PROGRAMME and BOOK of ABSTRACTS

JOCLAD 2023

20 - 22 APRIL

VIANA DO CASTELO, PORTUGAL

Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Viana do Castelo

CLAD
Associação Portuguesa de Classificação e Análise de Dados

XXX MEETING OF THE PORTUGUESE ASSOCIATION FOR CLASSIFICATION AND DATA ANALYSIS
XXX JORNADAS DE CLASSIFICAÇÃO E ANÁLISE DE DADOS

Programme and Book of Abstracts

XXX Meeting of the Portuguese Association for Classification and Data Analysis (CLAD)

20–22 April 2023

Viana do Castelo, Portugal

www.joclad.ipt.pt/joclad2023/

Sponsors

Banco de Portugal
Caixa de Crédito Agrícola
Câmara Municipal de Viana do Castelo
Centro de Matemática da Universidade do Minho
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viana do Castelo
Instituto Nacional de Estatística/Statistics Portugal
Instituto Politécnico de Viana do Castelo
PSE – Produtos e Serviços de Estatística

Organisers

Associação Portuguesa de Classificação e Análise de Dados (CLAD)
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viana do Castelo

Programme and Book of Abstracts
XXX Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2023)

Editors: A. Manuela Gonçalves, Conceição Rocha, Paula Brito, Sónia Dias

Publisher: CLAD

Printed: Statistics Portugal

ISBN 978-989-35097-0-8

Depósito legal: 514193/23

Preface

We warmly welcome you to JOCLAD 2023! JOCLAD 2023 – Meeting of the Portuguese Association for Classification and Data Analysis (CLAD) – aims at bringing together researchers and practitioners interested in Data Science. This year, to celebrate the thirty meeting of CLAD, JOCLAD 2023 is being held in the Portuguese northern coastal city of Viana do Castelo. After many meetings held throughout Portugal (2018 in Almada, 2019 in Viseu, 2020 in Lisbon, 2021 in Covilhã, 2022 in Porto, in a reduced format), this year JOCLAD 2023 takes place from 20 to 22 April at the High School of Technology and Management of the Polytechnic Institute of Viana do Castelo (ESTG–IPVC), who co-organises the event. ESTG was created in 1985, to respond to the growing need of specialized labor from industries and other organisations of the region. ESTG is situated in the beautiful city of Viana do Castelo, next to the North beach, overlooking the iconic Monte de Santa Luzia.

Viana do Castelo is a city in the north of Portugal that brings together the natural beauty provided by the river, the ocean and the mountains. It combines tradition (with an emphasis on traditional costumes and gold filigree art), modernity and innovation. The symbol of Viana is the "Heart of Viana", a heart-shaped jewelry item that is based on the filigree technique. Viana is also a city where many important names of the contemporary Portuguese architecture are present, marking some of the city's spaces and buildings. This is the case of Praça da Liberdade by Fernando Távora, the Library by Álvaro Siza Vieira, the Youth Hostel by Carrilho da Graça, the Hotel Axis by Jorge Albuquerque or the Cultural Center of Viana do Castelo, by Souto Moura, among many others. We would like to welcome all participants of JOCLAD 2023 to Viana do Castelo, and wish that you are able to enjoy one of the most beautiful cities of Portugal.

This volume is one of the main outcomes of JOCLAD 2023 and documents the meeting contents, particularly its programme. The JOCLAD 2023 programme includes a mini course on April 20 on "Data Science for health – Concepts and methods for learning on temporal health-related data" by Myra Spiliopoulou (Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, Germany). The diversity and liveliness of research in Data Science is also illustrated by the invited plenary talks, for which we thank Myra Spiliopoulou, Francisco de A. T. de Carvalho (Computer Science Center of the Federal University of Pernambuco – Brazil), and José G. Dias (University Institute of Lisbon – ISCTE-IUL, Portugal). We also thank the organisers of the Thematic Sessions, Bank of Portugal, Statistics Portugal, the Portuguese Statistical Society (SPE), ENERCON GmbH, and Sonae MC.

The programme includes the Fernando Nicolau Award, whose evaluation panel comprises Helena Bacelar Nicolau (Honorary President), Gilbert Saporta, Mário Figueiredo, and Paulo Gomes. In this edition, student participation has been strongly encouraged, and this year witnessed the largest number of student applications. A Thematic Session is devoted to the students granted with a 2023 CLAD scholarship: two Master's students, four PhD students (including an Honourable Mention). The awards have been kindly provided by CLAD, the evaluation committee includes A. Manuela Gonçalves (Chair), Adelaide Freitas, and Paulo Infante.

Additionally, this volume contains all the abstracts of talks and posters presented at regular oral and poster sessions. This is going to be a particularly well attended, and hopefully successful, edition of JOCLAD, with a final outcome of 27 contributions accepted for oral presentation and 25 for poster presentation. Each abstract published in this volume has been evaluated by at least one anonymous member of the scientific committee. We thank all the authors who submitted an abstract to our meeting and the reviewers who supported the editorial process with their fast and constructive commitment. We thank the members of the Scientific Committee, A. Manuela Gonçalves, Irene Oliveira, Isabel Silva, João Cordeiro, Margarida G. M. S. Cardoso, and M. Rosário Oliveira, whose work definitely contributed to ensure the overall quality of the JOCLAD 2023 programme. We also thank all the chairs of the sessions. Last but not least, it is a pleasure to thank all sponsors for helping the organisation of this meeting. Our institutional sponsors deserve a special mention – Statistics Portugal and Bank of Portugal – whom we thank for their lasting and generous support.

A successful meeting involves more than just the presentation of talks and posters, it is also a meeting of people and exchange of research ideas and collaborations. Thus, a social program has been arranged – a social dinner and a visit to Gil Eannes Naval Museum in Viana do Castelo – to promote and facilitate this desired networking. Our deep thanks extend to the local organising committee, Conceição Rocha, Paula Cheira, Pedro Pinto, Sandra Silva, and Sónia Dias, as well as to our local sponsors, who made it possible for participants to become more involved with Viana do Castelo, creating opportunities to know more about this region through social and gastronomic moments. We wish you a productive and stimulating meeting and a memorable stay in Viana do Castelo. Finally, thank you all for your support for JOCLAD 2023, and for helping us to make this a successful meeting. Your high-quality work is essential for CLAD to continue its tradition of excellence in advancing Data Science. We do hope to meet you again for JOCLAD 2024!

Viana do Castelo, April 2023

Chair of the Scientific Programme Committee
A. Manuela Gonçalves

Conference Chair
Sónia Dias

President of CLAD
Paula Brito

Organisation

President of CLAD

Paula Brito

Chair of JOCLAD 2023

Sónia Dias (Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viana do Castelo)

Local Organising Committee

Conceição Rocha (CPES - INESC-TEC)
Paula Cheira (ESTG - IPVC & INESC-TEC)
Pedro Pinto (ESTG - IPVC & INESC-TEC)
Sandra Silva (ESA - IPVC, Prometheus & INESCC)
Sónia Dias (ESTG - IPVC & INESC-TEC)

Chair of the Scientific Programme Committee

A. Manuela Gonçalves (Universidade do Minho)

Scientific Programme Committee

A. Manuela Gonçalves (Universidade do Minho)
Irene Oliveira (Universidade de Trás-os-Montes e Alto Douro)
Isabel Silva (Universidade do Porto)
João Cordeiro (Universidade da Beira Interior)
Margarida G. M. S. Cardoso (Instituto Universitário de Lisboa)
M. Rosário Oliveira (Universidade de Lisboa)

Contents

Programme Overview

Thursday, 20 April

8:30	Registration	Hall of ESTG
9:00	**Mini Course**	Room A1.2
10:30	Coffee Break	Hall of Library
11:00	**Mini Course** (cont.)	Room A1.2
12:30	Lunch Time	Scala Restaurant
13:30	Registration	Hall of ESTG
14:00	**Opening Session**	Auditorium
14:30	**Keynote Lecture I – Francisco de A. T. de Carvalho**	Auditorium
15:30	Coffee Break	Hall of Library
16:00	**Parallel Sessions I**	Auditorium & Room A1.2
18:00	**Visit to Gil Eannes Naval Museum**	
19:00	**Reception: Verde de Honra**	Centro de Mar

Friday, 21 April

8:30	Registration	Hall of ESTG
9:00	**Parallel Sessions II**	Auditorium & Room A1.2
10:20	Coffee Break	Hall of Library
10:40	**Poster Session I**	Hall of Library
11:00	**Keynote Lecture II – José G. Dias**	Auditorium
12:00	**Thematic Session I - CLAD Corporate**	Auditorium
13:00	Lunch Time	Scala Restaurant
14:30	**Thematic Session II - Bank of Portugal**	Auditorium
15:30	**Thematic Session III - Statistics Portugal**	Auditorium
16:30	Coffee Break	Hall of Library
17:00	**Thematic Session IV - CLAD 2023 Scholarships**	Auditorium
18:40	General Assembly of CLAD	Auditorium
20:30	**Social Dinner**	Dona Aninhas Hotel

Saturday, 22 April

9:00	**Parallel Sessions III**	Auditorium & Room A1.2
10:00	**Parallel Sessions IV**	Auditorium & Room A1.2
11:00	Coffee Break	Hall of Library
11:20	**Poster Session II**	Hall of Library
11:40	**Thematic Session V - SPE**	Auditorium
12:40	**Vianackathon Awards**	Auditorium
13:10	Lunch Time	Scala Restaurant
14:45	**Fernando Nicolau Award**	Auditorium
15:15	**Keynote Lecture III – Myra Spiliopoulou**	Auditorium
16:15	**Closing Session**	Auditorium

Auditorium - Auditorium Francisco Sampaio
Hall of Library - Hall of Library Barbosa Romero

Programme

Thursday, 20 April

8:30 Registration - Hall of ESTG

9:00 **Mini Course** - Room A1.2
Data Science for health – Concepts and methods for learning on temporal health-related data
Myra Spiliopoulou, p. 5
Chair: Adelaide Freitas

10:30 **Coffee Break**

11:00 **Mini Course** (cont.)

12:30 **Lunch Time**

13:30 Registration - Hall of ESTG

14:00 **Opening Session** - Auditorium Francisco Sampaio

14:30 **Keynote Session I** - Auditorium Francisco Sampaio
Learning from symbolic data
Francisco de A. T. de Carvalho, p. 9
Chair: Sónia Dias

15:30 **Coffee Break**

16:00 **Parallel Sessions I**

	Auditorium Francisco Sampaio **Data analysis applications I** Chair: Irene Oliveira	Room A1.2 **Time series modelling** Chair: Maria Eduarda Silva
16:00	**SLE-DAS performance in phase 3 clinical trials**, Ana Matos, Carla Henriques, Diogo Jesus, Luis Inês, p. 61	**Improving short-term forecasts of environmental time series via state-space modeling**, F. Catarina Pereira, A. Manuela Gonçalves, Marco Costa, p. 69
16:20	**Designing experiments for use in agriculture: the example of large field trials for grapevine selection**, Elsa Gonçalves, Antero Martins, p. 63	**Stationary and non-stationary state-space models in the presence of outliers: a simulation study**, F. Catarina Pereira, A. Manuela Gonçalves, Marco Costa, p. 71
16:40	**Bootstrap confidence intervals for association measures in sparse contingency tables**, João Rocha, Adelaide Freitas, Isabel Pereira, p. 65	**Modeling the fuel consumption of a NRP ship using a Kalman filter approach**, M. Filomena Teodoro, Pedro Carvalho, Ana Trindade, p. 73
17:00	**The link between internal social responsibility, work culture and innovative behavior: a statistical approach**, Mara Cunha, Helena Sofia Rodrigues, Ana Teresa Oliveira, p. 67	**Bayesian approach to modelling time series of counts under censoring**, Isabel Silva, Maria Eduarda Silva, Isabel Pereira, Brendan McCabe, p. 75

18:00 **Visit to Gil Eannes Naval Museum**

19:00 **Reception: Verde de Honra** – Centro de Mar

Friday, 21 April

8:30 Registration - Hall of ESTG

9:00 **Parallel Sessions II**

	Auditorium Francisco Sampaio **Data analysis applications II** Chair: Luís M. Grilo	Room A1.2 **Clustering** Chair: Margarida G. M. S. Cardoso
9:00	**Inventories discretionary management through accounting choices – The case of small and medium-sized Portuguese companies in commercial sector**, M. Filipa Nogueira, Augusta Ferreira, Carlos Ferreira, p. 77	**A comparison of some methods for clustering of variables of mixed types**, Ndèye Niang, Mory Ouattara, Gilbert Saporta, p. 85
9:20	**Identifying characteristics of marketing-influenced eating vulnerability**, Carla Henriques, Raquel Guiné, Ana Matos, Madalena Malva, p. 79	**Clustering of pediatric hospitalizations by hospital resources use**, Daniel Cordeiro, Ana Azevedo, Bárbara Peleteiro, Lucybell Moreira, Elsa Guimarães, Raquel Cadilhe, Rita Gaio, p. 87
9:40	**Are the European countries well prepared for the new technological challenges?**, Fernanda Figueiredo, Adelaide Figueiredo, p. 81	**Clustering ECG time series for the quantification of physiological reactions to emotional stimuli**, Beatriz Henriques, Susana Brás, Sónia Gouveia, p. 89
10:00	**A risk model for classifying stocks**, Irene Brito, p. 83	**Clustering analysis for household week-daily water consumption profiles characterization**, João Bastos, Flora Ferreira, Duarte Silva, Wolfram Erlhagen, Estela Bicho, p. 91
10:20	**Coffee Break**	

10:40 **Poster Session I** - Hall of Library Barbosa Romero

Chair: Paula Cheira

Linear regression for symbolic density-valued data
Rui Nunes, Paula Brito, Sónia Dias, p. 117

Analyzing compositional data using distributions defined on the hypersphere
Adelaide Figueiredo, p. 119

Survival forests in lifetime analysis
Cecília Castro, Ana Paula Amorim, p. 121

How much time do we spend on the sofa?
António Balau, Fábio Rodrigues, Sofia Ribeiro, Cristina Lopes, Cristina Torres, Lurdes Babo, Isabel Vieira, p. 123

Identification of potential causes in the number of beneficiaries of social disability pension in small municipalities in the northern region of Portugal
Cristina Torres, Lurdes Babo, Isabel Vieira, Isabel Cristina Lopes, Rui Monteiro, Carla Ferreira, Inês Bem-Haja, p. 125

Evolution of mean sea level: particular case of the port of Viana de Castelo
Dora Carinhas, Miguel Picoto, Paulo Infante, p. 127

Statistical analysis of humanoids' arm movements
Eliana Costa e Silva, Gianpaolo Gulletta, Estela Bicho, Wolfram Erlhagen, p. 129

Application of data reduction methods in the creation of SoResilere – Social resilience index applied to flood affected municipalities
Rita Jacinto, Fernando Sebastião, João Ferrão, Eusébio Reis, p. 131

Alto Minho regional performance through SDG11: a cluster analysis
Helena Sofia Rodrigues, Ângela Silva, Jorge Esparteiro Garcia, p. 133

Gender equality in wages in Portugal, between 1994 and 2020
Ana Freitas, Elenice Santos, Marileide Silva, Vanessa Lima, Isabel Vieira, Cristina Lopes, Cristina Torres, Lurdes Babo, p. 135

Characterization of mobbing in Portuguese accounting professionals using Leymann inventory of psychological terror scale items
Irene Oliveira, António Dias, Margarida Simões, Ana Paula Monteiro, Rui Silva, p. 137

Birth rate in Portugal
Maria Guerra, Ana Pessoa, Ana Rodrigues, Maria Mesquita, Lurdes Babo, Isabel Vieira, Cristina Lopes, Cristina Torres, p. 139

Assessment of exhaustion, cognitive weariness and physical fatigue of security services workers: PLSc-SEM approach
Luís M. Grilo, Tiago F. Braz, Helena L. Grilo, p. 141

11:00 **Keynote Lecture II** - Auditorium Francisco Sampaio
Finite mixture models: an overview
José G. Dias, p. 11

Chair: Paula Brito

12:00 **Thematic Session I - CLAD Corporate** - Auditorium Francisco Sampaio

Chair: Carlos Ferreira

12:00 **Data analysis in wind industry – Power performance measurement**, Fernando Barroso, Márcio Ferreira, p. 17
12:30 **From data to execution**, Ana Freitas, p. 19

13:00 **Lunch Time**

14:30 **Thematic Session II - Bank of Portugal** - Auditorium Francisco Sampaio
Robots in Official Statistics: are they helping or replacing us?

Chair: Luís Teles Dias

14:30 **A robot which counts robots: knowing better BPstat users**, Filipa Oliveira, Leonardo Almeida, p. 23
14:50 **From people to Python: a new approach on securities holdings statistics quality assurance**, Diogo Nobre, Ludgero Glórias, Pedro Silva, p. 25
15:10 **Give me a guess on the purpose of this transaction: diving in microdata on external money transfers**, Martha Düker, Helena M. Marques, p. 27

15:30 **Thematic Session III - Statistics Portugal** - Auditorium Francisco Sampaio
Challenges of Data Science in Official Statistics
Chair: Carlos Marcelo

15:30 **Enterprise mortality – A prediction model**, Alexandre Cunha, Rui Gouveia, Sandra Lagarto, Vasco Cordeiro, Vera Dias, p. 31

15:45 **AI-based prediction and identification of errors in the geographic location of buildings**, Aldina Piedade, Carmen Costa, Bartholomeus Schoenmakers, p. 33

16:00 **More data sources, more information, more quality**, Sofia Rodrigues, Almiro Moreira, Paulo Saraiva, João Poças, Bruno Lima, p. 35

16:15 **Access to official statistical information for scientific research purposes**, Pinto Martins, p. 37

16:30 **Coffee Break**

17:00 **Thematic Session IV - CLAD 2023 Scholarships** - Auditorium Francisco Sampaio
Chair: A. Manuela Gonçalves

17:00 **Fisher discriminant analysis for interval data**, Diogo Pinheiro, M. Rosário Oliveira, Igor Kravchenko, Lina Oliveira, p. 41

17:20 **Handling missing data in the prediction of childhood obesity: a simulation study**, Mafalda Oliveira, Susana Santos, Rita Gaio, p. 43

17:40 **Nowcasting the Portuguese unemployment rate with Google Trends**, Eduardo Andre Costa, Maria Eduarda Silva, p. 45

18:00 **Using wavelets to denoise sound data and identify patterns: a narwhal example**, Carolina S. Marques, Emmanuel Dufourq, Carl Donovan, Marianne Marcoux, Tiago A. Marques, p. 47

18:20 **Are multilayer networks useful for mining multivariate time series?**, Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, Fernando Silva, p. 49

18:40 General Assembly of CLAD - Auditorium Francisco Sampaio

20:30 **Social Dinner** - Dona Aninhas Hotel

Saturday, 22 April

9:00 **Parallel Sessions III**

	Auditorium Francisco Sampaio **Machine learning and classification** Chair: João Cordeiro	Room A1.2 **Models with latent variables** Chair: José G. Dias
9:00	**A valuation model for lab-grown diamonds**, Margarida G. M. S. Cardoso, Luís Chambel, p. 93	**Experiences at home during the COVID-19 quarantine – A cluster analysis**, Eulália Santos, Vasco Tavares, Fernando Tavares, Margarida Oliveira, p. 99
9:20	**Decomposed mutual information maximization: a feature selection method based on mutual information**, M. Rosário Oliveira, Francisco Macedo, Rui Valadas, Eunice Carrasquinha, António Pacheco, p. 95	**How to measure the fit of a structural equation model with omissions by design**, Paula C. R. Vicente, p. 101
9:40	**Probabilistic Vector Machines**, Pedro Duarte Silva, p. 97	**Students' burnout at a Portuguese polytechnic: PLSc-SEM approach**, Luís M. Grilo, M. Cristina Costa, p. 103

10:00 **Parallel Sessions IV**

	Auditorium Francisco Sampaio **Spatial data analysis** Chair: Isabel Silva	Room A1.2 **Compositional and symbolic data analysis** Chair: Pedro Duarte Silva
10:00	**Counterfactual impact evaluation – An exploratory study on urban revitalization projects in Aveiro and Ilhavo municipalities**, André Lima, Paulo Batista, João Marques, p. 105	**Compositional data vectors: how useful they can be?**, Adelaide Freitas, Marta Maltez, Marco Costa, p. 111
10:20	**Black scabbardfish species distribution: geostatistical inference and bayesian sampling design under preferential sampling**, Paula Simões, M. Lucília Carvalho, Ivone Figueiredo, Andreia Monteiro, Isabel Natário, p. 107	**A supervised clustering algorithm for preventing fraud in edge attributed network components**, Pedro Campos, Célia Carvalho, p. 113
10:40	**An approach to estimate residential real estate prices with scarce information**, Marco Marto, Catarina Duarte Ribeiro, João Barrias, Paulo Batista, p. 109	

11:00 **Coffe Break**

11:20 **Poster Session II** - Hall of Library Barbosa Romero

Chair: Sandra Silva

11:40 **Thematic Session V - SPE** - Auditorium Francisco Sampaio
Three Statistical Strategies to Deal with Health Data
Chair: Clara Cordeiro

11:40 **Statistical modelling methods applied to the diagnosis of temporomandibular disorders**, Ricardo São João, Henrique J. Cardoso, David F. Ângelo, p. 53
12:00 **Multivariate statistical quality control methods for monitoring the concentration of particulate matter in an occupational contex**, M. Rosário Ramos, Carla Viegas, Susana Viegas, Elisabete Carolino, p. 55
12:20 **Predicting hospital patients' no-show through statistical and machine learning techniques**, Ana Borges, Mariana Carvalho, p. 57

12:40 **Vianackathon Awards**
Chairs: Sónia Dias and Pedro Campos

13:10 **Lunch Time**

14:45 **Fernando Nicolau Award**
Chair: Paulo Gomes

15:15 **Keynote Lecture III** - Auditorium Francisco Sampaio
Data Science in support of wellbeing: analysing sparse medical data
Myra Spiliopoulou, p. 13
Chair: M. Rosário Oliveira

16:15 **Closing Session** - Auditorium Francisco Sampaio

Abstracts

Mini Course

20 April, 9:00 - 10:30, 11:00 - 12:30, Room A1.2

Data Science for health – Concepts and methods for learning on temporal health-related data

Myra Spiliopoulou[1]
[1] Research Lab KMD: "Knowledge Management & Discovery", Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, Germany, myra@ovgu.de

---

Medical research and medical decision support increasingly rely on data sciences advances. The research domain that encompasses advances on learning for health is vast and fragmented. In this tutorial, we will discuss following issues:

1. Forms of medical data used in medical research vs medical decision support;
2. Learning on multidimensional clinical data – prediction and phenotyping;
3. Taking time into account during learning on clinical data;
4. Time and Ecological Momentary Assessments (EMA) - learning on mHealth data.

This tutorial is for data scientists who want to apply and extend their methods for tasks in the medical domain. The focus is less on elaborate learning algorithms and more on the tasks of specifying the learning problems, preparing the data and making assumptions before applying simple and elaborate learning algorithms. Some of the approaches we see have been applied on public domain data, but most of them have been designed for tasks in concrete clinical or public health settings.

Keynote Lectures

20 April, 14:30 - 15:30, Auditorium Francisco Sampaio

Learning from symbolic data

Francisco de A. T. de Carvalho[1]
[1] Centro de Informática da Universidade Federal de Pernambuco - UFPE (https://portal.cin.ufpe.br/), fatc@cin.ufpe.br

In this presentation we discuss some challenges of extending supervised and unsupervised machine learning and data science methods to symbolic data described by new variable types (set-valued, interval-valued, distributional-valued) aiming to describe concepts (groups of items) where taking into account the variability is needed. In the first part we will present robust linear regression methods for interval-valued data based on the ordinary least squares and kernel functions. Then, in the second part we will present a suitable extension of fuzzy k-means clustering algorithm to interval-valued data.

