the observations was randomly sampled for this task, while the test set consists of the remaining observations, i.e. 20% of the initial dataset. Three supervised learning algorithms were implemented: decision tree [9], random forest [10, 11], and gradient boosting [12]. In particular, decision trees have been implemented by *rpart* R package, random forest by *randomForest* R package, and boosting by the *xgboost* algorithm of [13] implemented in the *xgboost* R package.

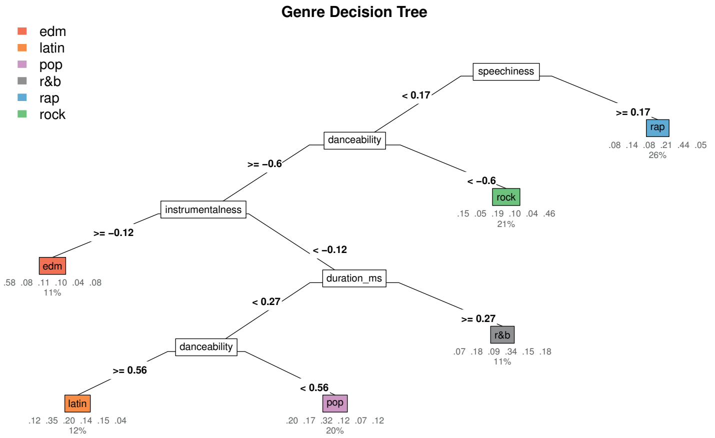


Fig. 2. Optimally pruned genre decision tree. Below each leaf, we report the distribution of true genre labels (in the order, EDM, latin, pop, R&B, rap, and rock) among the songs assigned to that leaf, and the proportion of songs assigned to that leaf.

Decision tree operates by iteratively partitioning the songs into subsets (called leaves) which are the most internally homogeneous possible. The optimally pruned tree is reported in Figure 2. Looking at that, we can note that speechiness is the most important feature by classification power, followed by danceability, instrumentalness and duration. In particular, the feature speechiness makes an initial separation of rap from other classes: tracks with values greater than or equal to 0.17 fall within the rap genre (25% of the total). Concerning the remaining observations, we observe that tracks with low danceability are classified as rock (21% of total); with high instrumentalness are classified as EDM (11%); with high duration are classified as R&B (11%); with high danceability are classified as latin (12%); with low danceability are classified as pop (20%). The algorithm managed to correctly classify 44% of rap songs, 46% of rock songs and 58% of EDM songs, thus achieving fairly high percentages. For the remaining classes, however, the percentages were lower: 34% for R&B tracks, 32% for pop tracks and 35% for latin tracks. Overall, the model achieved an accuracy of 40.67%.

In order to improve prediction accuracy, we decided to employ two ensemble learning algorithms: random forests and gradient boosting. Random forests are bagging algorithms, where many weak learners are trained on as many bootstrap sets (see Chapter 15 in [14]). In this case, 100 decision trees were trained on as many bootstrap sets generated from the initial training set and then the obtained predictions were combined into a single classifier. The random forest model structured in this way shows an increased level of accuracy compared to decision tree: 56.10% of tracks were classified in the correct genre.

Gradient boosting falls into the category of boosting techniques, in which a single weak learner is trained and a new one is added at each iteration (see Chapter 10 in [14]). The added learners consider the errors of the previous learners to improve prediction accuracy. This was performed by *eXtreme Gradient Boosting* (*xgboost*): it is a scalable machine learning system for tree boosting, designed to be highly efficient and flexible. It provides a parallel tree boosting, a regularised model that avoids overfitting. In the present case, using this type of algorithm, a classification accuracy level of 58.18% was attained.

When shifting from decision tree to random forest and gradient boosting, we observe an increase in classification accuracy which is differentiated by genre (see Figure 3): higher increases are reached for rock, rap and EDM (accuracy more than doubled), indicating that songs falling into these classes have characteristics which make them easier to discriminate; on the contrary, genres such as R&B, pop and latin are more difficult to classify: in fact, even by the *xgboost* classifier, which turned out to be the most accurate method, they do not even reach a 50% accuracy.

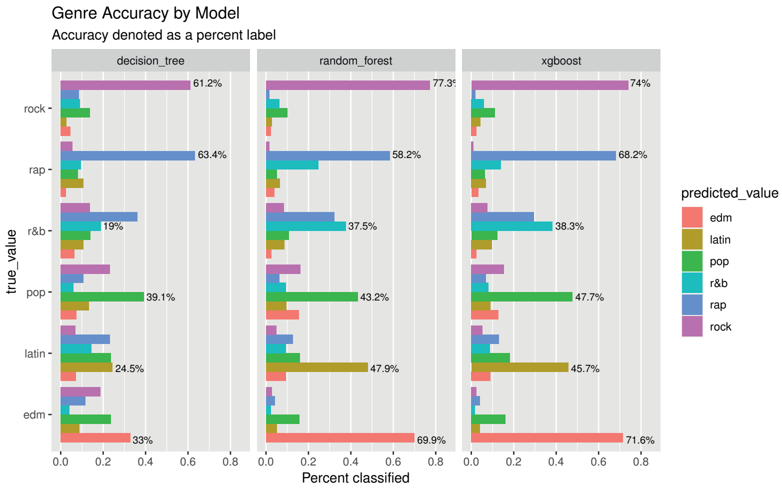


Fig. 3. Genre prediction accuracy for each of the three statistical learning methods.

Analysing the optimal decision tree in Figure 2, speechiness clearly was the most important variable in terms of explanation power. In the end, the degree of importance of each audio feature for classifying genres was studied for all the three final outputs produced. It turns out that the most important feature is speechiness also in the random forest. In *xgboost* the most important variable is instead duration. In general, the features with some usefulness for the purposes of genre classification of music tracks are speechiness, danceability, tempo, instrumentalness, duration, valence and energy. Features such as key and mode do not instead contribute substantially to classification.

4 Conclusions

Three machine learning methods have been employed to classify the musical genres of a large sample of songs drawn by Spotify: decision tree, random forest and machine learning. We empirically verified that the latter two methods clearly outperform the first one, we have learnt that speechiness is the audio feature which has the largest discriminatory power in terms of genre classification, followed by danceability, and we have observed that rock and rap songs are definitely the easiest to be classified.

References

1. Russell, S. Artificial intelligence a modern approach. (Pearson Education, Inc., 2010).
2. Eriksson, M., Fleischer, R., Johansson, A., Snickars, P. & Vonderau, P. Spotify teardown: Inside the black box of streaming music. (Mit Press, 2019).
3. Hulaud, S. Identification of taste attributes from an audio signal. (Google Patents, 2018), US Patent 9, 934,785.
4. How Spotify Uses ML to Create the Future of Personalization, https://engineering.atspotify.com/2021/12/how-spotify-uses-ml-to-create-the-future-of-personalization/, Accessed: 2023-01-27.
5. Understanding + classifying genres using Spotify audio features, https://www.kaylinpavlik.com/classifying-songs-genres/, Accessed: 2023-01-27.
6. Spotify genre classification, https://github.com/walkerkq/spotify_genre_classification, Accessed: 2023-01-27.
7. Thompson, C., Parry, J., Phipps, D. & Wolff, T. spotifyr: R Wrapper for the' Spotify' Web API. R package version 2.1. 1. (2019).
8. Spotify for Developers, https://developer.spotify.com/documentation/web-api/, Accessed: 2023-01-27.
9. Breiman, L., Friedman, J., Olshen, R. & Stone, C. Classification and regression trees. (Routledge, 2017).
10. Ho, T. Random decision forests. *Proceedings Of 3rd International Conference On Document Analysis And Recognition.* **1** pp. 278-282 (1995).
11. Breiman, L. Random forests. *Machine Learning.* **45**, 5-32 (2001).
12. Breiman, L. Arcing classifier (with discussion and a rejoinder by the author). *The Annals Of Statistics.* **26**, 801-849 (1998).
13. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining.* pp. 785-794 (2016).
14. Hastie, T., Tibshirani, R., Friedman, J. & Friedman, J. The elements of statistical learning: data mining, inference, and prediction. (Springer,2009)

The use of Bradley-Terry comparisons in statistical and machine learning models to predict football results

Roberto Macrì Demartino, Nicola Torelli and Leonardo Egidi

Abstract Over the last few years, there has been a growing interest in the prediction and modeling of competitive sports outcomes, with particular emphasis placed on this area by the Bayesian statistics and machine learning communities. In this paper, we have carried out a comparative evaluation of statistical and machine learning models to assess their predictive performance by evaluating alternative summaries of past performances of the involved teams. More specifically we consider the (Bayesian) Bradley-Terry model, which is a widely used statistical framework for ranking items based on paired comparisons that have been applied successfully in various domains, including football. The analysis was performed including in some canonical goal-based models both the Bradley-Terry-derived ranking and the widely recognized FIFA ranking commonly adopted by football fans and amateurs.

1 Introduction

The development of statistical and machine learning models to forecast the outcomes of international football competitions, such as FIFA World Cups or UEFA Champion's Leagues, has always attracted the interest of several statisticians and data scientists. From a statistical viewpoint, the outcome of a football match may be predicted using two different approaches. The *goal-based* approach entails modelling the number of goals scored and conceded by the teams competing in each match

---

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

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

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

using an appropriate count distribution. The underlying distribution for goal scoring is a double-Poisson distribution (Maher, 1982; Baio and Blangiardo, 2010; Egidi et al, 2018), with the expected goals depending on team-specific strengths, such as the attack and defense abilities. However, some bivariate-Poisson models designed to capture the goals' correlation have been developed throughout the years (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003).

The *result-based* approach consists of explicitly modelling the three-way process: home win, draw, away win. An ordered probit model was proposed by Koning (2000) for football results in the Netherlands, whereas Carpita et al (2019) explored and modelled the European Soccer database using a logit regression model. The increasing popularity of huge volumes of data encouraged the development of fundamentally different Machine Learning approaches such as Classification and Regression trees (CART) and ensemble learning methods such as the Random Forests (Breiman, 2001). Schauberger and Groll (2018) examined and evaluated the predictive performance of several Random Forests configurations in the context of international football matches

This paper aims to compare several statistical models and some result-based Machine Learning algorithms to investigate prediction performance for future matches using the outcomes of the FIFA World Cup 2022 held in Qatar. Typically, an additional source of information is considered using the FIFA ranking as a further model covariate. However, the aforementioned ranking does not take into account certain characteristics and has often been accused of merely being a political indicator. As a consequence, a modification of the Bradley-Terry model (Bradley and Terry, 1952) introduced by Leonard (1977) can be used to determine a novel world team ranking, and the latter could then be included in the above models as a further predictor. To assess the global predictive efficacy, we compared the inclusion of the two alternative rankings in the models for both the group stage and knockout stage from the FIFA World Cup 2022. Although not conclusive about the choice of a final world team ranking, we strongly believe this study could help discriminate between alternative ranking formulations and, possibly, drive the use of better ranking measures.

2 Statistical modelling and Machine Learning framework

We assess the predictive performance between Poisson statistical models and five Machine Learning procedures such as Random Forest, Classification and Regression Trees (CART), Bagged CART, Multivariate Adaptive Regression Splines (MARS), and Neural Networks. We consider the inclusion of two Word Cup rankings, one based on the predictions of a Bayesian Bradley-Terry model (Leonard, 1977), and the other one based on the usual FIFA rankings, available at `https://www.fifa.com/fifa-world-ranking/men`.

2.1 Goal-Based approach

Goal-based models assume that the number of goals scored in a game by each team follows a discrete distribution, typically two independent Poisson or a bivariate-Poisson accounting for positive correlation. Thus, for each game, we need to consider the pair of counts (X_{in}, Y_{jn}), for $i \neq j = 1, \ldots, N_T$ and $n = 1, \ldots, N$. The first count X_{in} denotes the non-negative number of goals scored by the home team i and the second count Y_{jn} denotes the number of goals scored by the visiting team j, both in the n-th game. A general bivariate Poisson model for each pair of counts is

$$\begin{aligned} \left(X_{in}, Y_{jn} \mid \lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right) &\sim \text{BivPoisson}\left(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right) \\ \log\left(\lambda_{1n}\right) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2}\omega_n, \\ \log\left(\lambda_{2n}\right) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2}\omega_n, \\ \log\left(\lambda_{3n}\right) &= \beta_0, \end{aligned} \tag{1}$$

where λ_{1n} and λ_{2n} describe the scoring rate for the home team and the away team, respectively. In particular, θ denotes a common baseline parameter; the parameters *att* and *def* denote the unknown attack and defense abilities for the home team h_n and the away team a_n in the n-th game. The coefficient λ_{3n} describes the dependence between the two random counts. When $\lambda_{3n} = 0$, the two components are independent, then the bivariate Poisson model reduces to a double-Poisson model. Furthermore, $\omega_n = (rank_{h_n} - rank_{a_n})$ is the difference between the chosen World Rankings of the home and the away team in the n-th game. Some sum-to-zero constraints for the attack and defense parameters' abilities is introduced for identifiability.

2.2 The Bradley-Terry model

The Bradley-Terry model (Bradley and Terry, 1952) is typically applied to rank teams or players in a pairwise comparison context. Let consider a match between team T_i and team T_j, with $i \neq j = 1, \ldots, N_T$, we can assume that the winning probability for T_i over T_j in the n-th match, with $n = 1, \ldots, N$, is

$$\mathbb{P}(T_{in} \text{ beats } T_{jn}) = \frac{\exp(\alpha_i - \alpha_j)}{1 + \exp(\alpha_i - \alpha_j)}, \tag{2}$$

where each team's ability is represented by parameters denoted by α_i and α_j. Parameter identifiability is obtained by imposing a constraint such as $\sum_{k=1}^{N_T} \alpha_k = 0$. In case of multiple matches between the same two teams, the results are assumed independent.

However, the classical Bradley-Terry model above does not account for a three-way process since the ties are not allowed. Davidson (1970) extended the previous model by including a novel parameter, ν, which affects the probability of a draw. When $\nu = 0$, we move back to a standard Bradley-Terry model. Leonard (1977) proposed a Bayesian formulation of the Bradley–Terry model using non-conjugated prior distributions. The advantages include greater flexibility to extend hierarchical models, the ability to specify the prior parameters, and the ability to specify alternative prior distribution families. The ability parameters estimates can be then used to create a novel team's ranking.

3 An application to the prediction of World Cup 2022 results

The selection of the training and test sets is crucial and is likely to influence the predictions. As a consequence, we opt to train our statistical models and machine learning algorithms on a starting training set that is iteratively updated including the incoming outcomes during the World Cup for two different prediction scenarios, the group stage and the knockout stage.

We consider the results from more than 3000 international matches played between the years 2018 and 2022 as the initial training set. The matches vary from the FIFA World Cup through the UEFA Euro Championship to normal friendly matches. The data excludes Olympic Games and matches in which at least one of the teams was the national B-team or a U-23 lineup. The aforementioned starting training set is available on Kaggle at the following link `https://t.ly/87nP`.

3.1 Predictive performance

We compare the predictive accuracy between some Poisson-based models, fitted using the `footBayes` package (Egidi and Palaskas., 2022), and five Machine Learning techniques outlined in Section 2, provided by the `caret` package (Kuhn, 2022).

Table 1 shows the accuracy of the predictions for the considered methods in all of the prediction scenarios. The Diagonal Inflated Bivariate Poisson models, allowing for draw inflation, perform quite well in the group-stage scenario, whereas most of the Machine Learning algorithms over-perform the statistical models in the knockout-stage scenario. Furthermore, Machine Learning models using the Bradley-Terry rank tend to perform generally better than those with the traditional FIFA rankings in the first scenario.

Table 2 shows the Brier Score (Brier, 1950) for the considered models in both the prediction scenarios. Despite the results presented in Table 1, the Poisson models employed with the Bradley-Terry rank show a better accuracy as compared to the models that used the FIFA ranking in the first scenario. However, in the knockout scenario, the models with the FIFA ranking were more accurate. It can be supposed that models based on the Bradley-Terry ranking exhibit optimal performance in scenarios where there is a higher degree of heterogeneity between team strengths.

	group-stage		knockout-stage	
	BT Rank	FIFA Rank	BT Rank	FIFA Rank
Diag. Infl. Biv. Pois.	0.521	0.562	0.500	0.562
Biv. Pois.	0.458	0.479	0.500	0.562
Double Pois.	0.458	0.479	0.562	0.562
Random Forest	0.500	0.333	0.687	0.687
Bagged CART	0.521	0.375	0.625	0.750
CART	0.396	0.396	0.437	0.625
MARS	0.458	0.396	0.562	0.625
NN	0.500	0.396	0.625	0.625

Table 1: Prediction accuracy for the selected methods: two prediction scenarios.

	group-stage		knockout-stage	
	BT Rank	FIFA Rank	BT Rank	FIFA Rank
Diag. Infl. Biv. Pois.	0.606	0.621	0.542	0.523
Biv. Pois.	0.616	0.630	0.530	0.497
Double Pois.	0.623	0.631	0.521	0.493
Random Forest	0.748	0.867	0.497	0.466
Bagged CART	0.772	0.851	0.472	0.480
CART	0.682	0.663	0.671	0.586
MARS	0.634	0.663	0.561	0.586
NN	0.618	0.664	0.549	0.585

Table 2: Brier Score for the selected methods: two prediction scenarios.

4 Discussion

The Bradley-Terry model and its related approaches have been demonstrated to be effective in ranking football teams. However, there is potential to further develop these models to enhance their accuracy and capabilities in predicting the outcomes of football games. The development of more advanced techniques, such as incorporating additional variables or a dynamic methodology that accounts for changes in teams' strength over time, could improve the model's performance and lead to more accurate predictions. Moreover, the investigation of the alternative use of Bradley-Terry-derived rankings in place of FIFA or similar rankings remains open.

References

Baio G, Blangiardo M (2010) Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics 37(2):253–264, DOI 10.1080/02664760802684177

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3/4):324–345

Breiman L (2001) Random forests. Machine Learning 45(1):5–32, DOI 10.1023/A:1010933404324

Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthey Weather Review, 78(1):1–3

Carpita M, Ciavolino E, Pasca P (2019) Exploring and modelling team performances of the kaggle european soccer database. Statistical Modelling 19(1):74–101, DOI 10.1177/1471082X18810971

Davidson RR (1970) On extending the bradley-terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association 65(329):317–328

Dixon MJ, Coles SG (1997) Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46(2):265–280, DOI https://doi.org/10.1111/1467-9876.00065

Egidi L, Palaskas V (2022) footBayes: Fitting Bayesian and MLE Football Models. URL `https://github.com/leoegidi/footbayes`, r package version 0.2.0

Egidi L, Pauli F, Torelli N (2018) Combining historical data and bookmakers' odds in modelling football scores. Statistical Modelling 18(5-6):436–459, DOI 10.1177/1471082X18798414

Karlis D, Ntzoufras I (2003) Analysis of sports data by using bivariate poisson models. Journal of the Royal Statistical Society: Series D (The Statistician) 52(3):381–393, DOI https://doi.org/10.1111/1467-9884.00366

Koning RH (2000) Balance in competition in dutch soccer. Journal of the Royal Statistical Society: Series D (The Statistician) 49(3):419–431, DOI https://doi.org/10.1111/1467-9884.00244

Kuhn M (2022) caret: Classification and Regression Training. URL `https://CRAN.R-project.org/package=caret`, r package version 6.0-93

Leonard T (1977) An alternative bayesian approach to the bradley-terry model for paired comparisons. Biometrics 33(1):121–132

Maher MJ (1982) Modelling association football scores. Statistica Neerlandica 36(3):109–118, DOI https://doi.org/10.1111/j.1467-9574.1982.tb00782.x

Schauberger G, Groll A (2018) Predicting matches in international football tournaments with random forests. Statistical Modelling 18(5-6):460–482, DOI 10.1177/1471082X18799934

A new approach for quantum phase estimation based algorithms for machine learning

Oumayma Ouedrhiri, Oumayma Banouar, Salah El Hadaj and Said Raghay

Abstract One of the greatest developments in computer science is undoubtedly quantum computing. It has demonstrated to give various benefits over the classical algorithms, particularly in the significant reduction of processing time, because to the parallelism and entanglement properties. One of the most crucial quantum computing algorithms is quantum phase estimation (QPE). It is called the eigenvalue finding module for unitary operators. The key to this process is the Fourier transform. Numerous issues, including the order finding problem and the factoring problem, have been studied and solved using it. It was also used to calculate the eigenvalues of unitary matrices and quantum sampling methods. In this paper, we study and discuss recent improved versions for the QPE procedure, their advantages and experimentation. We also propose a new approach for QPE based algorithms for machine learning (ML). These algorithms are the HarrowHassidim-Lloyd (HHL) algorithm for solving linear systems, the quantum singular value thresholding (QSVT) algorithm for matrix completion in recommender systems, and the quantum principal components analysis (QPCA) for data visualization.

Key words: Artificial intelligence, Machine learning, Quantum computing, Quantum machine learning, Quantum phase estimation.

Oumayma Ouedrhiri
Oumayma Ouedrhiri, Faculty of sciences and techniques Marrakech, e-mail: oumayma.ouedrhiri@ced.uca.ma

Oumayma Banouar
Oumayma Banouar, Faculty of sciences and techniques Marrakech, e-mail: o.banouar@uca.ac.ma

Salah El Hadaj
Salah El Hadaj, Faculty of sciences and techniques Marrakech, e-mail: elhadajs@yahoo.fr

Said Raghay
Said Raghay, Faculty of sciences and techniques Marrakech, e-mail: s.raghay@uca.ma

Introduction

One of the essential quantum computing [1] subroutines is quantum phase estimation (QPE). For many quantum machine learning (QML) algorithms [8], it serves as a fundamental building block [12]. Its main objective is to find the eigenvalues of a unitary matrix with an unchanging eigenvector (QPE). This routine plays a crucial role in the creation of quantum algorithms across various disciplines.

The QPE procedure is considered a ”subroutine” that contributes to the completion of complex computational tasks when combined with other subroutines. As an example, the QPCA algorithm [7] serves to reveal properties of a quantum density matrix and enable dimensionality reduction. The HHL algorithm [5, 4] also utilizes the QPE subroutine to solve linear systems and inpired many QML algorithms. The QSVT algorithm [3] uses the QPE algorithm along with a rotation module for matrix completion.

In this paper, we explain the procedure of the QPE algorithm, along with mathematical details. Then, we study and discuss some recent proposed improvements for the QPE algorithm. Indeed, the advantages of each method are depicted as well as its experimentation. Finally, a new approach for the QPCA, the HHL and the QSVT algorithms is proposed.

1 Quantum phase estimation

The objective of the phase estimation process [11] is to determine the phase θ for a unitary operator U that holds an eigenvector $|\psi\rangle$ and its corresponding eigenvalue $e^{2\pi i\theta}$. The QPE algorithm employs two registers: one to store the eigenstate $|\psi\rangle$ and the necessary number of qubits for representation, and another to store t qubits initialized to $|0\rangle$. The number of counting qubits t is dependent on the desired level of accuracy in the estimation of θ and the success rate required from the method. The whole process of the algorithm can be described by the following:

$$
\begin{aligned}
|0\rangle^{\otimes n} &\longmapsto \frac{1}{2^{n/2}}(|0\rangle+|1\rangle)\otimes(|0\rangle+|1\rangle)\otimes\cdots\otimes(|0\rangle+|1\rangle) \;\; \text{(after the Hadamards)} \\
&\longmapsto \frac{1}{2^{n/2}}\left(|0\rangle+e^{2\pi i 0\cdot\phi_1\phi_2\cdots\phi_n}|1\rangle\right)\otimes\left(|0\rangle+e^{2\pi i 0\cdot\phi_2\phi_3\cdots\phi_n}|1\rangle\right)\otimes\cdots \\
&\otimes\left(|0\rangle+e^{2\pi i 0.\phi_n}|1\rangle\right) \text{(after the CUs)} \\
&\longmapsto |\phi_n\rangle\otimes|\phi_{n-1}\rangle\otimes\cdots\otimes|\phi_1\rangle \;\; \text{(after the QFT †)}
\end{aligned} \tag{1}
$$

We can define the phase and then the eigenvalue by measuring the end state. Below is the quantum circuit for phase estimation [14]. Qubits in the state $|\psi\rangle$ are found at the bottom of the register, and qubits in the t counting qubits are found at the top.

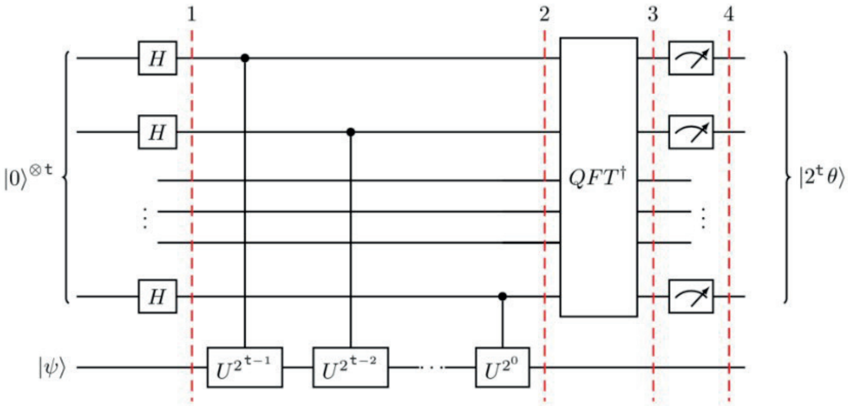


Fig. 1 Quantum Phase Estimation circuit

2 Proposed improvements for Quantum Phase Estimation

The QPE algorithm has a lot of limitations. The circuit's input matrix must be Hermitian, unitary, and 2n-dimensional. It should also be normalized to make the unitary matrix. And choosing the amount of counting qubits affects the accuracy of the circuit. Also, it can be difficult to retrieve information from the circuit's output. And finally, reducing depth and controlled-rotation gates is essential to improving the accuracy of phase finding on quantum computers.

Recently, many approaches were proposed [9, 2, 10, 13, 6, 15] to improve the QPE algorithm and adress its limits.