Keywords: machine learning, symbolic data, regression, clustering, interval-valued data

The basic units of machine learning and data analysis are single individuals. Usually, they are described by variables (quantitative, qualitative, binary) that are single-valued, i.e., that assume as value a number or a category. The individuals are described by a vector of quantitative, qualitative and binary values.
However, when analyzing groups of individuals, take into account the *variability* inherent to the data is needed. Variability occurs when we have, e.g., data about patients but our aim is to describe and analyse hospitals. The traditional way to take into account the variability when describing a group of individuals is aggregating data from its individuals using the mean (or the median) for quantitative data or the mode for qualitative data.
E. Diday [1] argued that to improve the way to take into account the variability when describing groups of individual, new variable types, that assume as value a set of categories, an interval or even an empirical distribution (histogram) are needed. He termed these new variables as *symbolic variables* and the values that they assume as *symbolic data*.
Thus, the basic units of machine learning and data analysis becomes groups of individuals described by symbolic data. Because symbolic data cannot be easily represented in a vector space as real-valued data is, the development of machine learning and data analysis tools able to manage symbolic data are very much challenging and needed. Some methods have been already developed (see Ref. [2] for a survey), but many works are still to be done.
In this presentation we discuss some challenges of extending supervised and unsupervised machine learning and data analysis methods to symbolic data.
In the first part we will present two robust linear regression methods for interval-valued data based on the ordinary least squares and kernel functions [4, 6]. In both methods, the sum of squared errors is defined in a high dimensional space, by means of a non-linear

mapping applied on the observed response variable and on its corresponding mean vector, and computed in the original space thanks to the kernel trick. The first method penalizes the presence of outliers in the centers and/or radius of the intervals through the use of Gaussian kernel functions. The centers (the radius) outliers are penalized on the centers (on the radius) regression. The second method also penalizes the presence of outliers in the centers and/or radius through the use of Gaussian kernel functions. However, both the centers and the radius outliers penalize both the center and the radius regressions. Moreover, the observations with outliers on both center and radius are more penalized than those observations with outliers only in the center or only in the radius. Besides, for each method we provide a suitable iterative algorithm to estimate the regression parameters.
In the second part we will present an extension of the fuzzy K-means clustering algorithm that uses suitable adaptive distances with the purpose to cluster interval-valued data [3, 5]. From an initial solution, this fuzzy K-means clustering algorithm optimizes explicitly a suitable objective function by alternating three steps aiming to compute the fuzzy cluster representatives, the fuzzy partition, as well as relevance weights for the interval-valued variables. Indeed, most often conventional fuzzy K-means clustering algorithms consider that all variables are equally important for the clustering task. However, in real situations, some variables may be more or less important or even irrelevant for clustering. Due to the use of adaptive distances, this fuzzy K-means clustering algorithm tackles this problem with an additional step where a relevance weight is automatically learned for each interval-valued variable. Besides, this fuzzy K-means clustering algorithm can use either the Euclidean distance or the robust City-Block and Hausdorff distances to compare the objects and the cluster representatives. Each particular distance results in a different cluster representative that is obtained from the optimization of the objective function either algebraically in the case of the Euclidean distance or by means of an algorithmic solution in the case of City-Block and Hausdorff distances.

References

[1] H.-H. Bock and E. Diday. *Analysis of Symbolic Data.* Springer, Berlin, Heidelberg, 2000.

[2] Paula Brito. Symbolic data analysis: another look at the interaction of data mining and statistics. *WIREs Data Mining Knowl. Discov.*, 4(4):281–295, 2014.

[3] Francisco de A. T. de Carvalho. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognit. Lett.*, 28(4):423–437, 2007.

[4] Francisco de A. T. de Carvalho, Eufrasio de Andrade Lima Neto, and Ullysses da N. Rosendo. Interval joint robust regression method. *Neurocomputing*, 465:265–288, 2021.

[5] Francisco de A. T. de Carvalho and Eduardo C. Simões. Fuzzy clustering of interval-valued data with city-block and hausdorff distances. *Neurocomputing*, 266:659–673, 2017.

[6] Eufrasio de Andrade Lima Neto and Francisco de A. T. de Carvalho. An exponential-type kernel robust regression model for interval-valued variables. *Inf. Sci.*, 454-455:419–442, 2018.

21 April, 11:00 - 12:00, Auditorium Francisco Sampaio

Finite mixture models: an overview

José G. Dias

Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, jose.dias@iscte-iul.pt

---

Finite mixture models have been used in many fields for different purposes and under different names. Other non-exact names are model-based clustering and latent class models. This presentation gives an overview of this area. In particular, it summarizes the research that has been published both theoretical papers and in applications. A systematic literature review using the PRISMA methodology was used. The text mining analysis then identifies topics in the literature. Results show an explosion of the research in the field since 2000. Social and health sciences are the most prominent application areas, mainly focused on the detection of unobserved heterogeneity.

Keywords: finite mixture models, latent class models, discrete latent variables, model-based clustering, systematic literature review

---

Finite mixture (FM) models and related latent variable models are over one hundred years old. The origin of the FM model is usually attributed to Newcomb and Pearson [4, 5]. Stigler [7], however, at least traces its origin back to the analysis of conviction rates by Poisson in the second quarter of the nineteenth century. Since 2000, the use of these models has grown exponentially. In the past few decades, advances in computer technology, FM modeling has proven to be a powerful tool for the analysis of a wide range of empirical problems. For instance, in the social sciences, which have a long tradition of latent class (LC) models, following the seminal work by Lazarsfeld and refinements notably by Goodman and Clogg (see, e.g., [2] and [1]), more sophisticated models are gaining popularity. McLachlan and Peel [3] provide a good overview of the field until 2000. The exponential growth of these models over the past two decades clearly shows that they are directly related to the democratization of statistical computation using fast personal computers (PCs) and increasing availability of software for their estimation.

This work presents an overview of the field using a systematic literature review. In addition to searching for articles using keywords to retrieve papers, we also used papers citing well-known references in the field (e.g., [8, 3, 6]). The extraction and selection of papers from the Web of Science follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. A total of 38,997 papers were included in the analysis. Topic analysis, a special case of text mining, is used to identify topic clusters in the corpus.

Results show the diverse use of FMs in the literature. Most publications use FMs to identify clusters. However, in other applications and contexts, topics cover density estimation, defining prior probabilities in Bayesian statistics, discrete latent variables, the golden standard problem, speech modeling, imagine analysis, longitudinal and trajectory analysis, or social class analysis. This research establishes a typology in the field of FM methodology and shows its wide range and flexible use in statistical modeling.

References

[1] C. C. Clogg. Latent class models. In G. Arminger, C.C. Clogg, and M.E. Sobel, editors, *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pages 311–359. Plenum, New York, 1995.

[2] L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[3] G.J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[4] S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4):343–366, 1886.

[5] K. Pearson. Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894.

[6] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8(1):289–317, 2016.

[7] S.M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belkap Press of Harvard University Press, Cambridge, MA and London, 1986.

[8] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York, 1985.

22 April, 15:15 - 16:15, Auditorium Francisco Sampaio

Data Science in support of wellbeing: analysing sparse medical data

Myra Spiliopoulou[1]
[1] Research Lab KMD: "Knowledge Management & Discovery", Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, Germany, myra@ovgu.de

---

Smartphone apps have great potential in healthcare and in patient empowerment. They give to patients with chronic conditions the opportunity of monitoring body signals such as blood pressure and insulin levels, keeping a diary of their condition, and acquiring reminders on health-related tasks such as medications and physical exercises. Not only do the patient learn to live better with their condition; the caring physician acquires valuable information on how the condition is perceived day by day. While some of these apps record signals such as paces or sleep hours unobtrusively, others demand active patient involvement, eg for filling information about nutrition and momentary assessments. The latter require self-discipline, otherwise the assessments would give a partial, incomplete picture of the patient's wellbeing. Data science can contribute in two ways: by exploiting information from patients with many data to learn about patients with few data, and by interpreting the gaps – the missing data. We see methods for both on the example of patients filling daily a questionnaire with an mHealth app.

References

[1] M. Schleicher, R. Pryss, W. Schlee, and M. Spiliopoulou. When Can I Expect the mHealth User to Return? Prediction Meets Time Series with Gaps. In Martin Michalowski, Syed Sibte Raza Abidi, and Samina Abidi (Eds.), Artificial Intelligence in Medicine (AIME 2022), Springer International Publishing, 2022.

[2] M. Schleicher, R. Pryss, J. Schobel, W. Schlee, and M. Spiliopoulou. Expect the gap: A recommender approach to estimate the absenteeism of self-monitoring mHealth app users. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2022.

[3] V. Unnikrishnan, M. Schleicher, C. Puga, R. Pryss, C. Vogel, W. Schlee, and M. Spiliopoulou. A Similarity-Guided Framework for Error-Driven Discovery of Patient Neighbourhoods in EMA Data. Proc. of Intelligent Data Analysis (IDA 2023) - To appear, 2023.

[4] V. Unnikrishnan, Y. Shah, M. Schleicher, C. Fernández-Viadero, M. Strandzheva, D. Velikova, P. Dimitrov, R. Pryss, J. Schobel, W. Schlee, and M. Spiliopoulou. Love thy Neighbours: A Framework for Error-Driven Discovery of Useful Neighbourhoods for One-Step Forecasts on EMA data. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 295–300, 2021.

Thematic Session
CLAD Corporate

21 April, 12:00 - 12:30, Auditorium Francisco Sampaio

Data analysis in wind industry – Power performance measurement

Fernando Barroso[1], **Márcio Ferreira**[2]
[1] ENERCON GmbH Sucursal em Portugal, fernando.barroso@enercon.de
[2] ENERCON GmbH Sucursal em Portugal, marcio.ferreira@enercon.de

---

This presentation intends to present a practical case on data handling and analysis in wind industry.
The measurement of power performance of a wind turbine generator covers several areas of Data Science, like ETL, EDA or ML models.
The purpose of a power performance measurement is to assess the level of compliance of a given wind turbine generator to it's predefined/contracted power curve, by means of external wind and power measurement.
These measurements are defined and must follow the procedure described in IEC 61400-12-1 [1].

Keywords: data analysis, wind energy, renewables, power performance

---

As a Data Analysis practical case in wind industry, we will present a wind turbine power performance measurement with site calibration. We will focus on the data analysis aspects of this process.
In complex terrains, a power performance measurement starts before wind turbine installation with a process called site calibration.

Figure 1: Wind Turbine with Meteorological Mast for Power Performance Measurement

This process consists on placing two meteorological mast to create a wind speed relational model between the future turbine location and the measurement meteorological mast. This model will later be used to estimate the wind speed at the turbine position.

To create this model, data from both masts must be collected, filtered and synchronized. [1] defines two models for wind speed prediction based on wind flow conditions: linear regression or wind direction and shear matrix.
This calibration is performed for specific sectors of 10^{o} of direction around the turbine position. To be considered calibrated, sectors must fulfill a series of requirements according to [1].
After turbine installation and commissioning, the measurement can take place, preferably in the same season as the site calibration process.
For this procedure, a power reading datalogger shall be installed at the turbine. Power data must be collected together with the meteorological mast data and turbine data.
Again, data must be collected from the different sources, filtered and synchronized. Data from calibrated sectors can then be binned by wind speed and averaged to create a power curve.
This measured power curve can then be compared with the contracted power curve and its performance assessed.

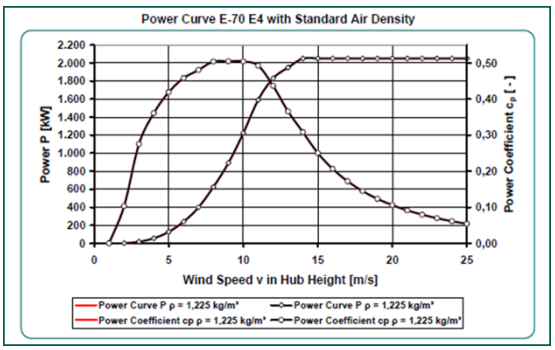

Figure 2: Power Curve Example

To have a better figure of the performance assessment, the measured data is converted to a performance indicator: the Annual Energy Production (AEP).
AEP will provide a figure on the expected energy produced over an year for the measured power curve. This is obtained by applying the power curve to the yearly wind distribution of that specific localization. The AEP of the contracted power curve is also calculated for the same wind distribution. The ratio of the measured and contracted AEP will give an indication on the turbine performance level.

References

[1] International Electrotechnical Commission. Iec 61400-12-1 wind energy generation systems - power performance measurement of electricity producing wind turbines. International Electrotechnical Commission, 2017.

21 April, 12:30 - 13:00, Auditorium Francisco Sampaio

From data to execution

Ana Freitas[1]
[1] Sonae MC, ancfreitas@mc.pt

Sonae MC is a Portuguese retail company that stresses the valuable impact of data-driven business decisions with a focus on its Cartão Continente loyalty card program. The organization has been able to leverage machine learning and artificial intelligence techniques to automate and streamline decisions and obtain a thorough understanding of its consumers' actual and predicted behavior, preferences, and purchase habits by analyzing and interpreting data from the loyalty card program, that feed the customer journey strategy. One concludes that data-driven decision-making, powered by the insights from Cartão Continente, is a critical component of Sonae MC's current and future strategy.

Keywords: data, loyalty, impact, science, retail

Sonae MC is a leading Portuguese retail company that recognizes the strategic importance of data-driven decision-making.
The retail industry is facing intense competition and increasingly more complex and demanding consumers, and companies that have access to data are better positioned to gain a competitive advantage. The rapid expansion of data and the development of new technologies have made it possible for retail enterprises to gain valuable insights from their data, which they can use to make more informed and precise decisions.
Sonae MC has recognized the importance of data-driven decision-making and has invested in creating a data-driven culture and a robust data infrastructure.
Cartão Continente loyalty program's mission is to create an ecosystem of brands that fully covers all moments and consumption areas of the Portuguese families (e.g. retail food, fashion, health and wellness, food chains, travel, entertainment, financial services). The company has leveraged its loyalty card program as a crucial data source to support its decision-making processes along the customer journey in order to foster customer engagement.
Customer data includes:
- customer DNA (socio-demographic information)
- transactional data detailing which products are bought, when, where, and the interaction with promotions and digital services (e.g. App Cartão Continente, personalized discounts, digital wallet).
In the customer journey, data science plays an important role.
Creating awareness on a specific brand from the ecosystem, customer acquisition models enable the identification of the right customers to contact based on their behavior in other

partner areas. Explainable models such as logistic regression models are a good fit for this sort of problem.
Designing baseline customer strategies relies on segmenting customers on a macro (recency, frequency and monetary segmentation e.g. Loyal, Frequent, Occasional) or product oriented (e.g. Healthy, Traditional, Economic) frameworks. Segmenting transactions is also valuable to understand the combination of shopping missions performed by each customer and the inherent store format, supporting strategic decisions on store layout and services. K-means, Ward clustering with Euclidean distances and Ward clustering with Jaccard distances are three different approaches with characteristics that make them more suitable for each of the clustering challenges above, respectively.
On customer retention, lifetime value and churn prediction can be tackled by fitting a Pareto/NBD model.
Fostering loyalty, personalization is crucial, combining predictive and prescriptive methods. Personalized leaflets, targeted discount coupons and delivering weekly shopping lists are examples of analytical products with impact in customer loyalty and customer satisfaction.
Ultimately, transforming data into wisdom is not only a journey towards achieving business goals, but it is also an ethical journey. It is essential to keep in mind the ethical implications of data science projects in retail, such as data privacy and fairness, and to strive for transparency and accountability throughout the entire process.

Acknowledgements The present research was supported by Sonae MC, SGPS, S.A.

Thematic Session
Bank of Portugal

21 April, 14:30 - 14:50, Auditorium Francisco Sampaio

A robot which counts robots: knowing better BP*stat* users

Filipa Oliveira[1], **Leonardo Almeida**[2]
[1] Banco de Portugal, fnoliveira@bportugal.pt
[2] Banco de Portugal, llalmeida@bportugal.pt

BP*stat* provides statistical information on the Portuguese and euro area economies through over 270,000 published time-series. To better understand users, logs generated on BP*stat* were analysed to identify users missed by traditional tools. By using Python, a layered data collection approach was employed to reduce and clean over 200,000 daily logs, enabling efficient analysis and classification using an internal link following approach. This automation revealed previously unknown user behaviours and provided real-time insights for the Banco de Portugal.

Keywords: classification, segmentation, big data, layer approach, python

With over 270,000 published time-series, BP*stat*, the statistical website of the Banco de Portugal, offers users statistical information on the Portuguese and the euro area economies. Data can be exported in CSV or Excel format, as well as image formats (JPG, PNG, SVG, and PDF) and it is also possible to collect data automatically online using BP*stat* Application Programming Interface (API).With this wide range of possibilities, knowing our users, their preferred tools and formats " [...] is crucial to decide the type of content to create, the level of technical complexity or simplification of language that must be assumed, and the channel that should be privileged" [1]. By using Google Analytics, we can analyse website traffic and user behaviour, if users accept cookies. However, a deeper knowledge of our users also means to understand all the segments they explore in BP*stat*. But how to detect those who do not accept cookies or just extract our information automatically without accessing BP*stat* directly? Where there is a record, technology can reach it. The web server hosting BP*stat* records all events or activities that occur on the website over time (logs). These logs contain information such as the user's IP address, type of browser used, pages visited, date and time of visits, among other data, which allow us to characterize users. However, big data not only provides solutions but also brings challenges. The generated information – more than 200,000 logs per day – is humanly impossible to analyse and contains a lot of noise, making analysis and interpretation difficult. The challenge in automating this process is how to collect this information and automate the analysis to determine which of these logs refer to downloads and how many are performed by robots (automated processes)?

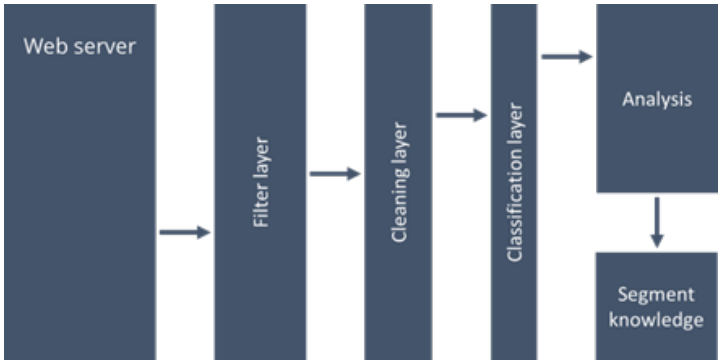

Figure 1: Layer approach

To solve the big data problem, we adopted a layered approach, where the collected information goes through 3 different layers before being analysed. The first layer aims to reduce the amount of information to be processed, reducing the number of logs by more than 97%. The second layer involves cleaning the logs by converting their text format into different columns according to their attributes, such as IP, date, or generated link. The third layer involves the classification of the generated link recorded in each log to identify global attributes, file download format, country of origin, among others. To solve the main problem of identifying who the robots really are, we adopted internal link following techniques. The link is traced to verify that its origin was a BP*stat* page where it is possible to generate downloads. If the download link does not have an associated BP*stat* page, it was generated automatically. Additionally, it is also possible to use the logs to count the number of API downloads, which can only be performed via programming. These links are constructed differently, allowing them to be separated from web-based automations. This analysis is carried out for the different segments identified in layer 3, based on the previously identified attributes, allowing to segment the analysis and obtain more accurate results.

Acknowledgements The process of automatic classification is carried out using Python and allows the identification of new user segments that were previously unknown. With this new robot, it was possible to account that in the last quarter of 2022, BP*stat* had over 6,000 robots from 70 different countries that performed more than 300,000 downloads of the time series data available on BP*stat*. This allowed us to identify robots that collected the same information every 20 seconds, drawing our attention to new types of risks and reminding us of the need for real-time quality controls.

References

[1] Lígia Maria Nunes, Ana Colaço, Rita Marques, and Mariana Oliveira. When reaching is no longer enough: 8 tips to engage with central banks' data users. In *UNECE Conference of European Statisticians*, 2021.

21 April, 14:50 - 15:10, Auditorium Francisco Sampaio

From people to Python: a new approach on securities holdings statistics quality assurance

Diogo Nobre[1], **Ludgero Glórias**[2], **Pedro Silva**[3]
[1] Banco de Portugal, danobre@bportugal.pt
[2] Banco de Portugal, lglorias@bportugal.pt
[3] Banco de Portugal, pmssilva@bportugal.pt

In the context of quality control of information reported for securities holding statistics, a Python solution was implemented to identify data issues that needed to be further addressed by the reporting agents, according to a set of criteria. With this solution we saw significant efficiency gains, by increasing the number of issues identified and, at the same time, drastically reducing the time spent in the process.

Keywords: python, automation, securities

Technological developments have led to an increase in the usage of computer capacity to automate work tasks. Through these automation mechanisms, some tasks can be performed with the same degree of quality as if they were performed by a human being, and, in many cases, with gains in speed and accuracy, resulting in more efficient processes. Python currently stands out as one of the most used programming languages for automation processes for its simplicity and flexibility. By automating these tasks, organizations can improve their productivity and focus on higher-value activities that require human intervention.

This paper presents a new automated process that is able to assist in the quality control of the data reported monthly by the Portuguese custodians (or investors) on securities held in custody by them or in their portfolio, capture possible reporting errors and prepare e-mails with questions to be sent to the respective reporting agent for further analysis.

Granular data on securities holdings and transactions for all the institutional sectors of the Portuguese economy are collected by Banco de Portugal on a monthly basis since 1999. In this process, more than fifty institutions report information about transactions and holdings of all securities held in custody or held in their own portfolio, on a security-by-security and investor-by-investor basis. Subsequently, an analysis process is carried out at the micro data level (transaction-by-transaction, security-by-security, investor-by-investor) to find the main highlights, different relevant events (custodian changes, share price changes, etc.), and potential reporting errors. Reporting agents are then questioned whenever inconsistencies are detected in the data. This overall process was extremely time consuming and repetitive and, to make it more efficient, we have developed an automatic

tool for detecting possible inconsistencies. In addition to this, the new tool also automatically creates the e-mail with the questions to be sent to the reporting agents for further analysis. We used Python for the development of the automation mechanism, in particular the following packages: Urllib; Sqlalchemy; Openpyxl and Pandas.
To detect possible inconsistencies, the following quality criteria were defined: securities transacted at zero value; evolution in market prices seem insufficient to explain the reported values; for unlisted securities, changes in stocks are not sufficiently explained by transactions; other changes in price and volume of at least 1M€.
Based on such criteria the questions to be sent are separated into two groups: zero-valued transactions, and market value questions. The new automated process brought a set of advantages, namely (1) questions are sent to the reporting agents on a more timely manner; (2) files are organized by set of questions, which makes it easier for the institutions to analyse the issues and provide an answer; (3) it allow us more time to focus on the statistical treatment of more complex transactions.

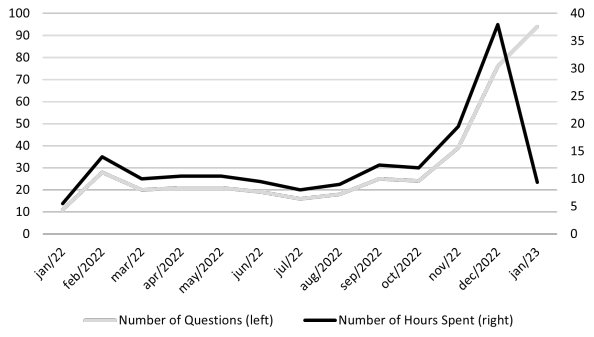

Figure 1: Number of questions (left) and hours spent (right) by production round

The process was first implemented in the January 2023 production round and the analysis of the results led us to conclude that there are expressive gains of productivity. Applying the new tool made it possible to increase the number of questions raised with a significant reduction in the associated effort (Figure 1). In particular, we were able to the reduce the time spent in this step of the statistical production process (analysing the data, preparing the questions, and sending the e-mails to the reporting agents) from 2-3 full days of work of 2 people to only 1 full day of work of 1 person, i.e., a decrease of around 80% of time spent. It is also worth mentioning that the new tool was able to identify more than 200 questions, which were then prioritized and reduced to around 100 questions (nevertheless, an increase of around 23% vis-à-vis December 2022) in order to give reporting agents the possibility to analyse and answer in due time and, ultimately, adapt to this new system.
In the future, we are considering the development of further enhancements to the quality control process. Firstly, we intend to analyse the definition of different thresholds for the quality criteria and develop new consistency tests. Finally, we expect to develop a machine learning process to assist in distinguishing between true errors, that need to be corrected, and other type of anomalous, but correct, observations, such as outliers.

Disclaimer: The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

21 April, 15:10 - 15:30, Auditorium Francisco Sampaio

Give me a guess on the purpose of this transaction: diving in microdata on external money transfers

Martha Düker[1], **Helena M. Marques**[2]
[1] Deutsche Bundesbank, martha.dueker@bundesbank.de;
[2] Banco de Portugal, hmmarques@bportugal.pt

On a monthly basis, banks send to Banco de Portugal data on all cross-border settlements made daily on behalf of their individual customers. Former investigations on these data applied different machine learning methods to indicate first patterns. Those findings, as for example specific country or bank groups combined with the new developed heuristics from the dashboard approach and the analysis of external data sources are applied on the cross-border settlements to flag some transactions as being related with a higher probability to a specific balance of payments class and, if possible, item.

Keywords: dashboard approach, search-match approach, heuristics, institutional sector of households, balance of payments

On a monthly basis, banks send to Banco de Portugal data on all cross-border settlements made daily on behalf of their individual customers. This big volume of information[1] is reported with no statistical classification, but its potential is undeniable for external statistics related to the institutional sector of households. However, because of the absence of information about the purpose of these transfers, the classification of each transaction to the balance of payments (b.o.p.) items is not straightforward. Currently, these data, aggregated by month and country of origin/destination, combined with other sources, are used for some estimations. Still, there might be patterns within the settlement data of natural persons which could be used to classify those transfers on an individual basis.
First, to get a better understanding of the settlements' data for 2021 and 2022 stored in a SQL database, an ethnographic analysis[2] was performed using a Power BI dashboard, where summary statistics for the inflows and outflows of 37 financial institutions and 223 territories are included. In a second step, other data sources were evaluated to obtain insights on specific characteristics for households cross-border settlements: i) information about subscription fees and European financial headquarters of digital service providers[3],

[1] Average inflows per month for 2022: 2 388 million euro (2 631 210 settlements); average outflows per month for 2022: 1 174 million euro (3 919 210 settlements)

[2] Ethnography is one of many methods in social sciences and often deals with unstructured data, where the data is not yet finally categorized, and the categories are often unknown at the beginning [1].