Table 1: Comparative study of the mentioned proposed QPE algorithm improvements

Date	Proposed method	Experimentation	Advantages
2019	arbitrary precision QPE	IBM Q (Python)	• Improved accuracy • Reduced circuit depth • Less noise • Doesn't require use of ancillary qubits
2021	QPE with squeezed quazi-Bell states	No experimentation	• Increased accuracy • Advantageous under non-ideal conditions (perturbation, dissipation)

2021	SPEA	IBM Q (Python)	• Hardware requirement reduction • Can be used in a variational approach • Considers probabilistic outputs
2021	QPE with GSSs	No experimentation	• Exponentially faster than the QPE algorithm • Using noise-resilient quantum states • The phase estimation at each step of the algorithm based on a single measurement
2022	QPESIM	C++ environment	• Allow for scalable simulations of previously inaccessible atomic and molecular systems • Take advantage of modest computational resources • Implementation allowed for the use of significantly larger active spaces than previously possible • Reduction of error in correlation energy obtained in simulations
2022	two-step protocol QPE	No experimentation	• Lower error when compared to the QPE algorithm • The method's circuits are shallower • Independent of the failure probability • Do not require multi-qubit entanglement

3 A new approach for QPE based algorithms for machine learning

Enhancing the QPE subroutine can directly help to improve the QPCA, HHL and QSVT algorithms. The idea is to look for the QPE algorithm that offers better accuracy while requiring minimal hardware, and can be implemented on an accessible quantum simulator. The SPEA algorithm is the best, since It utilizes a variational approach and can determine any unknown eigenphase-eigenstate pair, does not require any new quantum hardware and can be implemented on IBM Q platform.

Adopting the SPEA protocol enhances the accuracy of the QPCA algorithm and allows to extract the unknown eigenphases-eigenstates pairs without requiring significant hardware development. The improved algorithm will be as follows:

Algorithm 1 The improved QPCA algorithm using SPEA

```
Input: covariance matrix of the dataset $\Sigma$ ;
Encode $\Sigma$ in a quantum density matrix $\rho$;
Perform the SPEA procedure to the density matrix;
Output: Extract the output eigenvalues
```

Also, improving the QPE algorithm and using SPEA instead can lead to a better HHL algorithm in terms of accuracy and hardware requirements. The HHL process is described in the following algorithm : Finally, using the SPEA module will help the QSVT algorithm to have better outcomes. The proposed improved QSVT version will go as the following.

Algorithm 2 The improved HHL algorithm using SPEA

```
Input: matrix $A$ and vector $\vec{b}$ ;
State preparation: Load the state $|b\rangle \in \mathbb{C}^N$ ;
Perform the SPEA procedure to $U = e^{iAt}$ to estimate the eigenvalues of $A$, the input matrix associated to the system we aim to solve;
Eigenvalue inversion : Add an ancilla qubit and perform a rotation conditioned on the eigenvalues extracted from the QPE block;
Apply the reverse phase estimation algorithm SPEA†;
Measure the ancilla qubit in the computational basis to get the $x$ vector's entries.
```

Algorithm 3 The improved QSVT algorithm using SPEA

```
Input: A quantum state $|\psi_{\mathbf{A}_0}\rangle$, a unitary $e^{i\mathbf{A}t_0}$, and a constant $\tau$ ;
Quantum registers preparation in the state $|\psi_1\rangle = |0\rangle|0\rangle^L|0\rangle^C|\psi_{\mathbf{A}_0}\rangle^B$,
Perform the SPEA procedure to the matrix $A$ on the state $|0\rangle^C|\psi_{\mathbf{A}_0}\rangle^B$.
Apply the unitary $\mathbf{U}_{\sigma,\tau}$ to convert the eigenvalues $|\sigma_k^2\rangle$ to the intermediate result $|y_k\rangle$, with $y_k = (1-\tau/\sigma_k)_+ \in [0,1)$;
Apply rotations $R_y$ to retrieve the value of $y_k$;
Uncompute: apply $\mathbf{U}^\dagger$ which represents the inverse of all the operations before of $\mathbf{R}_y$;
Measure the top ancilla qubit to get the output state $|\psi_S\rangle$
```

Conclusion

In this paper, we explained the computational and mathematical details of the QPE algorithm. Afterwards, a comparison between different approaches to improve the algorithm was conducted, enabling us to select the most appropriate version. This was then integrated into the process of the QPE-based algorithms instead of using the conventional method. The proposed methods hold the potential for improved performance and efficiency. However, it is necessary to test them on available quantum simulators or even actual quantum devices and experiment on practical issues such as recommender systems and image processing problems. This will show how quantum computing can enhance machine learning algorithms and contribute to the advancement of artificial intelligence applications.

References

1. Bera, R.K.: The Amazing World of Quantum Computing. Springer Nature Singapore Pte Ltd, Singapore (2020).
2. Delgado de Souza, D. and Vidiella-Barranco, A. :Quantum phase estimation with squeezed quasi-Bell states, Optik, Vol 244, 167532, (2021).
3. Duan, B., Yuan, J., Liu, Y. and Li, D. : Efficient quantum circuit for singular value thresholding, Phys. Rev. A, 98 (1), (2018).
4. Duan, B., Yuan, J., Yu, C.H., Huang, J. and Hsieh, C.Y. : A survey on HHL algorithm: From theory to application in quantum machine learning. Physics Letters A 384, (2020).
5. Harrow,A.W., Hassidim, A. and Lloyd, S.: Quantum algorithm for solving linear systems of equations, Physical Review Letters. 103 (15): 150502 (2008).
6. Kang, C., Bauman, N., Krishnamoorthy, S. and Kowalski, K. : Optimized Quantum Phase Estimation for Simulating Electronic States in Various Energy Regimes. arXiv:2206.00802. (2022).
7. Lloyd, S., Mohseni, M. and Rebentrost, P. : Quantum principal component analysis. Nature Phys 10, 631–633 (2014).
8. Mishra, N. and al: Quantum Machine Learning: A Review and Current Status, Data Management, Analytics and Innovation pp 101-145, (2020).
9. Mohammadbagherpoor, H., Oh, Y., Dreher, P., Singh, A., Yu, X. and Rindos, A. : An Improved Implementation Approach for Quantum Phase Estimation on Quantum Computers. IEEE International Conference on Rebooting Computing (ICRC), 1-9. (2019).
10. Moore, A., Wang, Y., Hu, Z., Kais, S. and Weiner, A. : Statistical approach to quantum phase estimation. New Journal of Physics. 23. (2021).
11. Nielsen, M.A. and Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, New York, USA (2000).
12. Ouedrhiri, O., Banouar, O., Raghay, S. and Elhadaj, S.: Quantum phase estimation based algorithms for machine learning, 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6, (2021).
13. Pezzè, L. and Smerzi, A. : Quantum Phase Estimation Algorithm with Gaussian Spin States, PRX QUANTUM 2, 040301, (2021).
14. Quantum phase estimation in qiskit: https://qiskit.org/textbook/ch-algorithms/quantum-phase-estimation.html
15. Smith, J., Barnes, C. and Shukur, D. : An iterative quantum-phase-estimation protocol for near-term quantum hardware. (2022). arXiv:2206.06392v1.

A comparison of ensemble algorithms for item-weighted Label Ranking

Un confronto tra algoritmi pesati per la previsione di dati in classifica (Label Ranking)

Alessandro Albano, Mariangela Sciandra, Antonella Plaia

Abstract Label Ranking (LR) is a non-standard supervised classification method with the aim of ranking a finite collection of labels according to a set of predictor variables. Traditional LR models assume indifference among alternatives. However, misassigning the ranking position of a highly relevant label is frequently regarded as more severe than failing to predict a trivial label. Moreover, switching two similar alternatives should be considered less severe than switching two different ones. Therefore, efficient LR classifiers should be able to take into account the similarities and individual weights of the items to be ranked. The contribution of this paper is to formulate and compare flexible item-weighted Label Ranking algorithms using bagging, random forest, and boosting ensemble methods.

Abstract Il Label Ranking (LR) è una tecnica non standard di classificazione supervisionata che consiste nell'ordinare un insieme finito di alternative in base a un insieme di variabili esplicative. I modelli LR tradizionali considerano tutte le alternative come ugualmente importanti. Tuttavia, in molti casi, non riuscire a prevedere la corretta posizione in classifica di un'alternativa altamente rilevante può essere considerato più grave di non riuscire a prevederne una banale. Inoltre, scambiare due alternative simili dovrebbe essere considerato meno grave di scambiare di due alternative fortemente diverse. Pertanto, un classificatore LR efficace dovrebbe essere in grado di prendere in considerazione sia la somiglianza tra gli elementi che i loro pesi individuali. Il contributo di questo articolo consiste nel formulare degli algoritmi di Label Ranking pesati per l'importanza delle alternative da ordinare, che fanno uso di noti modelli di ensemble di alberi decisionali, rispettivamente: bagging, random forest e boosting.

Key words: Ensemble Methods, Label Ranking, Boosting, Bagging, Random Forest

---

Alessandro Albano, Mariangela Sciandra, Antonella Plaia
Department of Economics, Business and Statistics, University of Palermo, e-mail: alessandro.albano@unipa.it, mariangela.sciandra@unipa.it, antonella.plaia@unipa.it

1 Introduction

When a group of n people express their preferences on a limited set of objects (m alternatives), preference data are generated. When alternatives are placed in any predetermined sequence and preferences are described by using integers to indicate the rank of each alternative, preference data are called rankings.
In many real-world cases, the ranking responses are paired with additional features that characterize the judges, e.g., socioeconomic and socio-demographic characteristics. In these cases, the goal is to learn a function that predicts the ranking responses of new instances based on the set of covariates. In this paper, we undertake a weighted distance-based approach to Label Ranking, employing three ensemble algorithms. This paper extends the work of Albano et al. (2022), where they propose an item-weighted boosting algorithm for Label Ranking. Bagging and Random Forest procedures are used for the same task here, and they are compared to Boosting methods. The paper is organized as follows: Section 2 introduces item-weighted distance and consensus measures for ranking data. The three ensemble algorithms are presented in Section 3, and compared employing a simulation study in Section 4. Finally, the conclusions are presented.

2 Ranking data: an item-weighted approach

Formally, given the finite set of alternatives (or class labels) $Y = \{y_1, \ldots, y_m\}$, the ranking π is a mapping from Y to the set of ranks $\{1, \ldots, m\}$, endowed with the natural ordering of integers; $\pi = \{P_\pi(y_1), \ldots, P_\pi(y_i), \ldots, P_\pi(y_m)\}$, where $P_\pi : Y \longrightarrow \{1, \ldots, m\}$ is the rank of each alternative, being 1 for the alternative ranked first, 2 for the alternative ranked second, and so on.
Measuring the gap between rankings using dissimilarity or distance measures is an intriguing problem in the context of preference data. Distances between ranks generally regard all items as equally important and do not take the point of disagreement into account. Albano and Plaia (2021) proposed an item-weighted version of the Kemeny distance (Kemeny, 1959) by considering a weighting vector $w = (w_1, w_2, ..., w_m)$, where $w_i \geq 0$ is the importance given to the i-th item in a ranking. The item-weighted Kemeny distance $d_{K,e}$ between two m-size rankings, π_1 and π_2 is:

$$d_{K,e}(\pi_1, \pi_2) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j \left| a_{ij} - b_{ij} \right|, \tag{1}$$

where a_{ij} and b_{ij} are the generic elements of the score matrices defined by Emond and Mason (2002). The corresponding item-weighted rank correlation coefficient (defined as an extension of τ_x in Albano and Plaia (2021)) is:

$$\tau_{x,e}(\pi_1, \pi_2) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j a_{ij} b_{ij}}{Max[d_{K,e}]}, \tag{2}$$

where the denominator represents the maximum value of the weighted Kemeny distance $\text{Max}[\text{d}_{K,e}] = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j$.
In some cases, the weights could be determined following the principle that switching two items that are comparable in some way should carry a smaller penalty than switching two unrelated ones. In this setting, a symmetric penalization matrix **P**, reflecting the dissimilarity among the elements, is needed.
Therefore, Eqs (1) (2) are modified by replacing $w_i w_j$ with p_{ij}, where p_{ij} (≥ 0) is the generic element of the **P** matrix.
The definition of item-weighted measures between rankings allows us to find the weighted consensus ranking $\hat{\pi}$:

$$\hat{\pi} = \arg\max_{\pi \in S^m} \sum_{l=i}^{n} \tau_{x,e}(\pi^{(l)}, \pi),$$

where S^m is the universe of all rankings with m objects. The weighted consensus is considered the ranking with the highest level of agreement with the whole set of rankings.

3 Item-weighted ensemble methods for ranking data: Bagging, Boosting and Random Forest

The three algorithms presented in this section are based on three well-known ensemble methods: boosting, bagging and random forest. They all use decision trees as weak learners, extended to ranking data. To build a ranking decision tree, the root node, which contains all observations, is split into a nested sequence of subtrees. The process follows a splitting criterion consisting of maximising the impurity reduction at each step. To relax the assumption of indifference among alternatives, the impurity is based on the item-weighted Kemeny distance (Eq. 1).
The item-weighted boosting algorithm (Algorithm 1), introduced by Albano et al. (2022), repeatedly applies decision trees to weighted versions of the n-size training sample T. It updates the inclusion probability of each instance so that previously misclassified instances receive an higher change to be sampled .
In this research we compare it with two new algorithms: Algorithm 2 (the item-weighted bagging algorithm), which starts from the training set T of size n, and generates B new training sets size n_b, by sampling from T uniformly and with replacement. A decision tree is fit at each replicate of the dataset, and the results are aggregated in the final step; and, Algorithm 3 (the item-weighted random forest algorithm), that modifies algorithm 2, by considering only a small but consistent number of unique features in the splitting process.

4 Experimental evaluation

The data simulation was carried out using a two-step method. Firstly, we generated the predictor space, $X = \{X_1, \ldots, X_{20}\}$, using the **R** package `MixSim` (Melnykov

Algorithm 1 AdaBoost.R - Item-weighted boosting for ranking data (Albano et al., 2022)

```
Input: A training set $T$, a number of iterations $B$, a vector of weights $w$
Output: a ranker $C_f(.)$ that maps a given $x$ to a ranking of the labels

initialize $p_b(i) = 1/n \ \forall i = 1, 2, ...n$
for $b \leftarrow 1$ to $B$ do
    take a sample $T_b$, drawn from the training set $T$ using weights $p_b(i)$
    fit a ranking tree $C_b(.)$ to $T_b$
    $e_b = \sum_{i \in T_b} p_b(i) \left(1 - \frac{\tau_{x,e}(i)+1}{2}\right)$ where $\tau_{x,e}(i) = \tau_{x,e}(C_b(x_i), y_i)$
    $\alpha_b = \frac{1}{2} \ln((1 - e_b)/e_b)$
    update the weights $p_{b+1}(i) = p_b(i) exp \left(\alpha_b \left(1 - \frac{\tau_{x,e}(i)+1}{2}\right)\right)$ and normalize them
end for
$C_f(x_i) = \text{arg max}_{y_i \in S^m} \sum_{b=1}^{B} \alpha_b \tau_{x,e}(C_b(x_i), y_i))$
```

Algorithm 2 Item-weighted Bagging ensemble algorithm

```
Input: A training set $T$, a number of iterations $B$, a vector of weights $w$
Output: a ranker $C_f(.)$ that maps a given $x$ to a ranking of the labels

for $b \leftarrow 1$ to $B$ do
    take a sample $T_b$, drawn from the training set $T$
    fit a ranking tree $C_b(.)$ to $T_b$
end for
$C_f(x_l) = \text{arg max}_{y_l \in S^m} \sum_{b=1}^{B} \tau_{x,e}(C_b(x_l), y_l))$
```

Algorithm 3 Item-weighted Random Forest algorithm

```
Input: A training set $T$, a number of iterations $B$, a vector of weights $w$
Output: a ranker $C_f(.)$ that maps a given $x$ to a ranking of the labels

for $b \leftarrow 1$ to $B$ do
    take a sample $T_b$, drawn from the training set $T$
    fit a ranking tree $C_b(.)$ to $T_b$ by randomly selecting a sample of the covariates for each split
end for
$C_f(x_l) = \text{arg max}_{y_l \in S^m} \sum_{b=1}^{B} \tau_{x,e}(C_b(x_l), y_l))$
```

et al., 2012). It was used to create clustered data given a finite mixture model with K Gaussian components. We generated $K = 5$ components, resulting in 5 clusters (Fig. 1). The total number of data-points is $N = 500$, while the mixing proportions (of length K) are uniform, $\zeta = (0.2, 0.2, 0.2, 0.2, 0.2)$.

In a second step, 4-dimensional ($m = 4$) ranking responses were generated within each cluster using the Mallows (1957) model, an exponential model defined by a central permutation π_0 and a dispersion parameter θ

$$Pr(\theta) = \frac{\exp(-\theta d_{K,e}(\pi, \pi_0))}{\psi(\theta)}. \tag{3}$$

When $\theta > 0$, π_0 represents the mode of the distribution, i.e., the permutation with the highest probability to be generated. In our case, each of the five clusters was assigned a different mode $\Pi_0 = \{\pi_0^{(1)} = (2,3,1,3), \pi_0^{(2)} = (3,1,2,3), \pi_0^{(3)} = (1,1,1,2), \pi_0^{(4)} = (1,1,2,2), \pi_0^{(5)} = (1,3,2,1)\}$. For each cluster, n_k rankings are generated.

We generated three scenarios by varying the dispersion parameter: ($\theta = 2$ low noise, $\theta = 0.7$ medium noise, $\theta = 0.4$ high noise). Finally, the weighting vector, used both for generating rankings (Eq. 3) and for running the label ranking algorithms is set to $\mathrm{w} = (2,5,5,2)$.

Fig. 1 reports the results of a Principal Component Analysis (PCA) applied to the predictor space. It should be noted that the item-weighted distances between the cluster modal rankings reflect the euclidean distances between the cluster centroids. The methods are experimentally evaluated through a five-fold cross validation procedure. That is, each dataset (three different scenarios based on θ) was randomly partitioned into five separate folds. Four folds were used as the training set in each branch, and the last fold (a different one for each branch) was used as the test set.

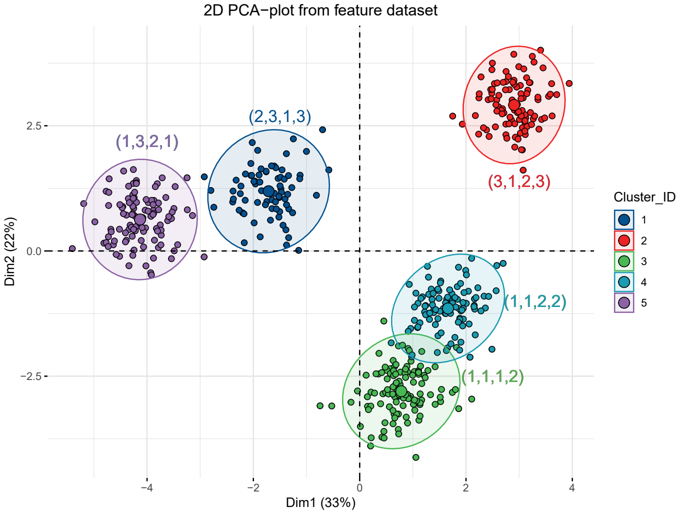


Fig. 1: PCA results in the predictor space

Fig. 2 reports the cross-validation error vs the number of trees. The results show a general improvement of the classification error as the number of trees increases, stabilising progressively in each scenario. The predictive performances of the classification algorithms are strongly related to the value of the dispersion parameter. In particular, for a low value of θ, the boosting method is the best, while for a medium level of noise, bagging and boosting achieve comparable results. Finally, the bagging ensemble method performs best for a high value of θ. Therefore, the degree of label noise is crucial to choose the best algorithm.

5 Conclusions

In this paper, we investigated the role of ensemble methods based on decision trees for item-weighted Label Ranking. Three item-weighted ensemble algorithms

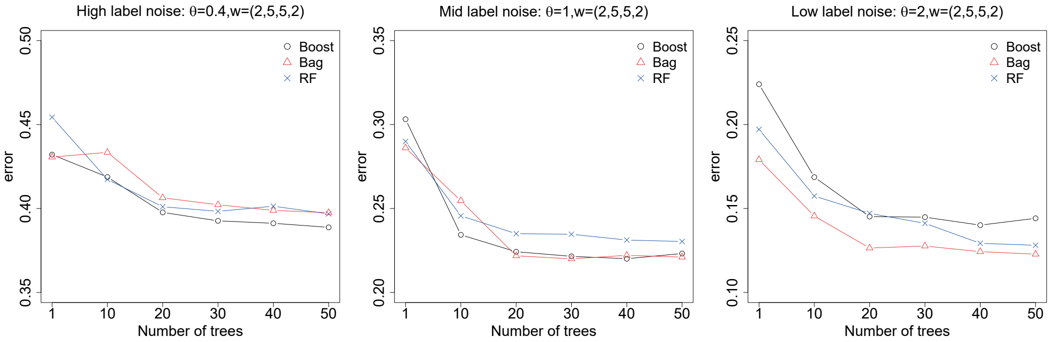


Fig. 2: Comparison of the three weighted algorithms in the three scenarios

(boosting, bagging and random forest) have been employed and compared in tackling the task of predicting rankings given a set of explanatory variables. The predictive performance of the proposed methods is investigated through a simulation study. The results show that bagging is the best method for low noise levels, hence in scenarios that are easier to predict since response rankings are well separated between clusters. Conversely, boosting achieves better results in scenarios with high noise and great confusion between the responses of different clusters. Future developments should focus on extending the set of simulations by investigating differences in prediction errors when increasing the number of noise explanatory variables and outliers.

References

Albano, A. and Plaia, A. (2021). Element weighted Kemeny distance for ranking data. *Electronic Journal of Applied Statistical Analysis, 14(1)*, pages 117–145.s.

Albano, A., Sciandra, M., and Plaia, A. (2022). A weighted distance-based approach with boosted decision trees for label ranking. *Expert Systems with Applications*, 213:119000.

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1):17–28.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, 88(4):577–591.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika*, 44(1/2):114–130.

Melnykov, V., Chen, W.-C., and Maitra, R. (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51:1–25.

Unsupervised Learning of Option Price in a Controlled Environment: a Neural Network Approach

Federico Gatta, Vincenzo Schiano Di Cola, Francesco Piccialli, and Salvatore Cuomo

Abstract Option pricing is one of the most intense research areas in Finance. In this work, we deal with the American call option. We approximate the solution of this free boundary problem through a meshless unsupervised approach based on Deep Learning. The general framework of this work is the Physics-Informed Neural Networks. Thanks to the analytical solution for this problem, we can compare different approximator configurations and understand how convergence and approximation errors are related. So, even if the problem we currently face has restricted practical applications, the primary purpose of this preliminary work is to extract valuable knowledge about the convergence of such an approach. The final aim is to apply a similar methodology in other contexts where it can be instrumental in solving problems with heavy real-world relevance. In this perspective, we carry out experiments allowing changes in model conditions. So, we can do parametric studies that could help to calibrate pricing models.

Key words: unsupervised models, option pricing, deep learning, physics-informed neural network

---

Federico Gatta
Scuola Normale Superiore, Pisa, e-mail: federico.gatta@sns.it

Vincenzo Schiano Di Cola
Department of Mathematics and Applications, University of Naples Federico II, Naples, Italy e-mail: vincenzo.schianodicola@unina.it

Francesco Piccialli
Department of Mathematics and Applications, University of Naples Federico II, Naples, Italy e-mail: francesco.piccialli@unina.it

Salvatore Cuomo
Department of Mathematics and Applications, University of Naples Federico II, Naples, Italy e-mail: salcuomo@unina.it

1 Introduction

Artificial Intelligence and Deep Learning have greatly improved the field of finance. Among the various topics discussed in recent years, a focus has been placed on numerical methods for determining the value of options. *Option pricing* involves examining and creating various methods to determine the fair value of *financial options*. This study focuses specifically on *American call* options, which give the holder the right to purchase a specific asset at a fixed price at any time prior to maturity. In contrast, European call options can only be exercised at maturity. Although the two types of options are similar in theory, the mathematical treatment is different. The *Black-Scholes model* provides an explicit pricing formula for European options, but no such solution exists for American options [1]. However, it has been concluded that it is not advantageous to exercise an American call option before maturity, so its price is expected to be equal to that of a European call option.

Our problem We focus on pricing American call options as this contract is a helpful reference point for evaluating approximation errors. In the future, we aim to extend the proposed method to other contracts and more complex models with no exact solutions. Evaluating the effectiveness of our approach in this context will give us insight into its potential for use in option pricing. This requires having a clear understanding of the ”ground truth.”

Our approach In this work, we employ a *Partial Differential Equations* (PDE) approach to achieve our objective: determine the price function as the solution to a PDE. We operate within the framework of the Black-Scholes model. Conventional methods for solving this PDE have several drawbacks [2]. We aim to overcome these limitations by using a recent meshless Deep Learning strategy based on Physics-Informed Neural Networks (PINNs) [3, 4]. PINNs use the power of neural networks as a universal approximator [5, 6] to find the PDE solution by incorporating the underlying physics in the loss function. Moreover, the problem is approached in an unsupervised manner, meaning the network learns the problem directly without the guidance of other numerical solvers.

Moreover, we also investigate a parametric model, i.e., a model that takes in input the market conditions, which are usually treated as a parameter. In this way, the network can learn how to price the option as the volatility or the strike price change. This approach dramatically benefits by allowing a parametric study and several simulations with different market conditions with just one PINN training. The development of reliable parametric models could help to calibrate pricing models, which is critical in modern financial engineering [7].

Roadmap The structure of this paper is as follows. Section 2 provides a mathematical overview of the option pricing problem and the model we have developed. Section 3 presents the experiments and results we obtained. Finally, Section 4 concludes this work and proposes potential areas for future research.

2 PINN Model for Black-Scholes Equation

In this section, we outline the Black-Scholes equation for pricing American call options, as well as the initial and boundary conditions and the assumptions we make.
Mathematical formulation It should be noted that an American call option gives the holder the right to sell a specific underlying asset at a fixed *strike* price K at any time between 0 and the option's *maturity* T. The option price C is usually expressed as a function of the underlying asset price, S, and time t, while the other parameters are usually treated as fixed. We stress that American options can be exercised every time before maturity T. This implies a free boundary domain mathematically represented by a function $B(t)$, also known as *free boundary*. The Black-Scholes model describes the option price dynamic as follows:

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC + \frac{\partial P}{\partial t} = f(S,t) = 0 \qquad S < B(t); \quad 0 \leq t \leq T \tag{1}$$

where r is the risk-free rate, and σ is the asset volatility. The initial and Dirichlet boundary conditions are formalized by Equation 2, the free boundary initial and boundary (Dirichlet and Neumann) conditions are in Equation 3.

$$C(S,T) = (S-K)_+ \quad S \geq 0 \qquad \& \qquad C(0,t) = 0 \quad 0 \leq t \leq T \tag{2}$$

As for the free boundary initial and boundary (Dirichlet and Neumann) conditions:

$$B(T) = K \qquad \& \qquad C\big(B(t),t\big) = B(t) - K, \quad \frac{\partial C}{\partial S}\big(B(t),t\big) = 1, \; 0 \leq t \leq T \tag{3}$$

We work in the interval $\Omega = [B(t), 3K] \times [0,T]$ for a computational reason ([8]).
Now, we look at the general model developed for solving the American call option pricing problem. The two unknown functions are approximated by two Feedforward Neural Networks (FNN), as in [9]. An FNN can be thought of as a collection of vectors, named *layers* l_j, *weights matrices* and *bias vectors* W_j and b_j, and usually non-linear *activation functions* ϕ_j. More formally, given $n_j \in \mathbb{N}$ and $j \in \{0,1,\cdots,L\}$, the layers are linked to each other in this way:

$$l_j \in \mathbb{R}^{n_j}, \quad W_j \in \mathbb{R}^{n_j \times n_{j-1}}, \quad b_j \in \mathbb{R}^{n_j}, \quad l_j = \phi_j(W_j l_{j-1} + b_j) \tag{4}$$

The first layer, also known as the *input layer*, contains the problem input. The last layer, or *output layer*, returns model outputs. For the solution network FNN^{sol} and the free boundary network FNN^{fb}, the input and output layers are as follows:

$$l_0^{sol} = (S,t) \in \mathbb{R}^2 \quad l_{L^{sol}}^{sol} \approx P(S,t) \in \mathbb{R} \qquad l_0^{fb} = (t) \in \mathbb{R} \quad l_{L^{fb}}^{fb} \approx B(t) \in \mathbb{R}^1 \tag{5}$$

The approximation $\approx$ is intended to be after the convergence of the learning process by minimizing a loss function, namely $\mathscr{L}^{sol}$ for FNN^{sol} and $\mathscr{L}^{fb}$ for FNN^{fb}.
Collocation points and loss function The loss function plays a crucial role in adequately training an FNN. This is done by creating an appropriate training set

$\mathscr{X} = \{x_i\}_{i \in N}$ made up of several *collocation points* belonging to different types, accordingly to the initial and boundary conditions:

$$\mathscr{X}^{sol} = \mathscr{X}^{sol}_{PDE} \cup \mathscr{X}^{sol}_{init} \cup \mathscr{X}^{sol}_{Dir} \cup \mathscr{X}^{sol}_{Fb\_Dir} \cup \mathscr{X}^{sol}_{Fb\_Neu} \qquad \mathscr{X}^{fb} = \mathscr{X}^{fb}_{init} \cup \mathscr{X}^{fb}_{Dir} \cup \mathscr{X}^{fb}_{Neu} \tag{6}$$

Similarly, the loss functions can be viewed as the sum of the contribution derived from differential formulation, initial and boundary condition:

$$\mathscr{L}^{sol} = \mathscr{L}^{sol}_{PDE} + \mathscr{L}^{sol}_{init} + \mathscr{L}^{sol}_{Dir} + \mathscr{L}^{fb}_{Dir} + \mathscr{L}^{fb}_{Neu} \qquad \mathscr{L}^{fb} = \mathscr{L}^{fb}_{init} + \mathscr{L}^{fb}_{Dir} + \mathscr{L}^{fb}_{Neu} \tag{7}$$

The losses are intended to be the *mean square error* (that is, the square of $l2$ norm) of the difference between the evaluation of the networks in the collocation points and the target vector given by the differential operator and initial and boundary conditions. Finally, we remark that this kind of problem can be viewed as an example of Unsupervised Learning. Indeed, no external knowledge is injected into the network, and the loss functions only look at the mathematical formulation of the problem.

Parametric Model As already mentioned, there are four parameters in the Equations considered till now. These parameters are fixed. In this work, we also propose a *parametric model*, i.e. a model that contains the market conditions and works on them as input for option pricing. So, the input layers of FNN^{sol} and FNN^{fb} defined in Equation 5 are modified as in Equation 8. No changes occur in the outputs.

$$Solution \quad l_0^{sol} = (S, t, \sigma, K) \in \mathbb{R}^4 \qquad Free\ Boundary \quad l_0^{fb} = (t, \sigma, K) \in \mathbb{R}^3 \tag{8}$$

3 Experimental Stage

In the experimental stage, we carry out experiments in the standard setting, that is, $C = C(S,t)$ and σ, K, T, r treated as parameters (i.e., the 2D model); we carry out experiments with the parametric model with $C = C(S,t;\sigma,K)$, that is, the volatility and strike price are allowed to vary (thus, this is a 4D model). In both cases, we consider 8 hidden layers for FNN^{sol} and 4 for FNN^{fb}. As for the activation functions, we use *tanh*. The training exploits 6000 epochs with 30000 and 100000 collocation points in $\mathscr{X}^{sol}_{PDE}$ for 2D and 4D models, respectively; 300 and 10000 for the other losses, drawn uniformly in Ω. The optimizer is *RMSprop* [10], and the learning rate is 0.01 with exponential scheduling of rate 0.975 every 100 iterations. For every step of FNN^{fb}, five steps of FNN^{sol} are performed.

The convergence results expressed in terms of PINN losses are described in Table 3. In the table, we report training and testing losses. The first ones are referred to as the training dataset described above. The latter are computed on a test set with points that are not used in the train set (i.e., unseen from the network). In particular, 10000 points are drawn for $\mathscr{L}^{sol}_{PDE}$ and 1000 for the other losses. As the table shows, the networks have a good generalization capability. Indeed, the loss values in the train and test sets are very close.

Model	$\mathscr{L}^{sol}$	$\mathscr{L}^{fb}$	$\mathscr{L}^{sol}_{PDE}$	$\mathscr{L}^{sol}_{init}$	$\mathscr{L}^{sol}_{Dir}$	$\mathscr{L}^{fb}_{init}$	$\mathscr{L}^{fb}_{Dir}$	$\mathscr{L}^{fb}_{Neu}$
2D - Train	4.965	0.314	3.848	0.907	0.000	0.104	0.180	0.180
2D - Test	4.692	0.308	3.662	0.826	0.000	0.104	0.174	0.030
4D - Train	3.693	0.855	2.517	0.351	0.000	0.030	0.217	0.217
4D - Test	2.764	0.838	1.676	0.282	0.000	0.031	0.219	0.587

Table 1 PINN losses in the 2D and 4D models. All the losses are intended to be in the scale $10^{-}3$.

As for the approximation error, it is obtained by comparing the solution with the output of the Black-Scholes formula. The comparison is made by using 10000 points (unseen from the approximators during the train). The $l1$ and $l2$ norms of the error vectors are:

$$2D\ Model: \begin{cases} l1 & 0.0052 \\ l2 & 0.0072 \end{cases} \qquad 4D\ Model: \begin{cases} l1 & 0.0564 \\ l2 & 0.0890 \end{cases} \tag{9}$$

Finally, Figure 1 contains an empirical analysis of the behavior of approximation error as the network's width varies. In particular, 9 different widths are compared, from 8 neurons for each hidden layer in FNN^{sol} and 4 for FNN^{fb}, up to 110 and 45 neurons, respectively. As the figure shows, the error in the 4D model is more variable, with higher variance and wider oscillations. However, this point deserves a further and more accurate study.

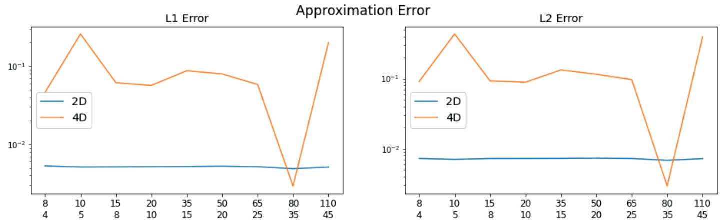


Fig. 1 Approximation error as a function of the layer's width. The left and right panels show $l1$ and $l2$ losses, respectively. The blue line is for the 2D model, and the orange one is for the 4D. The error is reported on a logarithmic scale. The network's width is on the x-axis. The upper number is concerned with FNN^{sol}, and the lower one with FNN^{fb}.

4 Conclusion

In this preliminary work, we have analyzed the behavior of unsupervised networks when approximating the solution of a free boundary problem. Knowing the analytical solution gives us a clear picture of the error behavior, even if it annuls practical usability. One possible direction for further analysis would be a more comprehensive study about the potentiality of PINNs approximator. This investigation can be done by considering the source of randomness derived from random weight initialization. In this context, it is helpful to understand how the error and its variance are related to the number of network parameters and problem dimensionality. Another improvement would be the application of such a methodology in more realistic and complex pricing frameworks, where there is a need for approximation strategies as no closed-form solutions are available.

References

1. F. Black, M. Scholes, The pricing of options and corporate liabilities, The Journal of Political Economy 81 (3) (1973) 637–654.
2. A. Quarteroni, Numerical models for differential problems, Vol. 2, Springer, 2009.
3. M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational physics 378 (2019) 686–707.
4. S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, F. Piccialli, Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next, Journal of Scientific Computing 92 (3) (2022) 88. doi:10.1007/s10915-022-01939-z. URL https://doi.org/10.1007/s10915-022-01939-z
5. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural networks 2 (5) (1989) 359–366.
6. K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural networks 4 (2) (1991) 251–257.
7. S. Liu, A. Borovykh, L. A. Grzelak, C. W. Oosterlee, A neural network-based framework for financial model calibration, Journal of Mathematics in Industry 9 (1) (2019) 1–28.
8. S. Shuka, Finite-difference methods for pricing the american put option, Advanced Mathemathics if finance Honours Project.
9. S. Wang, P. Perdikaris, Deep learning of free boundary and stefan problems, Journal of Computational Physics 428 (2021) 109914.
10. G. Hinton, N. Srivastava, K. Swersky, Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, Cited on 14 (8) (2012) 2.

`SEMgraph`: An R Package for Causal Network Inference of High-Throughput Data with Structural Equation Models

`SEMgraph`: un pacchetto R per inferenza di reti causali di dati ad alta intesità di elaborazione con modelli di equazioni strutturali

Mario Grassi and Tarantino Barbara

Abstract The development of high-throughput sequencing (HTS) in molecular biology and medicine has emphasized the urgent need for scalable statistical methods for modeling complex biological systems. We created the R package `SEMgraph` to combine network analysis and causal inference within the context of structural equation modeling (SEM). It offers a fully automated toolkit for managing complex biological systems as multivariate networks, ensuring robustness and reproducibility through data-driven model architecture and perturbation evaluation, and making it simple to understand in terms of causal relationships between system components. `SEMgraph` is available at https://cran.r-project.org/web/packages/SEMgraph.

Abstract *Lo sviluppo del sequenziamento ad alto rendimento (HTS) nella biologia molecolare e nella medicina ha sottolineato l'urgente necessità di metodi statistici scalabili per modellare sistemi biologici complessi. Abbiamo creato il pacchetto R* `SEMgraph` *per combinare l'analisi di rete e l'inferenza causale nel contesto della modellazione di equazioni strutturali (SEM). Offre un toolkit completamente automatizzato per la gestione di sistemi biologici complessi come reti multivariate, garantendo robustezza e riproducibilità attraverso l'architettura del modello basata sui dati e la valutazione delle perturbazioni e semplificando la comprensione in termini di relazioni causali tra i componenti del sistema.*
`SEMgraph` *è disponibile su https://cran.r-project.org/web/packages/SEMgraph.*

Key words: SEM, Network analysis, Causal Inference, HTS

Mario Grassi
Department of Brain and Behavioral Sciences, University of Pavia, Pavia, 27100, Italy, e-mail: `mario.grassi@unipv.it`

Barbara Tarantino
Department of Brain and Behavioral Sciences, University of Pavia, Pavia, 27100, Italy, e-mail: `barbara.tarantino@unipv.it`

1 Introduction

In biomedical research, figuring out and comprehending the mechanisms underlying intricate phenotypic features is of primary importance. High-throughput sequencing (HTS) technology made it clear how complex disease (and generally phenotypical) features actually were, bringing molecular biology and medicine inside the big data age. Signaling pathways, metabolic reaction chains, or extremely vast protein-protein interaction networks are frequently used to depict biological processes (also called interactomes). Complexity in biological systems results from interactions and reactions between its individual components. Structured biochemical and biomedical data may easily be transformed into networks and statistical models, which we commonly refer to as knowledge-based models, thanks to the abundance of publicly accessible biomedical resources. Network models should be updated and tested using a straightforward and transparent workflow, starting from current knowledge. From a computational perspective, the goal is to free the user from initial setup selection by immediately estimating the algorithm and model parameters from quantitative data using effective and parallelizable methods. This problem inspired us to create the R package `SEMgraph`, which uses structural equation modeling (SEM) [1] to enable causal inference on complex biological networks.

2 The `SEMgraph` package

Model syntax and graph analysis are combined in the `SEMgraph` package, which offers a R environment for automated and data-driven causal inference for system biology and molecular medicine. HTS data is often structured into pathways or large networks, enabling either confirmatory or exploratory analysis of salient biological properties. Within `SEMgraph`, this is practically achieved through algorithm-assisted search for the optimal trade-off between best model fitting (i.e., the optimal context) and perturbation (i.e., exogenous influence) given data, in which knowledge is used as supplementary confirmatory information. The input network and the underlying statistical model are interchangeable representations of the same object: a set of interacting variables linked by causal relationships.

A typical SEMgraph workflow involves four main processes, which are represented in Figure 1: (i) data import and graph pre-processing; (ii) causal architecture learning; (iii) looking for (perturbed) network communities and pathways; and (iv) model fitting. SEMgraph's primary objective is to identify important participants in the most appropriate causal model as defined by three contextual sources of data that are simultaneously used in model construction and analysis: graph architecture, quantitative data, and potential perturbing causes.

To achieve this goal, SEMgraph integrates popular R packages for model management and causal inference. Packages `igraph` [2] and `lavaan` [3] provide the basic environment for graph manipulation and model fitting, while `glmnet` [4] and `ggm` [5] constitute the backbone for Directed Acyclic Graph (DAG) estimation and Bow-free Acyclic Paths (BAP) hidden variables deconfounding. The employed methodologies are general enough to accept different graph types (e.g., directed, undirected, or mixed) and any kind of quantitative data, including bio-molecular, sequencing, and clinical data.

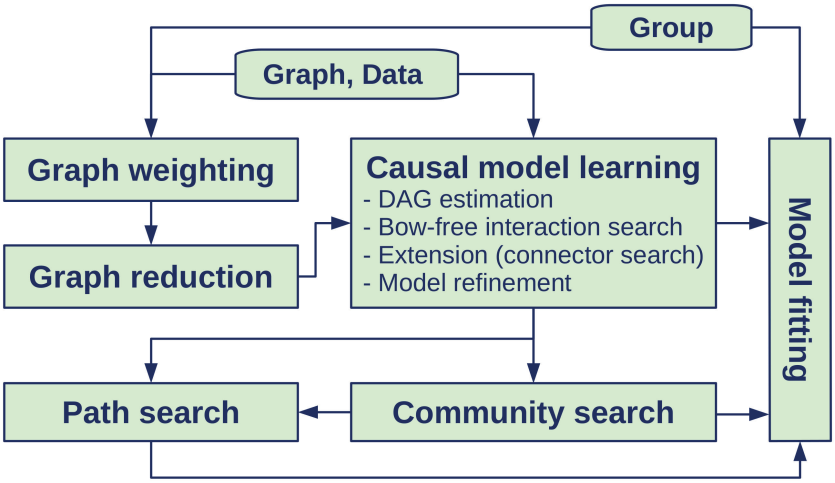


Fig. 1 SEMgraph basic analysis workflow.

3 SEM fitting functions

The three fundamental SEMgraph arguments are:

- a graph G = (V, E), where V is the set of nodes and E is the set of edges;
- continuous high-throughput data, $Y(\mathrm{n} \times \mathrm{p})$;
- a binary group, X = $\{0,1\}$ that represents the experimental condition and control groups.

`SEMrun()` is the primary model evaluation function. This function tests (estimates) the positive definiteness of the covariance matrix, transfers data onto the input graph (removing potential identifier discrepancies), turns the igraph object into a lavaan syntax, then fits the model. Results of the model fitting are recorded inside a lavaan object together with the output graph as (colored) igraph object.

`SEMrun()` will only produce estimates for effects of the path coefficient if the group argument is not provided. On the other hand, if specified, group influence can be modeled as an exogenous variable (`fit=1`) acting on every node and affecting their activity. By setting `fit=2`, the two-groups SEM provided in `SEMrun()` enables estimate of edge perturbation in addition to node perturbation. Standard error

(SE) computation will be disabled for large graphs ($|V| > 100$), and parameter estimates will instead be computed using package `ggm` [5] through residual iterative conditional fitting (RICF) [6], with permutation-based P-values, or for two-group SEM, a constrained gaussian graphical model (CGGM) [7], and de-sparsified (debiased) P-values, will be processed.

A classic illustration of a perturbed route propagating inside the cell regulatory network is provided by biological signaling pathways. When a ligand interacts with a cell surface receptor (*source*), information begins to flow through the cytoplasm to the cell nucleus, where specific factors (*sinks*) are either activated or inhibited, controlling transcription, replication, cell development, and fate. Second messengers, enzymes, and chaperones act as *connectors* and modulators of this information flow. An average causal effect (ACE) can be used to analyze this directional information flow, which can be computationally represented by a directed acyclic graph (DAG) [8]. `SEMace()` function turns the input graph into a DAG and calculates ACEs between each potential source-sink node pair using a back-door set [9], the dagitty minimum set [10], or the ”optimal” O-set with the smallest asymptotic variance [11]. For every ACE, standard errors (SE) are calculated using either the traditional MLE or bootstrap-based methods.

Additionally, SEMgraph offers a collection of tools for causal inference that are readily accessible to users with little statistical knowledge. With the goal of estimating an improved DAG model, we suggest four predefined techniques that are implemented in the `modelSearch()` function by combining the `SEMdag()` (for LASSO structure discovery), `SEMbap()` (for latent variables deconfounding) and `resizeGraph()` (for knowledge-based graph refinement) functions. Model search can be controlled only with two arguments: α (i.e., the significance level for the false discovery rate correction) and β (i.e., the LASSO coefficient threshold).
A companion `SEMdata` R package, which contains a variety of working datasets as `RData` objects and interactomes as `igraph` objects from widely used biological databases, including KEGG, STRING, Reactome, and hiPathia, is accessible at https://github.com/fernandoPalluzzi.

4 Conclusions

SEMgraph is a quick, simple-to-use, yet effective tool for causal network analysis. It integrates causal inference and structure learning within the framework of SEM combining data-driven discovery and confounding correction to model interpretability, and bridging graph theory with statistical inference. We will illustrate the utility of our SEMgraph in empirical examples with RNA-seq expression data and methylation (CpG) data on Amyotrophic Lateral Sclerosis (ALS), and the Coronavirus disease (COVID-19), respectively.

References

1. Bollen, K. A. Structural Equations with Latent Variables. John Wiley Sons, Hoboken, NJ, USA, 1st edition (1989).
2. Csardi, G., Nepusz, T. The igraph software package for complex network research. InterJournal, Complex Systems, 1695 (2006).
3. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1–36 (2012).
4. Tibshirani, Robert et al. "Strong rules for discarding predictors in lasso-type problems." Journal of the Royal Statistical Society. Series B, Statistical methodology vol. 74,2 (2012): 245-266. doi:10.1111/j.1467-9868.2011.01004.x
5. Marchetti, G.M., Drton, M., Sadeghi, K. ggm: Graphical Markov Models with Mixed Graphs. R package version 2.5. (2020) URL: https://CRAN.R-project.org/package=ggm.
6. Drton, M., Eichler, M., Richardson, T.S. Computing Maximum Likelihood Estimated in Recursive Linear Models with Correlated Errors. Journal of Machine Learning Research, 10(81), 2329–2348, (2009).
7. Hastie, T., Tibshirani, R., Friedman, J. H. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, Springer (2009).
8. Pearl, J. Causality: Models, reasoning, and inference. Cambridge University Press, New York, NY, USA, 2nd edition, (2009).
9. Pearl, J. Graphs, Causality, and Structural Equation Models. Sociological Methods Research, 27(2), 226–284, (1998).
10. Perkovic, E., Textor, J., Kalisch, M., Maathuis, M.H. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. Journal of Machine Learning Research, 18:1-62, (2018).
11. Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. On efficient adjustment in causal graphs. Journal of Machine Learning Research, 21(246), 1–45 (2020).

Dynamic models based on stochastic differential equations for biomarkers and treatment adherence in heart failure patients

Modelli dinamici basati su equazioni differenziali stocastiche per biomarcatori e aderenza al trattamento in pazienti affetti da scompenso cardiaco

Caterina Gregorio, Nicola Rares Franco, Francesca Ieva

Abstract Studying the temporal relashionships between biomarkers and drug use can be the key to optimize treatment and improve healthcare. Information contained in eletronic health records is still not exploited enough for this purpoise. We believe that multi-subjects models based on stochastic differential equations can be the answer to understand temporal patters from intensive longitudinal observational health-care data regarding biomarkers and treatments. In this work we show how both fixed and random effects models can be used in this framework using a motivating example coming from a real-world application involving heart failure patients.

Key words: temporal stochastic differential equations, mixed-effects models, electronic health records, heart failure

1 Introduction

Thanks to electronic health records, longitudinal data regarding biomarkers and drug use are now available in real-world clinical settings at a granular and individual level. This source could provide valuable information about possible patterns and interactions between these elements over time. With this aim, statistical models enabling the study of time-dependent dynamical systems are needed. In this respect,

---

Caterina Gregorio
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy;
Biostatistics Unit - Department of Medical Science, University of Trieste, Trieste, Italy
e-mail: caterina.gregorio@polimi.it

Nicola Rares Franco,
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy;
e-mail: nicolarares.franco@polimi.it

Francesca Ieva,
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy;
CHDS, Center for Health Data Science, Human Technopole, Milan, Italy
e-mail: francesca.ieva@polimi.it

models based on differential equations are a standard approach in many fields, such as biology, economics, and engineering, as they provide a natural way to describe the temporal behaviour of phenomenons. However, these models are often framed in the single-subject scenario, that is $N = 1$. On the contrary, in social and medical sciences it is of interest to consider multi-subjects models. Recently, several approaches have been developed towards this direction, in an attempt to propose differential models that can handle repeated measures data coming from multiple subjects. Another important distinction is the one between models featuring either fixed or random effects. The first assume the existence of a system common to all subjects, while the latter relax this assumption, allowing for heterogeneity in model parameters across subjects due to both measured and unobserved factors. Hierarchical Stochastic Differential Equations (SDEs) model have been proposed under a Bayesian framework in the psychometric literature, see the work by Driver & Voelkle in [1]. However, since complex interactions between biomarkers cannot be captured by linear systems, nonlinear models have also become of interest. In [2], Sun et al. proceed along this direction by proposing a semiparametric nonlinear model with fixed effects based on ordinary differential equations, while, in [3], Chow et al. propose the use of nonlinear ordinary equation models with random effects. Another interesting aspect concerns higher order systems, which have been discussed in the framework of hierarchical stochastic dynamical models by Oud et al. in [4]. The aim of this work is to propose a continuous-time dynamic modeling approach, grounded on the use of electronic healthcare data, for studying the relationship between biomarkers and drug use. As a motivating example, we consider a problem concerning heart failure patients, where one is interested in studying the joint dynamics of potassium, renal function and treatment adherence.

2 Motivating problem and data

Heart failure is a consequence of many cardiovascular diseases and, despite improvements in treatments, mortality, and hospitalization rates remain high. To this day, the effective management and treatment of heart failure patients mostly relies on the monitoring of biomarkers over time: among these, two of key importance are potassium and estimated Glomerular Filtration Rate (eGFR). Potassium is essential for the correct functioning of the heart, being one of the substances regulating its electrical activity. Normal potassium levels are maintained through healthy eating habits and well functioning kidneys, which help eliminate the potassium resources in excess. Traditionally, kidney function is measured through the eGFR. Moreover, potassium levels can also be influenced by substances such as drugs and diseases. For instance, patients with heart failure may experience alterations of potassium levels due to the heart failure and renal diseases, but also as a consequence of medical treatments. Among heart failure treatments, Mineralocorticoid Receptor Antagonists (MRAs) are at the same time considered as one of the cornerstone of therapy in heart failure, but they are also thought to be the main ones responsible for potassium abnormalities. Therefore, understanding the dynamics regulating the use of MRAs, potassium and eGFR at the patient level, is an important step towards the improvement of medical healthcare and the personalization of clinical treatments. However, suitable models that can properly integrate real-world data in the descriptionof the relationship between these three actors, have not been developed yet. Among the several challenges encountered when addressing this problem we mention: 1) the interest in modeling the overall temporal evolution while also accounting for subjects heterogeneity, 2) the complex interactions between biomarkers and treatments, which are, in general, unknown and possibly regulated by nonlinear mechanisms.

Data were obtained by the interrogation of the administrative regional health data of Friuli Venezia Giulia Region in the Northern part of Italy, integrated with data derived from the Outpatient and Inpatient Clinic E-chart (Cardionet ®). This integrated database constitutes the Trieste Observatory of Cardiovascular Diseases. Specifically, this was a cohort observational, non-interventional study involving patients living in Trieste who had a Heart Failure diagnosis between January 2009 and December 2020 and had initiated treatment with MRAs with at least one cardiological evaluation, two years of observation, and thirty blood measurements over the two years period. Longitudinal laboratory measurement data, as well drug purchases data, were used for the analysis.