[3] The research done in the course of this work allowed for the systematization of sparse and decentralized information about digital service providers present in Portugal, like the segment where they operate, type of contract and the country of the European financial headquarters, which can be assumed to be the country of destination of the customers' payments.

ii) data on real estate sales to non-residents; iii) payments' data where the merchant category codes[4] are included, and iv) stocks on loans and deposits of private households. Additionally, knowledge about the specialization in specific lines of business of some banks was considered (for instance, consumer credit). Third, the findings on specific country or bank groups from former investigations on these data at Banco de Portugal combined with the new developed heuristics from the dashboard approach and the analysis of external data sources were applied to flag some transactions as being related with a higher probability to a specific balance of payments class and, if possible, item. The dashboard will be automatically updated on a monthly basis through SQL-queries and can be used as a tool to monitor bank, country, or time-related characteristics of the information received. The flags will be applied through pre-defined rules based on the developed heuristics with different SQL-queries to the individual transactions, signaling those that are more likely to be assigned to a specific b.o.p. class. If the so far developed heuristics are applied on the settlements data with respect to the total volume by year and direction, for the inflows are flagged 10 percent for 2021 and 7 percent for 2022 and for the outflows are flagged 5 percent for 2021 and 4 percent for 2022 (Figure 1 left, shows the respective amounts).

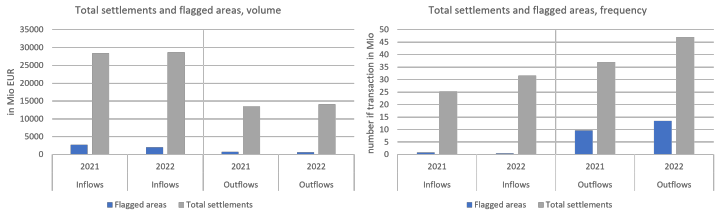

Figure 1: Applied heuristics on settlements' data, 2021 and 2022, volume in EUR and number of transactions

For the total number of transactions by year and direction, this results in flagged inflows being 3 percent for 2021 and 1 percent for 2022, and flagged outflows being 26 percent for 2021 and 29 percent for 2022 (Figure 1 right, shows the respective number of transactions). It may appear that the numbers of flagged transactions are relatively low. Therefore, further research should focus on additional data which can give supplementary information on specific characteristics of transactions to develop additional heuristics which can be applied through search-match-methods on the settlement's data.

References

[1] M. Hammersley and P. Atkinson. Ethnography. principles in practice. Third edition, 1983.

[4]Merchant Category Codes (MCCs) are used in the paayment processing industry to identify the type of business in which a merchant is engaged.

Thematic Session
Statistics Portugal

21 April, 15:30 - 15:45, Auditorium Francisco Sampaio

Enterprise mortality - A prediction model

Alexandre Cunha[1], **Rui Gouveia**[2], **Sandra Lagarto**[3], **Vasco Cordeiro**[4], **Vera Dias**[5]
[1] Statistics Portugal, alexandre.cunha@ine.pt
[2] Statistics Portugal, rui.gouveia@ine.pt
[3] Statistics Portugal, sandra.lagarto@ine.pt
[4] Statistics Portugal, vasco.cordeiro@ine.pt
[5] Statistics Portugal, vera.dias@ine.pt

A machine learning model was developed to predict which companies are at risk of closing their activities on a monthly basis using four data sources. The objective is to identify economically dead companies earlier than the current annual method. The model was evaluated using various metrics, and the gradient boosted machines model performed the best, although false positives were identified. Further analysis is needed to improve the model's precision and reliability, but this model seems to have a significant potential to anticipate company deaths.

Keywords: classification, gradient boosted machines, enterprise mortality

This study aims to develop a classification/prediction model based on machine learning techniques that can identify companies that are at risk of closing their activities on a monthly basis. Currently, the identification of economically dead companies is done annually, if they do not belong to the populations of companies in the Sistema de Contas Integradas de Empresas (SCIE) in the following two years. The use of monthly administrative data sources may allow earlier identification of dead companies.

The classification/prediction model was developed based on four data sources: the SCIE, the e-Fatura of Autoridade Tributária (e-Fatura), the Declaração Mensal de Remunerações of Segurança Social (DMR-SS) and the Ficheiro de Unidades Estatísticas (FUE). The objective is to predict monthly whether a company is dead or not based on historical data, thus enabling the anticipation of company deaths for up to 11 months and, at any time, without having any information about the activity of the companies in the next 24 months, predict their death.

Different Machine Learning models were used, including logistic regression, K-nearest neighbors, classification trees, random forests, and gradient boosted machines. To evaluate the accuracy of the model, various metrics such as accuracy, sensitivity, specificity, and ROC curve were analyzed. An analysis of the confusion matrices was also performed to

identify false positives and false negatives.

The results show that the model with the best results is the gradient boosted machines, with evaluation metrics above 90%. However, the precision values were low, resulting in a high percentage of false positives. Further analysis revealed that some of these false positives became true positives in the following months, indicating that the model is anticipating the death of some companies.

To improve the model and avoid false positives, it is necessary to conduct an exploratory study of misclassified companies, especially regarding their economic and financial characteristics. This can help explain the obtained results and refine the model for more precise and reliable results. In summary, this classification/prediction model has great potential to help anticipate the death of companies and enable more efficient management of business resources.

Acknowledgements
We would like to express our sincere gratitude to Miguel Godinho de Matos, Full Professor of Information Systems and Management at Católica Lisbon School of Business & Economics. We thank him for his guidance and support throughout the project. Providing valuable technical support, advice, helpful comments and suggestions, he was extremely important during all the process.

21 April, 15:45 - 16:00, Auditorium Francisco Sampaio

AI-based prediction and identification of errors in the geographic location of buildings

Aldina Piedade[1], **Carmen Costa**[2], **Bartholomeus Schoenmakers**[3]
[1] Statistics Portugal, aldina.piedade@ine.pt
[2] Statistics Portugal, carmen.costa@ine.pt
[3] Statistics Portugal, bart.schoen@ine.pt

One of the challenges for on-site data collection is to assure an accurate location of the buildings. During the last Census round Statistics Portugal the census fieldworkers have been updating the building dataset, however we know problems exists with the geographic location they identified. To correct these errors manually is a labour-intensive task, automatized methods can be a solution to minimize this.

Keywords: errors forecasting, errors analysis, building contours, random forest algorithm, artificial intelligence

---

This work presents a proposal for predicting and identifying errors in the geographic position of buildings in the Geographical Building Base(BGE), which is the geographic location component of different statistical units' databases of Statics Portugal (SP) as the National Dwellings Register (FNA) or National Base of Buildings (BNE).

From some of the BGE building we know the problems of the positional accuracy, we pretend to contribute with a methodological exercise using Artificial Intelligence (AI) techniques, that can predict which buildings are poorly positioned (off the roof), but also in the future are able to identify the existence of new buildings based on Aerial Photography (e.g., Orthophotomaps, Satellite Images, etc).

To proceed the development of this work, a test geographic area (north of Portugal) was defined, where a set of buildings point (within Geographic Information System software's) were considered "Well located" and another set was classified as "Badly located". The buildings footprints available online, such as BING and the Open Street Map (OSM) databases, were used to classify the building location.

The overlapping of the points of the buildings with the images (Orthophotomaps, 2018), thus allowed to obtain images of the 4 Ortho bands (red, green, blue and infrared) from a 10-meter buffer around the point and enabled the conversion of these images into numerical matrices.

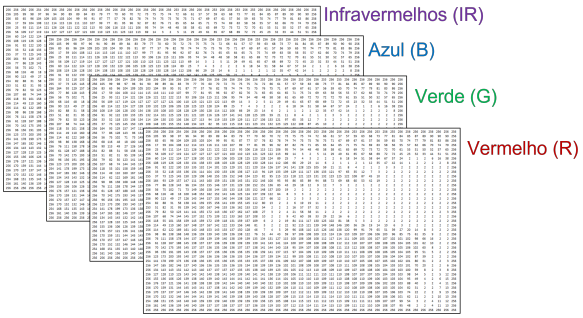

Figure 1: Example of numerical matrix for one building

Subsequently, a Machine Learning Model was applied using the generated matrices, as well as other information that characterizes the buildings, such as the number of entrances, number of floors, construction period, among others. The Random Forest algorithm was applied to this set of information to generate error prediction models, using the R Software. The model was subsequently evaluated with the ROC Curve method, obtaining an Area Under the Curve of 0.97.

This result showed that the applied methodology proved to be robust for the purpose, with technical and scientific room for improvement.

21 April, 16:00 - 16:15, Auditorium Francisco Sampaio

More data sources, more information, more quality

Sofia Rodrigues[1], **Almiro Moreira**[2], **Paulo Saraiva**[3], **João Poças**[4], **Bruno Lima**[5]

[1] Statistics Portugal, sofia.rodrigues@ine.pt
[2] Statistics Portugal, almiro.moreira@ine.pt
[3] Statistics Portugal, paulo.saraiva@ine.pt
[4] Statistics Portugal, joao.pocas@ine.pt
[5] Statistics Portugal, bruno.lima@ine.pt

The increasing and intensive use of non-statistical data sources (administrative and other) brings new and important challenges to Statistics Portugal and the official statistics production processes. In the context of Statistics Portugal's strategic objective for the creation of the National Data Infrastructure, the operating logic of the statistics production process in an integrated manner has changed. Recent organizational rearrangements at Statistics Portugal facilitate a change of the production chain towards a data-driven perspective. E-invoice data can be considered as the proof of concept of this strategy; therefore, its example is presented.

Keywords: national data infrastructure, administrative data, data integration

---

In the pursuit of the strategy of strengthening statistical production through the appropriation of a wide range of administrative data, Statistics Portugal (SP) has developed in the last years several initiatives, in which we highlight the strategic objective of the creation of the National Data Infrastructure (IND), which main objective is to create a single point of access to the various types of data and make it available to serve multiple purposes or projects, either to produce official statistics by Statistics Portugal or for research purposes. Statistics Portugal has made a significant investment in learning new skills, tools, and techniques, to process and analyse (massive) sets of data, to be internally available to produce statistics, in a very short period.

The adjustments made in the internal organization of the Institute, reinforcing the competences of Methodology and Information Systems Department and the Management and Data Collection Department in the data management and data analysis were an important tool to the exploitation of new data sources. Following the strategy of intensive use of administrative and other data, in March 2020, SP created a new organic unit, devoted exclusively to the collection and analysis of administrative data.

Concerning the analysis and processing of the various types of data, a top-down logic is put in place, in which processes such as the identification and treatment of outliers are

carried out in a phased manner, starting the analysis at the activity (NACE) level. The figure below represents the outlier treatment in the e-Invoice data:

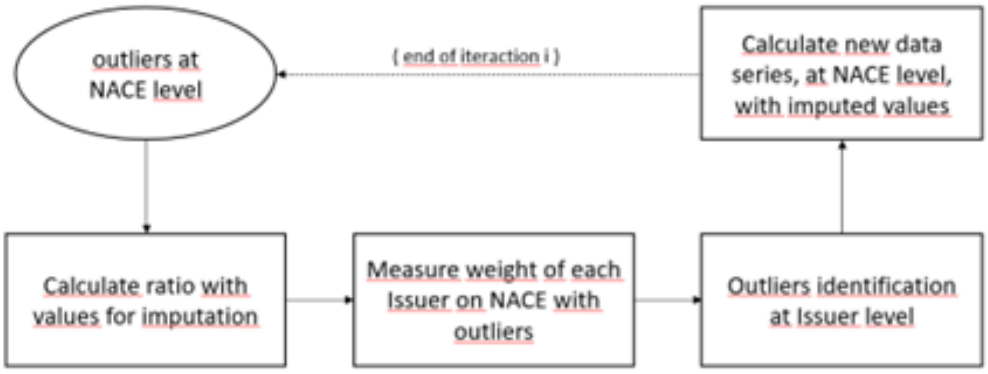

Figure 1: Outliers treatment

The treatment of the e-invoice data was a very important experience, rapidly making available huge data sets already processed, transformed, and enriched proved to be an important contribution to building and meeting IND' objectives. The example of the e-invoice data treatment played an important role in applying a set of procedures already used in traditional sources (as surveys) and to be followed for all the other administrative sources.

References

[1] Barteld Braaksma and Gert Buiten. Redesign of the chain of economic statistics in the netherlands. Technical report, Statistics Netherlands, 2012.

[2] ESTP. Methods and selected topics. Technical report, Eurostat, 2018.

[3] UNECE. The impact of globalization on national accounts - annex 2.2 – a consistency unit at statistics netherlands: reducing asymmetries in national accounts and related statistics. Technical report, United Nations, New York and Geneva, 2011.

[4] Thom Werkhoven. Improving cross domain coherency in dutch business statistics. Technical report, Dutch Statistics, 2013.

21 April, 16:15 - 16:30, Auditorium Francisco Sampaio

Access to official statistical information for scientific research purposes

Pinto Martins[1]
[1] Statistics Portugal, pinto.martins@ine.pt

According to the Law on the National Statistical System (Law No. 22/2008), individual statistical data on individuals and organizations for scientific purposes can be provided in an anonymized form.

To comply with this regulation, the conditions and procedures required for access by researchers to anonymized individual statistical data from the databases held at Statistics Portugal, produced by Statistics Portugal and by the entities with delegated powers, for scientific purposes, are presented.

Keywords: free and privileged access, accreditation of researchers, anonymized individual statistical data, official statistics, public use files

---

The access to anonymized official microdata by researchers in Portugal is guaranteed by Statistics Portugal (SP).

Statistics Portugal, aware of the fact that the academic community has special needs regarding statistical information, namely for the development of research work and for the elaboration of Master and PhD theses, established a Protocol with the Portuguese Foundation for Science and Technology (FCT) and the Directorate-General of Education and Science Statistics (DGEEC), with the purpose of facilitating access by (accredited) researchers to official statistical information needed to carry out their activity.

The protocol concerns researchers from universities and other legally recognized higher education and research institutions. Researchers with proven affiliation in international organisations such as: specialised agencies of the United Nations (International Labour Organisation (ILO), Food and Agriculture Organisation (FAO), United Nation Educational, Scientific and Cultural Organisation (UNESCO), The World Bank Group and the International Monetary Fund (IMF), and OECD, are also eligible for accreditation concerning their expertise and reputation in quality scientific research.

Each researcher must sign a form and a Statement of Confidentiality Commitment. The accreditation is valid during the declared length of the research project and only for the data identified in the request. It requires signature of a Code of Conduct by the applicant and the research institution of affiliation.

Under the protocol three access modes are possible to be authorized, including provision of fully anonymised data files and ready-made tables that allow no form of re-identification of statistical units; and exceptionally, on-site access in a safe environment, allowing the use of indirectly identifiable microdata under strictly controlled conditions (subject to a previous additional assessment by SP and an external group of experts in the statistical domain of the request).

PhD students have access under the same conditions as other researchers. Master's students need to fulfil an additional condition: the request and the statement of commitment must be also signed by the supervisor.

Non-resident researchers can access statistical data under the same conditions as the Portuguese ones in case they participate in a Foundation for Science and Technology Portuguese training scholarship or in cooperation programmes in R&D with Portugal.
Access to this information is free of charge for duly accredited researchers. In addition, and to meet the needs of other users, but also of researchers, to access more detailed information, Statistics Portugal has prepared some files with information at the observation unit level - the so-called Public Use Files (FUPs). These files (data and metadata) contain anonymised records, treated and prepared in such a way that the observation unit cannot be identified directly or indirectly, except when it is individual statistical data on Public Administration. They are freely accessible and comply with the principle of statistical confidentiality and protection of personal data.
This access implies prior acceptance of the conditions of use and is directly available for download from the Statistics Portugal's website (www.ine.pt).

Thematic Session
CLAD 2023 Scholarships

21 April, 17:00 - 17:20, Auditorium Francisco Sampaio

Fisher discriminant analysis for interval data

Diogo Pinheiro[1], **M. Rosário Oliveira**[2], **Igor Kravchenko**[3], **Lina Oliveira**[4]
[1] Instituto Superior Técnico, Univ. Lisboa, diogo.pinheiro.99@tecnico.ulisboa.pt
[2] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, rosario.oliveira@tecnico.ulisboa.pt
[3] Instituto Superior Técnico, Univ. Lisboa, igor.kravchenko@tecnico.ulisboa.pt
[4] CAMGSD and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, lina.oliveira@tecnico.ulisboa.pt

---

In Data Science, entities are usually described by vectors of single-valued measurements. Symbolic Data Analysis can model more complex data structures such as intervals and histograms that possess internal variability. In this work, we propose an extension of multi-class Fisher Discriminant Analysis to the interval scenario based on Mallows' distance and Moore's algebraic structure. We illustrate this symbolic approach in the context of classification of a real financial dataset with several classes.

Keywords: symbolic data analysis, interval-valued data, multi-class classification, symbolic fisher discriminant analysis, mallows' distance

---

Fisher Discriminant Analysis (FDA) is a method that uses data reduction to separate classes of multivariate data. The goal is to find low-dimensional subspaces where the data can be projected to make the distinction between classes more perceptible. Finding the directions that best characterize these subspaces can be formulated as a maximization problem of the ratio between a measure of the variability between classes and a measure of the variability within classes [2].

FDA and most classification methods focus on the analysis of points in $\mathbb{R}^p$, $p \in \mathbb{N}$, as it is the most common type of data available. We refer to it as Conventional Data. We may, however, be interested in more complex data structures, such as lists, histograms, or distributions, able to capture variation or explain phenomenons that conventional data cannot. All of these types of data are examples of Symbolic Data and are the subject of study of Symbolic Data Analysis. These structures may result from the aggregation of conventional data according to the research interests or may exist in their own right. We focus on interval-valued data, that is, objects whose measurements are intervals, uniquely identified by their center and range (i.e., width). We refer to the intervals as *macro-data* and to the underlying distribution of conventional data inside them as *micro-data*.

In this work, we propose an extension of the conventional Fisher Discriminant Analysis based on the Mallows' distance [1] and the Moore's interval algebraic structure [3], where the latter generalizes the usual definition of linear combination of vectors, and the former depends on the quantile functions of the interval micro-data.

By considering a model that links the macro-data with the underlying micro-data [4], we are able to show that the squared Mallows' distance between two intervals can be expressed as the sum of the squares of the Euclidean distance between the intervals' centers and the weighted Euclidean distance between the intervals' ranges. The contribution of the ranges is weighted according to the micro-data distribution within the intervals, extending previous work on this topic.
The Huygens theorem [1] allows for a decomposition of inertia based on the squared Mallows' distance that we adopt to generalize Fisher's objective function to interval data. Similarly to the conventional case, the directions of the projections can be found by solving a maximization problem. This now requires measures of the between and within variabilities from both the centers and ranges appearing in the summands of the squared Mallows' distance.
We illustrate this symbolic approach with a financial example that consists in using different stock price indicators, as a means to characterize a company and automatically classify it in one of the considered industrial classes of primary business activity. Results suggest the existence of differences among the classes, a conclusion that is not obvious from the variations of the stock prices alone.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia, Portugal, through the projects UIDB/04621/2020, PTDC/EGE-ECO/30535/2017, and UID/MAT/04459/2020.

References

[1] A. Irpino and R. Verde. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Data Science and Classification*, pages 185–192, 2006.

[2] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 2007.

[3] R. Moore, R. Kearfott, and M. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, 2009.

[4] M. Oliveira, M. Azeitona, A. Pacheco, and R. Valadas. Association measures for interval variables. *Advances in Data Analysis and Classification*, 16:491–520, 2021.

21 April, 17:20 - 17:40, Auditorium Francisco Sampaio

Handling missing data in the prediction of childhood obesity: a simulation study

Mafalda Oliveira[1,2], **Susana Santos**[2], **Rita Gaio**[3,1]
[1] Faculdade de Ciências da Universidade do Porto mafalda.ines.oliveira@gmail.com
[2] EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal & Laboratório para a Investigação Integrativa e Translacional em Saúde Populacional (ITR), Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal, susana.m.santos@ispup.up.pt
[3] Centro de Matemática da Universidade do Porto, argaio@fc.up.pt

Missings can lead to bias and incorrect conclusions. On the complete dataset from GENERATION XXI, initially with 5.39% missings, a simulation study concerning the same percentage of missings was carried out, involving six imputation methods and a logistic regression model predicting childhood obesity. The best-performing method (MICE with predictive mean matching) was then applied to the original entire dataset.

Keywords: missings values, single imputation, mice, simulation, logistic regression

Missing values refer to the lack of information in the dataset resulting from a missing or incorrect value for a particular variable in a specific observation. The existence of missing values is one of the most inconvenient problems in data analysis; as it may introduce bias and lead to incorrect conclusions, a thorough analysis should be conducted. This study performs a simulation to compare the performance of six imputation methods in logistic regression. The most appropriate methodology is then applied to the dataset under analysis and a logistic regression model to predict obesity at age 13 was fitted to the imputed dataset.

We used data from the cohort study GENERATION XXI - G21, including variables collected at birth regarding the mother and the newborn. The initial dataset had 5.39% missings and its complete version comprises 3504 mothers and 17 variables. The comparison of the performances of the six missing imputation methods applied (Mean/ Mode, Random, Hot Deck, MICE with predictive mean matching (MICE-Pmm), MICE with linear models (MICE-Lm), and MICE with random forests (MICE-Rf)) consisted of a simulation study, carried out in the following way. Firstly, a subset (n=3504) of the complete cases of the entire dataset (n=4637) was considered, and a logistic regression model predicting obesity was performed. In this subset, a proportion of missing values equal to the proportion of missing values in the initial dataset (5.39%) was randomly generated, giving rise to a dataset with an MCAR mechanism. The process was repeated 150 times. For every new

dataset with missing values, the six imputation methods were performed followed by the logistic regression model predicting the outcome variable (obesity at age 13). For each imputation method and generated dataset, the model's coefficients, respective standard errors, and corresponding p-values were retained. Since the simulation was repeated 150 times, there were 150 values for each parameter in each imputation method. The sample percentile of each actual parameter value was then computed, in every model. To evaluate the overall results, the mean and standard error of the percentiles are displayed in table 1, including coefficients, p-values, and standard deviations.

Table 1: Mean (standard error) for the sample percentiles for each method.

	MICE-Lm	MICE-Pmm	MICE-Rf	Hot Deck	Mean	Random
Coefficients	0.47 (0.11)	0.49 (0.05)	0.52 (0.14)	0.49 (0.21)	0.51 (0.16)	0.55 (0.24)
P-Values	0.44 (0.16)	0.40 (0.15)	0.36 (0.19)	0.52 (0.18)	0.45 (0.19)	0.42 (0.21)
Standard Errors	0.02 (0.05)	0.02 (0.04)	0.04 (0.05)	0.92 (0.23)	0.33 (0.41)	0.73 (0.13)

The values presented in table 1 indicate that MICE-Pmm was the most consistent method across the values of the coefficients since the mean was 0.49 and the values had a standard error of 0.05; this means that all coefficients had a percentile value near 0.5 which is desired. Although the other methods returned a mean close to 0.5, the standard errors were greater suggesting that there were several coefficients in which the percentile was far from 0.5. The percentile values for the P-values estimates were similar across all methods except for the MICE-Rf in which mean values were further from 0.5. Therefore, the MICE-Pmm is the best imputation model for the dataset under analysis. The values for the standard errors are unsatisfactory for all the imputation models.
Multiple Imputation with Chained Equations (MICE) with predictive mean matching (for continuous variables), binary logistic regression model (for binary variables), and multinomial logistic regression model (for polytomous variables) was therefore applied to the entire dataset, and a logistic regression model to predict obesity was carried out. Backward elimination was used as the method for the selection of the variables in which the least significant variable was removed from the model until all variables were significant. The final model identified the mother's age, years of education, marital status, smoking habits during pregnancy, gestational weight, mother's BMI, weight gain during pregnancy, and the newborn's weight and length as factors associated with obesity at age 13.

Acknowledgements G21 was funded by Programa Operacional de Saúde – Saúde XXI, Quadro Comunitário de Apoio III and Administração Regional de Saúde Norte (Regional Department of Ministry of Health). It has support from the Portuguese Foundation for Science and Technology and from Calouste Gulbenkian. This study was supported by the European Union Horizon 2020 Research and Innovation Programme under Grant Agreement 824989 (EUCAN-Connect) and 874583 (ATHLETE).

21 April, 17:40 - 18:00, Auditorium Francisco Sampaio

Nowcasting the Portuguese unemployment rate with Google Trends

Eduardo Andre Costa[1], **Maria Eduarda Silva**[2]
[1] Faculdade de Economia da Universidade do Porto, ecosta.phd@fep.up.pt
[2] Faculdade de Economia da Universidade do Porto, INESC TEC - LIADD and CEF.UP, mesilva@fep.up.pt

This research nowcasts Portuguese unemployment rates, from 2019 to 2021, by examining Google Trends daily and monthly series related to the labour market and explores their prediction quality during the COVID-19 outbreak. The results demonstrate the impressive prediction quality of GT daily indicators and how the pandemic adversely influences the predictions.

Keywords: nowcasts, unemployment rate, google trends, covid-19

The unemployment rate is a crucial indicator of an economy's health. However, delays in monthly announcements of such official statistics are usual. In addition, worldwide events such as the COVID-19 pandemic and the 2007-2008 global financial crisis prevented the timely issue of labour health estimates. The delays hamper policymakers in evaluating an economy's present or recent past state, which challenges researchers to find relevant predictors to accurately measure the current state of an economy.

The widespread availability of data generated from online activities, as well as the cost-free and promptness, have enabled researchers to use these data to improve estimates [3], anticipate announcements [3], complement traditional data sources [1], and gain insights into macroeconomic indicators [2]. Specifically, search engine data is becoming increasingly important for economic research and analysis, as it can supply insights into economic trends, analyse markets and anticipate future economic activity. In this context, Google Trends (GT) is a well-known search engine data source which provides readily available time series at various frequencies, from hourly to monthly, based on the demand of users of Google's search engine.

Researchers have used GT data to track labour market changes and gain insight into the impacts of different economic events. Furthermore, previous studies have demonstrated that series from GT can effectively predict unemployment rates for countries such as the USA, France, Spain, Italy, the UK, Turkey, Canada, the Czech Republic, Hungary, Poland, Slovakia and Portugal.

Specific to the Portuguese case study, the National Statistical Office publicises monthly unemployment rates delayed 60 days from the end of a reference month on average. Accordingly, this study nowcasts the Portuguese monthly unemployment rate from Jan-19 to Nov-21 using GT daily and monthly series as predictors and reproduces the integration

between official rate publication and GT data availability. The aims include generating two sets of nowcasts, one based on lengthy daily frequency GT series predictors using Mixed Data Sampling (MIDAS) regressions [4] and the other based on monthly GT series as predictors via ARIMAX. The nowcast prediction accuracy is compared between each set of nowcasts and against benchmarks based exclusively on historical values of the official unemployment rate. Additionally, the accuracy is analysed for the whole prediction period and divided into civil years (2019, 2020 and 2021), allowing the predictions assessment under the coronavirus outbreak situation.
The results show that GT data supply relevant information for predicting the Portuguese unemployment rate. Before the COVID-19 pandemic (2019), predictions were notably the most accurate forecasts, followed by 2021 and 2020, when the pandemic peaked. Nowcasts using GT daily data produced more precise predictions than GT monthly data and benchmarks, indicating the capacity of this sampling frequency to extract sharper insights into unemployment dynamics.

Acknowledgements The authors gratefully acknowledge support from CLAD and CEF.UP. This research has been financed by Portuguese public funds via FCT (Fundação para a Ciência e a Tecnologia) and ESF (European Social Fund) through individual PhD scholarship with reference 2021.07583.BD, and in the framework of the FCT project with reference UIDB/50014/2020.

References

[1] Roberto Barcellan, Peter Bøegh Nielsen, Caterina Calsamiglia, Colin Camerer, Estelle Cantillon, Bruno Crépon, Bram De Rock, László Halpern, Arie Kapteyn, Asim I. Khwaja, and et al. Developments in data for economic research. In Laszlo Matyas, Richard Blundell, Estelle Cantillon, Barbara Chizzolini, Marc Ivaldi, Wolfgang Leininger, Ramon Marimon, and FrodeEditors Steen, editors, *Economics without Borders: Economic Research for European Policy Challenges*, pages 568–611. Cambridge University Press, Cambridge, 2017.

[2] Daniel Borup and Erik Christian Montes Schütte. In Search of a Job: Forecasting Employment Growth Using Google Trends. *Journal of Business & Economic Statistics*, pages 1–15, August 2020.

[3] Dario Buono, Gian Luigi Mazzi, George Kapetanios, and Massimiliano Marcellino. Big Data Types for Macroeconomic Nowcasting. In Paul Konijn and Domenico Sartore, editors, *Eurostat Review on National and Macroeconomic Indicators*, pages 93–145. Publications Office of the European Union, Luxembourg, 2017.

[4] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. The MIDAS Touch: Mixed Data Sampling Regression Models. Mimeo, UNC and UCLA, 2004.

21 April, 18:00 - 18:20, Auditorium Francisco Sampaio

Using wavelets to denoise sound data and identify patterns: a narwhal example

Carolina S. Marques[1], **Emmanuel Dufourq**[2], **Carl Donovan**[3,4], **Marianne Marcoux**[5], **Tiago A. Marques**[1,3,6]
[1] Centro de Estatística e Aplicações, Universidade de Lisboa, Lisbon, Portugal, csmarques@fc.ul.pt
[2] African Institute for Mathematical Sciences, Muizenberg, South Africa, dufourq@aims.ac.za
[3] Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, Scotland, crd2@st-andrews.ac.uk
[4] DMP Statistical Solutions, crd2@st-andrews.ac.uk
[5] Arctic Aquatic Research Division, Fisheries and Oceans Canada, Winnipeg, MB R3T 2N6, Canada, marianne.marcoux@dfo-mpo.gc.ca
[6] Departamento de Biologia Animal, Universidade de Lisboa, Lisbon, Portugal, tiago.marques@st-andrews.ac.uk

---

Sound files containing wildlife sounds for identification have background noise. One way to denoise the data is by using wavelets. We consider a discrete wavelet transform (DWT), which approximates signals with a series of wavelets at different scales. DWT are used to identify features of interest. We use data from narwhal tags and seek to extract their vocalizations. We identified around 20000 sounds and use a validation dataset of 19 minutes; we estimate a sensitivity of 0.747. This suggests DWT can reliably clean sound data and open doors for more advanced deep learning methods.

Keywords: DWT, denoise, narwhal, pattern identification, signal processing

---

For some species it is easier to study them by their vocalizations. A common tool to obtain their vocalizations are animal borne tags with acoustic sensors. Manually detecting their vocalizations can be extremely time consuming, hence, there is a need to find methods that increase its efficiency.
One of the most common used methods to denoise signals is the discrete wavelet transform (DWT). The DWT needs the "mother wavelet" (MW), which refers to the wavelet function that is used to perform the wavelet transform. The choice of the MW affects the frequency decomposition of the signal and the coefficients obtained. In this study we evaluate the performance of 8 different MW (Figure 1).
The first level of the decomposition (LD) corresponds to the approximation coefficients and contains the low-frequency components of the signal. The following LD are detail coefficients, which contain the high-frequency components of the signal [1]. The length of

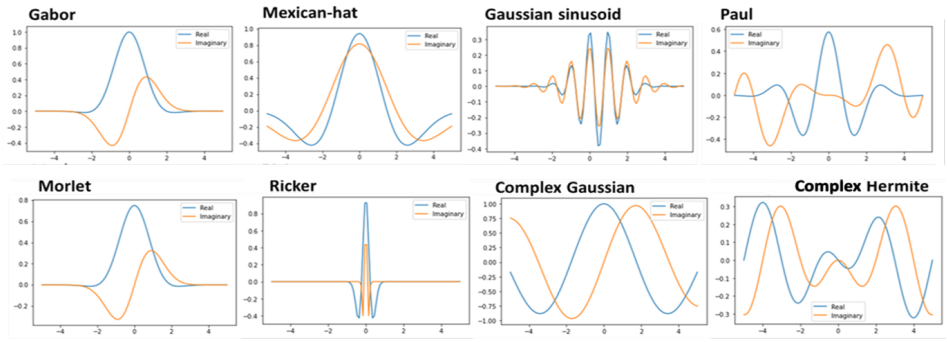

Figure 1: Different MW used for this study.

the LD will be smaller than the original signal, and each LD corresponds to a different frequency band [1]. We compare the results obtained by the different MW. The Gabor wavelet with the third LD gave the best result based on the DWT results (Figure 2).

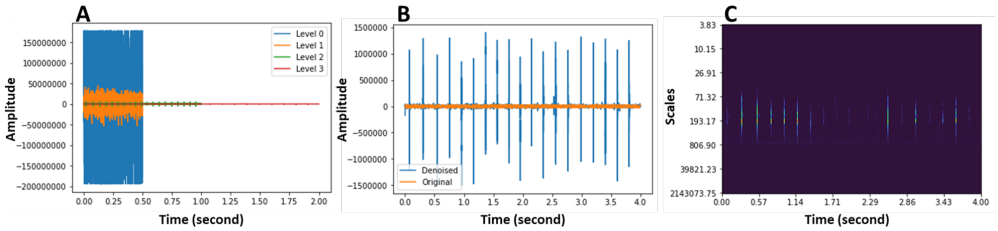

Figure 2: **A**- The results from the DWT. **B**- The comparison between the third LD and the original signal. **C**- Scalogram for the third LD.

For the automatic detection we find the peaks in the clean signal, with certain characteristics, like height, width and distance to the previous peak. To obtain a validation on the results we manually analysed the first 60 seconds on each of the segments of the data (around 20% of the data) and then compared to the results. We get a precision of 0.852, a sensitivity of 0.747 and an F-1 score of 0.796. These results show that this is an easy and clean way to denoise the sound, which then allows an easier identification of vocalizations.

Acknowledgements This research was supported by FCT through funding of Centro de Estatística e Aplicações (CEAUL), ref. UI/BD/154258/2022 and it was only possible thanks to the ACCURATE project, funded by the US Navy Living Marine Resources program (contract no. N3943019C2176).

References

[1] P. S. Addison, J. Walker, and R. C. Guido. Time–frequency analysis of biosignals. *IEEE Engineering in Medicine and Biology Magazine*, 28(5):14–29, 2009.

21 April, 18:20 - 18:40, Auditorium Francisco Sampaio

Are multilayer networks useful for mining multivariate time series?

Vanessa Freitas Silva[1], **Maria Eduarda Silva**[2], **Pedro Ribeiro**[3], **Fernando Silva**[4]

[1] CRACS-INESC TEC, Faculdade de Ciências, Universidade do Porto, vanessa.silva@fc.up.pt
[2] LIAAD-INESC TEC, Faculdade de Economia, Universidade do Porto, mesilva@fep.up.pt
[3] CRACS-INESC TEC, Faculdade de Ciências, Universidade do Porto, pribeiro@fc.up.pt
[4] CRACS-INESC TEC, Faculdade de Ciências, Universidade do Porto, fmsilva@fc.up.pt

Summary features from time-indexed data have proved to be a promising approach in time series mining tasks, such as classification and clustering problems. For univariate time series, there are several sets of features available in the literature. However, that is not the case for multivariate time series. In this work, we propose a set of features for multivariate time series based on multilayer time series networks. The results indicate that these features capture characteristics of multivariate time series data useful for mining tasks.

Keywords: multivariate, time series, multilayer network, topological features, clustering

Nowadays, mobile and sensing devices enable collecting data over time with multiple variables, creating multivariate time series (MTS) datasets. Traditional univariate time series (UTS) analysis tools can be extended to multivariate settings, but the presence of serial and cross-dependence pose additional challenges. Therefore, new and improved approaches are needed for effective analysis of MTS.

Summary features (or statistics) have proved to be an important task in time series applications such as classification, clustering and forecasting [1]. Traditional methods for computing such statistics on UTS data involve conventional statistical and non-linear measures of time series analysis which are often complex calculations and can be impractical for certain datasets. For MTS, the available features are limited. Common approaches involve measuring correlation and causality between time series variables or extending UTS statistics to each variable. The latter involves concatenating feature vectors of each time series into a unique vector or calculate the correlation between features. However, this approach fails to capture cross-dependence in the MTS data.

Network science studies connections and relationships in a system using graph theory and topological features and has potentiated the emergence of mapping methods that transform and represent time series as graphs that model the dependencies between the data [3]. Multilayer networks are complex and high-level graphs that model multidimensional data

while preserving important properties [2]. These networks include intra-layer connections within the same graph (dimension) and inter-layer connections between different graphs (dimensions). Network science offers a wide range of topological features for characterizing single-layer networks, which can be extended and adapted to multilayer networks.
Motivated by the above, this work proposes novel MTS features based on a new mapping method that transforms the MTS into a multilayer network. This network is called the *multilayer horizontal visibility graph* (MHVG) and is based on a new visibility concept, *cross-horizontal visibility*. MHVG represents both serial and cross dependencies of the MTS through intra-layer and inter-layer edges [4]. From the MHVG we can extract a diverse set of features (at different level dimensions), namely, intra-, inter- and all-layer features and relational features, that result in a new set of MTS features.
This study aims to evaluate the effectiveness of MTS network features for clustering synthetic datasets generated from well-established MTS models. The topological features considered are based on average degree, average path length, number of communities, modularity, a new feature called average ratio degree, and Jensen-Shannon divergence. To this end, 100 samples of size $T = 10000$ from six bivariate models (white noise, VAR, GARCH models), in a total of 600 bivariate time series, are generated. These time series are then mapped into MHVG using horizontal and cross-horizontal visibility methods and topological features are extracted from the resulting MHVGs.
The results indicate that the features extracted from MHVG, which use both intra-layer and inter-layer edges between lagged nodes, preserve information about the MTS data (serial and cross-dependencies) achieving better accuracy when compared to conventional time series features. Different features of MHVGs capture different properties of the time series showing that these features are useful to MTS mining problems, such as clustering.

Acknowledgements The authors gratefully acknowledge support from FCT within project UIDB/50014/2020 and research grant PD/BD/139630/2018.

References

[1] Yanfei Kang, Rob J Hyndman, and Feng Li. GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2020.

[2] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[3] Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, and Fernando Silva. Time series analysis via network science: Concepts and algorithms. *WIREs Data Mining and Knowledge Discovery*, 11(3):e1404, 2021.

[4] Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, and Fernando Silva. MHVG2MTS: Multilayer horizontal visibility graphs for multivariate time series analysis, 2023.

Thematic Session
SPE

22 April, 11:40 - 12:00, Auditorium Francisco Sampaio

Statistical modelling methods applied to the diagnosis of temporomandibular disorders

Ricardo São João[1], **Henrique J. Cardoso**[2], **David F. Ângelo**[3]
[1] Polytechnic Institute of Santarém, Portugal; CEAUL, Portugal; CINDUR, Portugal; CEG-Uab, Portugal, ricardo.sjoao@esg.ipsantarem.pt
[2] Instituto Português da Face, Portugal, henrique.cardoso@ipface.pt
[3] Instituto Português da Face, Portugal; Centre for Rapid and Sustainable Product Development, Polytechnic Institute of Leiria, Portugal; Faculty of Medicine of Lisbon University, Portugal, david.angelo@ipface.pt

---

The aim of this presentation is to illustrate statistical modelling applied to temporomandibular disorders (TMD). The multifactorial aetiology and pathogenesis of TMD, its complexity and consequently its correct diagnosis are challenges in the medical field. Statistics can make an additional contribution to its better understanding. A diversified range of statistical approaches will be used based on the records present in the EUROTMJ database, considering patients diagnosed with TMD in the last three years.

Keywords: temporomandibular disorders, diagnosis, logistic regression, statistical modelling

---

Temporomandibular disorders (TMD) are a set of musculoskeletal and neuromuscular disorders that involve the temporomandibular joint, the masticatory muscles and all associated structures. Aetiologically, TMD are associated with various risk factors that, individually or together, contribute to triggering the development of the disease. Anatomical, pathophysiological, psychosocial, hormonal, traumatic and gender-related aspects are identified as risk factors. The most frequent symptom of TMD is orofacial pain that affects a large part of the population. Other symptoms are also frequently reported: sounds in the joint, limitation in opening the mouth, masticatory and cervical muscle tension, headaches, otalgia, ear fullness, tinnitus and vertigo. The origin of TMD symptoms may be related to muscular or intra-articular changes, or both.

Epidemiologically, it is not easy to estimate the prevalence of TMD due to the underdiagnosis performed by several health professionals. Due to its multifactorial aetiology and pathogenesis, it serves as motivation for the application of statistical modelling, justifying the topic of the present communication. To this extent, a 3-year prospective study was designed with patients diagnosed with TMD from 2019 to 2022 at the Instituto Português da Face. Clinical data, registered in the EUROTMJ database, was considered for its implementation.

The initial statistical approach analyzed the prevalence of clinical signs and symptoms of TMD and their association with sociodemographic characteristics, risk factors, comorbidities, most frequent complaints and parafunctional habits. Given the nature of some of these variables, measured on a nominal/ordinal scale, the contingency tables were used, namely to perform independence tests and association measures. On the other hand, when variables measured on an interval/ratio scale are considered, Pearson's parametric correlation coefficient was used or, alternatively, Spearman's non-parametric coefficient. Another relevant aspect concerns the distribution of different variables according to their TMD classification, which requires the use of the parametric ANOVA test or the non-parametric Kruskal Wallis test and, if necessary, the *post hoc* tests.

Subsequently, a relevant question for clinical practice and statistics will be to what extent the adoption of self-report questionnaires (e.g. Fonseca Anamnestic Index-FAI and Visual Analog Scale-VAS) can contribute to a correct classification of the TMD diagnosis. Definitive TMD diagnosis is made through clinical and imaging analysis, namely through computed tomography, magnetic resonance imaging or minimally invasive methods. We know that these procedures are expensive and time-consuming when compared to self-report questionnaires that are easy to answer, quick and inexpensive. However, it is not known to what extent the identification of certain patterns in these questionnaires together with other characteristics of the patients can contribute to a correct diagnosis of TMD. Therefore, we used Generalized Linear Models with logit link function, specifically the Multinomial and Ordinal Logistic Regression models, given the nature of the TMD response variable. The subdivision of the sample to determine the training model (adjustment) and the test model (validation and prediction), the use of discriminant measures and measures of quality of adjustment, as well as the selection of the covariates (manual vs. automatic choice) to be included in the final model, are important when combined with the remaining assumptions (absence of complete separation, absence of multicollinearity, proportional hazards, independence of irrelevant alternatives, etc.). The methodology now implemented is very well portrayed in the Frank Harrell's book [1]. The level of accuracy of the previous model(s) dictates the pertinence of using this type of scales obtained through self-report questionnaires or by seeking other alternatives.

Acknowledgements This work was supported by Research Grant SORG 2019-project: "EUROTMJ –recording TMJ outcomes in a central database". This work was co-financed by Fundação para a Ciência e Tecnologia (FCT), project UIDB/00006/2020

References

[1] F. E. Jr. Harrell. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer Series in Statistics, Switzerland, 2016.

22 April, 12:00 - 12:20, Auditorium Francisco Sampaio

Multivariate statistical quality control methods for monitoring the concentration of particulate matter in an occupational context

M. Rosário Ramos[1], **Carla Viegas**[2], **Susana Viegas**[3], **Elisabete Carolino**[4]
[1] CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, Universidade Aberta, Centro de Estudos Globais(CEG-UAb), mariar.ramos@uab.pt
[2] H&TRC – Health & Technology Research Center, ESTeSL – Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa(IPL), Comprehensive Health Research Center(CHRC) carla.viegas@estesl.ipl.pt
[3] H&TRC, ESTeSL, IPL, NOVA National School of Public Health, Public Health Research Centre, CHRC, susana.viegas@ensp.unl.pt
[4] H&TRC, ESTeSL, IPL; ISAMB – Instituto de Saúde Ambiental, Faculdade de Medicina da Universidade de Lisboa, Portugal, etcarolino@estesl.ipl.pt

---

Some adverse effects of exposure to particulate matter in an occupational context are already known, such as infectious allergies and respiratory symptoms. The aim of this study was to propose a methodology based on Statistical Process Control for monitoring particulate matter in occupational environments. Data are mass concentration of five particles (PM0.5, PM1.0, PM2.5, PM5.0 and PM10.0) corresponding to different impacts on workers' health. To monitor the concentration of particles, parametric and nonparametric multivariate statistical quality control charts were adopted to cover different data characteristics.

Keywords: multivariate control charts, Hotelling's T^2, principal components, bootstrap, occupational exposure

---

Due to adverse health effects it's important to ensure a monitoring system of concentration of particles in occupational environments [3], in support of the implementation of corrective and preventive measures, individual or collective. A methodology for monitoring concentration of this particles through a multivariate approach is proposed.
Classical control systems using Shewhart charts are often difficult to manage in real time. Moreover, the use of univariate control charts on measurements taken simultaneously for variables possibly correlated, can distorts type I error and wrongly estimate the probability of being within control limits. This distortion of the control procedure increases with the number of variables to be controlled. Two case studies, each in specific working conditions were analysed, taking mass concentration of particles of five diameters (PM0.5, PM1, PM2.5, PM5 and PM10) in $\mu g/m^3$. On a data-driven basis parametric and nonparametric approaches were selected.

Case study 1 is located in an animal feed production industry, variables under study are approximately multivariate normally distributed. Multivariate charts are used based on Hotelling's T^2 supplemented with Shewhart charts, assuming that the p-correlated variables are derived from measurements performed simultaneously and follow a multivariate normal distribution. For unknown parameters the test statistic becomes $T^2 = n(\overline{X} - \overline{\overline{X}})' S^{-1} (\overline{X} - \overline{\overline{X}})$. The Hotelling's chart shows two distinct phases of the review process. The first stage is used for establishing the control, often referred to as a retrospective process stage, where it evaluates its stability and calculates the values $\overline{\overline{X}}$ and S. The second stage has the function of monitoring the stability of the process. Upper (UCL) and lower (LCL) control limits for each stage are derived. The natural procedure would be to run multiple comparison tests to identify out-of-control particles. The goal was to propose a simpler method, reducing the number of variables to analyse. A principal component analysis was applied to reduce dimensionality and univariate control charts were calculated on each component scores. When a chart indicates out of control, the component loadings where considered to identify the particles with greater contribution to that state.
Case study 2 occurs in a horse stable environment, multivariate normal distribution can not be assumed. Three nonparametric types of control charts were proposed in [2]. The main idea was to reduce each multivariate measure to the univariate index, that is, its relative center-exterior classification induced by a depth of data. This approach is completely nonparametric. The new r, Q, and S charts can be considered as data-depth-based multivariate generalizations of the univariate X, $\bar{x}$, and CUSUM charts respectively. The author also shows how bootstrap can be applied in statistical quality control to obtain valid control charts for stationary weakly dependent data, as well as for iid data that are not normally distributed. They are easily applicable because they are completely nonparametric and were implemented in the R package qcr for Statistical Quality Control [1].
In both case studies it was found that the process was out of control in specific workplaces. It is important to ensure a monitoring system of particles with adverse health effects in these specific occupational environments, to support decisions on protection measures.

Acknowledgements: M. Rosário Ramos was partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. Elisabete Carolino was supported by FCT/MCTES (UIDB/05608/2020 and UIDP/05608/2020).

References

[1] M. Flores, R. Fernández-Casal, S. Naya, and J. Tarrío-Saavedra. Statistical quality control with the qcr package. *R Journal*, 13:194, 2021.

[2] R.Y. Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387, 1995.

[3] P. Thangave, P. Duckshin, and L. Young-Chul. Recent insights into particulate matter (pm2.5)-mediated toxicity in humans: An overview. *International Journal of Environmental Research and Public Health*, 19(12), 2022.

22 April, 12:20 - 12:40, Auditorium Francisco Sampaio

Predicting hospital patients' no-show through statistical and machine learning techniques

Ana Borges[1], **Mariana Carvalho**[2]
[1] CIICESI, Escola Superior de Tecnologia e Gestão, Politécnico do Porto, aib@estg.ipp.pt
[2] CIICESI, Escola Superior de Tecnologia e Gestão, Politécnico do Porto, mrc@estg.ipp.pt
.

Patients' uninformed absenteeism can have negative consequences for both patients and healthcare providers. By understanding the reasons behind missed appointments, healthcare systems can develop strategies to reduce the rate of absenteeism mitigating its harmful impact and improving patient outcomes. Statistical and machine learning approaches have been successfully applied to predict patients' uninformed absenteeism in healthcare settings. This study exemplifies the application of models in that context, namely: Logistic Regression, K-Nearest Neighbour, Random Forest and Gaussian Naïve Bayes. For that, real data on the speciality of Physical Medicine and Rehabilitation is analysed.

Keywords: patients' absenteeism, logistic regression,k-nearest neighbour, random forest, gaussian naïve bayes

---

Uninformed absenteeism occurs when a patient misses an appointment without notifying the healthcare provider in advance. It is proven that it can lead to several problems, both for the patient and for the healthcare provider, resulting in the seeming wastage of resources, both material and human [1, 2]. One way to mitigate the harmful impact of the patient's no-show is to predict the patient's behaviour.