3 Methods

For each subject i, let m_i be the number of discrete time points $t_{i,k} = 0, \dots, t_{i,m_i}$ at which v longitudinal processes of interest are measured. Note that we do not make any assumption about the regularity of such observations, which might as well be taken at irregular time intervals. We model the measurement process as

$$\mathbf{y}_i(t) = \boldsymbol{\eta}_i(t) + \varepsilon_i(t), \tag{1}$$

where $\boldsymbol{\eta}_i(t)$ is the v-dimensional vector containing the latent values of the longitudinal processes at time t, for subject i, while $\varepsilon_i(t) \sim \mathcal{N}(0_v, \Sigma)$ models possible measurement errors with a suitable covariance matrix $\Sigma \in \mathbb{R}^{v \times v}$. As a first step, we shall prescribe the dynamics of the latent variables, $\boldsymbol{\eta}$, through a suitable model based on SDEs. In doing so, we consider both the case of fixed and random effects.

3.1 Dynamic continuous time models

Following Driver & Voelkle [1], we model the measurement process dynamics through the following multivariate SDE,

$$d\boldsymbol{\eta}_i(t) = \Big(\mathbf{A}\boldsymbol{\eta}_i(t) + \psi + \mathbf{M}\chi_i(t)\Big)dt + \mathbf{G}d\mathbf{W}_i(t). \tag{2}$$

Here, $\mathbf{A} \in \mathbb{R}^{v \times v}$ is the drift matrix, containing the auto- and cross-effects, ψ_i is a v-vector describing long term behaviors, $\chi_i(t)$ is a d-vector of fixed and external time-dependent covariates, which are assumed to be independent of the fluctuations in the system, and whose action is dictated by the rectangular matrix $\mathbf{M} \in \mathbb{R}^{v \times d}$, wherease, $\mathbf{W}_i(t)$ is an independent Wiener process modeling noise, whose action is modulated by the lower triangular matrix $\mathbf{G} \in \mathbb{R}^{v \times v}$. The estimation of the model parameters, $\mathbf{A}, \psi, \mathbf{M}, \mathbf{G}$, is obtained under a Bayesian framework, while we exploit a hybrid Kalman filter for the computation of subjects specific likelihoods.

3.2 Mixed effects dynamic continuous time models

Under the hierarchical Bayesian framework, one can extend (2) in the natural way to include random effects:

$$d\eta_i(t) = \Big(\mathbf{A}_i \eta_i(t) + \psi_i + \mathbf{M}_i \chi_i(t)\Big) dt + \mathbf{G}_i d\mathbf{W}_i(t). \tag{3}$$

Let $\Phi_i := [\mathbf{A}_i, \psi_i, \mathbf{M}_i, \mathbf{G}_i,]$ be the s-dimensional vector of subject-specific parameters, $s = v^2 + v + vd + v(v-1)/2$. We model the latter as

$$\Phi_i = f(\mu + \mathbf{R}\mathbf{h}_i + \beta \mathbf{z}_i) \quad \text{with} \quad \mu, \mathbf{h}_i, \beta \in \mathcal{N}(\mathbf{0}, \mathbf{1}). \tag{4}$$

Here, $\mu \in \mathbb{R}^s$ is a common mean vector, $\mathbf{R} \in \mathbb{R}^{s \times s}$ is the matrix square root of the populations covariance matrix, which parametrizes the effect of subject specific deviations $\mathbf{h}_i \in \mathbb{R}^s$, wherease, $\beta \in \mathbb{R}^{s \times w}$ describes the effect of the fixed w predictors $\mathbf{z}_i$, and, finally, $f : \mathbb{R}^s \to \mathbb{R}^s$ is a suitable operator that makes (4) consistent with the parameters support. More detail on the models can be found in Driver & Voelkle, see [1]. Here, we have described the general model in which all parameters are subject-specific; however, depending on the problem under study, the model may as well include random effects only for a subset of the model parameters.

4 Clinical application

For our analysis, we consider three variables: potassium, eGFR and adherence to the MRAs treatment. Potassium levels are measured directly through blood samples, while eGFR values are derived from creatinine measurements via the CKD-EPI formula [5], which also accounts for subjects' age and gender. To measure treatment adherence, instead, we use a variation of the so-called Proportion of Days Covered (PDC), which, in pharmacoepidemiology, is a common approach to quantify pharmacological consumption over a predefined period of observation, cf. [6].

Recently, this definition has been extended to take into account its time-dependent nature by Spreafico and Ieva [7]. Here, we consider a time-dependent weighted version of the PDC that can be tuned to capture short-term behaviors, namely

$$\text{wPDC}_i(t) := \left[\sum_{t_{i,j} \le t} e^{-\gamma(t - t_{i,j})} a_i(t_{i,j})\right] \cdot \left[\sum_{t_{i,j} \le t} e^{-\gamma(t - t_{i,j})}\right]^{-1} \in [0,1] \tag{5}$$

where $a_i(t) \in \{0,1\}$ is a binary variable indicating whether, at time t, subject i was covered or not by the MRAs treatment. Here, γ is a suitable hyperparameter that we use to tune the sensitivity of the wPDC to local oscillations: note in fact that, for $\gamma \downarrow 0^+$ we recover the time-dependent definition in [7] while, for $\gamma \uparrow +\infty$ one has $\mathrm{wPDC}_i(t) = a_i(t)$. Finally, in order for us to treat the wPDC as an (unconstrained) real-valued variable, we consider its transformation via the logit transform, that is $p \to \log(p) - \log(1-p)$.

4.1 Results

In the analysis, 73 subjects were included, 62% of which are male. The median age is 76 years old (IQR: 72 - 82)26% patients reported advanced symptoms, while 63% of them experienced chronic renal disease. Figure 1 reports the trajectories of the biomarkers and treatment adherence as observed in 6 subjects (right panel), as well as a network representation of the dynamical system under study (left panel).

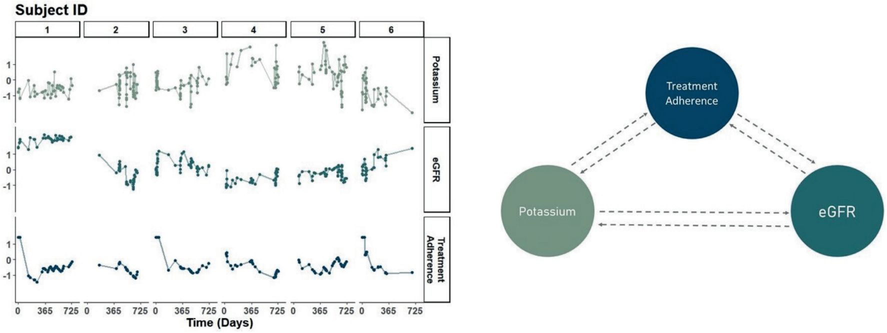


Fig. 1 Observed longitudinal trajectories for six subjects (left panel). Dynamic Network under study (right panel).

To estimate the dynamics of the continuous models, we relied on the R package `ctsem`. For our study, we considered three different models, namely:

- *Model 1*: A fixed effect model;
- *Model 2*: A hierarchical model with subject-specific intercepts;
- *Model 3*: A hierarchical model with random effects and drift parameters;

First, assuming the existence of a common dynamics and intercepts across subjects, it was possible to estimate both the direct and indirect effects of each variable in the network, together with their total effect over different time intervals. Once this assumption was relaxed with the introduction of random intercepts, it was possible to estimate the auto and cross-effects taking into account for possible heterogeneity in the average levels of the process due both to fixed covariates such as age, sex and, Heart Failure Severity and unobserved factors. Moreover, thanks to *Model 3* we were also able to estimate subject-specific networks. Finally, to assess the validity of our results , we compared the predictions obtained through the different models with the observed clinical data.

4.2 Conclusions

The proposed approach based on SDEs and time-dependent continuous models has proven to being able to capture statistically significant temporal interactions among potassium, renal function and treatment adherence with MRAs in health failure patients. In the future, it would be of interest to study extensions of these models involving more complex dynamics described by higher order models or nonlinear ones, for instance by adapting the ideas in [2] or [9] to our framework.

Acknowledgements The authors thank the Biostatistic Unit of the University of Trieste, Prof. Giulia Barbati and the Trieste Observatory of Cardiovascular Disease, Dr.Andrea Di Lenarda and Dr.Arjuna Scagnetto.

References

1. Driver, C. C., & Voelkle, M. C. (2018). Hierarchical Bayesian continuous time dynamic modeling. Psychological methods, 23(4), 774.
2. Sun, M., Zeng, D., & Wang, Y. (2021). Modelling temporal biomarkers with semiparametric nonlinear dynamical systems. Biometrika, 108(1), 199-214.
3. Chow, S., Lu, Z., Sherwood, A., & Zhu, H. (2016). Fitting Nonlinear Ordinary Differential Equation Models with Random Effects and Unknown Initial Conditions Using the Stochastic Approximation Expectation–Maximization (SAEM) Algorithm. Psychometrika, 81, 102-134.
4. Oud, J.H.L., Voelkle, M.C., Driver, C.C. (2018). First- and Higher-Order Continuous Time Models for Arbitrary N Using SEM. In: van Montfort, K., Oud, J.H.L., Voelkle, M.C. (eds) Continuous Time Modeling in the Behavioral and Related Sciences. Springer, Cham.
5. Levey, A. S., & Stevens, L. A. (2010). Estimating GFR using the CKD Epidemiology Collaboration (CKD-EPI) creatinine equation: more accurate GFR estimates, lower CKD prevalence estimates, and better risk predictions. American journal of kidney diseases : the official journal of the National Kidney Foundation, 55(4), 622–627.
6. Andrade, S. E., Kahler, K. H., Frech, F., & Chan, K. A. (2006). Methods for evaluation of medication adherence and persistence using automated databases. Pharmacoepidemiology and drug safety, 15(8), 565–577.
7. Spreafico, M., & Ieva, F. (2020). Dynamic monitoring of the effects of adherence to medication on survival in heart failure patients: A joint modeling approach exploiting time-varying covariates. Biometrical Journal, 63, 305 - 322.
8. Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous Time Structural Equation Modeling with R Package ctsem. Journal of Statistical Software, 77(5), 1–35.
9. Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences, 113(15), 3932-3937.

Detecting anomalies in time series categorical data: a conformal prediction approach

Rilevamento di anomalie in serie temporali di dati categorici: un approccio basato su conformal prediction

Matteo Landrò, Aymeric Stamm and Simone Vantini

Abstract Despite large portions of the information consist of qualitative features, only a limited number of methods have been developed to identify anomalous observations in time-dependant multivariate categorical data. Derived from the Conformal Prediction framework, we propose an anomaly detection procedure based on k-nearest neighbors algorithm and applied in a time series context. The method is able to successfully detect anomalies in multivariate categorical data either at global level (the new observation as a whole) or at local level (its single categorical features), producing a false alarm rate always bounded by the specified significance level.

Key words: Anomaly detection, conformal prediction, categorical time series

1 Introduction

An *anomaly*, or outlier, is an observation which deviates so much from rest of the data so that to raise suspicions that it was generated by a different mechanism [8]. That is, anomalies are the results of a shift in the data generating process. Most literature on anomaly detection has been devoted to the study of quantitative or nu-

---

Matteo Landrò
MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. Now at SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, USA, e-mail: matteo.landro@polimi.it

Aymeric Stamm
Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, Nantes, France e-mail: aymeric.stamm@cnrs.fr

Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy e-mail: simone.vantini@polimi.it

merical data [1, 15]. By contrast, despite large portions of the information consist of qualitative features, significantly little attention has been paid toward detecting outliers in categorical data [17] and only recently we have seen of such procedures being discussed. [15] present the first review of anomaly detection algorithms for categorical data. Among all the solutions proposed, a limited number of methods have been developed to identify anomalous observations in time-dependant multivariate categorical data or categorical data streams. Time series data are typically generated by continuous measurements over time, with values that do not change significantly or change in a smooth way. It is therefore interesting for the user to depict sudden changes in the underlying data [1]. [3] provide a first review of anomaly detection methods on univariate categorical data streams. [13] focuses specifically on transaction databases and propose a method to detect transactions that are likely to be outliers. To the best of our knowledge, no other research has been published on the identification of anomalies for categorical time series data, in particular in the multivariate setting.

In recent times, a new method for anomaly detection based on Conformal Prediction (CP) framework, namely Conformal Anomaly Detector (CAD) has been developed [9]. Conformal Prediction was introduced in [16]. It allows to produce prediction intervals with guaranteed error rate, with the sole assumption of exchangeable data. The promising results of this procedure has generated a plethora of new articles where Conformal Prediction is applied, from simple multivariate data [10], to time series data [4] and functional data [11, 5]. For a detailed review on conformal prediction application refer to [7]. Conformal Anomaly Detector is a general non-parametric algorithm for anomaly detection. It has a well-calibrated false alarm rate and a principle way of defining the anomaly threshold, which does not make any assumption on the nature of frequency of anomalies. As for standard conformal prediction, CAD requires a suitable non-conformity measure to be specified that accurately quantifies how different each observation is from the rest of the data. Conformal Anomaly Detector has already been applied to a variety of data types [9, 14, 6, 12], showing good performances compared to traditional methods.

We propose a Conformal Anomaly Detector based on k-nearest neighbors algorithm to detect anomalous observations in multivariate categorical data, which we further extend to account for temporal dependency.

2 Conformal prediction and conformal anomaly detection

Conformal prediction deals with the definition of prediction intervals for regression and classification problems. Assume a sequence of n i.i.d. (or exchangeable) examples $z_1, z_2, ..., z_n \in Z$ and a new example $z_{n+1} \in Z$. We are interested in identifying a set of elements $\Gamma^{\varepsilon} \subseteq Z$ to contain z_{n+1} with a certain level of confidence $1-\varepsilon$, with $\varepsilon \in [0,1]$ defined as the *significance level*. The idea is to consider each possible realization z of z_{n+1} and estimate its corresponding *p-value* p_z. The prediction set consists of all z having $p_z > \varepsilon$. This is analogous with statistical hypothesis testing where the most unlikely values are rejected at significance level ε. In order to estimate the p-values, the concept of *nonconformity measure* is introduced. It is a real function $A(B,z) : Z^n \times Z \rightarrow \mathbb{R}$ that allows to score how different an example z is from a set B. For a given $z_i \in Z$ we define the corresponding p-value as:

$$p_i := \frac{|\{j = 1, ..., n+1 : \alpha_j \geq \alpha_i\}|}{n+1} \tag{1}$$

with $\alpha_j = A(B, z_j)$ representing the nonconformity score for the element j. One can construct a prediction set Γ^{ε} for the new observation z_{n+1} including all the possible realization $z \in Z$ of z_{n+1} with p-value $> \varepsilon$. The resulting prediction sets show exact finite-sample validity under the sole exchangeability assumption.

Conformal anomaly detection stems from the conformal prediction procedure in a very simple way. If are only interesting in detecting anomalies we can limit the CP procedure to check only the p-value for the observed label instead of calculating p-values for all possible labels. In other words, a new example will be labelled as anomaly if its p-value is lower that the specified significance level ε. Since the probability that the conformal predictor erroneously excludes the true label is less or equal to ε, the expected false alarm rate of a Conformal Anomaly Detector is bounded by ε.

3 Proposed methodology

Analogously to a conformal predictor, the choice of nonconformity measure is central to the performance of a Conformal Anomaly Detector. The more suitable the nonconformity measure, the more sensitive the algorithm is to detect anomalies. The notion of similarity or distance for categorical data is not as straightforward as for continuous data. In the former, the different values that a categorical attribute takes are not inherently ordered. The simplest way to find distance between two categorical attributes is by using the *overlap* measure, which assigns a similarity of 1 if the values are identical and 0 if the values are not identical. The clear drawback of this measure is that all matches, as well as mismatches, are treated as equal, leading towards inaccurate distance metrics. To overcome this limitation, one could use data-driven similarity functions for categorical attributes that take into account the frequency distribution of different attribute values in a given data set [2]. Examples of such metrics are the *Goodall* or the *Lin* measures.

Another challenge posed by categorical streams is the time dependency between subsequent examples. Conformal prediction is based upon the notion of excheangeable data. When data are serially dependent this assumption is violated. In this case, the conformal procedure can be designed to have a block structure to preserve dependency in order to generate approximately valid results [4]. Moreover, the nonconformity measure has to somehow take into account the dependence structure in the data.

To construct a valid CAD for categorical time series we propose a nonconformity measure based on a k-nearest neighbors (k-nn) forecasting approach to construct the p-value for a new observation z_{n+1}. In this case, the target y_i of a generic example z_i correspond to z_i itself and its associated feature vector x_i is described by the l lagged examples of the target. For every element of the time series, and under a block structure as in [4], a k-nn regressor searches among all the historical data the k closest feature vectors, thus identifying the corresponding target which should be the closest to y_i, and predicts the new example's target $\hat{y}_i$ by their majority class. Distances between feature vectors are evaluated through a categorical similarity metric. The nonconformity score α_i is then calculated as the distance between the actual target y_i of the example z_i and its associated predicted target $\hat{y}_i$. Once defined the nonconformity scores for all the element of the time series, the p-value for z_{n+1} is calculated according to (Equation 1). Lastly, z_{n+1} is labelled as anomalous if associated with a $p_{n+1} < \varepsilon$.

The procedure presented as such allows to identify anomalies at a global level, that is, it associates an anomaly score to the observation as a whole regardless if the anomaly is caused by a certain group of features or another. We also propose an extension of the global CAD method which can detect anomalies at local level, specifically the single categorical features of the observation. Assume each example of the sequence $z_1, z_2, ..., z_n \in Z^m$ to be a multivariate random vector $z = (a_1, ..., a_m)$ whose m components are categorical variables. Given a new example $z_{n+1} \in Z^m$, the local CAD will produce m p-values, one for every attribute of the example, allowing to identify the specific feature(s) causing the anomaly.

4 Application and main results

The proposed anomaly detection algorithm is applied to a number of simulated datasets of multivariate categorical time series. The datasets are generated starting from a 2D Haar wavelet which takes a multivariate vector of real values as an input (i.e. the baseline). The Haar wavelet coefficients are then perturbed according to an autoregressive (AR) process to add a temporal dependence in the dataset. Finally, for each example of the time series, the continuous values of each attribute are binned to form equal sized categories. Anomalies are then inserted in the dataset by changing the baseline of the Haar wavelet at specific time t of the series.

The proposed method is able to successfully detect the simulated anomalies, showing a false alarm rate always bounded by ε, both in the global and the local approach. Different categorical distance metrics and other hyper-parameters of the procedure are tested and results are compared.

In addition, we show the use of the proposed CAD on a real-world dataset about land crop monitoring.

References

1. C. C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2017.
2. S. Boriah, V. Chandola, and V. Kumar. Similarity Measures for Categorical Data: A Comparative Evaluation. In *Proc. 2008 SIAM Int. Conf. Data Min.*, volume 1, pages 243–254, 2008.
3. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24:823–839, 2012.
4. V. Chernozhukov, K. Wüthrich, and Z. Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, pages 732–749. PMLR, 2018.
5. J. Diquigiovanni, M. Fontana, and S. Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.
6. S. Farouq, S. Byttner, M.-R. Bouguelia, and H. Gadd. Mondrian conformal anomaly detection for fault sequence identification in heterogeneous fleets. *Neurocomputing*, 462:591–606, 2021.
7. M. Fontana, G. Zeni, and S. Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.
8. D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.
9. R. Laxhammar and G. Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):67–94, 2015.
10. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
11. J. Lei, A. Rinaldo, and L. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence 2013 74:1*, 74(1):29–43, 2013.
12. C. Liu, M. Liang, Jingwen Hou, Z. Gu, and Z. Wang. LogCAD: An Efficient and Robust Model for Log-Based Conformal Anomaly Detection. *Security and Communication Networks*, 2022:1–13, 2022.
13. N. Mishra, Manoj Gupta. To detect outlier for categorical data streaming. *Int. J. Sci. Eng. Res*, 6(5):1–5, 2015.
14. X. Pan, H. Wang, X. Cheng, X. Peng, and Y. He. Online detection of anomaly behaviors based on multidimensional trajectories. *Information Fusion*, 58:40–51, 2020.
15. A. Taha and A. S. Hadi. Anomaly Detection Methods for Categorical Data. *ACM Computing Surveys*, 52(2):1–35, 2020.
16. V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science Business Media., 2005.
17. S. Wu and S. Wang. Parameter-free anomaly detection for categorical data. In P. Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 112–126, 2011.

The structural behavior of Santa Maria del Fiore Dome: an analysis with machine learning techniques

Il comportamento strutturale della Cupola di Santa Maria del Fiore: un'analisi con tecniche di machine learning

Stefano Masini and Silvia Bacci and Fabrizio Cipollini and Bruno Bertaccini

Abstract The Brunelleschi's Dome overlooking the cathedral of Santa Maria del Fiore in Florence is a symbol of the Italian Renaissance. Because of the presence of numerous cracks distributed on its entire surface, the Dome is subjected to a continuous monitoring activity that relies, among others, on electronic sensors, mainly deformometers, to measure the movements of the cracks, and thermometers, to measure the masonry temperatures. These instruments are active since more than 30 years and take measures more times a day, thus producing a huge amount of data. In this contribution, we aim at applying some machine learning techniques (i) to describe the overall movement of Dome surface through a suitable synthesis of the measures of the sensors and (ii) to make medium- and long-term predictions about the evolution of the Dome.

Abstract *La Cupola del Brunelleschi sovrastante la cattedrale di Santa Maria del Fiore a Firenze è un simbolo del Rinascimento italiano. A causa della presenza di numerose crepe distribuite sull'intera superficie, la Cupola è sottoposta a una continua attività di monitoraggio che si basa, tra gli altri, su sensori elettronici, principalmente deformometri per misurare i movimenti delle crepe e termometri per misurare la temperatura dei muri. Questi strumenti sono attivi da oltre 30 anni e rilevano le misure più volte al giorno, producendo così un'enorme mole di dati.*

Stefano Masini
Dept. of Computer Science, University of Pisa, Largo B. Pontecorvo 3, I-56127 Pisa, e-mail: stefano.masini@unifi.it

Silvia Bacci
Dept. of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni 59, I-50134 Firenze e-mail: silvia.bacci@unifi.it

Fabrizio Cipollini
Dept. of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni 59, I-50134 Firenze e-mail: fabrizio.cipollini@unifi.it

Bruno Bertaccini
Dept. of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni 59, I-50134 Firenze e-mail: bruno.bertaccini@unifi.it

In questo contributo, il nostro scopo è l'applicazione di alcune tecniche di machine learning per (i) descrivere il movimento complessivo della Cupola tramite un'opportuna sintesi delle misure dei sensori e (ii) fare previsioni a medio e lungo termini riguardo all'evoluzione della Cupola.

Key words: Artificial Intelligence, Cultural heritage preservation, Dimensionality reduction techniques, Forecasting, Multivariate time series data, Sensor data

1 Introduction

The cathedral of Santa Maria del Fiore in Florence (IT) with its Dome is one of the most famous buildings of the Italian Reinassance. The Dome was built by Filippo Brunelleschi in the period 1420-1436 adopting a special technique (with bricks disposed as an "herringbone pattern") that allowed setting up the construction site without shoring. The result was impressive: nowadays, the Dome is still of the largest masonry domes in the world, weighing more than 43,000 tons. Unfortunately, from the beginning some cracks appeared on the surface of the Dome, thus the building has always been subject to careful monitoring.

The monitoring system of Brunelleschi's Dome is made up of a multiplicity of instruments, such as piezometers, plumb lines, tele-coordinometers, thermometers, and mechanical and electronic deformometers. In particular, in 1987 were installed several electronic deformometers devoted to measuring the movements of the single cracks at least four times a day. Thus, a huge amount of data has been accumulated since the late of 1980s. The complex nature of relations among variables (mainly, movements of cracks and seasonal and daily changes of the masonry temperatures) together with the limits of the computational resources and competencies available in the scientific community have meant that to date these data have not yet been subjected to a systematic study. Indeed, the analyses carried out in previous works usually focused on a single device or a limited set of them [1, 4, 2]; a more recent work [3] took into account the entire set of electronic deformometers, but limited to a one-year period.

In this contribution, we aim at applying some machine learning techniques (i) to describe the overall movement of Dome surface through a suitable synthesis of the measures of the sensors and (ii) to make medium- and long-term predictions about the evolution of the Dome.

Section 2 provides some more details on data, Section 3 describes the machine learning methods used in the analysis, Section 4 illustrates some preliminary results, and Section 5 concludes with some final remarks.

2 Data

In the following we focus on data coming from the 57 electronic deformometers. A deformometer is a sensor installed on the walls across a crack to measure the changes of its width: at installation the instrument is set on value 0, so that positive measures denote a dilatation of the masonry structure and, then, a shrink of the crack, while negative measures refer to a contraction of the walls and, then, a widening of the crack. Deformometers are allocated on the entire surface of the Dome, with a major concentration on those sectors where there is a major presence of cracks. Here we consider the measurements of the complete set of 57 deformometers collected from 1997 to 2017.

Together with the measures of the deformometers, we also take into account the measures of the 47 masonry thermometers installed upon the Dome, as previous studies [1, 4, 2, 3] outlined a strong association between temperatures and movements of the cracks.

To account for gaps and outliers present in the data due to blackouts that periodically put electronic sensors out of action, producing anomalous oscillations, full scale values, or missing observations, we have to pre-treat data. For this aim, we followed the approach proposed in [2], based on the estimation of a quadratic-sinusoidal regression model per each sensor, thus obtaining a complete data matrix.

3 Methods

The first part of the analysis aims at synthesizing the measures of the entire set of sensors to describe the overall behavior of the Dome (and not of its single cracks). This typical problem of dimensionality reduction is addressed through the Kernel Principal Component Analysis (KPCA) [5].

Compared to traditional PCA, which combines observations in a linear way, KPCA allows us to make a non-linear projection of the observations preserving the relative distances between data points. In KPCA we use a function (kernel) to map the data from the original space in a new high-dimensional features space in order to verify whether, in the new space, the data are linearly separable. The algorithm requires to set the kernel type (linear, polynomial, gaussian rbf or sigmoid) and the gamma parameter (which is a space regularization parameter). In order to find their best combination, we use ScikitLearn's GridSearchCV with cross-validation function and, since KPCA is an unsupervised learning algorithm, we use the distance between the original point and the pre-image calculated on the new high-dimensional feature space as reconstruction error.