To demonstrate the potential of statistical and machine learning methods in predicting patient absenteeism, we compare the performance of the Logistic Regression, K-Nearest Neighbour, Random Forest and Gaussian Naïve Bayes algorithms, using data from a hospital located in the north of Portugal. The choice of the mentioned methods is justified by the fact that we are dealing with a binary outcome (show versus no-show).

The collected data consists of information regarding 2001 patients and 61.522 appointments of the Physical Medicine and Rehabilitation specialty between 3rd November 2018 and 15th March 2020. Alongside the dependent binary variable with the value 0 if the patients did not attend the appointment and 1 otherwise, the following variables were considered: Gender, Age, Marital Status, the month of the scheduled appointment, weekday of the scheduled appointment, waiting time between appointment creation and its date, number of prior no-shows, the distance between residence and hospital, mean temperature

Table 1: Precision, recall, f1-score, accuracy and ROC AUC of each model.

	Logistic Regression				**K Nearest Neighbourd**			
	precision	**recall**	**f1-score**	**accuracy**	**precision**	**recall**	**f1-score**	**accuracy**
Train	0.568	0.629	0.597	0.576	0.747	0.729	0.738	0.741
Test	0.565	0.629	0.596	0.573	0.692	0.621	0.658	0.631
ROC AUC	0.663				0.568			
	Random Forest				**Gaussian Naive Bayes**			
Train	1.0	1.0	1.0	1.0	0.555	0.386	0.455	0.538
Test	0.702	0.861	0.774	0.748	0.564	0.392	0.462	0.544
ROC AUC	0.709				0.542			

and season (winter, summer, spring and autumn). Since the data has class imbalance (approximately only 28.8% of patients did not attend the appointment), there was the need to balance the data using SMOTE (Synthetic Minority Oversampling Technique) to optimize the results by over-sampling the minority class with synthetic data. The data was divided into 80% for training and 20% for testing and the four models were trained using a set of default parameter settings. The performance of the models was validated using 5-fold cross-validation and a set of performance metrics were analyzed, including accuracy, precision, recall and f1-score. Additionally, the Area Under The Curve of Receiver Operating Characteristics (AUC ROC) was calculated for each model. Table 1 presents a detailed overview of these metrics.

From the results presented in Table 1, we can detect overfitting for both Random Forest and K-Nearest Neighbour, since the models had a significantly better performance on the training set than on the test set. Therefore, in this case, the choice should fall on the Logistic Regression model since it presents higher values in the values of precision, recall, f1-score and accuracy, compared to Gaussian Naïve Bayes algorithm. For this model, the AUC ROC obtained was 0,663. From the results from the Logistic Regression, we can detect that female patients, and younger patients present a higher chance of a no-show. Appointments scheduled in the summer season or in August and September month present a higher risk of no-show, as days with higher mean temperatures. Also, a higher number of prior no-shows and higher waiting times are related to a higher risk of absenteeism.

Acknowledgements This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through project UIDB/04728/2020.

References

[1] J Dunstan, F Villena, JP Hoyos, V Riquelme, M Royer, H Ramírez, and J Peypouquet. Predicting no-show appointments in a pediatric hospital in chile using machine learning. *Health Care Management Science*, pages 1–17, 2023.

[2] Miguel Maia, Ana I Borges, and Mariana Cavalho. Risk factors associated with hospital unwarned appointment absenteeism: A logistic binary regression approach. In *Proceedings of the 2022 5th International Conference on Mathematics and Statistics*, pages 52–58, 2022.

Contributed Sessions

20 April, 16:00 - 16:20, Auditorium Francisco Sampaio

SLE-DAS performance in phase 3 clinical trials

Ana Matos[1], **Carla Henriques**[2], **Diogo Jesus**[3], **Luis Inês**[4]
[1]Polytechnic Institute of Viseu and CISeD, Portugal, amatos@estgv.ipv.pt
[2]Polytechnic Institute of Viseu and CMUC, Portugal, carlahenriq@estgv.ipv.pt
[3] Centro Hospitalarde Leiria, Rheumatology Department and Faculty of Health Sciences, University of Beira Interior, Covilhã, Portugal, jesus.p.diogo@gmail.com
[4] Faculty of Health Sciences, University of Beira Interior and Centro Hospitalar e Universitário de Coimbra, Rheumatology Department, Portugal, luisines@gmail.com

---

Using effective quantifiers to assess new drug efficacy in phase 3 clinical trials is essential. In Systemic Lupus Erythematosus (SLE) clinical trials, the currently used primary endpoint measures are believed to have poor discriminatory capacity to differentiate responders from non-responders. A total of 1684 SLE patients were included in this investigation based on a post-hoc analysis of the combined BLISS-52 and -76 trials study population. This study gives evidence that the new definitions of SLE-DAS remission and low disease activity have discriminatory ability to identify patients receiving active drug from placebo and is associated with positive impact in patients quality of life.

Keywords: statistical inference, SLE-DAS, clinical trials

Systemic Lupus Erythematosus (SLE) is a prototypic autoimmune disease with various clinical and serological manifestations. The presentation of the disease and its course over time is highly variable both within and between individuals. SLE disease activity can be classified into different categories based on the severity of the clinical manifestations. Remission and low disease activity refer to states in which disease activity is controlled or acceptable, respectively.

The Systemic Lupus Erythematosus Disease Activity Score (SLE-DAS) is a composite score that incorporates multiple domains of disease activity. This score was derived by Jesus et al., 2019 [2] to create a comprehensive tool with high sensitivity in detecting clinically significant changes in disease activity. It is implemented in a user-friendly online calculator and has been validated in numerous studies. Through multicenter studies, accurate definitions of SLE-DAS low disease activity (SLE-DAS LDA) and SLE-DAS remission are attached [1].

In this study, a total of 1684 SLE patients (merged study population in the BLISS-52 and -76 trials (NCT00424476; NCT00410384)) were included: 562 on placebo, 559 on belimumab 1mg/Kg and 563 on belimumab 10mg/Kg. Significantly more patients attained SLE-DAS LDA on belimumab 1mg/Kg and 10mg/Kg as compared with placebo, at week 52 (13.0% vs. 17.9%, OR=1.459, p=0.023, and 13.0% vs. 21.7%, OR=1.853, $p < 0.001$,

respectively). Additionally, more patients on belimumab 10mg/Kg achieved SLE-DAS remission as compared to placebo (10.1% vs. 14.7%, OR= 1.532, $p = 0.019$).
The individual's perceived physical and mental health (health-related quality of life - HRQoL) and fatigue at the time was assessed using the 36-Item Short Form Survey (SF-36) and Functional Assessment of Chronic Illness Therapy fatigue scale (FACIT-F), respectively. Comparing SF-36 and the FACIT-F scores between patients achieving SLE-DAS remission vs. non-remission and SLE-DAS LDA vs. non-LDA, we observe that patients attaining SLE-DAS remission and SLE-DAS LDA presented better scores at week 52 (all $p < 0.001$) (Figure 1).

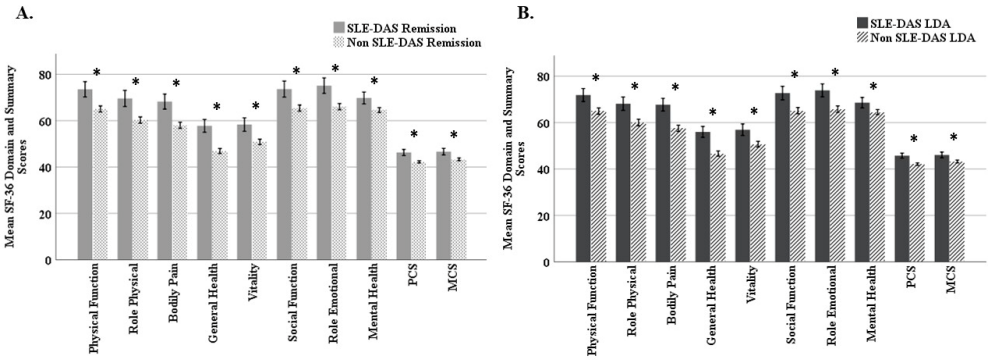

Figure 1: Mean SF-36 domain and summary scores, ($*p < 0.001$).

In conclusion, the SLE-DAS remission and LDA states are attainable treatment targets and have discriminant validity for identifying patients receiving active drug from placebo in clinical trials. The attainment of these SLE-DAS states is associated with positive impact in patient's quality of life.

Acknowledgements This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Ref[a] UIDB/05583/2020. Furthermore, we would like to thank the Research Centre in Digital Services (CISeD) and the Polytechnic of Viseu for their support and to thank GlaxoSmithKline (Uxbridge, UK) for granting access to the data from the BLISS-52 and 76 trials through the Clinical Study Data Request consortium.

References

[1] H. Assunção et al. Definition of low disease activity state based on the sle-das: derivation and validation in a multicentre real-life cohorty. *Rheumatology (Oxford)*, 61(8):3309–3316, 2022.

[2] D. Jesus et al. Derivation and validation of the sle disease activity score (sle-das): a new sle continuous measure with high sensitivity for changes in disease activity. *Annals of the Rheumatic Diseases*, 78:365–371, 2019.

20 April, 16:20 - 16:40, Auditorium Francisco Sampaio

Designing experiments for use in agriculture: the example of large field trials for grapevine selection

Elsa Gonçalves[1], **Antero Martins**[1]
[1] LEAF—Linking Landscape, Environment, Agriculture and Food Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal, elsagoncalves@isa.ulisboa.pt, anteromart@isa.ulisboa.pt

In initial phases of plant breeding, experiments with a large number of genotypes are evaluated; therefore, large field trials are required. The efficiency of the experimental designs of those field trials depends on the randomization process used to control environmental variation and the number of replicates. This work is focused on the importance of the number of replicates in large grapevine field trials for grapevine selection designed according to resolvable row-column designs.

Keywords: resolvable row-column design, linear mixed model, experimental designs, number of replicates, grapevine selection

To evaluate the most important economic traits, which are quantitative traits, well-designed experiments, which rely on the well-known principles of randomization, replication, and blocking, are needed. Randomization ensures that all experimental units are equally likely to receive any genotype, minimizing systematic errors or bias induced by the experimenter. Replication enables estimation of experimental error variance, and the precision of estimates increases with the number of replications. Finally, blocking controls the different sources of natural variation among experimental units and, when applied appropriately, controls field variation and helps to reduce background noise. These principles are strictly followed in agronomic experiments.

The efficiency of the experimental designs of large field trials depends on the randomization process used to control environmental variation. Blocking plays a key role in controlling spatial variability, water regimes, and farming operations. The need to control all these sources of variation led to establishing a two-dimensional layout of field trials, which have a strong tradition in agricultural experiments. Row-column designs [1] impose incomplete blocks in both row and column directions. Generating efficient row-column designs is a very important topic in experimental design research [2]. The overall objective is optimizing neighbour balance, ensuring that the number of pairwise adjacencies is as equal as possible across pairs of treatments over the field layout.

In the context of grapevine field trials for quantifying intravarietal variability and performing polyclonal selection (selection of a superior group of genotypes), simulation studies revealed that row-column designs are the most efficient when many genotypes are used and

the number of plots per incomplete block is greater than or equal to 10 [3]. Additionally, with real field data of yield, the importance of resolvable row-column designs to improve the control of spatial variation and background noise and the efficiency of grapevine selection was also proved [4].
Another factor in the success of an experiment is the number of replications. As is well known, the number of replicates is essential to allow a valid estimation of the error variance and to reduce its estimate. The methodological study presented in this work aims to demonstrate the importance of the number of replicates in the accuracy and precision of quantification of intra-varietal variability and genetic selection in grapevine. For this purpose, an exhaustive study comprising real yield data of several years from field trials of several autochthonous Portuguese grapevine varieties, designed according to resolvable row-column designs with 6 replicates, were considered.
Linear mixed model including genotypic and all design effects were fitted. For covariance parameters estimation, the residual maximum likelihood (REML) estimation method was used. Using as reference row-column designs with 6 repetitions, the relative bias and the relative mean square error associated with the estimates of the coefficient of genotypic variation and broad sense heritability for resolvable row-column designs with 5, 4, 3, and 2 repetitions were obtained to evaluate the accuracy and the precision of the quantification of the intra-varietal variability and selection.
The number of resolvable replicates of the field trial proved to be very important for efficient quantification of the intra-varietal genetic variability and selection in grapevine varieties. According to the results obtained, this last goal implies establishing resolvable row-column designs with no less than 4 replicates.

Acknowledgements This research was funded by the projects "Conservation and selection of ancient grapevine varieties" (PDR2020-784-042704) and "Save the intra-varietal diversity of autochthonous grapevine varieties (PRR-C005-i03–000016).

References

[1] Wiliams, E. R., John, J. A. Construction of row and column designs with contiguous replicates. *Applied Statistics*, 38, 149–154, 1989.

[2] Piepho, H.P., Williams, E.R. and Michel, V. Generating row–column field experimental designs with good neighbour balance and even distribution of treatment replications. *J Agro Crop Sci*, 207, 745–753, 2021.

[3] Gonçalves, E., St.Aubyn, A. and Martins, A. Experimental designs for evaluation of genetic variability and selection of ancient grapevine varieties: a simulation study. *Heredity* 104, 552–562, 2010.

[4] Gonçalves, E., Carrasquinho, I., Martins, A. Fully and Partially Replicated Experimental Designs for Evaluating Intravarietal Variability in Grapevine. *Australian Journal of Grape and Wine Research*, Article ID 5293298, 12 pages, 2022.

20 April, 16:40 - 17:00, Auditorium Francisco Sampaio

Bootstrap confidence intervals for association measures in sparse contingency tables

João Rocha[1], **Adelaide Freitas**[2], **Isabel Pereira**[3]
[1] Department of Mathematics & CIDMA, University of Aveiro, joao.corca@ua.pt
[2] Department of Mathematics & CIDMA, University of Aveiro, adelaide@ua.pt
[3] Department of Mathematics & CIDMA, University of Aveiro, isabel.pereira@ua.pt

---

When there is a small sample or lack of entries in a contingency table, Pearson's statistic may not be approximated by the chi-square distribution, conditioning inference about the association between the two nominal variables through Cramér's V. In this case, non-parametric bootstrap methods can be used. Based in a simulation study, we compared the coverage probabilities of various bootstrap confidence interval methods, namely, basic, percentile, BCa, and bootstrap-t, for V and an alternative association W. The methodology was illustrated with a real and small data set.

Keywords: chi-square, Cramér's V, bootstrap, confidence intervals, coverage probability

---

Contingency tables that have very small or no cell counts are said to be sparse, and provide both theoretical and computational challenges. This sparseness can happen either due to a small sample, or due to a large sample along with a relatively larger number of cells. The chi-square approximation tends to be poor for sparse tables containing both small and moderately large expected cell frequencies [1], conditioning inference about the association between two nominal categorical variables through Cramér's V association measure.
This work proposes to compare the performance of bootstrap confidence intervals for V - the statistic of interest - when a non-parametric approach is used. Besides that, since V suffers from some important limitations [3], bootstrap confidence intervals for a distance-based measure W, proposed in [3] as alternative association measure to V, will be also studied.
Bootstrap is a re-sample technique. In case of non-parametric bootstrap, no assumptions are made concerning the model distribution of data [2]. Let n be the number of observations of a sample Y. The general algorithm for a non-parametric bootstrap is the following [2]: 1. sample n observations randomly with replacement from Y to obtain a bootstrap sample Y^*; 2. calculate the bootstrap version of the statistic of interest (in our case: V and W); 3. repeat steps 1 and 2 a large number of times.
Confidence intervals for the association measures V and W were calculated using bootstrap samples, through both pivotal methods - basic and bootstrap-t - and non-pivotal methods - percentile and bias corrected accelerated (BCa).

In order to study the performance of these four bootstrap methods, a simulation study was carried out. Contingency tables were generated considering different marginal distributions. Coverage probabilities provided by these four bootstrap confidence intervals methods, for several values of V and W, were calculated and compared.
Then, this methodology was applied to a small data set, collected from an online survey, applied to Mathematics students from a public university in Portugal, regarding their perception on the learning process in an on-line environment.

Acknowledgements This work was supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. João Rocha was supported by a FCT research grant with reference BII-MATEAS-4-2021 in the scope of the Thematic Line MATEAS of CIDMA.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley Interscience, New York, 2nd edition, 2002.

[2] J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, 19(9):1141–1164, 2000.

[3] O. T. Kvålseth. An alternative to Cramér's coefficient of association. *Communications in Statistics - Theory and Methods*, 47(23):5662–5674, 2018.

20 April, 17:00 - 17:20, Auditorium Francisco Sampaio

The link between internal social responsibility, work culture and innovative behavior: a statistical approach

Mara Cunha[1], **Helena Sofia Rodrigues**[2], **Ana Teresa Oliveira**[3]
[1] Instituto Politécnico de Viana do Castelo, maracunha@ipvc.pt
[2] Instituto Politécnico de Viana do Castelo and CIDMA - Center for Research & Development in Mathematics and Applications, sofiarodrigues@esce.ipvc.pt
[3] Instituto Politécnico de Viana do Castelo and CISAS - Center for Research and Development in Agrifood Systems and Sustainability, ateresaoliveira@estg.ipvc.pt

Companies have become increasingly aware of social responsibility due to advances in globalization and the increase in business competitiveness. By rethinking their behaviors and conducts, give rise to different market positioning, translating into specific investments capable of conditioning labor relations. The purpose of this work is to study the relationship between policies and practices of internal social responsibility in the logistics department of Jerónimo Martins Group and their association with the work culture and innovative behavior. The results show that there is a strong correlation between internal social responsibility, highlighting corporate policies and organizational justice and the work culture.

Keywords: internal social responsibility, work culture, innovative behavior, correlation

Social responsibility and innovative behavior are interrelated concepts in the corporate world. Companies that prioritize social responsibility tend to exhibit more innovative behavior because they are motivated to make a positive impact on the world. Social responsibility initiatives often require companies to think creatively and outside the box to address social and environmental issues, which can lead to the development of new and innovative products, services, and processes. Innovative companies are also more likely to be socially responsible because they understand the importance of sustainability and the role they play in creating a better future for all stakeholders. This way, otfen results in cost savings and improvement of company's reputation, leading to increased consumer loyalty and brand recognition.
A survey was prepared, based on Turker work [2], related to the measurement of Corporate Social Responsability, and applied to the logistics department of Jerónimo Martins group. The survey was developed with the aim of perceiving how Internal Social Responsibility (ISR) interferes, positively or negatively, in the work culture (WC) and Innovative Behavior (IB). Additionally, this work intended to understand the influence of ISR on business

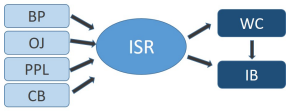

Figure 1: Conceptual model of the research

policies (BP), organizational justice (OJ), training and reconciliation of professional and personal life (PPL) and complementary benefits (CB), as the conpetual model describes (Figure 1).
For each dimension of the study, a set of items was selected, inquiring the level of agreement (on a 5-point Likert scale, where 1 represents strongly disagree and 5 strongly agree) of a set of setences. The detailed survey is provided in [1]. The Cronbach's alpha values were calculated, and they are all above of 0.888. Synthetic indices were produced for each dimension, using the average values of the respondent's responses. Both, inter-item correlations and inter-item composite-scale correlations, are higher than 0.3 and 0.5, respectively.
In Table 1, only the synthetic indices of each dimension are presented. It is possible to observe that the level of agreement with politics of social responsability of the company is positive, with values above the neutral point 3.

Table 1: Descriptive statistics for synthetic indices

Dimension	BP	OJ	PPL	CB	WC	IB
Mean	3.677	3.212	3.660	3.759	3.311	3.606
ST. Dev.	0.662	0.934	0.638	0.608	0.849	0.585

Spearman's correlation was calculated for each dimension of internal social responsibility, relative to work culture and innovative behavior (Table 2). The values obtained show that there are significative correlations between the internal social responsability dimensions and the work culture and innovative behavior.

Table 2: Spearman's correlation (** $p - value < 0.01$)

Dimension	BP	OJ	PPL	CB
WC	0.725**	0.813**	0.556**	0.412**
IB	0.480**	0.464**	0.421**	0.395**

Acknowledgements This work is supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020 (Rodrigues). This research was part of Logistics Master thesis of Mara Cunha.

References

[1] M. Cunha. *Políticas e Práticas de Responsabilidade Social Interna e a sua Associação com o Clima do Trabalho e com o Comportamento Inovador das Equipas de Logística: estudo de caso do Grupo Jerónimo Martins.* Tese de mestrado.

20 April, 16:00 - 16:20, Room A1.2

Improving short-term forecasts of environmental time series via state-space modeling

F. Catarina Pereira[1], **A. Manuela Gonçalves**[2], **Marco Costa**[3]
[1] Centre of Mathematics, University of Minho, Portugal, id9976@alunos.uminho.pt
[2] Department of Mathematics and Centre of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[3] Águeda School of Technology and Management and Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt

---

Within the scope of the TO CHAIR project, a state-space modeling approach is proposed to improve accuracy of forecasts obtained from the weatherstack.com website with a dataset of real environmental observations. The proposed methodology establishes a stochastic linear relationship between the observed environmental data and the h-step-ahead forecast brained from the website. This relation is modeled on a state space framework associated to the Kalman filter predictors. The proposed model linearly and stochastically relates the forecasts from the website (as a covariate) to the observations recorded at the study site.

Keywords: short-term forecasting, state-space models, Kalman filter, data assimilation, environmental time series

---

State space models have in their structure a latent process, the state, which is not observed, Harvey [2]. The Kalman filter is typically used to estimate it, as it is a recursive algorithm that, at each time, computes the optimal estimator, in the sense that it has the minimum mean squared error of the state when the model is fully specified, and the one-step-ahead predictions by updating and improving the predictions of the state vector in real time when new observations are available. This algorithm is applied in various areas of study (Gonçalves [1]).
This work aims to both establish a statistical modeling approach based on environmental variables and in state-space modeling and improve the forecasts obtained from an easily accessible online website, for different time horizons from 1 to 6 days.
This approach is developed for a case study — the maximum air temperature — in order to improve the accuracy of the forecasts obtained from the website https://weatherstack.com/. Improving the website's forecasts by combining accurate data from a portable station to minimize the forecasts' quality allows obtaining more accurate data that will serve as inputs to other mathematical models. In particular, corrected forecasts will be considered in the optimization models to better manage water availability for irrigation within the paradigm of sustainability, as advocated in the TO CHAIR project.

This approach aims to establish a model that can predict and calibrate or correct in real time the predictions obtained from the website. It is intended that a dynamic model (Petris [4]) calibrates the forecasts that are assumed to have a forecast error increased by the interpolation error of the website by incorporating the values observed at the study site via a portable weather station.
The framework proposed in this work can be seen as a method of Data Assimilation (DA), since the proposed approach combines forecasts across time and from different sources. DA is in general a sequential time-stepping procedure, in which forecasts from a source are compared with newly received observations to produce optimal forecasts.
The specification of the state-space model is performed using the maximum likelihood method under the assumption of normality of errors, where empirical confidence intervals are presented. In addition, this work also presents a treatment of outliers based on the ratios between the observed maximum temperature and the website's forecasts.
Overall, the proposed model significantly reduced the RMSE, MAE, and the MAPE, both in-sample and out-of-sample, compared to the website's initial forecasts, where it was observed that these forecasts underestimated, on average, the observed maximum temperature by about 13–14 %. Also, improved forecasts will lead to improved use of water resources, namely by planning irrigation more efficiently (Pereira [3]).

Acknowledgements This work has received funding from FEDER/COMPETE/NORTE 2020/POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/IEEA/00147/ 0069 33-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. F. Catarina Pereira was financed by national funds through FCT through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM. Marco Costa was partially supported by the Portuguese Foundation for Science and Technology (FCT) within the Projects UIDB/04106/2020 and UIDP/04106/2020 of CIDMA-UA.

References

[1] A.M. Gonçalves and M. Costa. Statistics on the computer. The bootstrap. *Stochastic Environmental Research and Risk Assessment*, 27:1021–1038, 2013.

[2] A.C. Harvey. *Forecasting, structural time series models and Kalman filter.* Cambridge University Press, New York, 2009.

[3] A.M. Pereira, F.C. Gonçalves and M. Costa. Short-term forecast improvement of maximum temperature by state-space model approach: the study case of the to chair project. *Stochastic Environmental Research and Risk Assessment*, 37:219–231, 2023.

[4] S. Petris, G. Petrone and P. Campagnoli. *Dynamic linear models with R.useR!* Springer, New York, 2009.

20 April, 16:20 - 16:40, Room A1.2

Stationary and non-stationary state-space models in the presence of outliers: a simulation study

F. Catarina Pereira[1], **A. Manuela Gonçalves**[2], **Marco Costa**[3]
[1] Centre of Mathematics, University of Minho, Portugal, id9976@alunos.uminho.pt
[2] Department of Mathematics and Centre of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[3] Águeda School of Technology and Management and Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt

---

There are several challenges when it comes to time series forecasting. Modeling non-stationary data can be a difficult task. In this work we propose a model with a state-space representation that can handle non-stationary data. The presence of outliers can impair parameter estimation and forecasting. Thus, this paper presents a simulation study to compare several methods of detecting and treating outliers in stationary and non-stationary time series with clean and contaminated data.

Keywords: state-space models, outliers, non-stationary time series, imputation, Kalman filter

---

Forecasting time series can be a challenging task. The presence of structural changes in the data can hinder model specification and thus impact the forecasts and inferential results [1].

Non-stationary time series are common in many areas of study, and therefore the choice of a forecasting model is limited, since most of the time series forecasting methods are not suitable for non-stationary series, and a data transformation is then required. One of the models that can deal with non-stationary time series is the state-space model, which, associated with the Kalman filter, is a powerful tool for forecasting, and is effective from a stochastic point of view.

Another common problem in time series analysis is the presence of outliers [3]. The presence of outliers can lead to misspecified models, large residuals, biased parameter estimation, and inaccurate forecasts. Therefore, detection and treatment of outliers play a key role in time series analysis and modeling. However, this task may not be so trivial, since an outlier may not necessarily be an extreme point and may only be a marginal point.

The objective of this paper is to compare, through a simulation study, several methods of detention and treatment of outliers in stationary and non-stationary time series with clean and contaminated data.

In this work, we considered the univariate stationary state-space model given by

$$Y_t = \beta_t + e_t, \tag{1}$$
$$\beta_t = \phi\beta_{t-1} + \varepsilon_t \tag{2}$$

with $t = 1, \ldots, n$, where Y_t is the observed variable, e_t is the observation error that is independent and identically distributed with Gaussian distribution of zero mean and variance σ_e^2, i.e. $e_t \sim N(0, \sigma_e^2)$ and $E(e_t, e_s') = 0$, for all, $t \neq s$; ε_t is the state error that is independent and identically distributed with Gaussian distribution of zero mean and variance σ_ε^2, i.e. $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and $E(\varepsilon_t, \varepsilon_s') = 0$, for all, $t \neq s$, and e_t and ε_t are uncorrelated. For the stationary scenario, we simulate time series generated by the model defined by equations (1)-(2), with a range of values for autoregressive parameter $\phi = (0.25, 0.50, 0.90)$, where 1,000 replicates with valid estimates were considered, i.e., estimates within the parameter space: $-1 < \phi < 1$, $\sigma_\varepsilon > 0$, and $\sigma_e > 0$. For the non-stationary scenario, we simulate time series generated by the local level model, which is a particular case of the model defined by equations (1)-(2), where $\phi = 1$. For each parameter combination, 1,000 replicates with valid estimates were considered, i.e., estimates within the parameter space: $-1 < \phi < 1$, $\sigma_\varepsilon > 0$, and $\sigma_e > 0$. For both stationary and non-stationary cases, we simulate time series of sample sizes of $n = (50, 200, 500)$ with a range of values for σ_ε^2 and σ_e^2 $(0.05, 0.10, 1.00)$. To compare the methods of detection and treatment of outliers, in each case for stationary and non-stationary time series, we also consider both clean and contaminated data, i.e., $e_t \sim N(0, \sigma_e^2)$; $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and $e_t \sim (1-p)N(0, \sigma_e^2) + pN(10\sigma_e, \sigma_e^2)$; $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, where $p = 0.05$ [2].

Acknowledgements F. Catarina Pereira was financed by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. Marco Costa was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

References

[1] F. Z. C. Rose, M. T. Ismail, and M. Tumin. Outliers detection in state-space model using indicator saturation approach. *Indonesian Journal of Electrical Engineering and Computer Science*, 22:1688–1696, 2021.

[2] R. Crevits and C. Croux. Robust estimation of linear state space models. *Communications in Statistics - Simulation and Computation*, 48(6):1694–1705, 2019.

[3] D. You, M. Hunter, M. Chen, and S. Chow. A diagnostic procedure for detecting outliers in linear state–space models. *Multivariate Behavioral Research*, 55:1–25, 2019.

20 April, 16:40 - 17:00, Room A1.2

Modeling the fuel consumption of a NRP ship using a Kalman filter approach

M. Filomena Teodoro[1,2], **Pedro Carvalho**[1], **Ana Trindade**[1]
[1] CINAV, Center of Naval Research, Naval Academy, Military Universitary Institute, Portuguese Navy, Portugal, mteodoro64@gmail.com
[2] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, 1048-001 Lisboa, Portugal, maria.teodoro@tecnico.ulisboa.pt

The Kalman filter can be applied in the most diverse areas of knowledge, e.g. medicine, agriculture, social sciences, computing, etc. The Kalman filter is a recursive tool that can be used under the aim of Navigation and Integration Systems. We make a brief approach to the derivation of a Kalman filter splitting the work into two parts. By first, a Kalman filter is used to simulate different situations analyzing the "response" of the filter considering distinct cases for distinct states of the shop motor; at second, a specific Kalman filter is built to filter the fuel consumption data collected directly from the on-board records of a ship from Portuguese Republic.

Keywords: Kalman filter, fuel consumption modeling, computational algorithms

In 1960, the engineer Rudolf Kalman published the article [3] in which he presented a new method of linear filtration. This method uses measurements of independent variables and the associated noise to filter the system signal and predict its next state through the use of statistical techniques. This new method introduced by Kalman in early sixty decade came to be known as Kalman Filter (KF) and had its first use aboard the spacecraft navigation computers of the APOLLO project. The KF is one of the most applied methods for tracking and estimation due to its simplicity, optimality, tractability and robustness [2].
Accordingly with the authors of [1], the KF can be seen as a sequential minimum mean square error (MMSE) estimator of a signal with noise. This signal is described by a state (or dynamical) model. When the errors are Gaussian distributed, the KF conduces to an optimal MMSE estimator; if the errors are not Gaussian distributed, the estimator still is a linear MMSE estimator. In [4] is illustrated in a simple way the basic concepts of FK.
This work presents a brief approach to the derivation of a KF when the input is a scalar quantity. It was considered a KF to estimate a first order system. Several simulations were carried out with these models and the results of applying the filter in different situations were analyzed. It ended with the analysis of a KF model, built specifically to filter NRP Douro consumption data. The approach that was made in this work considered only the discrete KF, once, in practice, observations and controls were carried out in discrete case.

The KF was used in the form of a linear system to obtain the best estimate of the state vector conditional to past observations. The estimate was calculated using the reconstruction of the state vector using the previous state vector estimate, the known inputs and measured outputs. The observed consumption value was registered hourly in NRP Douro, with AV3 machine regime, using an one-dimensional KF considering white random noise in the measure equation. In figure (1) we can find the observed and estimated consumption. It was possible to verify that the applied KF was effective and conduced to a good measures of estimation and prevision.

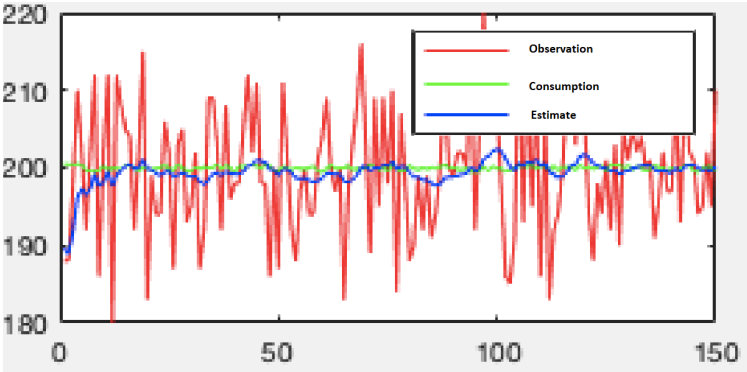

Figure 1: Consumption (Litre/hour) versus Time (hour). AV3 machine regime. Consumption estimate (blue), consumption observation (red), real consumption (green)). Measure error variance $Q = 0.5$.

Acknowledgements This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV),Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project `UID/Multi/04621/2019`.

References

[1] R. G. Brown and Y. C. Hwang. *Introduction to random signals and applied Kalman filtering: with MATLAB exercises. 4^{th} edition.* John Wiley & Sons, New Jersey, 2012.

[2] S. J. Julier and J. K. Uhlmann. New extension of the kalman filter to nonlinear systems. In Ivan Kadar, editor, *Processing, Sensor Fusion, and Target Signal Recognition, VISPIE Proceedings 3068*, 1997.

[3] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[4] G. Welch and G. Bishop. An introduction to kalman filter, 2006. Technical Report TR 95-041.

20 April, 17:00 - 17:20, Room A1.2

Bayesian approach to modelling time series of counts under censoring

Isabel Silva[1], **Maria Eduarda Silva**[2], **Isabel Pereira**[3], **Brendan McCabe**[4]
[1] Faculdade de Engenharia, Universidade do Porto and CIDMA, Portugal, ims@fe.up.pt
[2] Faculdade de Economia, Universidade do Porto and LIADD-INESC TEC, Portugal, mesilva@fep.up.pt
[3] Departamento de Matemática, Universidade de Aveiro and CIDMA, Portugal, isabel.pereira@ua.pt
[4] Management School, University of Liverpool, UK, Brendan.Mccabe@liverpool.ac.uk

---

In this work the problem of modelling time series of counts under censoring is approached from a Bayesian perspective based on the Gibbs sampler with data augmentation (GDA) and multiple sampling. The methodology is applied to synthetic and real data and the results indicate that the the estimates are consistent and present less bias than those obtained when the censoring is neglected.

Keywords: Bayesian estimation, censored count series, Gibbs sampler with data augmentation, Poisson INAR(1) model

---

Time series data may present irregularities such as missing values and detection limits. Common examples can be found in survey questionnaires when the highest category is x *or more* counts or when measuring devices are not able to detect above and/or below a certain detection point or threshold. This kind of data is said to be (Type I) censored data. Censoring is a characteristic of the data gathering procedure, frequently found in several fields including environmental science, epidemiology, business and social sciences. Disregarding censoring lead to inference problems such as biased parameter estimation and poor forecasts.

The case of independent censored data and of Gaussian ARMA models under censoring have been studied in the literature. In this work, the analysis of time series of counts under censoring through Poisson first-order integer-valued autoregressive (PoINAR) models [2] is considered in the Bayesian framework. Hence, a modified Gibbs sampler with Data Augmentation (GDA) [1] in which the data augmentation is carried out by multiple sampling of the latent variables [3] from the truncated conditional distributions (GDA-MMS) is adopted.

Figure 1 shows a synthetic dataset with $n = 350$ observations generated from an PoINAR(1) process with parameters $\alpha = 0.5$ and $\lambda = 5$ (X_t, blue line), the respective right censored dataset (Y_t, red line), at $N = 11$, corresponding to 30% of censoring, and an augmented dataset (Z_t, black line). If we disregard the censoring, the estimates for the parameters

(assuming an PoINAR(1) model without censoring) present a strong bias. For instance, the Gibbs sampler gives $\hat{\alpha}_{Bayes} = 0.6242$ and $\hat{\lambda}_{Bayes} = 3.3297$. On the other hand, if we assume an PoINAR(1) model under censoring, the parameter estimates given by the proposed approach are $\hat{\alpha}_{GDA-MMS} = 0.4834$ and $\hat{\lambda}_{GDA-MMS} = 4.9073$. Therefore, it is important to consider the censoring in data in order to avoid some inference issues that lead to poor time series analysis.

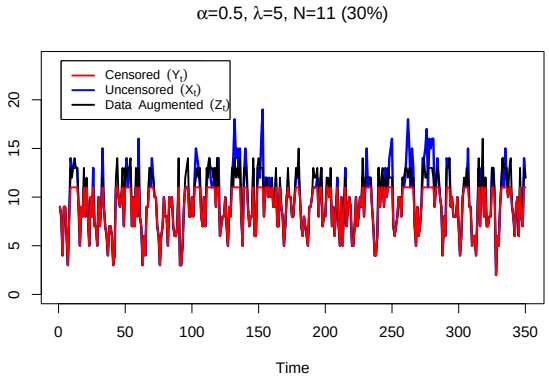

Figure 1: Synthetic dataset generated from an PoINAR(1) process, the respective right censored dataset and an example of data augmentation.

Experiments with synthetic data allow to conclude that the approach leads to estimates that present less bias than those obtained if censoring is neglected. Moreover, the GDA-MMS approach allows to obtain a complete data set, making it a valuable method in other situations such as missing data.

Acknowledgements The first and third authors were partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. The second author was partially supported by Portuguese National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UIDB/50014/2020.

References

[1] S. Chib. Bayes inference in the tobit censored regression model. *Journal of Econometrics*, 51:79–99, 1992.

[2] M. G. Scotto, C. H. Weiß, and S. Gouveia. Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, 15:590–618, 2015.

[3] R. Sousa, I. Pereira, M.E. Silva, and B. McCabe. Censored Regression with Serially Correlated Errors: a Bayesian approach. `https://arxiv.org/abs/2301.01852`, 2023.

21 April, 9:00 - 9:20, Auditorium Francisco Sampaio

Inventories discretionary management through accounting choices - The case of small and medium-sized Portuguese companies in commercial sector

M. Filipa Nogueira[1], **Augusta Ferreira**[2], **Carlos Ferreira**[3]
[1] ESGT, Politechnic Institute of Santarém, University of Aveiro, filipa.nogueira@esg.ip.santarem.pt
[2] CICF/ISCA-UA, University of Aveiro, augusta.ferreira@ua.pt
[3] IEETA/DEGEIT, University of Aveiro, carlosf@ua.pt

Discretionary management of inventories by managers causes deliberate changes in the financial statements that support decision making, thus deliberately causing asymmetry of information among stakeholders. This study, using 30,797 Portuguese SMEs from the commercial sector and not listed, comprising the period of 2013 to 2019, suggests the evidence of the discretionary management of real activity on inventories and accounting choices by managers, in order to present financial statements that meet their own interests. Future work includes the replication of this work to compare results in various activities of the commercial sector and the extension to the trading sector from other countries to analyze the similarity of results across different business realities.

Keywords: inventory, accrual, discretionary, estimation models

Managers have the ability to exercise opportunistic choices of accounting policies and real activities that enable them to achieve both accounting and managerial outcomes that match their objectives, pursuing various incentives for discretion according to Positive Accounting Theory and Agency Theory. There are several models to measure the practice of those types of discretionary actions in which managers may engage in acts that allow them to alter the value of assets, liabilities, income and expenses to achieve the desired asset value and results. On the other hand, most of the scientific literature on this subject focuses on listed companies [1] [2] [3] [4].
Considering that the Portuguese business fabric is composed of 99.9 % Small and Medium-sized Enterprises (SME), this work addresses empirically and with regression models, the theme in more than 30,000 Portuguese SME from the commercial sector and not listed (SABI database, version 91.00), comprising the period from 2013 to 2019, intending to answer the question "what is the effect of discretionary inventory management and accounting choices, through real activities, in the accounting information of Portuguese SMEs in the commercial sector?".

The results of the estimation models allow us to conclude that: (i) the managers of these SMEs tend to include discretion in the management of real activities and that the volume of inventories is one of the factors that weights; (ii) the discretion included in inventory management is related to commercial management. Managers can change the value of companies' financial reports to present statements that meet the requirements of stakeholders and their own interests; (iii) the companies included in the study use the increase or decrease in the volume of inventory to change the cost of goods sold and consequently the value of the earnings and the value of the company; (iv) the discretion included in inventory management is related to accruals, with managers choosing accounting policies that allow them to achieve their own objectives. This discretion causes deliberate changes in the economic and financial reality of the companies, causing asymmetry of information between the stakeholders and biasing supporting elements for decision making.
Concerning future work, to mitigate some limitations, we intend to replicate this study comparing the results in various activities of the commercial sector such as retailers, wholesalers of various activities such as: food products, fuel, technological and other household equipment because the motivations and forms of discretion may have different configurations. Also, the comparison of the behavior of Portuguese firms with SMEs in the trading sector from other countries would allow us to verify whether the Portuguese reality and the degree of discretion are similar across different business realities.

References

[1] S.G. Anton. The effect of discretionary accruals on firm growth. empirical evidence for smes from emerging europe. *Journal of Business Economics and Management*, 21(4):1128–1148, 2020.

[2] D. Cohen, S. Pandit, C.E. Wasley, and T. Zach. Measuring real activity management. *Contemporary Accounting Research*, 37(2):1172–1198, 2020.

[3] P.M. Dechow, A.P. Hutton, J.H. Kim, and R.G. Sloan. Detecting earnings management: A new approach. *Journal of Accounting Research*, 50(2):275–334, 2012.

[4] S. Roychowdhury. Earnings management through real activities manipulation. *Journal of Accounting and Economics*, 42(3):335–370, 2006.

21 April, 9:20 - 9:40, Auditorium Francisco Sampaio

Identifying characteristics of marketing-influenced eating vulnerability

Carla Henriques[1], **Raquel Guiné**[2], **Ana Matos**[3], **Madalena Malva**[4]
[1] Polytechnic Institute of Viseu and CMUC, Portugal, carlahenriq@estgv.ipv.pt
[2] Polytechnic Institute of Viseu and CERNAS, Portugal, raquelguine@esav.ipv.pt
[3] Polytechnic Institute of Viseu and CISeD, Portugal, amatos@estgv.ipv.pt
[4] Polytechnic Institute of Viseu, Portugal, malva@estgv.ipv.pt.pt

Marketing-driven choices in people's dietary is assessed in a cross country survey involving 16 countries. The sample comprises 11919 responses to a questionnaire developed for the EATMOT project as described in [3] . In this study, cluster analysis based on people's marketing motivations revealed two well differentiated groups: low and notably motivated consumers. These two groups were compared, outlining some characteristics of consumers who are more prone to commercial and marketing motivations.

Keywords: cluster analysis, logistic regression, marketing motivations, food choices

Numerous people's dietary decisions are influenced by commercial and marketing motivations. Advertising and marketing tactics are in fact intended to pique consumer attention and influence their purchasing decisions. Yong people are known to be particularly vulnerable [1], and it is of interest to identify some other characteristics that may differentiate most vulnerable consumers. In this study, based on a sample of size 11919, collected as part of the project EATMOT [3], marketing motivations in food choices are analyzed through seven items. Factor analysis was applied by country, looking for a factor structure common to all countries. Three items were consistently combined in one factor. The other items were studied individually. Five variables were, then, considered to measure marketing motivations in consumer's food choices and used in a cluster analysis. Ward's method, single linkage, and average linkage were three hierarchical techniques that were used; their results were considered as initial solutions for the k-means method. To find an optimal number of clusters, k-means method was applied to 50 bootstrap samples and the similarity of cluster solutions for different numbers of clusters was examined using the rand index [2]. The two cluster solution emerged as an optimal solution, distinguishing consumers more prone to the influence of marketing (Figure 1).
Then, using statistical tests and logistic regression analysis, these two clusters were compared. For example, consumers of the Notably Motivated cluster were significantly younger (32.7 vs 36.7 years old, $p < 0.005$), as expected, and this cluster also had a higher percentage of women (72.2% vs. 70.2%, p=0.016), a higher percentage of single individuals (51.9% vs. 37.9%, $p < 0.005$), a lower proportion of individuals with university education

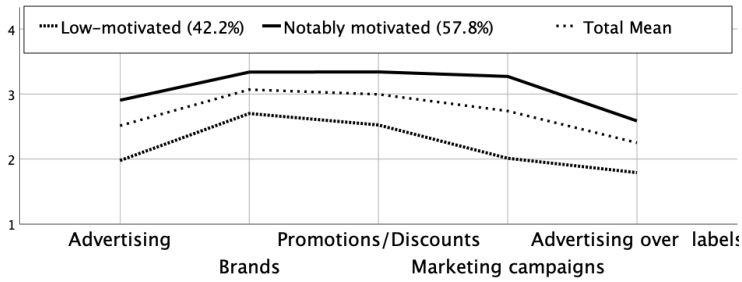

Figure 1: Clusters means in the five marketing motivations

(58.3% vs. 66.1%, $p < 0.005$), more consumers living in rural or suburban areas (37.2% vs. 27.6%, $p < 0.005$), and more consumers without an active professional activity, that is, unemployed, non-working students or retired (43.1% vs.31.1%, $p < 0.005$). Furthermore, higher BMI and less physical exercise revealed to be associated with a greater chance of belonging to the notably motivated group ($p < 0.005$).
We thus obtained evidence that the propensity for higher levels of commercial and marketing motivations is associated with socio-demographic, anthropometric, behavioural and health related characteristics of the consumer.

[3] **Acknowledgements** The authors would like to thank the CMUC for financial support. Additionally, the authors thank project EATMOT, the FCT, CERNAS and the Polytechnic Institute of Viseu.

References

[1] S.L. Calvert. Children as consumers: advertising and marketing. future children. *Future Children*, 18(1):205—234, 2020.

[2] S. Dolnicar and F. Leisch. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, 21(1):83–101, 2010.

[3] A. C. Ferrão, R. P. Guiné, P. Correia, M. Ferreira, J. Duarte, and J. Lima. Development of a questionnaire to assess people's food choices determinants. *Current Nutrition Food Science*, 15(3):281–295, 2019.

21 April, 9:40 - 10:00, Auditorium Francisco Sampaio

Are the European countries well prepared for the new technological challenges?

Fernanda Figueiredo[1], **Adelaide Figueiredo**[2]
[1] Faculdade de Economia da Universidade do Porto and CEAUL, Universidade de Lisboa, otilia@fep.up.pt
[2] Faculdade de Economia da Universidade do Porto and LIAAD- INESC TEC Porto, adelaide@fep.up.pt

---

A Double Principal Component Analysis is performed in this study to investigate in which European countries citizens are more prepared in using digital technologies, as well as the countries that up to now have been investing more financial resources in providing training on Information and Communications Technology. Several indicators associated with digital use and training were considered in this study, and the data for these variables during the period 2010-2020 were collected from the Eurostat database.

Keywords: digital, DPCA, ICT education, multivariate analysis

---

Information and Communication Technology (ICT) Education, in the general sense, means education to provide users with a diverse set of technological tools, definitions, and resources to create, store, communicate, manage and optimize the information. Nowadays these technical skills are very important in all professional activities, in particular, in the area of services, industry or education. Accessibility to the Internet and to other digital resources, and the acquisition of knowledge to use these tools, is also crucial in our lives. For instance, the use of Internet allows getting information and learning, facilitates communicating and socializing, and is also useful to buy or sell things. The motivation for this study lies in analyzing how these skills are more or less developed in the different European countries, understanding their evolution over the last decade and identifying the new needs in this emerging field.

To perform this multivariate analysis we applied the Double Principal Component Analysis (DPCA). This methodology was introduced by Bouroche [1] and it is an extension of the Principal Component Analysis (PCA). This method enables us to study several data tables with the same individuals and the same variables, obtained in different time instants or different circumstances. The main objectives of the Double Principal Component Analysis are to globally compare the different data tables, to study the evolution of the relations between the different variables, and study the evolution of the individuals.

DPCA was introduced to analyze quantitative data ([1]), but there are several adaptations of the DPCA to allow the analysis of categorical data (see, for instance, [2]).

Acknowledgements This work has received funding from National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia under the projects UIDB/00006/2020 (CEAUL) and LA/P/0063/2020 (INESC TEC).

References

[1] J. Bouroche. *Analyse des données ternaires: la double analyse en composantes principales.* PhD thesis, Université de Paris, 1975.

[2] Pérez R. Lera, L. and A. Boquet. El doble análisis en componentes principales para datos categóricos y su aplicación en un estudio de migración. *Revista Colombiana de Estadística*, 29(1):17–34, 2006.

21 April, 10:00 - 10:20, Auditorium Francisco Sampaio

A risk model for classifying stocks

Irene Brito

Centro de Matemática, Departamento de Matemática, Universidade do Minho
ireneb@math.uminho.pt

Classifying stock risks is an important task for investment decision-making problems, e.g. in the construction of portfolios. In this work, a risk model for the classification of stocks is proposed, that assesses risk using expected utility, entropy and variance. The model is applied to the PSI 20 index to form subsets of stocks with lower risk. Using the mean-variance model, the efficiencies of the subsets' portfolios are compared with the efficiency of the whole stock set. The results reveal that the risk model selects the relevant stocks for an optimal portfolio construction.