The principal components resulting from the application of the KPCA as well as the series of masonry temperatures are then used as inputs in a subsequent analysis aimed at providing predictions of the movements of the Dome at medium- and long-term. For this aim, we exploit the performance of some recurrent and convolutional neural network models [6], typically adopted for the prediction of multivariate time series data.

The above neural networks are used to make predictions of a certain number of steps (days) in the future (multiple-step forecasting). The model is trained using a sliding window of consecutive days (the further the future is, the wider the window), with the mean squared error as loss function and the mean absolute error as metric. The best performing model is a network composed of an initial convolutional layer with 40 (6x6) convolutional filters, followed by a bidirectional layer [8] with 20 Gated Recurrent Units (GRU) [7] and 2 more consecutive hidden layers.

4 Results

The results of the KPCA executed on the entire series of measures are displayed in Figure 1, where the first two principal components are plotted with points related to observations differently coloured according to the season.

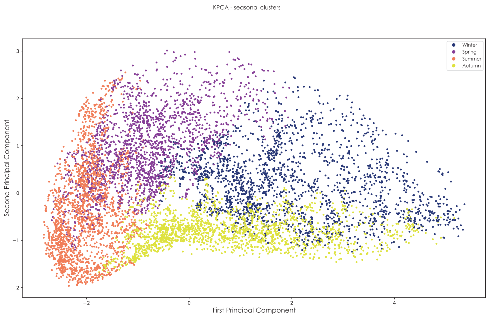


Fig. 1 Results of KPCA (best model: {’gamma’: 0.1, ’kernel’: ’poly’}): seasonal clustering of sensor data

The relation between observed cracks and seasonality emerges clearly from the figure. Namely, clusters of points relating to winter and summer seasons are well separated.

In light of these results, we execute again the KPCA on separate sets of observations, according to the location of the deformometers. We distinguish the deformometers into eight groups, corresponding to the eight slice webs that characterize the surface of the Dome, easily distinguishable with the naked eye thanks to the white marble cords. For the sake of clarity, the webs are numbered counterclockwise starting from the web that faces the nave (see [2] for the planning of the Dome and its webs). Figure 2 shows the trend of the first principal components for each web (top panel: odd webs; bottom panel: even webs), together with the trend of the daily average masonry temperatures (central panel). Note that the figure refers to a one-year time window, but the trend repeats with the same pattern throughout the entire period of observation (i.e., 1997-2017).

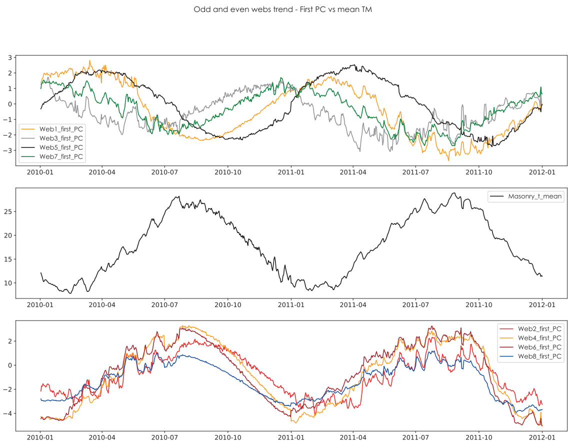


Fig. 2 First principal component of each web (top panel: odd webs, bottom panel: even webs) along a one-year window (January 1st, 2014 to January 1st, 2015)

Looking at Figure 2, we observe that movements of all webs follow a sinusoidal trend according to the temperature, with odd webs that move in the opposite direction with respect to even webs. These results provide evidence for a breathing mechanism of the entire Dome: when even webs shrink, odd webs widen, and vice-versa.

Finally, the first principal components obtained for each web through the KPCA are used in input in a neural network to make predictions. Figure 3 shows the next 100 days prediction results for web 2 with a window size equals to 300 ; results for other webs are similar.

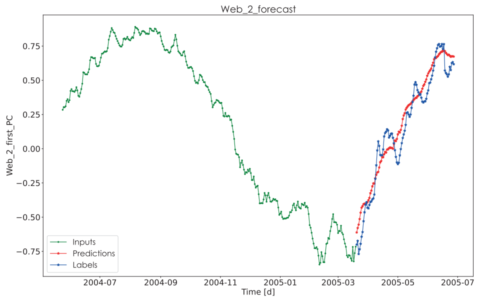


Fig. 3 Forecasting of web 2 trend (window size: 300 days, steps forward: 100 days)

5 Conclusions

The application of the machine learning techniques described in the contribution allowed us to achieve two important goals. First, we demonstrated the close correlation between the temperature and the behavior of the Dome over time, bringing to outlining a symmetry in the movements of the webs in even position and those in odd position. Second, we trained and tested some recurrent neural networks in order to predict the behavior of each web.

For the future, we will prosecute the work along the following main research lines. First, further variables will be taken into account, such as humidity, wind, solar exposition, and seismic measurements. Second, a software will be integrated in the current monitoring system to build a sort of “alarm system” in real-time.

Acknowledgements Authors thank the “Opera del Duomo Foundation” for having making available the data

References

1. Bartoli, G., Chiarugi, A., Gusella, V.: Monitoring systems on historic buildings: Brunelleschi Dome. Journal of Structural Engineering **122**, 663–673 (1996) doi: 10.1061/(ASCE)0733-9445(1996)122:6(663)
2. Bertaccini, B.: Santa Maria del Fiore Dome behavior: Statistical models for monitoring stability. International Journal of Architectural Heritage **9**, 25–37 (2015) doi: 10.1080/15583058.2013.774071
3. Bertaccini, B., Bacci, S., Crescenzi, F.: A Dynamic latent Variable Model for Monitoring the Santa Maria del Fiore Dome Behavior. In: Broy, M., Dener, E. (eds.) Lecture Notes in Computer Science - Computational Science and Its Applications ICCSA 2020, pp. 47-58. Springer professional (2020) doi: 10.1007/978-3-030-58811-3_4
4. Ottoni, F., Blasi, C., Coisson, E.: The crack pattern in Brunelleschi?s Dome in Florence: Damage evolution from historical to modern monitoring system analysis. Advanced Materials Research **133-134**, 53–64 (2010) doi: 10. 4028/www.scientific.net/AMR.133-134.53
5. Schölkopf B., Smola A., Müller K.: Kernel principal component analysis.Springer-Verlag Berlin Heidelberg 1997.; 7th International Conference on Artificial Neural Networks, ICANN 1997; doi: 10.1007/bfb0020217, isbn: 3540636315
6. Hochreiter S., Schmidhuber J. Long Short-Term Memory 1997. Neural Computation, nr.8, vol. 9, pp.1735-1780, doi: 10.1162/1997.9.8.1735
7. Cho K., Van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014, doi: 10.48550/ARXIV.1406.1078
8. Schuster M., Paliwal, Kuldip K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. November 1997. Vol. 45, pp 2673–2681, biburl: https://www.bibsonomy.org/bibtex/26026f083db110838a6b62c2eedfec9e9/nilsd

Statistics and Data Science for Arts and Culture: an Application to the City of Brescia

Statistica e Data science per l'Arte e la Cultura: un'Applicazione alla città di Brescia

Riccardo Ricciardi, Maurizio Carpita, Selene Perazzini, Paola Zuccolotto, Marica Manisera
Research group DS4BS [†]

Abstract This paper introduces the Project "Data Science for Brescia - Arts and Cultural Places". This project combines the use of big data, new technologies, and complex statistical methods to improve the understanding of how people visit cultural sites. The study examines the people's presence in cultural places and along cultural itineraries, and evaluates the visitors' opinions expressed both offline and online.
In addition, the research outputs are integrated with multimedia contents, to provide a holistic vision of artworks as smart objects.

Key words: Mobile Phone Networks, Sensory Museology, Opinion Mining, Smart Objects

1 Introduction

Statistics and Data Science can make an important contribution to mapping movements and experiences of people in cultural places, by supporting institutions and

---

[†] https://bodai.unibs.it/ds4bs/

Riccardo Ricciardi
University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: r.ricciardi@unibs.it

Maurizio Carpita
University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: maurizio.carpita@unibs.it

Selene Perazzini
University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: selene.perazzini@unibs.it

Paola Zuccolotto
University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: paola.zuccolotto@unibs.it

Marica Manisera
University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: marica.manisera@unibs.it

decision-makers in their efforts to improve public engagement and accessibility. These activities promote good practices in the spirit of the 2030 Agenda of the United Nations for the 17 Sustainable Development Goals [3].

In this context, the DMS StatLab and BODaI-Lab of the University of Brescia proposed the Project ”Data Science for Brescia - Arts and Cultural Places” (DS4BS), which aims to improve the understanding of how people visit cultural sites (museums, theaters, monuments, and historic buildings) in the Italian city of Brescia. The project combines the use of big data, new technologies, and complex statistical methods to achieve this goal; particularly, special consideration has been given to experimenting with new methods for public detection and engagement, investigating cultural attitudes and perceptions, and creating new forms of access to culture, especially in the context of cultural tourism.

In this paper, we introduce the main results achieved up to this moment.

2 Research lines

A Data Science approach is developed under the lens of two integrated perspectives, and the output of the research will be integrated with multimedia contents to define artworks as smart objects.

2.1 Line 1: People’s presences and movements

The first research line concerns the monitoring of presences and crowding in artistic and cultural places. We use mobile phone data, which are currently among the best data sources for the study of social phenomena in urban areas [2]. Indeed, they allow for observing the presence and movement of individuals at a high geographical (i.e., small area) and temporal (i.e., short time intervals) level of detail. The mobile phone data at our disposal have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it). The database refers to a selection of Points Of historical, artistic, cultural, or social Interest (POI hereafter) in the city of Brescia observed during the year 2022. Overall, we have information about 25 POIs, divided into 5 macro-categories (4 monuments, 6 museums, 4 squares, 3 theatres, 8 other typologies), and represented as circles with a 100-meter radius (see the left map of Figure 1). For each POI, the database reports the statistical presence, which is defined as the average number of individuals during a 15-minute interval, along with some information about the individuals (e.g., age and gender).

To further explore the multifaceted impacts of the events, we are analyzing people’s presences in conjunction with other data sources. In particular, some expenditure indices defined on Mastercard data have been provided by the Municipality of Brescia for the year 2022. The indices capture variations in different aspects of expenditure, such as the number of purchases, the total value of purchasing, and

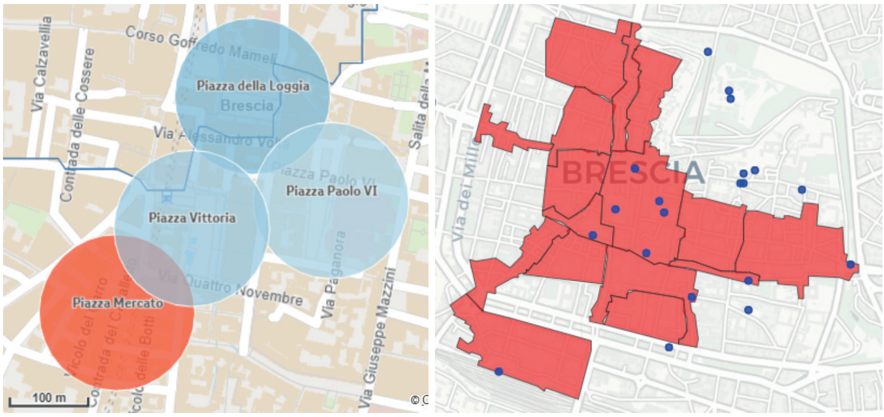


Fig. 1: Mobile phone and Mastercard data. Left: POIs of type ”squares”. Right: Map of the 16 commercial systems constituting the DUC (red) and overlapping or neighboring POIs (blue).

the average value of a purchase, and therefore allow us to investigate the economic impacts of cultural events. Our database refers to 9 industries (e.g., eating places, accommodations, total apparel, . . .) and 3 payment circuits (i.e., international, domestic, overall) in the 16 commercial systems constituting the “Distretto Urbano del Commercio” (DUC) of the city of Brescia (see the right map of Figure 1).

Various events of the year 2022 have been analyzed, such as the historical car race “Mille Miglia”. Nowadays, we are studying the recent cultural events in the context of “Bergamo-Brescia Capitale Italiana della Cultura 2023” (BGBS2023). In particular, we conducted some preliminary analysis on the inaugural events that took place over the weekend of 21-22 January 2023. Among these, results about the concert that took place in the POI “Piazza della Loggia” on Saturday 21st from 5 p.m. are shown in Figure 2. A general increase in the people’s presence with respect to the average value of Saturdays in 2022 emerges throughout the whole day: the number of people in the square has been on average 50% higher, with peeks up to +200%. Soon, the analysis of people’s presence will be integrated with data from footfall counters collected during the concert.

2.2 Line 2: Visitors’ experience

The second research line investigates the visitors’ experience and is based on both offline and online opinions.

The visitors’ experience is evaluated offline by means of questionnaires aimed at evaluating their visit, in terms of expectations and satisfaction with several aspects. In addition, visitors were asked to describe their experience in terms of sight, hearing, touch, smell and taste, even if the only sense concretely used is sight, or hearing. Acting in the field of “sensory museology”, the idea was to quantify the emotional

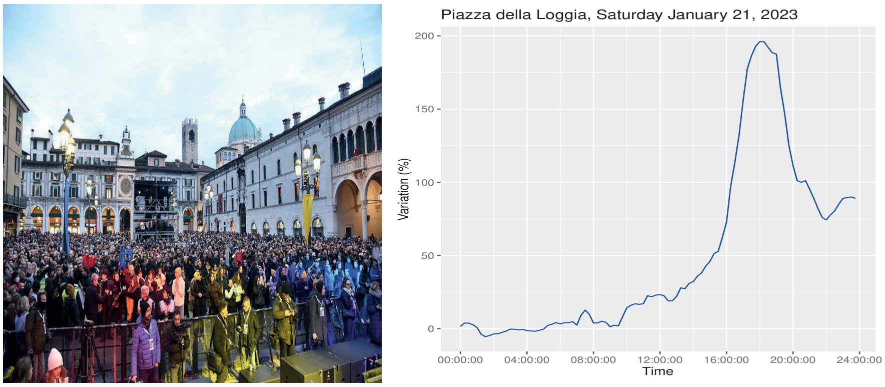


Fig. 2: Analysis of the people's presence in Piazza della Loggia during the inauguration of BGBS2023. Left: A concert on January 21st, 2023. Right: variation (%) in the average number of people in 15-minute time span on Saturday, 21st January 2023, with respect to the average Saturdays of the year 2022.

impact and the multisensory experience of visitors in front of artworks, resorting to phenomena like synesthesia and ideasthesia [7], in which activation of concepts evoke sensory feelings. Focusing on this part of the questionnaire, among different response scale formats, we proposed multi-point semantic differential scales, which required the respondent to position himself/herself on a rating between two bipolar adjectives.

In order to analyse those data with appropriate statistical models, we resorted to the CUM model (Combination of a discrete Uniform and a - linearly transformed - Multinomial random variable [6] [5]), recently proposed in the framework of the CUB (Combination of discrete Uniform and shifted Binomial random variables [8]) class of models. The same approach can be used whenever the response scale has an odd number of ordered categories with a midpoint that mean indifference, for example, "neutral," "neither positive nor negative," or "neither agree nor disagree", as often is the case of the Likert scales commonly used in several fields. For example, Figure 3 shows the results obtained from the application of the CUM model to the responses given by 665 visitors of the Santa Giulia museum in Brescia to a question investigating the easiness in visiting the museum. The response scale was a 7-point semantic differential scale ranging from "difficult" to "easy". Detailed results and interpretation can be found in [5].

Regarding online opinions, reviews on Google and Tripadvisor are provided by Fondazione Brescia Musei. Those will be the input of a mixed approach of topic modeling and sentiment analysis [1] [4], by comparing and integrating top-down and bottom-up strategies. About the first, an ontology of the attractions, i.e. a dictionary of aspects relevant for the decision-makers, is used as a *supervision* for the topic detection task, and then *awareness* and *sentiment* by attractions are estimated; in other words, the frequency and the opinion of people are evaluated for each predefined topic. The bottom-up approach aims, instead, at discovering what (and how)

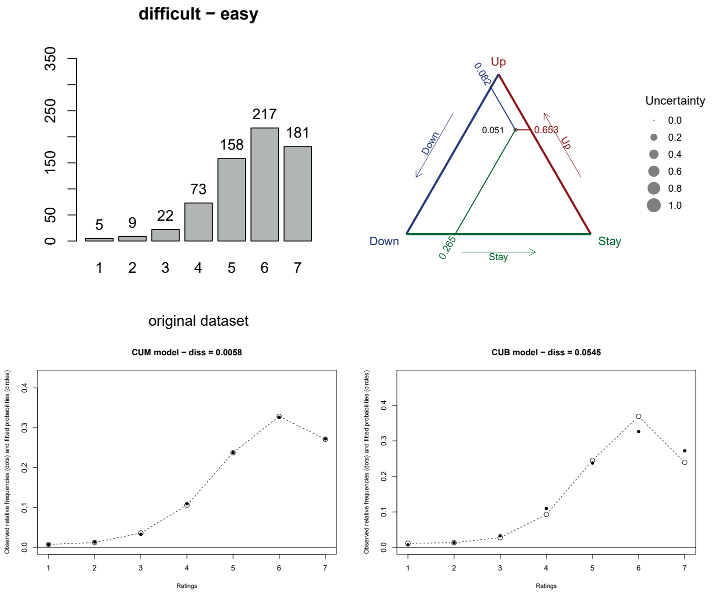


Fig. 3: Results from questionnaires about the Santa Giulia museum in Brescia. Top-left panel shows the frequency distribution, top-right panel the ternary plot, bottom panels show the observed versus fitted frequencies for the CUM model and the standard CUB model.

reviewers talk about, without any specific guidance about the topic. Both strategies are designed to apply traditional and state-of-the-art representations of text to process and understand the language used in the reviews.

2.3 Smart objects

Data from questionnaires will be integrated with data collected by mobile apps. In the computer science literature, several works aim to improve the visitors' experience by promoting the cultural heritage through mobile apps, IoT (Internet of Things) and gamification solutions. In our project, the idea is to define a data lake, able to link artworks to a rich variety of data, which - in combination with multimedia materials, visitors' movements, and artworks classification for different categories of visitors - will considerably improve the fruition of cultural heritage resources in a customized fashion. Moreover, the information equipment built around artworks as smart objects will enable the construction of diverse cultural itineraries by museum professionals, who need knowledge to design and re-organize the exhibition path, thinking about new communication languages that can be used to enhance all the facets of the visitors' experience.

3 Conclusions

This contribution focuses on the DS4BS Project and the city of Brescia, Italy, dedicated by the University of Brescia to support the important BGBS2023 event.

Our approach consists of (1) analyzing mobile phone data to investigate people's movements and presences; (2) evaluating the visitors' offline and online opinions; (3) integrating the information retrieved in (1) and (2) making smart objects of artworks, through a data lake that provides information to both a Content Management System for museums' professionals and mobile apps for the visitors.

An interesting innovation is the evaluation of tangible, emotional, and multisensory experiences at the museums.

All the stakeholders involved in the DS4BS Project took part to the activities of research and innovation, according to the principles of a Data science approach, where computer science, statistics, and expertise in the field intersect in defining the research questions, transform data into information and value and find agreed conclusions. Resorting to the most appropriate statistical methodology plays a crucial role and the multidisciplinary nature of the project can be useful to spread a sound statistical philosophy outside the community of statisticians.

References

1. Blei, D. M.:Probabilistic topic models. Communications of the ACM, **55**, 77-84 (2012)
2. Bibri, S.E., and Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable cities and society. **31**, 183-212 (2017)
3. Cf, O. D. D. S.: Transforming our world: the 2030 Agenda for Sustainable Development. United Nations. NY, USA, (2015)
4. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, **5**, 1-167 (2012)
5. Manisera, M., Migliorati, M., Zuccolotto, P.: A mixture model for the analysis of categorical variables measured on five-point semantic differential scales. Submitted for publication, (2023)
6. Manisera, M., Zuccolotto P.: A mixture model for ordinal variables measured on semantic differential scales. Econometrics and Statistics, **22**, 98–123 (2022)
7. Nikolić, D.: Ideasthesia and art. Digital Synesthesia. A Model for the Aesthetics of Digital Art, 41-52 (2016)
8. Piccolo, D., Simone, R.: The class of CUB models: statistical foundations, inferential issues and empirical evidence. Stat. Meth. Appl. **28**, 389–493 (2019)

Detecting Stance in Online Discussions about Vaccines

Francesco Pierri, Fabio Pizzo and Marco Brambilla

Abstract In 2019, the World Health Organization classified vaccine hesitancy as one of the 10 greatest threats to human health, underlining how about 1.5 million lives each year could be saved by preventing this phenomenon. It is therefore essential to analyze the behavior and ideas of the population in this regard. We study stance of discussants on the matter, by monitoring interactions on social media. We apply a complex stance detection model that manages to obtain good results even with a small amount of manually classified data (few-shot learning), by taking advantage of pre-trained BERT models upon which it applies transfer learning.

1 Introduction

The recent Sars-COV2 virus pandemic has once again focused the attention on one of the most controversial and debated issues ever, i.e., vaccines, focusing the spotlight on one of the greatest problems of the last years, defined as vaccine hesitancy. Misinformation spread online threatens to limit vaccine uptake. In this study, we try to interpret the phenomenon by applying stance detection to identify the position of discussants (in favor or against vaccines) on social media through a learning model that can effectively monitor the phenomenon at scale. We apply a complex stance detection model that manages to obtain good results even with a small amount of manually classified data (few-shot learning), by taking advantage of pre-trained BERT models upon which it applies transfer learning. Several studies addressed disinformation and vaccines: [2] explores the correlation between vaccination coverage and Internet searches, [6] studies the dynamics of the no-vax community on Twitter, and [3] analyzes its the echo chamber effect through Network Analysis, and [8] explores the case of information crisis in Italy on vaccines-related topics.

Francesco Pierri, Fabio Pizzo and Marco Brambilla
Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria. Milano, Italy.
e-mail: {firstname.lastname}@polimi.it

Stance Detection is a very complex task in the NLP scenario [1, 9]. The research in [5] explores the reproducibility of several existing stance detection models. With the work done in [7] the authors proposed a model based on similarity and statistical features, while in [4], different methods such as LR, SVM, CNN, Word2Vec, LSTM and FastText are proposed to solve the Stance Detection problem on vaccines-related Italian Tweets. Since this work regards our Italian scenario, we will take this and [1] as references in analyzing our results.

2 Data

To retrieve tweets of our interest, we exploited 2 types of Twitter's API, a set of programmatic endpoints used to engage with the conversation on Twitter: Search API and Stream API, by applying search terms related to vaccines in the Italian language. Once the data is collected, a label that indicates the stance of the tweet with respect to the vaccines topic is needed, since all the algorithms we employed for Stance Detection are Supervised and Semi-Supervised. Usually the approach used in Machine Learning for Data Annotation is a Human-in-the-Loop (HITL), in which people are involved in a virtuous circle of improvement of the label assignment. Since doing this in a completely manual way is very time-consuming, we used a semi-automatic HITL approach: we labelled a tweet based on the presence of some hashtags, defined as Gold Hashtags, that clearly indicates the stance of the tweets. This can be done only for ProVax and NoVax stances, because for the Neutral ones there are no hashtags that clearly show neutral behaviour. By applying this approach, we obtained a dataset of 20,806 labelled tweets, out of which 8,697 are tagged as pro-Vax, 1,898 are tagged as anti-Vax, and the remaining 10,211 are neutral. We performed a series of preprocessing steps on the collected tweets, including URL removal, standardization of some domain-specific words, punctuation and stopwords removal, lowercasing, lemmatization and stemming. Finally, each tweet was vectorized with a standard tf-idf feature weighting.

3 Methods

Baseline models. We trained Logistic regression, Support Vector Machines, and Naive Bayes models as baselines for our problem, in order to understand if different approaches lead to different results. Since the data inevitably presents dimensions that do not contribute to the prediction of the correct class, we applied dimensionality reduction both through threshold-based feature selection, and Principal Component Analysis (PCA).

Language models. We then applied a more complex model to better grasp the meaning and relationships between words. We chose BERT as a basic brick to construct our architecture. We seletected the most appropriate pre-trained models, from the *HuggingFaces* library. The set of candidates models must be pre-trained on Italian corpus and with no differences between uppercase and lowercase characters. BERT uses an algorithm called BPE (Byte Pair Encoding) to tokenize words, and the relevant aspect of this is that if a word is not in the training dataset or it has a very low number of occurrences, it is not considered as a single token but it is tokenized in the set of its subwords. After a few tests, we chose the `dbmdz/bert-base-italian-uncased` model.

Fine tuning. Once the most appropriate model was chosen, a fine-tuning step was performed. Fine-tuning a pre-trained language model (LM) has become the *de facto* standard for implementing transfer learning in natural language processing as it can be done comparatively inexpensively in a few epochs. In particular, we applied two kinds of fine tuning, standard and adaptive. We applied standard Fine Tuning by adding a softmax layer on top of BERT, in order to calculate a class-score c_i for each sample and normalize it so that $\sum_{i=1}^{3} c_i = 1$. In the Adaptive Fine Tuning setting instead, once ended the pretraining phase, a few epochs of training on the very same pre-training tasks were performed. Before that, the dataset was enriched with words tokenized as multi-token.

Few-shot learning. Since the way in which we collected data introduces some biases to the models, and furthermore we didn't have time and human resources to manually label tens of thousands of tweets, we needed an approach suitable for these low-resource settings. The amount of data that a Deep Learning model requires to generalize on unseen data is directly proportional to its number of parameters, making the task of training a huge model such as BERT directly on the Stance Detection task very difficult. We used a semi-supervised approach recently discovered for other tasks such as Sentiment Analysis or others and known as Iterative Pattern Exploiting Training, well described in [10].

Class imbalance. Our dataset is highly imbalanced. Imbalanced classes pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed on the assumption that every class has the same number of samples. To rebalance the classes, we applied Synthetic Minority Over-Sampling Technique (SMOTE) and data augmentation (in the particular interpretation of back-translation augmentation, which consists of taking a sentence in the original language, translating it into another one, and then re-translating back to the first one, using a Transformer model already trained for this task).