Keywords: risk model, risk classification, expected-utility, entropy, variance

---

The mean-variance model was proposed by Markowitz [3] to assess and construct portfolios (weighted combinations of stocks) by minimizing risk, expressed by variance, and maximizing the expected return. Several other models were presented, where entropy was used for measuring risk and that can also be combined with other measures (see e.g. [2], [4]). Recently, the expected utility, entropy and variance model (EU-EV model) was applied to the selection of stocks for portfolio constructions [1]. In that model, entropy and variance, as uncertainty risk factors, are combined with expected utility, as preference factor, using a trade-off parameter. The EU-EV model for classifying stock risks is defined as follows. Consider a set of stocks $S = \{S_1, \ldots, S_I\}$ and the action space $A = \{a_1, \ldots, a_I\}$, where $a_i = (x_{i1}, p_{i1}; x_{i2}, p_{i2}; \ldots; x_{iN}, p_{iN}) \in A$ is the action of selecting stock S_i, $i = 1, \ldots, I$, yielding the frequency distribution of stock returns, where x_{in} are the outcomes occurring with probabilities p_{in}, $n = 1, \ldots, N$, that are represented by the discrete random variable X_i. The EU-EV risk for the action a_i is defined by

$$R(a_i) = \frac{\lambda}{2}\left[H(X_i) + \frac{\mathrm{Var}[X_i]}{\max\limits_{a_i \in A}\{\mathrm{Var}[X_i]\}}\right] - (1-\lambda)\frac{\mathbb{E}[u(X_i)]}{\max\limits_{a_i \in A}\{|\mathbb{E}[u(X_i)]|\}},$$

where $0 \leq \lambda \leq 1$, $u(\cdot)$ is the utility function and $H(X_i) = -\sum_{n=1}^{N} p_{in} \ln p_{in}$ is the entropy. The stocks are ranked according to the EU-EV risk, where given two stocks S_{i_1} and S_{i_2}, $i_1, i_2 \in \{1, \ldots, I\}$, if $R(a_{i_1}) < R(a_{i_2})$, then the optimal stock is S_{i_1}.
The EU-EV risk model was applied in [1] to the Portuguese Stock Index PSI 20, considering the returns obtained from daily closing prices of 18 component stocks, $S = \{S_1, \ldots, S_{18}\}$,

from January 2019 to December 2020. The stocks were classified according to the EU-EV risk, using the utility function $u(x) = \ln(x+1)$, if $x \geq 0$, and $u(x) = -\ln(1-x)$, if $x < 0$, and the best 9 stocks with lower risk were selected for different ranges of λ to construct portfolios. Five stock subsets, $Q_1, \ldots, Q_5$, were obtained with λ belonging to the intervals $[0, 0.1260)$, $[0.1260, 0.4685)$, $[0.4685, 0.5311)$, $[0.5311, 0.7771)$, $[0.7771, 1]$, respectively, where at 0.1260, 0.4685, 0.5331, 0.7771 occur changes in the subset composition. The mean-variance optimization problem was then applied to the whole set of stocks S and to the subsets of 9 selected stocks $Q_1, \ldots, Q_5$. Comparing the efficient frontiers of S with those of the five subsets (see Figure 1) one can observe that the performance of the sets $Q_1, \ldots, Q_4$ corresponding to $\lambda \in [0, 0.7771)$ is similar to those of S obtained with all 18 stocks. As for Q_5, with λ close to 1 and therefore privileging stocks with lower uncertainty and almost ignoring expected utility, it performs less well than S. One can conclude that the EU-EV model, for certain ranges of the trade-off factor, can be used to classify stock risks in order to construct efficient portfolios with a reduced number of stocks.

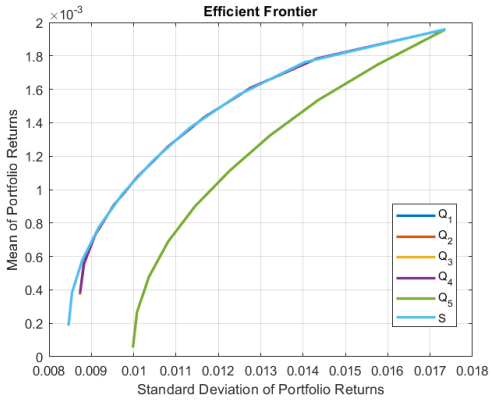

Figure 1: Efficient frontiers for $Q_1, \ldots, Q_5$ and for S

Acknowledgements The author thanks support from FCT through the Projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Irene Brito. A portfolio stock selection model based on expected utility, entropy and variance. *Expert Systems with Applications*, 213:118896, 2023.

[2] B. Li and R. Zhang. A new mean-variance-entropy model for uncertain portfolio optimization with liquidity and diversification. *Chaos, Solitons & Fractals*, 146:110842, 2021.

[3] H. Markowitz. *Mean-Variance Analysis in Portfolio Choice and Capital Markets.* Wiley, 2000.

[4] J. Yang, Y. Feng, and W. Qiu. Stock selection for portfolios using expected utility-entropy decision model. *Entropy*, 19:508, 2017.

21 April, 9:00 - 9:20, Room A1.2

A comparison of some methods for clustering of variables of mixed types

Ndèye Niang[1], **Mory Ouattara**[2], **Gilbert Saporta**[3]
[1] Cédric-CNAM, Paris, France, ndeye.niang_keita@cnam.fr
[2] Université de San Pedro, Côte d'Ivoire, ouattara.mory@usp.edu.ci
[3] Cédric-CNAM, Paris, France, gilbert.saporta@cnam.fr

We compare several old and recent methods for clustering a set of qualitative and quantitative variables.

Keywords: clustering of variables, mixed data, PCA, RV coefficient

The simultaneous treatment of a mixture of J quantitative variables $\mathbf{x}_j$ and Q qualitative variables $\tilde{\mathbf{x}}_q$ with m_q categories, whether in factorial analysis or clustering, is often based on the determination of one or more global or local (i.e. per class) synthetic variables optimizing the following criterion introduced in 1977 by Tenenhaus [6], reused by Escofier (1979), then Saporta [5], Kiers (1991) under the name of PCAMIX, and Pagès (2004):

$$\max_{\mathbf{c}} \left(\sum_{j=1}^{J} r^2 \left(\mathbf{c}, \mathbf{x}_j\right) + \sum_{q=1}^{Q} \eta^2 \left(\mathbf{c}, \tilde{\mathbf{x}}_q\right) \right) \tag{1}$$

where r^2 is the squared Pearson correlation coefficient between two quantitative variables and η^2 the squared correlation ratio between a quantitative and a qualitative variable. Both coefficients are equal to the proportion of variance of a dependent variable explained by an independent one.
The ClustOfVar algorithm [1] uses criterion (1) to perform a clustering of a set of variables of different nature around latent components in each group, extending the method of Vigneau and Qannari [7] introduced for exclusively quantitative variables.
Clustering variables around components is an interesting alternative to direct algorithms that start from the table of similarities, dissimilarities or distances between all variables, because it simultaneously optimizes the clustering and the representation of classes by a component as in a clusterwise approach.
A key issue is to use consistent and comparable similarity measures in the three cases : a pair of quantitative variables, a pair of categorical variables and a pair consisting in a quantitative variable and a categorical one. The association coefficients r^2 and η^2 are in common use, while various solutions have been proposed for the case of two categorical variables: chi-squared and its derivatives such as T^2, which is the square of the Tschuprow coefficient, or the largest eigenvalue of the Correspondence Analysis matrix derived from the cross-tabulation of two categorical variables [1].

Coefficients associated with categorical variables are not, however, comparable with each other or with r^2 because their distributions depend on their number of categories. In criterion (1) a qualitative variable plays a greater role the higher its number of categories m_q. The Escoufier RV coefficients [4] between tables generated by each quantitative variable and tables of indicators of the categories of the qualitative variables make it possible to define Euclidean similarities equal, according to the cases, to r^2,$\frac{\eta^2}{\sqrt{m_q-1}}$ or T^2 [3].

We can then perform hierarchical clustering with Ward's algorithm or k-means partitioning, either directly on the similarity matrix, or on the coordinates obtained by the Torgerson formula. This elegant but somewhat forgotten solution still suffers from a flaw: dividing by the square root of the degree of freedom does not completely correct the effect of the number of categories. For this, it may be wise to use as dissimilarity the p-value of the independence test in the spirit of the *likelihood linkage algorithm* [2]. However Euclidean properties are lost.

In addition, when the number of observations is very large, all p-values are close to zero (*paradox of large samples*) and are no longer usable. We propose to replace them by the corresponding fractiles of the standard normal distribution in the spirit of the *test values* used in the SPAD software. The larger the fractile, the greater the association between two variables. These different approaches are compared on real data sets.

References

[1] M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco. ClustOfVar: An R package for the Clustering of Variables. *Journal of Statistical Software*, 50(13):1–16, 2012.

[2] F. Costa Nicolau and H. Bacelar-Nicolau. Some trends in the classification of variables. In C. Hayashi, editor, *Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan*, pages 89–98. Springer, 1998.

[3] E.M. Qannari, E. Vigneau, and Ph. Courcoux. Une nouvelle distance entre variables. Application en classification. *Revue de Statistique Appliquée*, 46(2):21–32, 1998.

[4] P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3):257–265, 1976.

[5] G. Saporta. Simultaneous analysis of qualitative and quantitative data. In *Atti della XXXV Riunione Scientifica, Societa Italiana di Statistica, Padova, Italy*, volume 1, pages 62–72, 1990.

[6] M. Tenenhaus. Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, 25(2):39–56, 1977.

[7] E. Vigneau and E.M. Qannari. Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, 32(4):1131–1150, 2003.

21 April, 9:20 - 9:40, Room A1.2

Clustering of pediatric hospitalizations by hospital resources use

Daniel Cordeiro[1], Ana Azevedo[2,3,4], Bárbara Peleteiro[2,3,4], Lucybell Moreira[2], Elsa Guimarães[2], Raquel Cadilhe[2], Rita Gaio[1,5]

[1] Dep. Matemática, Fac. Ciências, Universidade do Porto up201506370@edu.fc.up.pt

[2] Centro de Epidemiologia Hospitalar, Centro Hospitalar Universitário de São João lucybell.moreira@chsj.min-saude.pt, elsa.guimaraes@chsj.min-saude.pt, raquel.cadilhe@chsj.min-saude.pt

[3] Instituto de Saúde Pública Universidade do Porto barbara.peleteiro@chsj.min-saude.pt

[4] Departamento de Ciências da Saúde Pública e Forenses, e Educação Médica Faculdade de Medicina da Universidade do Porto anazev@med.up.pt

[5] Centro de Matemática da Universidade do Porto argaio@fc.up.pt

---

This paper aims to cluster pediatric hospitalizations using hospital resources as input variables. As the data were of mixed-type, we considered the Gower's distance and used the Partition Around Medoids and the Ward's hierarchical clustering algorithms. We also modelled the data as a finite mixture. The agreement among the obtained groupings was evaluated by the Adjusted Rand Index. While the first two methods provided similar results, some of the mixture components did not resemble the previous cluster. Evaluation of the obtained solutions by the hospital experts favoured the 7-cluster structure identified by the Partition Around Medoids algorithm.

Keywords: clustering, partition around medoids, hierarchical clustering, finite mixture models, hospitalizations

---

The study analysed 3583 hospitalizations of children ($\leq$ 18 years old) at the Pediatric Department of Centro Hospitalar Universitário de São João (CHUSJ) between December 2021 and November 2022. There were essentially two types of variables: those related with hospital resources - Surgical Intervention (No/Yes), Intensive Care Unit (No/Yes), Admission Mode (Scheduled/Urgent), Number of Services (1, 2, ..., 8) and (total) Length of Stay (LoS) - and socio-demographic variables about the patients - Sex (Female/Male), Country of birth (Portugal/Other), District of residence (Porto/Other) and Age. The goal was to group patients according to similar resource use, thus providing meaningful and easy-to-read information to the stakeholders. To accomplish the task, clustering solutions were proposed.

As the data were of mixed-type, the Gower's distance was used to calculate a dissimilarity matrix and, based on that, the Partition Around Medoids (PAM) algorithm and Ward's Hierarchical Clustering were applied.

For PAM, the optimal number of clusters was determined by the Average Silhouette Width, suggesting 4, 7 or 8 groups.
The dendrogram associated with the Ward method also indicated 4, 7 or 8 groups. The 7-cluster structure showed the greatest agreement with the results from PAM (adjusted Rand Index for 4, 7 and 8 groups: 0.712, 0.995, 0.869, respectively). This structure consisted of the eight combinations of the three binary variables, with two of them grouped together.
For the Finite Mixture Model (FMM), different information criteria - Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Integrated Complete-data Likelihood (ICL) - provided different number of components, namely 4, 7 and 8. The 7-group structure was then compared with the previous algorithms, displaying a relevant similarity, although lower than the similarity between the first two (adjusted Rand Index FMM vs PAM: 0.455 FMM vs Ward: 0.456).
The 7-cluster PAM structure was found to be coherent by the hospital experts. Its description is presented in Figure 1. For a better understanding of the structure, its description using external patient-related variables is also given - Figure 2.
All statistical analyses were performed in R-version 4.1.1. and the package used were: *cluster*, *flexmix* and *fossil*.

PAM	Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6			Cluster 7			
	N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		p-value
Surgery	1035			737			93			157			906			583			72			<0,001
No	0	0%		0	0%		0	0%		8	5%		906	100%		583	100%		72	100%		
Yes	1035	100%		737	100%		93	100%		149	95%		0	0%		0	0%		0	0%		
Intensive Care Unit	1035			737			93			157			906			583			72			<0,001
No	1035	100%		737	100%		0	0%		0	0%		906	100%		568	97%		0	0%		
Yes	0	0%		0	0%		93	100%		157	100%		0	0%		15	3%		72	100%		
Admission Mode	1035			737			93			157			906			583			72			<0,001
Scheduled	0	0%		737	100%		0	0%		157	100%		0	0%		583	100%		11	15%		
Urgent	1035	100%		0	0%		93	100%		0	0%		906	100%		0	0%		61	85%		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	p-value
Number of Services	1035	1,4	0,5	737	1,0	0,2	93	3,5	1,4	157	3,2	0,9	906	1,2	0,4	583	1,0	0,3	72	2,3	0,9	<0,001
Length of Stay (hours)	1035	106,0	218,1	737	86,8	212,3	93	778,1	817,9	157	277,6	304,2	906	147,0	149,3	583	92,1	151,2	72	378,3	372,4	<0,001

Freq - Relative Frequency; N - Absolute Frequency; SD - Standard Deviation

Figure 1: Description of each cluster obtained from PAM

PAM	Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6			Cluster 7			
	N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		N	Freq		p-value
Sex	1035			737			93			157			906			583			72			<0,001
Female	344	33%		290	39%		39	42%		71	45%		416	46%		287	49%		35	49%		
Male	691	67%		447	61%		54	58%		86	55%		490	54%		296	51%		37	51%		
Country of birth	1035			737			93			157			906			583			72			<0,001
Portugal	982	95%		723	98%		90	97%		151	96%		875	97%		570	98%		63	88%		
Other	53	5%		14	2%		3	3%		6	4%		31	3%		13	2%		9	13%		
District of residence	1035			737			93			157			906			583			72			<0,001
Porto	761	74%		534	72%		55	59%		73	46%		668	74%		306	52%		53	74%		
Other	274	26%		203	28%		38	41%		84	54%		238	26%		277	48%		19	26%		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	p-value
Age (years)	1035	10,5	5,1	737	8,8	6,0	93	8,4	6,2	157	6,4	5,8	906	7,5	6,2	583	8,3	6,3	72	7,9	6,1	<0,001

Freq - Relative Frequency; N - Absolute Frequency; SD - Standard Deviation

Figure 2: Description of the clustering from PAM using external patient-related variables

Acknowledgements Rita Gaio was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the projects with reference UIDB/00144/2020 and UIDP/00144/2020.

21 April, 9:40 - 10:00, Room A1.2

Clustering ECG time series for the quantification of physiological reactions to emotional stimuli

Beatriz Henriques[1], **Susana Brás**[1,2], **Sónia Gouveia**[1,2]
[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Portugal.
[2] Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Portugal.
beatriz.henriques@ua.pt, susana.bras@ua.pt, sonia.gouveia@ua.pt

---

Emotion recognition systems aid in identifying emotions and improving the treatment of anxiety and depression. This work followed an experimental protocol for collecting physiological non-invasive data from participants watching emotional videos, provoking fear, joy, and neutrality. Data analysis and clustering clearly showed that the stimuli effectively provoked variations in the heart rhythm of the participants, exhibiting different degrees of intensity.

Keywords: emotion classification, ECG, cluster analysis, affective computing.

Good health and well-being are one of the United Nations' sustainable goals, in particular, the promotion of mental health. Mental health problems are usually associated with a magnification of negative feelings (fear, disgust, etc.), as well as with a difficulty in enjoying positive moments. In this context, the identification of emotional states may help when dealing with these difficulties. Emotion recognition systems can be one of the solutions to help people to identify and interpret their emotions, and are based on the information knowledge extracted while monitoring quantified body alterations through emotional stimulation. In these tasks, one of the most used physiological signals is the electrocardiogram (ECG) due to its non-invasive and non-intrusive nature, and its high informative content regarding spontaneous behaviour. The final goal is to incorporate the most relevant information conveyed in the features extracted from the acquired physiological signal into an algorithm capable of labelling the emotional states with a maximized performance.
This work enrolled 56 subjects in an experimental protocol at University of Aveiro, to visualize videos with different emotional content (Fear, Happy and Neutral) while monitoring the ECG signal. The work focused the analysis of time series of RR intervals i.e. the time interval between successive ECG R-wave occurrence times. The RR series were used to identify time instants exhibiting a significant physiological response based on a two-step procedure. Firstly, by identifying the events (time intervals) of the protocol with a significant response based on the normalized group average. Secondly, by inspecting if the individual response, at the identified events, is significant with respect to the corresponding

variations of the subject baseline. The intensity of the response to each stimuli was quantified as the area of each event normalized by its duration. The algorithms were implemented in Python via *Neurokit2* [1] (advanced biosignal processing tools), *SciPy* (statistics) and *scikit-learn* (clustering algorithms) modules.
After hierarchical clustering via dendrogram, six clusters were considered to discriminate several degrees of intensity in the response for the Fear stimuli. This number of clusters was considered into a K-means clustering algorithm to produce distinct non-overlapping clusters (0.62 Silhouette Coefficient, 2156.52 Calinski-Harabasz index and 0.42 Davies-Bouldin index). Figure 1 presents the distribution of the increase in heart rate as a reaction to the stimuli (averaged for all events) for each subject within each cluster and clearly highlights that as the intensity of the response increases so does the average increase of the heart rate of the subject with respect to its baseline pattern.

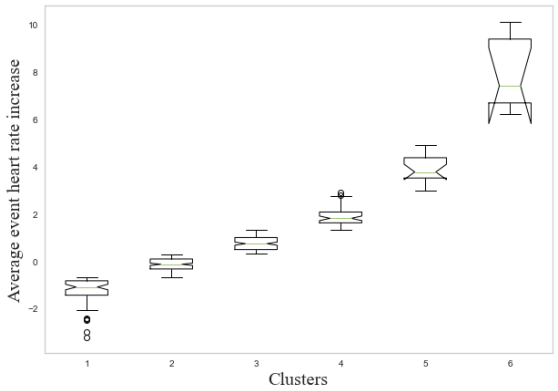

Figure 1: Boxplots showing the distribution of the averaged heart rate increase for the events of each subject within each cluster. The results displayed concern the Fear emotion.

Acknowledgements This work was partially funded by FCT - Fundação para a Ciência e a Tecnologia (FCT), I.P., through national funds, within the scope of the UIDB/00127/2020 project (IEETA/UA, http://www.ieeta.pt/). S. Brás acknowledges the support by national funds, European Regional Development Fund, FSE through COMPETE2020, through FCT, in the scope of the framework contract foreseen in the numbers 4, 5, and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19.

References

[1] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53:1689–1696, 2021.

21 April, 10:00 - 10:20, Room A1.2

Clustering analysis for household week-daily water consumption profiles characterization

João Bastos[1], **Flora Ferreira**[1,2,3], **Duarte Silva**[4], **Wolfram Erlhagen**[1,2], **Estela Bicho**[3]

[1] Department of Mathematics, University of Minho, Portugal, pg44585@alunos.uminho.pt
[2] Centre of Mathematics, University of Minho, Portugal fjferreira@math.uminho.pt, wolfram.erlhagen@math.uminho.pt
[3] Algoritmi Center, University of Minho, Portugal, estela.bicho@dei.uminho.pt
[4] Águas do Norte, Portugal duarte.silva@adp.pt

Cluster analysis is increasingly applied to identify patterns in water consumption in order to improve the planning and management of water demand. This study aims to detect the week-daily domestic water consumption profiles of a group with 342 individual customers from the north of Portugal. Time series k-means clustering was used to identify different profiles taking into account the seven days of a week. Three distinct patterns of water consumption over the day were identified, which could be used to increase the prediction accuracy of water consumption forecasting for network planning and operation.

Keywords: clustering, household water consumption, K-means, consumption variation

---

One of the current challenges in developing intelligent models for monitoring and planning water demand is the heterogeneity of the consumers served by the regional water system, including the possible differences in water consumption throughout the days of the week [1]. In this study, the objective is to detect daily domestic water consumption profiles taking into account each weekday.

The dataset used in this study contains hourly domestic water consumption measurements taken throughout 2021 by 342 individual customers. Of a total of 124 830 time series 312 with missing records were excluded, remaining between 360 to 365 days for each customer. The normalized hourly consumption for each day was obtained by dividing each hourly measurement by the sum of the 24 measurements for that day so that the sum of all normalized measurements for each day is equal to 1. To identify consumption patterns for each day of the week, seven time series were created for each customer, one for each day of the week. These time series were generated by computing the average consumption of each hour for a specific day of the week.

Time series K-Means clustering algorithm was selected to find the consumption patterns. To emphasize the importance of time-series shape, the Dynamic Time Warping (DTW) similarity measure [3, 4], which is sensitive to time shifts, was applied. However, to balance the time-shifting sensitivity desired for the model with a limited similarity span of the

time series, the Sakoe-Chiba constraint was applied with a radius of 2 [4]. This ensured a confined similarity to a temporal shift of 2 hours between two profiles. The Silhouette analysis [2] was utilized to optimize the number of clusters, k, by its ability to assess cluster coherence and distance from the others simultaneously, leading to the selection of $k = 3$ which had the highest silhouette score.
Applied the constrained Time-series K-Means with the selected number of clusters, one cluster stands out from the other two, given its close to constant consumption throughout the day. The time series belonging to this cluster are also evenly distributed throughout all the days of the week. Regarding the other two clusters, one cluster is comprised mainly of weekend time series, while the other one is more of weekdays. The first cluster had two peaks around 8 a.m. and 8 p.m., and a lower peak around 1 p.m., resembling a weekday consumption pattern. The latter cluster had three peaks at almost the same level, representing a more distributed consumption throughout the day, with the first two peaks occurring further ahead in the day being around 11 a.m. and 2 p.m.
The results showed that one general pattern is insufficient to define households' water use; three distinct patterns of daily water consumption associated with the weekday were identified. Future research will explore different sample aggregation approaches (e.g. household weekdays in the four seasons).

Acknowledgments Supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within the projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Noa Avni, Barak Fishbain, and Uri Shamir. Water consumption patterns as a basis for water demand modeling. *Water Resources Research*, 51(10):8165–8181, 2015.

[2] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[3] Hiroaki Sakoe. Dynamic-programming approach to continuous speech recognition. In *1971 Proc. the International Congress of Acoustics, Budapest*, 1971.

[4] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

22 April, 9:00 - 9:20, Auditorium Francisco Sampaio

A valuation model for lab-grown diamonds

Margarida G. M. S. Cardoso[1], **Luís Chambel**[2]
[1] ISCTE-IUL, BRU-IUL margarida.cardoso@iscte-iul.pt
[2] Sínese, luischambel@sinese.pt

This study aims to develop a valuation model for lab-grown cut diamonds based on data published on the Internet. Regression trees, Bayesian Networks, and K-Nearest Neighbors are used for this purpose. These different techniques have a complementary role in the application. The K-Nearest Neighbors has a good performance in prediction. Regression trees contribute to a better understanding of the relationship between predictors and the target. Bayesian Networks also add some insights in this respect. Finally, the models' results are compared to similar approaches when applied to natural diamonds.