4 Results

Once the models have been trained, we assessed their performance. We trained the baseline Machine Learning models mentioned in the previous section, evaluating them on the average of 5 different validation sets extracted with the k-Fold Cross-Validation approach for hyperparameter-tuning, looking at the accuracy and weighted F1-Score. The best models obtained maximum accuracy of 0.53 for Logistic Regression, 0.57 for SVM, and 0.50 for Naive Bayes.

VALIDATION Performance	**Accuracy**	**Recall**	**Precision**	**F1**	**ROC_AUC**
BERT	0.8976	0.8976	0.8940	0.8956	0.9766
BERT+OverSampling	0.9188	0.9188	0.8991	0.9088	0.9767
BERT+OS+AF	0.9201	0.9201	0.8940	0.8956	0.9769
BERT+OS+AF with new tokens	0.9476	0.9476	0.9040	0.9252	0.9803
TEST Performance	**Accuracy**	**Recall**	**Precision**	**F1**	**ROC_AUC**
BERT+OS+AF with new tokens	0.5944	0.5944	0.6240	0.6001	0.7329

Table 1 Result of the BERT models on the validation set and on the test set.

Starting from these, we removed some noise in the feature vectors by applying two well-known techniques: Principal Component Analysis (only for SVM and Logistic regression) and feature selection using log-odds ratio. These improvements yield to increases in performance of the models, up to an accuracy of 0.5932 for the SVM model.

BERT. We also applied a BERT-based model. We pre-processed data in the very same way as for baseline models, but without Hashtag Splitting, since the tokenization algorithm already embedded in BERT (BPE) works quite well. We started with our experiment searching for the best hyper-parameters with a Grid Search method, on the following: batch-size, max-length, learning rate, weight decay, padding, and truncation.

We obtained these best hyper-parameter values: Batch-Size = 64, Max-Length = 128, Lr = 0.002, Wd = 0.03. Subsequently we retried the experiment in 3 slightly modified version in such a way we could be able to conclude something about the contribution of each component:

1. Oversampling of the minority classes, obtaineing a training dataset composed by 6735 NoVax samples, 6735 proVax samples, 6735 Neutral samples;
2. Adaptive Fine-Tuning to slightly adapt the pre-trained model's weights to the words distribution before the Fine Tuning phase;
3. Adaptive Fine-Tuning with some additional tokens, that are the most important words.

Table 1 shows the results of the experiments on the Validation set of all the variations of the model, and on the test set for the best model.

We can noticed from the results previously presented that further steps of domain-adaptation in the fine-tuning phase may be useful for the model in order to apply its acquired knowledge to our scenario. We also noticed that the results on validation set are almost perfect, while the ones on the test set they are dramatically lower, even though the Validation Set is composed of unseen data too.

In this case, we concluded that the problem is not in the model but in the data. Indeed, the neutral class tweets are retrieved with an automatic procedure from newspaper account while in the test set tweets are manually labeled both from newspapers and from private users. This causes a bias that damage the test performance.

A semi-supervised approach: iPET. Although each of the previous experiments highlighted positive aspects to be taken into consideration, they didn't go beyond the state of the art's performance. Therefore we optimized them by applying PET and iPET semi-supervised approaches [10] that exploit an ensemble of BERT models and use the same task of the pre-training phase (MLM), and assign soft-labels to unlabelled samples in an iterative way, until obtaining large datasets. As a result, we generated a huge dataset of 150000 soft-labelled samples used to fine-tune another BERT model, through a procedure known as *knowledge distillation*.

To be sure that the obtained results are statistically relevant we repeated the training for 5 times with different seeds, saving the model for each repetitions and as it will be shown in the next Section the results remains almost the same.

We trained the model for 0 generations (PET) and then for 1 to 6 generations. Figure 1 shows the effect of the iterative construction of the soft-labeled dataset and the result in terms of accuracy with respect to more standard supervised methods for BERT fine tuning with such a small training set. As we can see the best model is obtained through 5 generation iterative training.

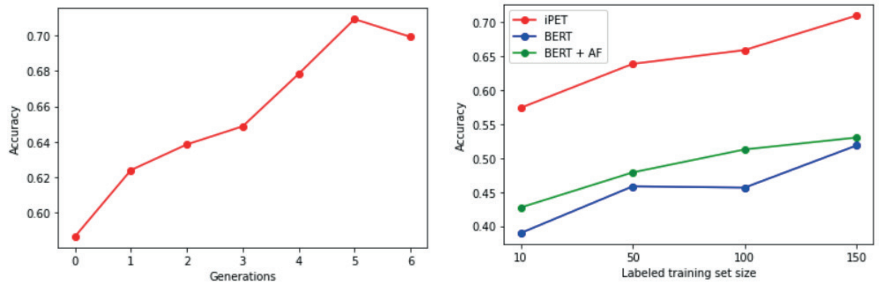


Fig. 1 The accuracy of the model trained with the iPET method in function of the generations (left) and a comparison of the different methods that use BERT as a model. In the second figure all the results of the iPET method come from a 5-generation training.

Table 2 reports the performance measures of the best model, considering 5 different 5-generations iterative training procedures.

	Accuracy	**Precision**	**Recall**	**F1_score**
Mean	0.7093889717	0.735213256	0.7093889717	0.7163866292
Std Dev	0.000231	0.000214	0.00094	0.0003

Table 2 Results of the model obtained with a iPET training for 5 generations and repeated for 5 times with different seeds.

5 Conclusions

In this study, we applied stance detection on social media interactions to study the role of disinformation on vaccination hesitancy. We confirmed that baseline ML models, even with aggressive data pre-processing and augmentation, are not suitable for the task. We therefore applied a few-shot transfer learning approach on pre-trained BERT models. To go beyond basic fine-tuning of the model, we adopted iPET, a semi-supervised algorithm applying knowledge distillation upon a model ensemble, obtaining much better performance.

References

[1] Abeer ALDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Proc. & Management*, 58(4):102597, 2021.

[2] Francesco Aquino, Gabriele Donzelli, Emanuela De Franco, Gaetano Privitera, Pier Luigi Lopalco, and Annalaura Carducci. The web and public confidence in mmr vaccination in italy. *Vaccine*, 35(35, Part B):4494–4498, 2017.

[3] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: The italian vaccination debate on twitter. *Proc. of the Int.l AAAI Conference on Web and Social Media*, 14(1):130–140, May 2020.

[4] Eleonora D'Andrea, Pietro Ducange, Alessio Bechini, Alessandro Renda, and Francesco Marcelloni. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116, 09 2018.

[5] Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. Stance detection in web and social media: A comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87, Cham, 2019. Springer International Publishing.

[6] Keith Gunaratne, Eric A. Coomes, and Hourmazd Haghbayan. Temporal trends in anti-vaccine discourse on twitter. *Vaccine*, 37(35):4867–4871, 2019.

[7] Fuad Mire Hassan and Mark Lee. In *Statistical Language and Speech Processing*, pages 273–285, Cham, 2019. Springer International Publishing.

[8] Alessandro Lovari, Valentina Martino, and Nicola Righetti. Blurred shots: Investigating the information crisis around vaccination in italy. *American Behavioral Scientist*, 65(2):351–370, 2021.

[9] Sara S. Mourad, Doaa M. Shawky, Hatem A. Fayed, and Ashraf H. Badawi. Stance detection in tweets using a majority vote classifier. In *The Int.l Conf. on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, pages 375–384, Cham, 2018. Springer International Publishing.

[10] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR*, abs/2001.07676, 2020.

Towards the specification of a self-exciting point process for modelling crimes in Valencia

Verso la definizione di un modello self-exciting per modellare i crimini a Valencia

Marcello Chiodi, Nicoletta D'Angelo, Giada Adelfio and Jorge Mateu

Abstract A number of papers have dealt with the analysis of crime data using self-exciting point process theory after the analogy drawn between aftershock ETAS models and crime rate. With the aim to describe crime events that occurred in Valencia in the last decade, in this paper, we justify the need for a self-exciting point process model through spatial and temporal exploratory analysis.

Abstract *Numerosi articoli si sono occupati dell'analisi dei dati sulla criminalità utilizzando la teoria dei processi di punto self-excting, dopo l'analogia tracciata tra i modelli ETAS e la criminalità. Con l'obiettivo di descrivere gli eventi criminali avvenuti a Valencia nell'ultimo decennio, in questo articolo giustifichiamo la necessità di un modello di processo puntuale self-exciting attraverso un'analisi esplorativa spaziale e temporale.*

Key words: Covariates; Crime data; Hawkes processes; Self- exciting point processes; Spatial Statistics; Spatio-temporal point processes

1 Introduction

Following the comparison made by Mohler et al. (2011) between crime and aftershock ETAS models, a number of articles have addressed the analysis of crime data using self-exciting point process theory. Given that crime data is typically clustered, several articles have specifically suggested a Hawkes-type point process modelling approach for these sort of data (Reinhart and Greenhouse, 2018). Examples of appli-

---

Marcello Chiodi, Nicoletta D'Angelo and Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy e-mail: `marcello.chiodi@unipa.it;nicoletta.dangelo@unipa.it;giada.adelfio@unipa.it`

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain e-mail: `mateu@uji.es`

cations incorporating the external information in Self-Exciting models can be found in Adelfio and Chiodi (2020). In particular, Park et al. (2021) modelled gang-related violent crimes in Los Angeles (California) using spatio-temporal Hawkes processes. Lastly, D'Angelo et al. (2022) has taken into account the self-exciting behaviour of points, proposing a spatio-temporal Hawkes point process model adapted to linear networks to analyse crime data in Bucaramanga (Colombia). In this work, we use temporal and spatial exploratory analysis to support the necessity for a self-exciting point process model in order to describe criminal occurrences that occurred in Valencia over the past ten years.

The outline of the paper is as follows. Section 2 presents the data. In Section 3, we introduce the self-exciting point process model. Section 4 contains the exploratory analyses to justify the need of such self-exciting model for the description of the crime data. Finally, future works are outlined in Section 5.

2 The data

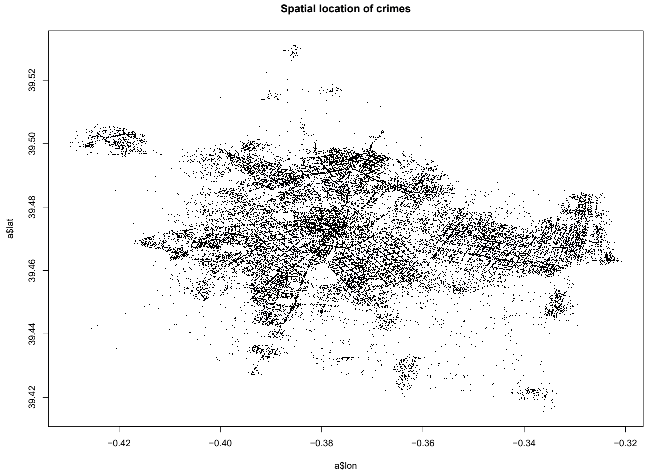


Fig. 1: Spatial location of the 78041 crimes occurred in Valencia from 2011 to 2019.

The analysed dataset is related to Valencia city, a town located in the Mediterranean sea with one million inhabitants. The original data time range is 2010-2020, but we restricted the analysis to the 2011-2019 period only, due to missing data and to avoid the lockdown period. The total number of events is, therefore, 78041 (out of the 90247 original ones). The crimes are categorised into four types: theft after hitting a person (55610), smooth theft with no force used (25342), theft to a woman with violence (454), and other thefts or robberies that can not be considered as in the previous categories (8841). Given the similarity between categories and the possible heterogeneity inside the latter category, we decided to merge theft to a woman with violence into the “other” category. Note that we used this classification for exploratory analysis only, but we do not aim at including the crime type information in the modelling strategy. The analysed events are displayed in Figure 1. Variables based on distances from events to “places of concentration” are also available, including the Euclidean distance from the nearest: atm, bank, bar, cafe, industrial site, marked, nightclub, police, pub, restaurant, or taxi. It is reasonable to assume that the proximity to some of these locations could influence the occurrence of crimes, and therefore we include such variables in our modelling strategy.

3 Methodology

Point processes can be formally specified in several ways, for instance, by considering the joint distribution of the counts of points in arbitrary sets or by defining a complete intensity function. To model events that are clustered, self-exciting point processes are often used. Examples include Hawkes models (Hawkes, 1971). The conditional intensity function of a linear self-exciting process is defined as the sum of two non-negative functions: a background that describes the large-scale variation of the intensity, and a triggered component, which describes its small-scale variation due to the interaction with the events in the past. Let N be a point process $\{(t_i, x_i, y_i) : i = 1, \ldots, n\}$ on a spatio-temporal domain $X = W \times T \subseteq \mathbb{R}^2 \times \mathbb{R}_+$, with area $|W| > 0$ and length $|T| > 0$, and with t representing the time, and x and y the two spatial coordinates. We aim at fitting a self-exciting model with a semi-parametric specification as follows

$$\lambda(t,x,y) = \mu_t(t)\mu_s(x,y) + \int_{-\infty}^{t-} \int_W g(t-\tau)h(x-u, y-v)N(\mathrm{d}u \times \mathrm{d}v \times \mathrm{d}\tau), \qquad (1)$$

where $\mu_t(t)$ represents the trend term in the temporal components of the background rate, $\mu_s(x,y)$ represents the spatial background rate, and $g(t-\tau)h(x-u, y-v)$ represents the sub-process triggered by an event previously occurring at time τ and location (u,v). Models of this kind are usually fitted by means of the E-M estimation algorithm. We employ a separable form of the spatial and temporal terms in equation (1) to simplify the underlying mathematics and to look for some pragmatic strategy. This formulation further allows interpreting separately the temporal and spatial components, peculiar to the description of our data. In the next section, along with some exploratory analysis, we show the data support our election.

4 Data analysis

We start the data analysis justifying the separability assumption between the spatial and temporal components. A common way to empirically confute this hypothesis is to assess graphically whether the spatial distribution of points changes conditionally to different time stamps. Figure 2 shows some kernel space intensity estimates by month.

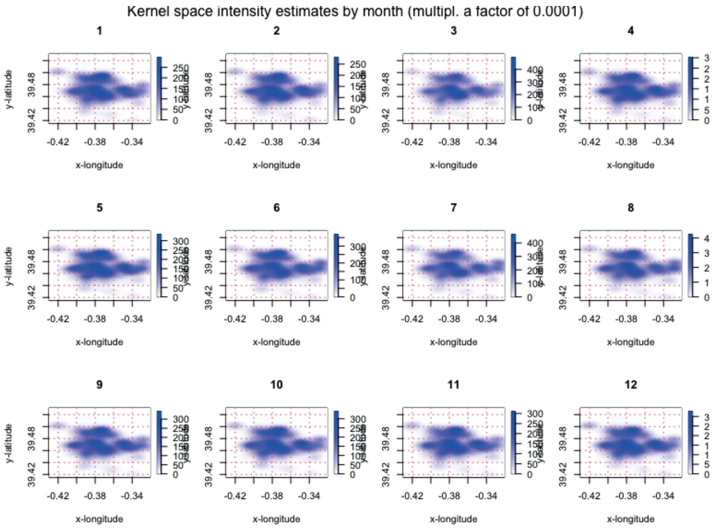


Fig. 2: Kernel spatial intensities estimated by month of the year.

Basically, we do not observe any relevant changes in the distribution of points when looking at different months of the year. The kernel spatial intensities by year, weekday and hour, also estimated but not shown here for brevity, showed the same behaviour. Therefore we proceed by assuming separability between the spatial and temporal occurrence of points.

Next, we perform some exploratory analyses by inspecting the temporal occurrence of crimes. Figure 3 displays the distribution of the number of crimes according to the type of crimes as a function of the year, month, weekday and hour of the day. It is evident that there is a prevalence in March, and in the summer months. This motivates us to think that the number of crimes can be affected by the different number of tourists as Valencia hosts the Las Fallas in March. At the time of writing, no tourist data at detailed space and time resolution are available. Their dependence, which is quite natural to expect, remains a key aspect to study in future work. Preliminary analyses at an aggregate level showed a correlation between crimes and tourist counts. Regarding the distribution by weekday, it looks evident that there is a prevalence on Friday and Saturday. Also, the hourly pattern appears relevant in determining the occurrence of crimes.

The temporal component $\mu_t(t)$ of equation (1) can be fitted by means of a Poisson log-linear model. After comparing different specifications with the AIC values, we found out that the best model contains all the temporal components (year, month,

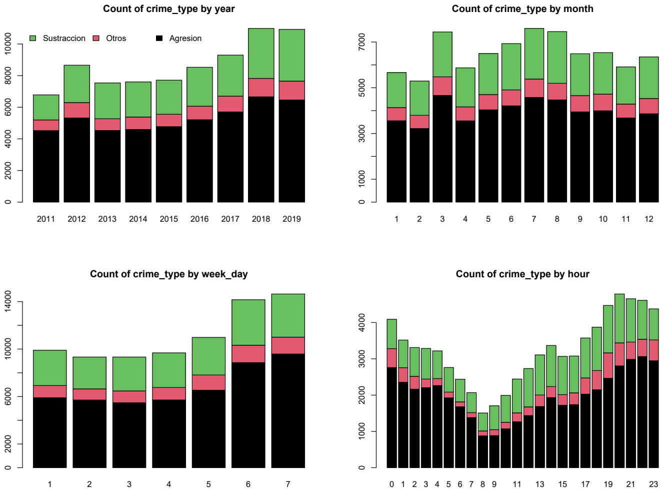


Fig. 3: Distribution of the number of crimes according to the type of crimes as a function of year, month, weekday and hour of the day.

weekday, and hour). Moreover, considering both the *serial month* and the *serial hour*, further decrease the information criterion used to select the best model. These new variables are obtained as: $\texttt{serial hour} = (\texttt{weekday} - 1) * 24 + \texttt{hour}$; $\texttt{serial month} = \texttt{month} + 12 * (\texttt{year} - 2011)$. Finally, the best model for the temporal trend is

$$\hat{\mu}_t = \exp(\hat{\alpha} + \hat{\beta}_1 * \texttt{serial month}_t + \hat{\beta}_2 * \texttt{serial hour}_t).$$

Finally, for justifying the need for a triggering component in our model, we compute the differences in the time occurrence for subsequent events (denoted `dt` henceforth). We also compute the spatial distance between such events (`ds`). We then fit a Gamma GLM with inverse link function, mainly to assess the effect of `ds` on `dt`. After comparing different specifications of such models, including some covariates, the best model is

$$\hat{\texttt{dt}}_i = (\hat{\alpha} + \hat{\beta}_1 * \texttt{weekday}_i + \hat{\beta}_2 * \texttt{hour}_i + \hat{\beta}_3 * \texttt{march}_i + \hat{\beta}_4 * \texttt{ind.dist}_i + \hat{\beta}_5 * \texttt{ds}_i)^{-1}. \tag{2}$$

The variable `march` is dichotomous, indicating the event occurrence in March. We found that including this variable improves the fitting of the model, than including the categorical `month` variable. The main result here is that `ds` is significant and with a negative estimated coefficient. This basically means that the closer the occurrence of subsequent events, the closer also their time occurrence. In other words, the spatial vicinity of crimes increases the vicinity in time. We also checked whether the contrary was true, that is to say, whether the vicinity in time could influence the vicinity in space. We basically fit the same Gamma glm with inverse link function, but with `ds` as the response variable. As expected, the best model among the competitor ones contains different covariates than (2), here listed: weekday, hour, march, the type of crime, and many other spatial covariates among which the distance from the nearest atm, bank, bar, cafe, industry, market, nightclub, police station, pub, restaurant, taxi. The main result is that also `dt` is significant in explaining `ds`, also with a negative coefficient. We take these results as the implicit evidence of the need for a triggering component in our model.

5 Future work

In this work, we have performed an exploratory analysis of the occurrence of crimes in Valencia from 2011 to 2019. With the aim of proposing a Hawkes-type self-exciting point process model, we first explore whether the separability assumption between the spatial and the temporal component is reasonable. Then, we have established that the self-exciting behaviour of points is actually needed, by means of a Gamma glm for modelling the time differences between subsequent events. The results motivate us to believe that a self-exciting model could be appropriate for both describing the data and predicting future events.

References

Adelfio, G. and Chiodi, M. (2020). Including covariates in a space-time point process with application to seismicity. *Statistical Methods & Applications*, pages 1–25.

D'Angelo, N., Payares, D., Adelfio, G., and Mateu, J. (2022). Self-exciting point process modelling of crimes on linear networks. *Statistical Modelling*, page 1471082X221094146.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Park, J., Schoenberg, F. P., Bertozzi, A. L., and Brantingham, P. J. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates. *Journal of the American Statistical Association*, pages 1–14.

Reinhart, A. and Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329.

A Clusterwise regression method for distributional data

Antonio Balzanella and Rosanna Verde and Francisco de A.T. de Carvalho

Abstract The focus of our work is the development of a clusterwise algorithm which utilizes a novel regression method for distributional data. Our approach involves creating a more effective regression model for distributional data by using a Logarithm transformation of the Derivative Quantile functions (LDQ) to map density functions into a Hilbert space. By using this new LDQ-based regression model, our algorithm can predict the response variable by dividing the set of objects into K clusters, where each cluster is associated with a local regression model that best fits the data.

Key words: Clusterwise regression, distributional data, quantile density functions

1 Introduction

This paper describes a new regression method and a clusterwise strategy for distributional data. The input data consists of distributional variables with a response variable and predictors represented by probability functions or empirical ones. The proposed approach involves partitioning the set of distributional-valued data into subgroups using a K-means-like clustering algorithm. The centroids of the clusters

---

Antonio Balzanella
Dept. of Mathematics and Physics, University of Campania Luigi Vanvitelli, 81100 Caserta, Italy
e-mail: antonio.balzanella@unicampania.it

Rosanna Verde
Dept. of Mathematics and Physics, University of Campania Luigi Vanvitelli, 81100 Caserta, Italy
e-mail: rosanna.verde@unicampania.it

Francisco de A.T. de Carvalho
Centro de Informatica, Universidade Federal de Pernambuco,
Av. Jornalista Anibal Fernandes s/n - Cidade Universitaria, Recife-PE, Brazil e-mail: fatc@cin.ufpe.br

are represented by linear regression models, and objects are assigned to the clusters based on the minimum sum of squared errors.

The first contribution of the paper consists in defining a more suitable regression model for distributional data. Previous methods, such as [7] and [4],used Non-Linear Least Squared method and the Wasserstein metric in a linear space, but imposed the constraint of non-negativity to ensure that the outcome remains a distributional variable. The proposed method addresses this issue by transforming the quantile functions into log quantile density functions, as described in [8], allowing for a mapping of density functions into a Hilbert space, resulting in a regression model that is better suited to Distributional Data Analysis (DDA).

The second contribution is a clusterwise method based on the new regression model using the LDQ transformation. This method predicts the response variable by partitioning the set of objects into K clusters, with each cluster represented by a regression model.

The clustering process involves two alternating steps. Fixed a predefined K number of clusters, the first steps involves representing the clusters with regression models; the second steps allocates the elements to clusters according to the minimum sum of the squared errors. The two steps are repeated until convergence to stable clusters is achieved.

2 Regression linear model for Distributional-valued Data

This section describes a new regression model that is used to analyze distributional-valued data. The main issue to address is that probability density functions, cumulative distribution functions and their inverse, quantile functions, are not in a Hilbert space.

Recently a suitable transformation has been introduced for mapping probability densities to a Hilbert space of functions through a continuous and invertible map [8]. A probability density function is firstly transformed by the derivative of its quantile function: $q(t) = \frac{dQ(t)}{dt} = \frac{dF^{-1}(t)}{dt}$, which is strictly positive and continuous in its domain $[0,1]$, and then, by the logarithmic transformation of $q(t)$:

$$l(t) = \ln q(t) = \ln \frac{1}{f(q(t))} = -\ln f(q(t))$$

where $l(t)$ denotes the logarithmic transform of the derivative of the quantile function (LDQ). On these functions it is possible to define the addition and scalar multiplication operations, the inner product and the Euclidean norm. The squared Euclidean distance between two LDQ functions $l_1(t)$ and $l_2(t)$ is given by:

$$d^2_{LDQ}(l_1(t), l_2(t)) = |||l_1(t) - l_2(t)\|_2 = \int_0^1 (l_1(t) - l_2(t))^2 \, dt$$

The LDQ transform has a limitation in that it does not preserve information on the location parameters of the distribution. This is because the derivatives of two quantile functions that differ by a constant term are equal.

A linear model on the LDQ transformation functions of $y_i(t)$ and x_{ij}(t) was proposed by [10]. However, this model did not effectively solve the problem of preserving information on the position of the density functions. The regression model, here proposed, attempts to overcome this issue by dividing the model into two regression models, as follows:

$$y_i^m = \beta_0^m + \sum_{j=1}^{p} \beta_j^m x_{i,j}^m + \varepsilon_i^m$$

$$y_i^l(t) = \beta_0^l(t) + \sum_{j=1}^{p} \beta_j^l x_{i,j}^l(t) + \varepsilon_i^l(t)$$

where: y_i^m is related to the minimum values x_{ij}^m; $y_i^l(t)$ and $x_{i,j}^l(t)$ are the LQD function of $y_i(t)$ and $x_{i,j}(t)$; ε_i^m is the i-th residual error of the minimum value; $\varepsilon_i^l(t)$ is the i-th residual error function of the LDQ function.

The parameters of the models are estimated by minimizing the sum of squared errors, by:

$$SSE_{OLS-LDQ} = \int_0^1 \left[\left(\tilde{\mathbf{y}}^l(t) - \tilde{\mathbf{X}}^l(t)\beta^l \right)^\top \left(\tilde{\mathbf{y}}^l(t) - \tilde{\mathbf{X}}^l(t)\beta^l \right) \right] dt + \\ + \left(\mathbf{y}^m - \mathbf{X}^m\beta^m \right)^\top \left(\mathbf{y}^m - \mathbf{X}^m\beta^m \right) \quad (1)$$

The least squares estimators for $\beta^l(t)$ and β^m represent the solutions of the two independent systems.

We propose the algorithm CRL-LDQ which is a method combining the Dynamic Clustering Algorithm [6] and the OLS-LDQ regression method for distributional-valued data. For a fixed number K of clusters, it looks for the best partition $P_k = C_1, \ldots, C_K$ and the best fitting models $\hat{y}^k$, for each cluster C_k, by minimising the $SSE_{OLS-LDQ}(P_k, \hat{y}^k)$:

$$SSE_{OLS-LDQ}(P_k, \hat{y}^k | \beta_{j(k)}^l, \beta_{j(k)}^m) = \sum_{k=1}^{K} \sum_{e_i \in P_k} \left[\|\tilde{\varepsilon}_{i(k)}^l(t)\|^2 + (\varepsilon_{i(k)}^m)^2 \right]$$

The algorithm is performed alternating two steps, until the convergence to a stable value.

- Set the number K of clusters.

- *Step 1 - Representation step* (best fitting):
  The local regression models are estimated by minimizing the objective function $SSE(P_k, \hat{y}^{k*})$ on the parameters $\beta_{(k)}^l(t)$ and $\beta_{(k)}^m$ $(1 \leq k \leq K)$.
  The OLS estimations are provided as solutions of the two independent systems.

- *Step 2 - Assignment step* (partitioning P_k):
  The optimal clusters P_k which minimize the criterion $SSE(P_k^*, \hat{y}^k)$, are obtained according to the following assignment rule:

$$P_k = \left\{ e_i \in E : \left[\|\hat{\varepsilon}^l_{i(k)}(t)\|^2 + \left(\hat{\varepsilon}^m_{i(k)}\right)^2 \right] = \min_{h=1}^{K} \left[\|\hat{\varepsilon}^l_{i(h)}(t)\|^2 + \left(\hat{\varepsilon}^m_{i(h)}\right)^2 \right] \right\}$$

  Thus, the observation e_i is assigned to cluster P_k if the sum-of-squared errors are minimal for this cluster regression model.

The convergence of the algorithm is guaranteed by the decreasing of the criterion according to the improvement of the cluster regression models fitting.

3 Preliminary results on simulated data

The proposed method's efficacy was tested using simulated data. The performance of the regression model was evaluated by measuring its ability to fit the data, which was determined by using values of K ranging from 1 to 4. A value of $K = 1$ represents the use of the model without partitioning the data into clusters. The simulated data consisted of 600 individuals and included one response variable and two predictors, which were generated from Gamma random variables with varying parameters. The response was a combination of the predictors with added Gaussian error. A pseudo R^2 index was calculated using a suitable decomposition of the deviance, and the result was 0.59 for $K = 1$. When the cluster-wise procedure was introduced, the results improved to 0.61, 0.74, and 0.95 for $K = 3$, which corresponded to the number of clusters in the data.

References

1. Bock H., Diday E.: Analysis of Symbolic Data. Springer Berlin, Heidelberg (2012)
2. Brito P., Dias S.: Analysis of Distributional Data. Chapman Hall (2022)
3. de Carvalho, F.d.A., Saporta, G., Queiroz, D.N. A: Clusterwise Center and Range Regression Model for Interval-Valued Data. In: Lechevallier, Y., Saporta, G. (eds) Proceedings of COMPSTAT'2010. Physica-Verlag HD. (2010)
4. Dias, S. and Brito, P.: Linear regression model with histogram-valued variables. Statistical Analy Data Mining, 8: 75-113 (2015) https://doi.org/10.1002/sam.11260
5. Diday E.: Introduction l'analyse factorielle typologique, Revue de Statistique Applique, XXII(4), 29-38, (1974)
6. Diday E.,Simon J.C.: Clustering analysis. Digital Pattern Recognition, 47-94 Springer (1980)
7. Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. Adv Data Anal Classif 9, 81106 (2015). https://doi.org/10.1007/s11634-015-0197-7
8. Petersen A., Mller H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics, Ann. Statist. 44(1), 183-218, (2016)

9. Spaeth, H.: Clusterwise Linear Regression. Computing 22 (4), 367373 (1979).
10. Zhao Q., Wang H., Lu S.: M-LDQ feature embedding and regression modeling for distribution-valued data. Information Sciences, Volume 609 (2022). https://doi.org/10.1016/j.ins.2022.07.064.

Increasing accuracy in classification models for the identification of plant species based on UAV images

Aumentare l'accuratezza nei modelli di classificaizone per l'identificazione delle species vegetali basati su immagini UAV

Anna Simonetto, Girma Tariku, Gianni Gilioli

Abstract The identification of plant species from RGB images is a challenge of growing importance in the field of biodiversity assessment. This study aims to develop an image pre-processing procedure that increases the accuracy of classification models applied to low-resolution plant images collected by RGB Unmanned aerial vehicles (UAVs). This procedure, based on contrast enhancement and super-resolution techniques, has been successfully tested on RGB images collected in agroecosystems.

Key words: Accuracy, Classification procedures, RGB images, Neural networks, K nearest neighbours

1 Introduction

Biodiversity assessment is a crucial aspect of the sustainable management of natural capital. In order to quickly and accurately analyse the biodiversity in wide areas, Unmanned aerial vehicles (UAVs) are increasingly being used due to their high mobility and ability to cover areas at different altitudes and locations with relatively lower costs [16]. The drawbacks is that at a high altitude, UAVs will have a low spatial resolution and it could make more difficult to detect features on plants [2].

Anna Simonetto
Department of Civil Engineering, Architecture, Land and Environment, and Mathematics (DICATAM), University of Brescia, via Branze 43, Brescia, e-mail: anna.simonetto@unibs.it

Girma Tariku
Department of Information Engineering (DII), University of Brescia, via Branze 38, Brescia e-mail: g.tariku@unibs.it

Gianni Gilioli
Department of Civil Engineering, Architecture, Land and Environment, and Mathematics (DICATAM), University of Brescia, via Branze 43, Brescia e-mail: gianni.gilioli@unibs.it

Furthermore, images taken by UAVs have shadows due to terrain factors [14]. The image pixel brightness of the shadow areas is compressed, and the information is deficient, which impacts the recognition of image information and thus limits the subsequent image application. Due to the development of different machine and deep learning topologies over the last few years, classification methods for plant identification from RGB images are focused on supervised learning techniques [7] by using machine learning tools [12] or transfer learning tools [13].

In our paper, we propose a pre-processing step based on image contrast enhancement [3] and super-resolution (SR) [15] image preprocessing technique for the training image dataset to improve the classification accuracy of low-resolution plant images of RGB UAVs. Although using high-quality images for image analysis would be ideal, this is not always possible in practice (e.g., attempting to identify relatively small objects in remote sensing imagery [10]). In these cases, performing transformations to increase image quality may prove useful in the attempt to identify and classify less salient objects in the imagery [3].

2 Procedure for plant image classification

The procedure used to generate the training image dataset is graphically illustrated in figure 1. The first step to produce an orthomosaic map of the studied area is collecting RGB images. We use DroneDeploy [9] to design an automated flight plan to take aerial pictures perfect for producing orthomosaic maps and 3D models. Then, we merge RGB images in the collection by looking for ”link points” and we use a digital surface model (DSM) to produce orthomosaic images from the drone photos. The orthophoto is segmented into useful single-picture objects. eCognition Developer v9.0.0 software [11] is used to segment the tree canopies applying the multi-resolution segmentation algorithm [1] to Orthomosaic image and DSM. Scale, shape, and compactness are the segmentation criteria. Following the segmentation process, plant species mapping is carried out for a limited set of pixels of image objects using the eCognition software, taking into account the spatial features of the image objects with regard to one another. A set of pixels would in this way act as a training sample for the classification technique known as supervised object-based classification [6]. We apply supervised learning using K nearest neighbours (KNN) to map plants. A class prediction for each group of pixels would then be generated by the KNN classification technique. Then, the training dataset plant pictures are extracted from the Orthomosaic image of the ground truth map with a class label. In the final step, we perform the classification using pre-trained transfer learning models.

The procedures proposed in this study focused on two steps (Image Contrast Enhancement and Super-Resolution preprocessing) as key elements to improve the classification accuracy of plant species. These steps are briefly described in the following sections.

2.1 Image Contrast Enhancement

Contrast enhancement is adjusting the dynamic range of pixel intensity distribution for good contrast enhancement in images facing low contrast concerns. we used an Image contrast enhancement technic for improving extracted UAV training dataset-images quality by decreasing the impact of distortions (i.e., blur, shadows, contrast issues, and noise) contained therein. The histogram of an image is an approximation of the pixel values distribution and can be calculated as:

$$h[i] = n_k/(M \times N) \qquad k \in [0,..,L-1]; \tag{1}$$

where I is the intensity of the pixel, L is the number of values (different intensities) that can be assumed by each pixel, n_k is the number of pixels with intensity k, $M \times N$ is the number of image pixels.

Starting from the consideration that good contrast images have a histogram close to uniform distribution U(0, L-1), we apply the following procedure:

- Compute the Image Histogram (probability distribution)
- Compute the Cumulative Distribution $F[k] = \sum_{i=0}^{k} h[i]$;
- Apply the point transformation $Ieq[i,j] = T[I[i,j]] = F[I[i,j]]$
- Rescale $Ieq[i,j]$ form $[0;1]$ to $[0;L-1]$.

where $Ieq[i,j]$ is the equalized image pixel intensity value at $(i = 0,...,M-1, j = 0,...,N-1)$, $T[I[i,j]]$ is point transformed image pixel intensity value at $(i = 0,...,M-1, j = 0,...,N-1)$ and $[I[i,j]]$ is the pixel intensity value K at $(i = 0,...,M-1, j = 0,...,N-1)$.

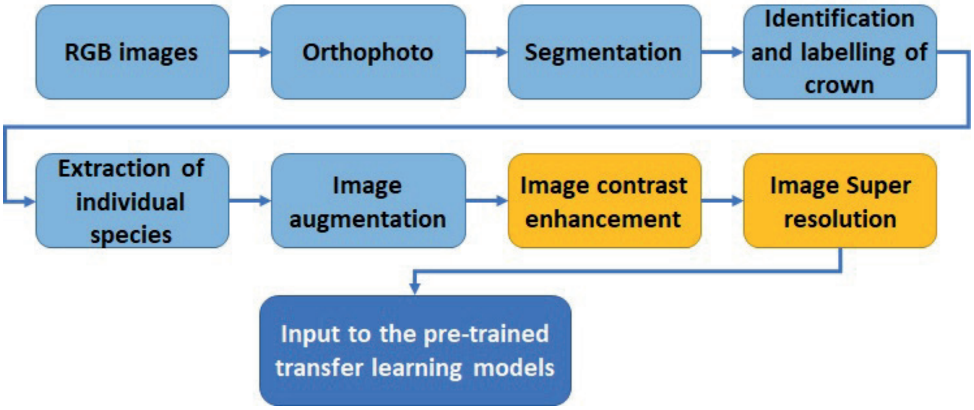

RGB images
Orthophoto
Segmentation
Identification and labelling of crown
Extraction of individual species
Image augmentation
Image contrast enhancement
Image Super resolution
Input to the pre-trained transfer learning models

Fig. 1 Processing procedure to classify plant species starting from RGB images. The steps analysed in this study (contrast enhancement and super-resolution image preprocessing techniques) are highlighted in orange.

2.2 Super-Resolution preprocessing

When training a super-resolution network with a per-pixel loss function, the goal is to minimize the per-pixel difference between the output and the ground truth image. When using the perceptual loss function, high-resolution images are generated by minimizing the difference between high-level image features of the output and ground truth, which are extracted from a pre-trained convolutional neural network. We propose an algorithm of single image super-resolution using a generative adversarial network based on the work of Ledig et al [4] and graphically described in Figure 2.

For the Generator Network, the input is a LR image with 9*9 kernels with 64 filters and ReLU. Then B residual blocks are applied and each block has 3*3 kernel with 64 filters followed by batch normalization and ReLU. Then two sub-pixel convolution layers are applied to the up-sample image to 4x. In the Discriminator Network, a discriminator will also discriminate real HR images from generated SR images. It contains eight convolution layers with an increasing number of 3x3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels. To reduce the image resolution, stride convolutions are applied each time the number of features is doubled. The resulting 512 kernel feature maps are followed by two dense layers and a final sigmoid activation function to obtain a probability for the real or fake image. For each layer of the neural network (orange blocks in Figure 2), we defined as activation function the Rectified Linear Unit (ReLu), and the value of each neuron needs to be calculated by the activation function to obtain a final value. In neural networks, the role of the activation function is to transform the neural network from linear to nonlinear, so that the neural network can better solve more complex problems.

Batch Normalization is applied to make neural networks faster and more stable adding extra layers in a deep neural network. The new layer performs the standardizing and normalizing operations on the input of a layer coming from a previous layer.

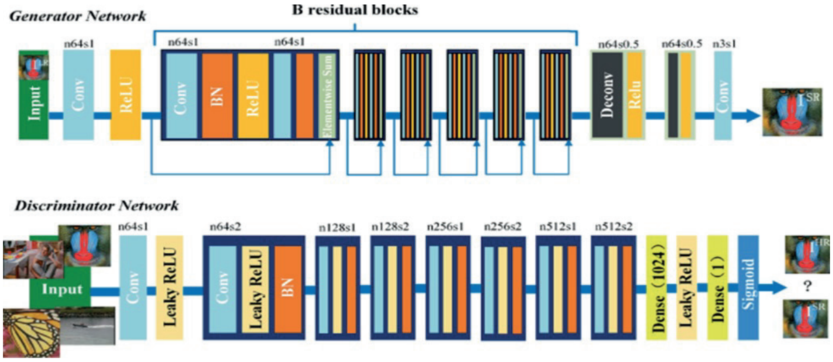


Fig. 2 Super resolution algorithm

We adopted mean squared error as loss function between content loss and adversarial loss. Adversarial loss tries to train the generator such that it produces natural-looking images which will be difficult for the discriminator to distinguish from real images.

3 Preliminary results

In Figure 3 we shown an application of the proposed algorithm to identify four plant species in an agroecosystem. We applied the ResNet50 pre-trained transfer learning model for four plant classes to 769 RGB UAV picture datasets. When we applied super-resolution and picture equalization, the classification accuracy jumped from 70.2% to 83.2%.

From our preliminary results, the proposed super-resolution and image contrast enhancement using histogram equalization methods improve the discriminative capacity of the applied classification model, helping to overcome the possible difficulties encountered when analysing low-resolution plant images collected by RGB UAVs.

References

1. Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M.: Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of photogrammetry and remote sensing, **58(3-4)**, 239–258 (2004)
2. Chen, J., Chen, Z., Huang, R., You, H., Han, X., Yue, T., Zhou, G.: The Effects of Spatial Resolution and Resampling on the Classification Accuracy of Wetland Vegetation Species

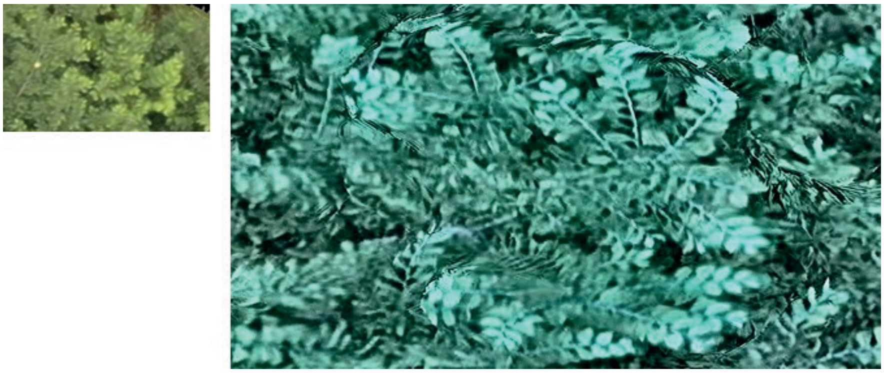


Fig. 3 Preliminary results of plant identification in an agroecosystem: on the left the original UAV image, on the right the equivalent super-resolution image using a generative adversarial network

and Ground Objects: A Study Based on High Spatial Resolution UAV Images. Drones. **7(1)**, 61 (2023)

3. González, D., Patricio, M. A., Berlanga, A., Molina, J. M.: A super-resolution enhancement of UAV images based on a convolutional neural network for mobile devices. Personal and Ubiquitous Computing. 1–12, (2019)
4. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4681–4690 (2017)
5. Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y.: A review of supervised object-based land-cover image classification. ISPRS Journal of Photogrammetry and Remote Sensing. **130**, 277–293 (2017)
6. Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y.: A review of supervised object-based land-cover image classification. ISPRS Journal of Photogrammetry and Remote Sensing. **130**, 277–293 (2017)
7. Nasteski, V.: An overview of the supervised machine learning methods. Horizons. b. **4**, 51–62 (2017)
8. Natesan, S., Armenakis, C., Vepakomma, U.: Resnet-Based Tree Species Classification Using Uav Images. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. **XLII-2/W13**, 475–481 (2019)
9. Putch, A.N.D.Y.: Linear measurement accuracy of DJI drone platforms and photogrammetry. San Francisco: DroneDeploy (2017)
10. Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., Chao, D.: Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. Remote Sensing. **12(9)**, 1432 (2020)
11. Trimble Germany GmbH: Trimble Documentation eCognition Developer 10.1 Reference Book; Trimble Germany GmbH: Munich, Germany (2021)
12. Vercio, L. L., Amador, K., Bannister, J. J., Crites, S., Gutierrez, A., MacDonald, M. E., ... Forkert, N. D.: Supervised machine learning tools: a tutorial for clinicians. Journal of Neural Engineering. **17(6)**, 062001 (2020)
13. Weiss, K., Khoshgoftaar, T. M., Wang, D.: A survey of transfer learning. Journal of Big data. **3(1)**, 1–40 (2016)
14. Xi, W., Zuo, X., Sangaiah, A. K.: Enhancement of Unmanned Aerial Vehicle Image with Shadow Removal Based on Optimized Retinex Algorithm. Wireless Communications and Mobile Computing. (2022)
15. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: The techniques, applications, and future. Signal processing. **128**, 389–408. (2016)
16. Zhang, C., Kovacs, J. M.: The application of small unmanned aerial systems for precision agriculture: a review. Precision agriculture. **13**, 693–712 (2012)

Travel time to university as determinant on students' performances

Arianna Burzacchi, Lidia Rossi, Tommaso Agasisti, Anna Maria Paganoni, and Simone Vantini

Abstract This research aims to investigate the impact of home-university travel time on university students' performance. The study uses a two-step process where, first, home-university travel times are estimated using a data-driven methodology and then evaluated for their impact on students' performance using statistical methods. The accessibility to university is measured using GPS data of anonymous individuals in the Comune di Milano collected by Cuebiq Inc. The estimated commuting time to university is used to create isochronous maps with Kernel Regression Estimation. The second phase uses personal and academic information of first-year engineering students from Politecnico di Milano to investigate the effect of travel time on performance. In this step Linear Mixed Effect methods are applied. The study is innovative as it is one of the first to focus on the topic and, in particular, it uses a new model for the estimation of commuting time.

Key words: University accessibility; Commuting time; GPS data; Higher education; Students' performance.

Arianna Burzacchi*, Anna Maria Paganoni, and Simone Vantini
MOX Laboratory, Department of Mathematics, Politecnico di Milano - Piazza Leonardo da Vinci 32, 20133 Milano, Italy
* arianna.burzacchi@polimi.it

Lidia Rossi**, and Tommaso Agasisti
Department of Management, Economics and Industrial Engineering, Politecnico di Milano - Via Lambruschini 4/B, 20156, Milano, Italy
** lidia.rossi@polimi.it

1 Introduction and Literature Review

Factors that affect university students' performance are multiple. These determinants range from personal characteristics of students, such as gender or nationality, to characteristics not observable, such as motivation or innate ability, but also factors of the context surrounding students, such as the school environment, family context and accessibility to educational institutions [4]. A possible measure of accessibility is the travel time, which refers to the time that a student takes to reach the university from home. The travel time is precisely the subject of this study, and the objective is to investigate the effect of home-university travel times on students' university performance.

In literature only few studies deal with assessing the impact of travel time on students' outcome and face the problem of how to estimate that time. The works by Kobus et al. (2019) [6] and Tigre et al. (2017) [10] use the most common and simple practice to collect this information that is via questionnaires. However, it has been shown that respondents tends to under-report their commuting time [9]. Falch et al. (2013) [3] estimate the travel time by assuming that journeys are made by car over the road network at the maximum velocity allowed by the road speed limits. In [2] Contreras et al. (2018) use a two-step process: firstly they compute the travel time from students' home and their university in three mode of transport (by car, by public transport and by foot) using Google Maps API, then, with Machine Learning algorithms and information about personal characteristics of students and home-school distance, they predict the mode of transport of the students. The main limitation of these two studies, however, is the need to make assumptions about the means of transport used, which may limit their validity for certain urban areas.

2 Methods and Data

The analytical framework develops in two phases. The first phase involves building isochronous maps of Milan starting from the GPS trajectories of anonymous users. Using the students' home addresses and the maps built, it's possible to estimate an arrival time at the university for each student. This information is combined with the student's collateral information to evaluate using statistical models if it's a determinant of the student's university performance (Fig. 1).
A detailed description of the method and its application can be found in Burzacchi et al. (2023) [1].

2.1 First phase: Estimate travel times

The aim of the phase is to create accessibility maps to the two university campuses of Politecnico di Milano, i.e. Leonardo and Bovisa campuses, and to use them as

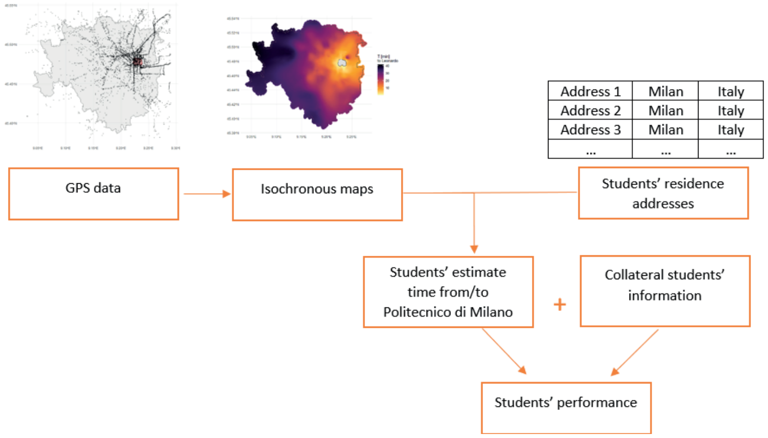

Address 1 | Milan | Italy
Address 2 | Milan | Italy
Address 3 | Milan | Italy
... | ... | ...
GPS data
Isochronous maps
Students' residence addresses
Students' estimate time from/to Politecnico di Milano
+
Collateral students' information
Students' performance

Fig. 1 A graphical representation of the conceptual framework

a proxy for the home-university travel time spent by students. The accessibility to university is evaluated using GPS data to create data-driven isochronous maps of travel times to university, i.e. to build travel time estimates from any point of origin within the city of Milan to the university sites of Politecnico di Milano.

The dataset of GPS traces was collected and provided by Cuebiq Inc., a firm specialised in location intelligence, and provided for research as part of their "Data for Good" programme. The dataset consists of records of geo-localisations of mobile phones that opted-in to anonymized data collection for research purposes through a GDPR compliant framework. Aligning with the project objectives, the original dataset is reduced and only the period and area of interest is considered. The selected observations are located in the Comune di Milano and its surroundings during one month of the first academic semester 2019/20, i.e. in the period from December 1st to December 20th 2019. A further reduction of the dataset is applied to consider only observations of working days (from Monday to Friday) and morning peak hours (arrival between 7:30 and 9:30), under the assumption that students' journeys to university are made for the first lesson of the day. Another necessary reduction is to consider GPS traces of users whose journey end at one of the two campuses of Politecnico di Milano, so that the final time estimation would be based GPS data of actual travellers to Politecnico di Milano. In this framework, a journey is defined as the representation of a user spatial movement through a sequence of GPS data points, ordered with respect to the timestamp, from an origin moving point to a stopping point. The selected identification criteria for journeys from GPS raw data set global constraints on distance, time and velocity between consecutive GPS measurements, applicable for all users, and additional device type-based constraints, which differ according to user's smartphone operating system.

Two reduced datasets are hence obtained, one for each university campus. Regarding the Leonardo campus, the reduced dataset comprises 10 458 GPS data points related to 1 074 journeys of 594 unique Cuebiq users. The number of journey per user ranges from 1 to 11, with an average of 1.8. Each journey collects a number of observations varying from 6 to 85, with 9.5 on average, describing trips of length from 330m to 63km, with 9.6km on average. Bovisa campus yields similar results.

Students' home-university travel times are finally estimated through the development of non-parametric geospatial statistical methodology for GPS data. The isochronous maps of travel times to Politecnico di Milano are created using Kernel Regression Estimation through Nadaraya-Watson Estimator ([7]; [11]) with Gaussian kernel and generalized nearest-neighbor bandwidth ([8]; [5]). The estimation performance is evaluated on a test set of 15% of journeys. For the case of Leonardo campus, the test set comprised 1 578 observations of 161 journeys, and the estimate achieved a Mean Absolute Error of $5.5min$ and a Mean Square Error of $77.9min^2$. Kernel regression is finally used to make predictions about the values of students' commuting time to university at the locations of their residence address, by looking at the values of commuting time observed at sampled locations of GPS traces.

2.2 Second phase: Investigate impact of travel time on students performance

The project focuses on first year engineering students at Politecnico di Milano in the academic year 2019/20. Only students resident in Milan are selected, finding a total of 669 observations. For each student, there are known the number of credits passed, the marks average obtained in the exams, and the number of exams passed at the end of the first semester of that academic year. Also known is the student's personal information, such as gender, age and family income; information regarding previous career, such as final grade obtained in high school and track followed; information regarding current career, i.e. admission grade and track followed in university; and, finally, through the students' address of residence and the analyses carried out in the first phase of the project, estimated commuting time to Politecnico di Milano. The descriptive statistics and description of the variables used in this phase of the study are reported in Table 1.

Linear Mixed Effect methods and Generalized Linear Mixed Effect methods are applied with the objective of investigating whether the travel time is one of the determinants of performance and, if so, to what extent. Students are nested in university track and the models proposed are two level models. Various models are analysed with different output variables: probability of passing an exam, marks average, number of credits passed and number of exams passed.

For instance, for each student i, $i = 1, \dots, n_l$; $n = \sum_l n_l$, in the track l, $l = 1, \dots, L$ the model for marks average can be written as:

$$y_{il} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{kil} + u_l + \varepsilon_{il} \quad \text{with } u_l \sim N(0, \sigma^2_{track}),\ \varepsilon_{il} \sim N(0, \sigma^2_{\varepsilon}) \tag{1}$$

where y_{il} is the output variable of student i, in track l; $\boldsymbol{\beta} = \{\beta_0, \dots, \beta_K\}$ is the

Table 1 Students' variables description and descriptive statistics

Variable	Description	Type	Values
Number of passed exams	Number of passed exams by the student during the first semester of 2019/20	Num.	[0;5]; mean=1.67; sd=1.32
Number of CFU passed	Number of CFU passed during the first semester of 2019/20	Num.	[0;37]; mean=15.09; sd=11.81
Passed exams	The student pass at least one exam during the first semester of 2019/20	Bin.	1= 507
Marks mean	Average of marks obtained by the student during the first semester of 2019/20 (baseline 18) [only for students that pass at least one exam]	Num.	[0-12]; mean=5.58; sd=3.04
Commuting time	Estimated commuting time in the first step expressed in hour	Num.	[7-2643]; mean=0.38; sd=0.17
Admission score	Score obtained by the student in the admission test	Num.	[0.34-1]; mean=0.68; sd=0.10
Admission age	Age difference of the student at the time of admission to the university with respect to the baseline of 18 years	Num.	[-1;14]; mean=0.67; sd=1.23
Gender	Student's gender	Cat.	M:490
High school grade	Grade obtained at the end of the high school by the student	Num.	[0.6-1]; mean=0.81; sd=0.12
High school track	High school track attended by the student	Cat.	Cl: 44; Sc: 565; Tec: 47; Other: 13
Income	Income of student's family	Cat.	High: 380; Middle: 115; Low: 48; Grant: 126
Bachelor track (level)	Track of student's bachelor	Cat.	

$(K+1)-$dimensional vector of parameters; x_{kil} is the value of the $k-$th predictor at student's level; u_l is the random effect of the track l; ε_{il} is the error; and it is assumed that $\boldsymbol{u}$ is independent of $\boldsymbol{\varepsilon}$.

3 Discussion and further developments

This research is innovative both in terms of the subject matter and the model used to estimate home-university travel times. This is the first study to assess the travel time to university estimated using GPS data and a fully data-driven approach as determinant of university students' performance.

A potential extension of this work would be to expand its reach beyond just students resident in Milan. By applying the same methodology to student commuters, this study could provide valuable insights into the challenges faced by students across a wider geographical area. This expansion would provide a more comprehensive understanding of the impact of home-university travel times on student performance. Additionally, it would be interesting to compare the results across different cities and identify any regional patterns or differences. Such an extension of the work would not only deepen our understanding of the subject matter, but it would also demonstrate the versatility and applicability of the innovative model developed in this study.

4 Aknowledgements

The authors thank Cuebiq for providing the dataset of GPS traces.

References

1. A. Burzacchi, L. Rossi, T. Agasisti, A. M. Paganoni, and S. Vantini. Commuting time as a determinant of higher education students' performance: the case of Politecnico di Milano. Technical report, Politecnico di Milano, 2023.
2. D. Contreras Guajardo, D. Hojman Trujillo, M. Matas, P. Rodríguez Perales, and N. Suárez. The impact of commuting time over educational achievement: A machine learning approach. Technical report, Universidad de Chile. Facultad de Economía y Negocios, 2018.
3. T. Falch, P. Lujala, and B. Strøm. Geographical constraints and educational attainment. *Regional Science and Urban Economics*, 43(1):164–176, 2013.
4. E. A. Hanushek. Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, pages 351–388, 1979.
5. T. Hayfield and J. S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.
6. M. B. Kobus, J. N. Van Ommeren, and P. Rietveld. Student commute time, university presence and academic achievement. *Regional Science and Urban Economics*, 52:129–140, 2015.
7. E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.
8. B. Silverman. *Density Estimation*. Chapman and Hall, 1986.
9. P. Stopher, C. FitzGerald, and M. Xu. Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34:723–741, 2007.
10. R. Tigre, B. Sampaio, and T. Menezes. The impact of commuting time on youth's school performance. *Journal of Regional Science*, 57(1):28–47, 2017.
11. G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

The FAITH project: integrated tools and methodologies for digital humanities

Alfio Ferrara, Sergio Picascia, Elisabetta Rocchetti, Gaia Varese

Abstract Integration of many different sources and expertise is a key factor to solve specific research problems, especially in areas such as social sciences. The FAITH (Fight Against Injustice Through Humanities) project's main objective is to provide common tools and methodology for the collection, digitization and integration of different historical sources. In particular, the proposed solution involves the employment of an unique meta-model gathering information from different artifacts. Moreover, the FAITH project aims at applying data analysis techniques (e.g. bayesian networks, natural language processing techniques, image processing techniques) to provide insights about social issues in a diachronic perspective.
Abstract *Abstract in Italian*

Key words: digital humanities, data integration, data analysis.

1 Introduction to the FAITH Project

Social sciences are characterized by the need of an interplay between different expertise to support research activities. In particular, anthropological studies lean on many types of historical artifacts such as ancient manuscripts and skeletons. In such cases, integration between different sources is fundamental, but its technical implementation could represent an obstacle hindering research activities, especially when there is a lack of data management skills. Moreover, each discipline has its own methodology to tackle research problems: this approach naturally leads to the separation among studies carried by different specializations.
In order to provide common tools and techniques to approach anthropological research scenarios, the FAITH (Fighting Against Injustice Through Humanities) project aims at providing a methodology for the collection, digitization and integration of sources that can allow scientists to address crucial social issues through a

---

Università degli Studi di Milano, e-mail: name.surname@unimi.it

real interdisciplinary research. In particular, the project focuses on the extrapolation and digitization of archaeological, anthropological, medical, genetic, environmental, geological, documentary, literary, legal and artistic data, across different historical periods (Roman, Middle Ages, modern and contemporary), in a diachronic perspective. To show the potential and usefulness of such approach, the FAITH project is developing a case study considering signs of violence, abuse and discrimination in Milan in the XIII century.

2 Collecting and Integrating Data in the Humanities

Ease the interaction between researchers of different fields and allowing them to have an exhaustive perspective on the historical period of interest, are the main reasons for having a central storage where all the data converges. Unfortunately, these circumstances rarely arise in practise, especially in the humanities, where a considerable amount of resources and time is required in order to perform even simple analysis. This is mainly due to the difficulties encountered in collecting and integrating such kind of data.

When it comes to collecting data, there are two main procedures followed, depending on the format they are in: usually the data acquisition process is still performed with analog tools or, when the data is already digitized, it is organized following a very rigid and domain oriented schema, which makes the integration with other data sources difficult. We are in strict collaboration with experts of the different fields in order to assist them in digitizing already collected information and providing them specific tools in order to facilitate the acquisition of new data.

The fundamental step performed towards the accomplishment of these goals has been the definition of a flexible and general meta-model, capable of adapting and supporting the heterogeneity that distinguishes data in humanities. The model revolves around the idea of an abstract *ENTITY*, which then materializes in one of the following main sub-classes:

- *SOURCE*: it represents an historical data source from which are extracted all the pieces of information described in the database. Apart from the title, we are interested in characteristics, such the conservation status, the completeness or the material of which it is made;
- *EVENT*: it represents an historical event in which other entities appear. It can be described by a short summary and include a degree of certainty;
- *PERSON*: it is used to record data regarding people;
- *PLACE*: it is used to record data regarding geographical places, with the possibility of specifying also their coordinates;
- *ANIMAL*: it is used to record data regarding animals;
- *INSTITUTION*: it is used to record data regarding institutions;
- *OBJECT*: it is used to record data regarding objects;
- *APPELLATION*: it is used to register the names (appellations) of other entities, namely: *PLACE*, *PERSON*, *ANIMAL*, *INSTITUTION* and *OBJECT*;

- *GROUP*: it used to define a set of entities showing common characteristics.

Each of the previous entities can be in relationship with any other entity, even one of the same class. These relationships are shaped using the *TYPE* entity, determining which is the kind of relationship between the objects, together with *PLACE* and *TIME SPAN* entities, specifying where and when the relationship holds. Entities can be associated with one or more attachments, defined by the *ATTACHED* entity, that can be represented by any kind of media file, such as images, video or audio. Finally, it is possible to report the measurement of an aspects of an entity using a *METRIC*, which is associated to the value it assumes and its error, recorded in a certain *UNIT OF MEASURE*.
A practical example is shown in Figure 1. One textual *SOURCE* we imported in the database is the *Liber Mortuorum*, which contains records of dead people in the area of Milan around the XV century. A single record from this source, the 'Death of a PERSON', is inserted as an *EVENT* in which the dead person appears as an instance of the *PERSON* entity. The dead *PERSON* has her name recorded using an *APPELLATION* (i.e. *Petra*), her biological sex registered using the corresponding *GROUP* (i.e. *Female*) of a specific *TYPE* (i.e. *Biological Sex*) and her age measured with the *METRIC* 'Age' using 'Years' as *UNIT OF MEASURE*. The place in which the event takes place is registered as an instance of the *PLACE* entity having its own APPELLATION (i.e., S. Bartolomeo intus), while the date in which it happens is recorded with a *TIME SPAN*.

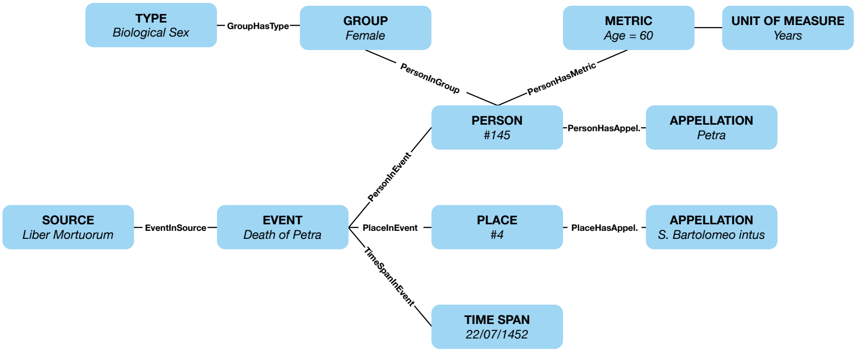


Fig. 1 The FAITH meta-model in practice: an example taken from the *Liber Mortuorum*.

3 Analysis Methods for the Digital Humanities

The application of data analysis techniques to digital humanities has a dual objective in our scenario: first of all, it allows experts to have an immediate high-level glimpse at the data at disposal, performing simple aggregation queries; on the other hand, complex statistical models can be built in order to model the intrinsic domain knowledge. Insights resulting from this last approach may also be included in the database, enriching even more the current available information.

Given the data at our disposal, we chose to focus mainly on probabilistic approaches. Indeed, our attention has been directed at integrating different data sources regarding the lifestyle of people in the area of Milan in the Late Middle Ages, in the attempt to reconstruct this specific historical context. We are trying to achieve this result employing Bayesian Networks, probabilistic graphical models that are able to exploit both tabular data and experts' knowledge in order to represent the information available. They are extremely effective at modeling the relationships elapsing between variables, computing the likelihood of certain scenarios, i.e. combination of evidences, and studying phenomena evolving over time with dynamic BNs [1].

They are particularly suitable for our task since the number of available tabular data is, at time, not sufficiently large for building an exhaustive network. BNs allow to support the phases of structure and parameter learning from data with the help of experts' knowledge: it is possible to define white and black lists in order to draw or avoid drawing edges between nodes, i.e. allowing for or denying dependencies relationships between variables; it is also possible to operate on the parameters, asking experts to estimate the values of conditional probabilities.

The expected outcome is a BN that allows us to generate 'what if' scenarios; in particular, we are in strict collaboration with anthropologists, studying how living conditions affected the modality with which people died at the time. This kind of analysis is interesting not only for inference purposes, but also in order to compare different time periods: for instance, we are currently observing how violent deaths distribute in Late Middle Ages with respect to now.

4 Ongoing Work and Future Prospects

At the present time, the FAITH project has built two data sources including, comprehensively:

- more than 40000 records representing events, people and relations between people occurred between 1452 and 1485 in Milan extracted from the *Liber Mortuorum*; specifically, these tuples record diseases (e.g. person X's disease), traumatic events (e.g. person Y is found dead at a specific place and time) and relationships (e.g. person X is married with person Y);
- 600 events from the *Liber sententiarum potestatis Mediolani* [2], which was written in 1385, and reports the criminal sentences pronounced by the chief magis-

trate of Milan Carlo Zen during his mandate. Each tuple stores information about different crimes decorated with the respective type of crime, people involved and type of condemnation or absolution.

We are currently working on an automated methodology to gather data from different sources and to organize it in the proposed meta-model. An experiment has been performed on partially structured data (e.g. Excel spreadsheets) to build the first mentioned data collection.

Based on the sources currently gathered, we have identified some research activities that will be pursued in the following months:

- we are going to employ probabilistic models, such as Bayesian Networks, in order to study the information recorded in the *Liber Mortuorum*. We would like to study how variables regarding personal data may effect the cause of death of an individual, and compare these relationships with the ones extracted analysing a dataset of a different time period;
- NLP techniques may be applied on the events extracted from the *Liber sententiarum potestatis Mediolani*: also in this case, a temporal comparison with data from another time period can be made, considering, for example, the different type of crimes committed and the characteristics of the people involved in them;
- finally, when it comes to image processing, we are considering the possibility of applying CNNs to medical images, such as CT scans and RMI, with the objective of developing unsupervised strategies for clustering and retrieving images.

References

1. Ghahramani, Z. (2006). Learning dynamic Bayesian networks. Adaptive Processing of Sequences and Data Structures: International Summer School on Neural Networks “ER Caianiello” Vietri sul Mare, Salerno, Italy September 6–13, 1997 Tutorial Lectures, 168-197.
2. Milano, Archivio Storico Civico e Biblioteca Trivulziana, Cimeli, 146.

Assessing the quality of Automatic Passenger Counter data for the analysis of mobility flows of local public transport systems

Valeria Maria Urbano, Arianna Burzacchi, Francesco Cherubini, Marika Arena, Giovanni Azzone, Piercesare Secchi, and Simone Vantini

Abstract Automated Passenger Counters (APC) represent a valuable source of information for the analysis of mobility patterns on public transport systems. Despite data gathered through APCs ensure higher accuracy and completeness compared to traditional sources of data (e.g. surveys), these data may present errors that might hinder the capability of understanding travel behavior. Addressing the need to quantitively assess the quality of APC data, this study aims at creating a high-quality dataset for supporting the decision-making of public transport operators. The study presents the data validation phase for the identification of the major quality problems, including noise, missing data, and outliers. Furthermore, a comparison of data from APC with manually collected data on transit passengers is presented with the aim of quantitatively evaluating the accuracy of APC data

Key words: APC, travel behavior, local public transport, data quality, APC's accuracy, manual counting comparison

---

Valeria Maria Urbano, Francesco Cherubini, Marika Arena, and Giovanni Azzone
Department of Management, Economics and Industrial Engineering, Politecnico di Milano - Via Lambruschini 4/B, 20156, Milano, Italy

Arianna Burzacchi, Piercesare Secchi, and Simone Vantini
MOX Laboratory, Department of Mathematics, Politecnico di Milano - Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Corresponding author: valeriamaria.urbano@polimi.it

1 Introduction

Understanding the dynamics of human travel behavior is pivotal for developing transportation strategies, both for the long term planning, the short-term, or even real-time, transportation management [8, 1]. Examples of the latter are the re-allocation of vehicles on specific routes adapting the service offer to the mobility demand in quasi-real time. Considering the long term planning, travel behavior provides insights on the current transportation needs. Moreover, it can anticipate future response to change in the infrastructure or in the transport service by enabling the identification of determinants of mobility [6].

The development of information sensing and communication technologies and their rapid diffusion in the last decades led to the proliferation of digital traces, including GPS trajectories, ticketing data, mobile phone traces. In this context, data from automatic passenger counting systems represent a valuable source of knowledge on mobility patterns of users using the public transport system. Automated Passenger Counters (APC) are people counting sensors installed on vehicles that are able to count the number of boardings and alightings at each stop of the network.

During the last decades, the analysis of travel behavior was based primarily on data collected through surveys. Authors underlined the limited quality of these researches, due to error proneness – as questionnaires collect self-reported average habits - and lack of scalability [3]. The latter is the result of the expensive cost and long time required for sampling, designing the questionnaire, and doing interviews. As a result, these data often cover less than 2 percent of the population [7], and there is no guarantee that they provide a statistically significant sample: according to Cools, Moons and Wets (2010) [2], an accurate understanding of human mobility patterns would require data from more than half of a population. Furthermore, surveys are made on a yearly base at most, while travel patterns show spatial and temporal variability as well as intrapersonal variability [4]. On the contrary, data from APCs provide a more precise understanding of human mobility patterns, enabling spatio-temporal analysis and ensuring higher accuracy and completeness. Moreover, APCs represent a cost-efficient data collection method. However, these data may present errors that might hinder the analysis of travel behavior. Addressing the need of evaluating the quality of APC data, this study aims at creating a high-quality dataset for subsequent decision-making for public transport operators, by identifying the major quality problems, including noise, missing data, and outliers. Moreover, the study provides a comparison of data from APC with data manually collected in order to evaluate the accuracy of the APC data.

The paper is organized as follows. Section 2 the methodology adopted for validating data is presented. More specifically, anomalies and inconsistencies of input data are identified and quantified, both in terms of frequency of occurrence and magnitude of anomalies. In section 3, the comparison between automatically collected data and manually collected data is presented.

2 Validation and anomaly detection of APC data

The company responsible for the management of public transport in Milan, which manages both the surface transport and the underground transport of the municipality, provided data from APC of vehicles moving on the surface of the city. The sample of data includes 789 rides on 16 different routes covering different areas of Milan. The analysis of the raw data enabled the identification of a set of anomalies that need to be considered when pre-processing the data. It is worth to highlight that, for the purpose of the study, pre-processing of raw data was performed with the final goal of analyzing mobility demand on a given route. Data from APC were integrated with information on the defined routes to validate data in order to identify anomalies and noise. This phase led to the recognition of three anomalies:

- noise: raw data may provide correct information on people counting but the information on the geo-localization could be inaccurate;
- detours: the vehicle is taking an unplanned detour with people on board;
- missing data: no signals registered for one or more stops along the route.

The noise registered affects 38% of the rides included in the sample, but it is not evenly distributed among routes. By analyzing the magnitude of noise in terms of the percentage of signals per ride that are affected by noise, we notice that the average value per route is lower than 2%. Moving to detours, the frequency measured in terms of re-routed rides is lower than 8% and affect two routes. The magnitude of the phenomenon is limited to 4 to 6% of the total number of observed signals per route. Missing data affects 18% of the rides. The link between missing data and noise is analyzed by means of an ANOVA test to determine if there is a statistically significant difference between the group of incomplete rides affected by noise and the group of incomplete rides not affected by noise. The test revealed that the two groups are statistically different (p value= 0.00025).

The detection of anomalies in the people counting data is performed, in a first step, by examining people counting data per vehicle in order to identify malfunctions in the automated people counting systems fitted to the vehicles. Two types of anomalies were identified:

- absence of both boarding and alighting passengers at each ride and every stop;
- absence of alighting passengers and presence of boarding passengers at each ride and every stop, and vice versa;

Anomalies in the APC data affected 102 vehicles. However, it is worth mentioning that the identified vehicles were equipped with sensors with low reliability. Moreover, outliers data in boarding and alighting passengers were identified by means of a bag-plot [5]. The identification of outliers data was performed by extracting the bi-variate sample (people boarded and people alighted) at each stop for a week of observations. If the sample does not present any anomalous measurement (both people boarded and alighted are never more than a certain threshold, fixed at 50), all the observations are classified as non-outliers. Otherwise, all the observations in the sample are classified according to a bagplot:

- bag: if the observation falls into the convex region that contains 50% of the data;
- outer: if the observation falls outside the bag but in the convex region (fence) that is obtained by inflating by a factor of three the bag;
- outlier: if the observation falls outside both the bag and the fence.

Additional domain knowledge was then applied to identify other outlies that were not captured in the bagplot. Outliers affect 0.1% of the analyzed rides and impact, on average, on less than 10% of stops per rides.

3 Comparing APC data with manual counting

The evaluation of the quality of measurements from APCs has been made possible by the provision of an additional data source, containing the manual measurements of transit passengers, including boardings, alighting and passengers on boards. Manual counting data were measured at the same stop of some surface lines and are related to subsequent rides. The matching of the provided datasets of manual counts and the corresponding APCs allows the comparison of 464 measurements of automatic and manual data on 10 surface lines. Each observation is also characterized by the measure time, ranging from 6am to 12pm, and by the type of APC employed for automatic counting. Following the guidelines of industry experts, sensors are classified according to their degree of reliability in high, mid, and low, getting the following proportion of measurements per sensor reliability: 17% high, 39%, and 45% low.

The first analysis performed on the quality of the APC data aims at evaluating the sensors' ability to detect the presence and absence of boarding and alighting passengers. Despite manual counting data cannot be considered ground truth, it can be assumed that manual measurement is completely reliable in identifying the presence and absence of transit passengers, and this information can therefore be used as a certain baseline to compare the APC data with. Table **??** shows these results in a misclassification table. For 82% of the measurements, we found consistency between automatic and manual measurements, showing good accuracy of the sensors in terms of identification of the presence/absence of transit passengers. It is also noted that the error of the sensor in counting passengers in their absence (1%) occurs in a lower percentage than the opposite error of not detecting passengers in their presence (5%), revealing the tendency of the sensors to underestimate the presence of transit passengers with respect to the absence. The analysis was repeated considering separately data from sensors of different reliability degree. The results show that high reliable sensors outperform the others the in terms of classification accuracy, increasing it from 80.6% of low reliable sensors and 80.1% of mid reliable sensors to the 89.6%.

Subsequently, the accuracy of APC data is evaluated by quantifying the magnitude of the error. The analysis is restricted to the set of data for which APCs are in agreement with manual counts with respect to absence and presence of transiting passengers, ending up with 380 measurements to be compared. Quantifying the

		Automatic			
		no boarding no alighting	only boarding	only alighting	both boarding and alighting
Manual	no boarding no alighting	15	3	1	1
	only boarding	4	40	0	11
	only alighting	5	1	21	22
	both boarding and alighting	15	6	15	304

Table 1 Classification table of measurements on presence/absence of transit passengers from manual counting (rows) and automatic counting (columns).

deviation between the number of passengers recorded manually and that measured automatically involves the determination of the following quantities: difference between manual and automatic boardings, difference between manual and automatic alightings, and difference between manual and automatic occupancy. The analysis reveals that:

- Differences in occupancy have the highest variability, with a range from -275 to $+96$ and a standard deviation of 29.56. This can be explained by observing that the automatic occupancy data are derived from the boardings and alightings at previous stops, therefore affected by the cumulative measure error. Additionally, it can be assumed that manual measurement is less reliable for occupancy, especially for high passenger capacities.
- Differences in boardings and alightings have a median of zero overall. Differences in boardings are smaller than the ones in alightings, as seen from the respective standard deviations of 4.90 and 7.58. This can be explained by observing that manual counting is probably more accurate when measuring boardings (already present at the stop and easy to be counted) than alightings (who are getting off the vehicle and are potentially missed in counting). It can be also noted that the variability of the differences increases proportionally to the number of passengers counted for both boarding and alighting measurements.

Finally, it was investigated whether the variability of APC accuracy differs between sensors of different reliability degree. The analysis is done considering both absolute deviations (Fig 1) and relative deviations, weighed by the number of passengers counted to take into account the proportionality of errors to the number of passengers. It was chosen to consider the pair of differences in boarding and alighting as a statistical unit and the sensor reliability level as a grouping factor. The analysis was carried out using a permutational one-way MANOVA. The p-value of the tests (0.0496 for absolute deviations and 0.0675 for relative deviations) confirms that we can state with statistical significance that there are differences in the three types of sensors.

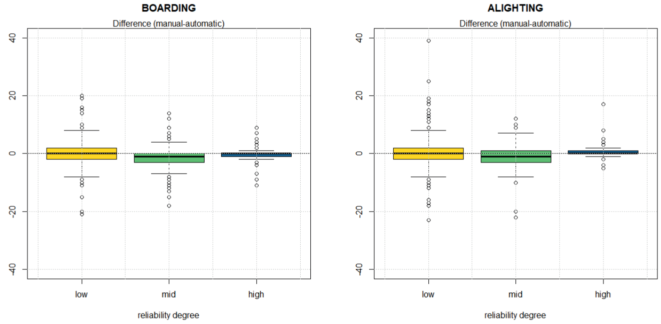


Fig. 1 Boxplots of boarding (left) and alighting (right) absolute deviations grouped by different sensor reliability degree

4 Acknowledgements

The authors thank Azienda Trasporti Milanesi S.p.A. both for providing the data and for the indications of technical expertise on APCs.

References

1. R. N. Buliung and P. S. Kanaroglou. Activity–travel behaviour research: conceptual issues, state of the art, and emerging perspectives on behavioural analysis and simulation modelling. *Transport Reviews*, 27(2):151–187, 2007.
2. M. Cools, E. Moons, and G. Wets. Assessing the quality of origin–destination matrices derived from activity travel surveys: Results from a monte carlo experiment. *Transportation research record*, 2183(1):49–59, 2010.
3. P. Fiadino, V. Ponce-Lopez, J. Antonio, M. Torrent-Moreno, and A. D'Alconzo. Call detail records for human mobility studies: Taking stock of the situation in the" always connected era". In *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, pages 43–48, 2017.
4. O. Järv, R. Ahas, and F. Witlox. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135, 2014.
5. R. I. Rousseeuw, P. and J. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.
6. Z. Wang, S. Y. He, and Y. Leung. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141–155, 2018.
7. L. Willumsen. Use of big data in transport modelling. 2021.
8. Y. Yue, T. Lan, A. G. Yeh, and Q.-Q. Li. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1(2):69–78, 2014.