Keywords: lab-grown diamonds, price, regression trees, K-nearest neighbors, Bayesian networks

The development of diamond synthesis technology in the 50's changed the industry, up to the point that almost all diamonds used in industrial applications are manufactured and, in recent years, gem-quality lab-grown (synthetic) diamonds captured an increasingly significant share of the jewelry market.
Previous works, referring to the use of Machine Learning techniques to predict diamonds' prices, generally consider natural diamonds - E.g. [1], [2]. Predictors include the 4Cs - Cut, Clarity, Carat, and Color - which are important determinants of a diamond's price.
We focus on predicting the prices of lab-grown diamonds. The data were collected from the website *https://www.1215diamonds.com*, on September 2022, and include 44443 observations (synthetic diamonds). In order to compare the performance of similar models as applied to natural diamonds, we collected data concerning natural diamonds from *https://belgiumdiamonds.net* (on April 2022) which comprises 34449 observations (natural diamonds). In both data sets, Train and Test samples were constituted, including 70% and 30% of the total number of observations, respectively.
Regression trees (Tree), Bayesian Networks (BN), and K-Nearest Neighbors (KNN) supervised approaches were used for predicting the prices of diamonds. They are implemented in the R packages "tree", "bnlearn" and "FNN". Following their parametrization, the referred approaches yielded the results presented in Table 1 and Table 2. Additional metrics were also considered.
The predictive capacity of KNN clearly surpasses the one obtained with Tree and BN. However, these latter algorithms help to uncover interesting relationships between the target and the predictors - E.g. see Figure 1. In the future, additional data sets, referring

Table 1: Lab-grown diamonds price prediction: R-Squared values

Sample	Tree	BN	KNN
Train	62.287	66.790	65.356
Test	59.835	59.977	60.120

Table 2: Natural diamonds price prediction: R-Squared values

Sample	Tree	BN	KNN
Train	88.080	93.952	90.508
Test	87.114	92.306	91.023

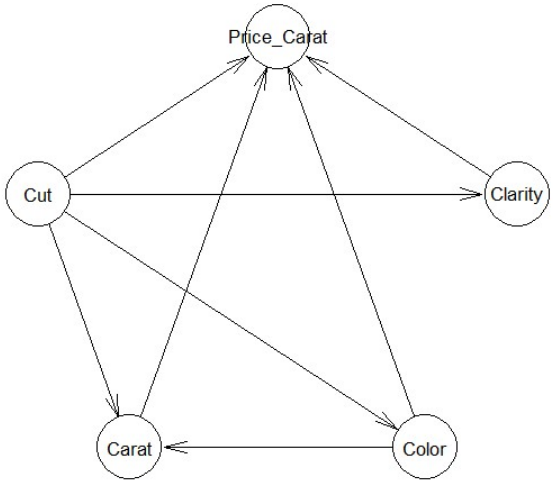

Figure 1: Bayesian Network structure learned from lab-grown diamonds data set

both to lab-grown and natural diamonds, should be used to assess the consistency of the relationships found.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, Grant UIDB/50021/2020.

References

[1] M. G. M. S. Cardoso and L. Chambel. A valuation model for cut diamonds. *International Transactions in Operational Research*, 12(4):417–436, 2005.

[2] Patel M.I. Jani S. Mihir, H. and R. Gajjar. Diamond price prediction using machine learning. 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4). IEEE, 2021.

22 April, 9:20 - 9:40, Auditorium Francisco Sampaio

Decomposed mutual information maximization: a feature selection method based on mutual information

M. Rosário Oliveira[1], **Francisco Macedo**[2], **Rui Valadas**[3], **Eunice Carrasquinha**[4], **António Pacheco**[5]
[1] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, rosario.oliveira@tecnico.ulisboa.pt
[2] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, francisco.macedo@tecnico.ulisboa.pt
[3] Instituto de Telecomunicações and Dep. Electrical and Computer Engineering, Instituto Superior Técnico, Univ. Lisboa, rui.valadas@tecnico.ulisboa.pt
[4] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, eunice.trigueirao@tecnico.ulisboa.pt
[5] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, antonio.pacheco.pires@tecnico.ulisboa.pt

---

Feature selection has been recognized for long as an important technique to reduce dimensionality and improve the performance of regression and classification tasks. The class of sequential forward feature selection methods based on Mutual Information (MI) is widely used in practice, mainly due to its computational efficiency and independence from the specific classifier. We propose the Decomposed Mutual Information Maximization (DMIM) method, which keeps the good theoretical properties of the best methods proposed so far but overcomes the complementarity penalization by applying the maximization separately to the inter-feature and class-relevant redundancies.

Keywords: mutual information, feature selection, classification

---

Feature selection is a preprocessing step that reduces the complexity and increases the interpretability of several statistical learning tasks, like classification and regression. Moreover, it avoids the performance degradation frequently occurring when redundant and irrelevant features are included in the learning process. In this talk, we proposed *Decomposed Mutual Information Maximization* (DMIM), a novel sequential forward feature selection method based on Mutual Information. The theoretical properties of DMIM were derived using the framework introduced in [1], which defines an optimal target objective function that the feature selection methods should try to approximate. DMIM, introduced in [2], shares the good theoretical properties of the best methods proposed so far, while avoiding the complementary penalization introduced by CMIM, its closest competitor.

We evaluated DMIM using two synthetic scenarios, for which an optimal feature ranking is available, and 20 publicly available real datasets analysed with kNN and C5.0 classifiers. The performance of the methods was assessed as the proportion of correct feature rankings, for the synthetic scenarios, and using seven different classifier performance measures, for the real datasets. The results obtained with the synthetic scenarios confirm the superiority of DMIM in overcoming the complementarity penalization. When addressing the real datasets, there is no global winner among the feature selection methods under study. The classification performance depends on the data and the classifier characteristics, and different methods adapt better to specific characteristics. Nevertheless, DMIM stands out in terms of the number of datasets where it is the best method. Moreover, in cases where it is not the best method, its performance is not far from the best one. Thus, given its superior theoretical properties and the results obtained with synthetic scenarios and real datasets, DMIM should be the preferred sequential forward feature selection method based on Mutual Information.

Acknowledgements This work was supported by FCT - Fundação para a Ciência e a Tecnologia, I.P., Portugal, through projects UIDB/04621/2020, UIDB/50008/2020, and PTDC/EEI-TEL/32454/2017 *Machine Learning based Profiling for Internet Security*.

References

[1] Francisco Macedo, M. Rosário Oliveira, António Pacheco, and Rui Valadas. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, 325:67–89, 2019.

[2] Francisco Macedo, Rui Valadas, Eunice Carrasquinha, M. Rosário Oliveira, and António Pacheco. Feature selection using decomposed mutual information maximization. *Neurocomputing*, 513:215–232, 2022.

22 April, 9:40 - 10:00, Auditorium Francisco Sampaio

Probabilistic Vector Machines

Pedro Duarte Silva[1]
[1] Católica Porto Business School & CEGE
Universidade Católica Portuguesa, psilva@ucp.pt

Support Vector Machines (SVMs) are among the most accurate classifiers in supervised problems. However, SVMs fail to complement these predictions with reliable estimates of class probabilities. A novel algorithm to estimate class probabilities from sequences of weighted SVMs will be presented. Numerical experiments will show that this algorithm, scales better than existing alternatives, and is competitive with both model free machine learning approaches, and model based statistical methodologies.

Keywords: support vector machines, kernel methods, multiclass classification, multiclass probability estimation

---

Kernel based Support Vector Machines (SVMs) were originally designed to handle two-class supervised classification problems, and quickly established themselves as one of the most accurate machine learning algorithms for class prediction. However, this success did not translate to the related task of deriving reliable probability estimates of class membership. In fact, Lin [3] has shown that, by targeting directly classification boundaries, standard SVMs do not carry much further information about class probabilities other than the predicted class by itself. Nevertheless, Lin, Lee and Wahba [4] showed that, by appropriately modifying (weighting) the loss function used in standard SVMs, nonstandard SVMs can estimate consistently a theoretical Bayes rule for any arbitrary setting of class probabilities. Based on this property, Wang, Shen and Liu [5] proposed to solve sequences of nonstandard SVMs with varying weight specifications, and to recover class probabilities from the frontiers between regions of the weights domain that lead to different predictions. The first proposal to extend this idea to the general k-class problems, is an *all-in-one* approach due to Wu, Zhang and Liu [7] (WZL). However, in this method the number of base weighted SVMs increases exponentially with the number of classes, and their training requires the optimisation of non-convex problems, making the method impractical for big, or even moderate, data problems. Multiclass probability estimation based on pairwise *one-against-the-rest* weighted SVMs were proposed in [8] and [6]. None of these two methods shares the computational difficulties of the *all-in-one* WZL method.
This work addresses the computational difficulties associated with the WZL *all-in-one* approach, and compares its statistical performance against competing alternatives. In particular, on the one hand, we will propose an improved method for recovering class probability estimates from weighted SVM predictions. In our approach, these estimates

will be based on the solutions of linear programming models that optimize an l_1-norm measure of the agreement between the predictions implied by probability estimates, and those made by weighted SVMs. One important advantage of this strategy is that, unlike in the original WZL method, the different weight specifications do not have to be uniformly distributed over a k-dimensional simplex, which allows for the creation of grids with satisfactory resolution, while ensuring that the number of required weighted SVMs only grows linearly with the number of different classes. On the other hand, we propose to replace the WZL SVM by an SVM based on an universal kernel without bias terms, using the weighted loss proposed by Lin, Lee and Wahba [2] (LLW). We note that the LLW loss leads to convex optimization problems and, as noted in [1], for multiclass SVMs based on universal kernels, dropping bias terms is statistically of minor importance, while allowing for the use of computationally efficient decomposition algorithms for SVM training. Based on these strategies, we were able the find reliable class probability estimates for problems with hundreds of examples, and more than a dozen different classes.
Numerical comparisons suggest that class probability estimation based on weighted SVMs are usually more accurate than competing distribution free machine learning approaches, and more reliable than model based statistical methodologies when their assumptions fail. Amongst the SVM based methods, no alternative is universally superior to the others, and the best method seems to depend on the particular data conditions at hand.

References

[1] U. Dogan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45):1–32, 2016.

[2] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[3] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):988–994, 2002.

[4] Y. Lin, Lee Y., and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

[5] J. Wang, X. Shen, and Y. Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, 2008.

[6] X. Wang, H.H. Zhang, and Y. Wu. Multiclass probability estimation with support vector machines. *Journal of Computational and Graphical Statistics*, 28(3):586–595, 2019.

[7] Y. Wu, H. Zhang, and Y. Liu. Robust model-free multiclass probability estimation. *Journal of the American Statistical Association*, 105(489):424–436, 2010.

[8] T. Xu and J. Wang. An efficient model-free estimation of multiclass conditional probability. *Journal of Statistical Planning and Inference*, 143:2079–2088, 2013.

22 April, 9:00 - 9:20, Room A1.2

Experiences at home during the COVID-19 quarantine - A cluster analysis

Eulália Santos[1], **Vasco Tavares**[2], **Fernando Tavares**[3], **Margarida Oliveira**[4]
[1] Escola Superior de Educação e Ciências Sociais, CI&DEI, Politécnico de Leiria, eulalia.santos@ipleiria.pt
[2] ISEG-UL—Lisbon School of Economics & Management, University of Lisbon, vctavares@aln.iseg.ulisboa.pt
[3] REMIT, Department of Economics and Management, Universidade Portucalense, ftavares@upt.pt
[4] ISCTE - Instituto Universitário de Lisboa, Business Research Unit (BRU-IUL), margarida.foliveira@hotmail.com

Experiences are a part of the daily life of any human being and allow acquiring a set of knowledge and skills. The study aims to analyze the experiences lived at home during the COVID-19 pandemic quarantine. A methodology based on a questionnaire survey was used to analyze the experiences lived by a sample of 726 Portuguese individuals during the quarantine period. The results show that through the application of exploratory and confirmatory factor analysis, a structure with four factors was obtained: Sense and Feel, Pandemic Feel, Pandemic Think, and Act. The application of the cluster analysis technique identified the presence of four clusters, showing that experiences are lived with different levels of intensity. These results can be useful for scientific knowledge in the behavioural area and for defining adequate strategies to improve the health, well-being and quality of life of individuals in pandemic contexts.

Keywords: experiences, confirmatory factor analysis, cluster analysis.

Each human being throughout his/her life acquires a set of experiences that allow him/her to gain knowledge and skills. An experience is a unique event or situation that can be experienced at home, at work, in school, on a travel, among other daily situations. But during the COVID-19 pandemic quarantine, experiences became limited to the "home" space, without contact with other people except for those living in the same space. Experiences have been approached in different areas of knowledge, such as in the fields of education, psychology, and marketing. Experience can be linked to emotions and subjectivity [5]. In marketing, experience is a key element, as it affects the way the consumer perceives feels, knows or does things [3]. The strategic experiential marketing modules are called Strategic Experiential Modules; according to Schmitt [4] they are based on the customer's experiences and are five: Sense (sensory experiences), Feel (affective experiences), Think (creative cognitive experiences), Act (physical behavioral and lifestyle experiences)

and Relate (social identity experiences related to a group or reference culture). In the context of the present work, the strategic experimental modules are the various types of experiences that individuals experience during the quarantine period, the consumers of the experiences are the individuals (726 Portuguese aged between 18 and 79) and the experience itself can be analysed as the brand or the company. To measure the experiences, a scale based on experiential marketing was used. It was adapted from the literature review of other experience scales used in the marketing and tourism areas.
In general terms, the surveyed individuals felt concerned about the delicate moment that was being experienced worldwide, the situation in which they found themselves made them think about the future of human life and the lifestyle they had. On the other hand, they considered that moment an opportunity for people living together to relate, they also thought about the country's economy and had a great desire to visit friends or family. In the confirmatory factorial analysis, the maximum likelihood estimation method was used and the following goodness of fit indices: the ratio of the Chi-square statistic by the degrees of freedom, Goodness of Fit Index, Comparative Fit Index and Root Mean Square Error of Approximation. The application of factor analysis to the scale of experiences lived during quarantine showed the existence of four factors: Sense and Feel, Pandemic Feel, Pandemic Think, and Act. Moreover, it manifests itself with greater intensity in the factors Pandemic Think and Sense and Feel. The fit indices of the model with 4 factors showed good fit quality, and the scale of experiences also showed evidence to be considered with adequate convergent, discriminant, and internal consistency validity ([1], [2]). In the cluster analysis, the squared Euclidean distance was used as a measure of dissimilarity and the Ward method to group individuals with homogeneous characteristics. This analysis allows defining four clusters. Cluster 1 is composed of individuals who live experiences intensely, cluster 2 is composed of more reflective individuals, cluster 3 is formed by individuals who were worried, and cluster 4 by those who live experiences with low intensity. This shows that experiences are lived at different levels of intensity.

References

[1] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate data analysis*. Pearson, 2014.

[2] R. B. Kline. *Principles and practice of structural equation modeling*. Guilford publications, 2016.

[3] S. Same and J. Larimo. Marketing theory: experience marketing and experiential marketing. In *7th International Scientific Conference "Business and Management*, pages 10–11, 2012.

[4] B. Schmitt. Experiential marketing. *Journal of Marketing Management*, 15(1-3):53–67, 1999.

[5] B. Schmitt. Experience marketing: Concepts, frameworks and consumer insights. *Foundations and Trends® in Marketing*, 5(2):55–112, 2011.

22 April, 9:20 - 9:40, Room A1.2

How to measure the fit of a structural equation model with omissions by design

Paula C. R. Vicente[1]
[1] Lusofona University, ECEO, COPELABS paula.vicente@ulusofona.pt

The adjustment of a structural equation model could be evaluated using different measures, such as, RMSEA, SRMR, CFI and TLI. This simulation study aims to understand how the number of indicators and of latent variables, as well as different sample sizes and parameter values could influence these indexes when analysing data from a planned missing design.

Keywords: fit measures, planned missing design, structural equation model, simulation study

The use of a planned missing design is important, because this kind of design allows the collection of high-quality data while reducing participant burden [1]. Used in different areas of knowledge, a particular case of a planned missing design is the 3-form design. This design could be used in cross-sectional and longitudinal research and consists in splitting the questions of the survey in four groups, X, A, B and C. All the participants must answer to the questions in the X block, and are then randomly assigned to answer to two other blocks, from A, B and C (see table 1). Consequently, 1/3 of the participants answer to questions in set XAB, 1/3 to XAC and 1/3 to XBC, instead of answering questions from all the four groups [2].

Table 1: Missing data pattern for a 3-form design

	Question set			
Form	X	A	B	C
1	O	O	O	-
2	O	O	-	O
3	O	-	O	O

On the other hand, in the adjustment of a Structural Equation Model (SEM) it is important to quantify this adjustment. As such, there are several different indexes, based in distinct criteria, to evaluate if the considered model is the right one. The most used indexes with this type of modeling are Root Mean Square Error of Approximation (RMSEA), Root Mean Square Residual (SRMR), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI).

The current study explores the effect of the non-responses due to a 3-form design on the SEM fit measures mentioned. The simulation conditions were obtained by manipulating five variables: sample size, model size, model type, factor loadings and correlation between factors. In figure 1, it is presented an example of a model used. For each simulated condition, 1000 replications were generated using the *simsem* package in R [3].
Results show that for medium and big sample sizes all the indices have acceptable values. However, in small samples, the index with the best performance is RMSEA.

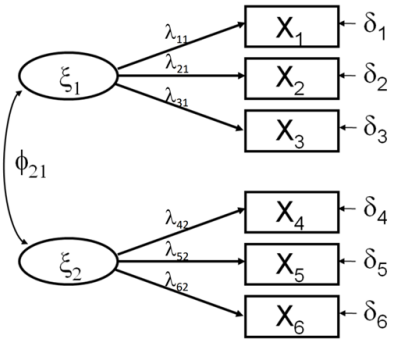

Figure 1: Structural equation model with 2 latent variables and 6 observed variables

References

[1] Enders C.K. *Applied Missing Data Analysis*. The Guilford Press, New York, 2010.

[2] Graham et al. Planned missing data designs in psychological research. *Psychological Methods*, 11:323–343, 2006.

[3] Pornprasertmanit et al. simsem: SIMulated structural equation modeling. Technical report, version 0.5-16, 2021.

22 April, 9:40 - 10:00, Room A1.2

Students' burnout at a Portuguese polytechnic: PLSc-SEM approach

Luís M. Grilo[1,2,4,5], **M. Cristina Costa**[3,5]
[1] Universidade de Évora, Portugal
[2] Universidade Aberta, Portugal
[3] Instituto Politécnico de Tomar (IPT), Portugal
[4] NOVA Math (Centro de Matemática e Aplicações), FCT NOVA, Universidade NOVA de Lisboa, Portugal
[5] Ci2 (Centro de Investigação em Cidades Inteligentes), IPT, Portugal

---

The well-being of college students is essential for them to succeed in their academic performance and future careers, as well as to prevent increasing dropout rates. A conceptual structural model was proposed considering "optimism" as an exogenous latent construct, "perceived stress" as a mediating construct, and "students' burnout" as the target latent construct. Based on data collected in a survey, we estimate a reflective model using the consistent Partial Least Squares estimator. We conclude that "optimism" has a direct negative effect on "perceived stress" (as expected) and also an indirect effect through this full mediator on "exhaustion", which is considered the central component of burnout.

Keywords: multigroup analysis, reflective model, survey, well-being, college students

---

During the years of the COVID-19 pandemic students saw their academic life being very limited, namely in terms of reduced social contacts (with family, colleagues and teachers). In addition, there is a lot of information about climate change and conflicts in the world that generates insecurity and uncertainty about the future, which in turn affects their well-being and leads to growing demotivation. Together with daily pressures to achieve a good academic performance with a view to entering the labour market, this generates high levels of stress that can result in burnout (considered a health condition that results from continuous and excessive stress that students are subjected to in their daily lives)[1]. In June 2022, we used an online questionnaire in a Portuguese Polytechnic, consisting of the Revised Life Orientation Test (LOT-R) to assess "optimism" [2], the Perceived Stress Scale (PSS) to assess "perceived stress" and the MBI-SS to assess "students' burnout" [1]. We considered the multivariate statistical technique of Structural Equation Modeling (SEM) [3] and a theoretical reflective model was proposed, where "optimism" is an exogenous construct, "perceived stress" is a mediating construct, and "burnout" (considering the three dimensions: "exhaustion", "cynicism" and "efficacy") is the endogenous construct [2, 4]. Using the complete sample, an estimated model was obtained using the consistent Partial

Least Squares (PLSc) estimator [3], which performs a correction of reflective constructs' correlations to make results consistent with a factor model. As expected, "optimism" has a negative direct effect on "perceived stress" and the latter has a positive direct effect on "exhaustion" and indirect effects on "cynicism" and "efficacy". Furthermore, "exhaustion" has a direct positive effect on "cynicism" and the latter has a direct negative effect on "efficacy". Considering the estimated model, two "submodels" by gender (male and female) were also estimated, and the results obtained in the multigroup analysis indicated that the biggest difference is in the value of the path coefficient (higher for men) between "perceived stress" and "exhaustion", which is statistically significant. The results obtained helped to better understand the state in which the institution's students find themselves in relation to the variables under study, and contributed to the consideration of some interventions in order to improve the well-being and, consequently, the performance of the students.

Acknowledgements
This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020.

References

[1] Schaufeli W. B., Maslach C. Leiter M. P., and Jackson S. E. *Maslach Burnout Inventory–General Survey.* Palo Alto, CA Consulting Psychologists Press, 1996.

[2] Chang E. C., Rand K. L., and Strunk D. R. Optimism and risk for job burnout among working college students: Stress as a mediator. *Personality and Individual Differences*, 29, 2:255–263, 2000.

[3] Hair J. F., Hult G. T. M., Ringle C. M., Sarstedt M., Danks N. P., and S. Ray. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R.* A Workbook, Springer, 2022.

[4] Grilo L. M. and Costa C. Optimism and stress as predictors of burnout in college students: Plsc-sem approach. *VIII Workshop on Comp. Data Analysis and Numerical Methods*, 2022.

22 April, 10:00 - 10:20, Auditorium Francisco Sampaio

Counterfactual impact evaluation - An exploratory study on urban revitalization projects in Aveiro and Ílhavo municipalities

André Lima[1], **Paulo Batista**[2], **João Marques**[3]
[1] Universidade de Aveiro, andre.lima@ua.pt
[2] Universidade de Aveiro, pauloricardolb@ua.pt
[3] Universidade de Aveiro, jjmarques@ua.pt

The study presents an approach of CIE (counterfactual impact estimation) of an urban revitalization project (Aveiro/Ílhavo municipalities), using the Difference-in-differences quasi-experimental method to estimate the impacts on housing prices. A framework to address the well-known spatial issues through a hedonic price modeling setting will be presented. In particular, we will define the spatial extent (georeferencing of the urban intervention), develop an approach to fix spatial heterogeneity and integrate spatial features in the definition of the counterfactual(s).

Keywords: CIE, urban revitalization, quasi-experimental, spatial heterogeneity

In the context of the Portugal 2020 Program, this project will focus on interventions co-financed under thematic objective 6 (preserve and protect the environment and promote the efficient use of resources), specifically, their translation into the Regional Operational Programmes (POs). The impact evaluation of public policies is critical to enable policymakers to make informed decisions and to improve the design of future public policies [1]. This study follows the CIE approach, answering questions like "How much difference does it make?" and, as expected, will try to analyze the conditions to claim the empirical evidence and (ideally) the possible causal interpretation.
In this study, it is intended to ensure a selection of projects with a clear spatial character (capable of being georeferenced) and adequate to the objectives of the study (interventions on the urban tissue). Thus, (spatial) Difference-in-difference quasi-experimental methodology is presented to estimate the impact of urban revitalization interventions of public rehabilitation projects funded by the Portugal2020 program. The Aveiro and Ílhavo Municipalities (Portugal) are selected as case studies and the housing prices were selected as our target feature where impacts will be evaluated. The selected projects follow a clear spatialized character (capable of being georeferenced) and are adequate to the objectives of the study (interventions on the urban tissue) and cover a period from June 2016 to March 2017.
To develop the analytical exercise, we addressed the challenges of collecting relevant data. Two datasets were obtained, converting the pre and post time periods of the interventions.

The data was provided on housing advertising portals and required a deep ETL (extracting, transforming, and loading) process to turn it adequate for the study purposes. From the model point of view, our CIE approach will rely on the hedonic price modeling framework for housing prices and their determinants [2]. On behalf of this methodology, we will embrace the spatial pitfalls highlighted by spatial econometrics literature and develop a time and spatial matching approach in order to build the required counterfactual setting. The data gathered was retrieved from two data sets, each one covering two different time frames. They will be provided by two different entities, despite both of them will be collected in the context of housing advertising operations (check Table 1 below).

Table 1: Data Sets Detail.

	Dataset CasaSapo	**Dataset Prime Yield**
Timeframe	2005-2010	2018-2021
Source / Method	Housing Portal	Enterprise
Data Points	67367	14026

As the datasets were not developed for CIE proposed, both were pre-processed such as: removal of data entries with missing data and which are not concerning housing transactions, comparable variables were created for both datasets (from the initial information).
The CIE framework presented here will take into account the spatial heterogeneity and dependence though a choice of a model (spatial) specification guided by the Elhorst approach [3]. Moreover, the counterfactual(s) will be defined by jointly consider the intrinsic housing attributes and its surrounding (spatial) features [4]

Acknowledgements The study (under development) is sponsored by the Portuguese Technical Assistance Operational Program (project reference POAT-01-6177-FEDER-000054). The authors thank Janela Digital SA and PrimeYield SA for providing data on housing listings.

References

[1] EU Commision. *EVALSED - The resource for the evaluation of Socio-Economic Development: Sourcebook - Method and techniques*. 2013.

[2] R. Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82, 1974.

[3] J. P. Elhorst. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. 2014.

[4] J. Dubé, M. AbdelHalim, and N. Devaux. Evaluating the impact of floods on housing price using a spatial matching difference-in-differences (sm-did) approach. *Sustainability*, 13, 2021.

22 April, 10:20 - 10:40, Auditorium Francisco Sampaio

Black scabbardfish species distribution: geostatistical inference and bayesian sampling design under preferential sampling

Paula Simões[1], **M. Lucília Carvalho**[2], **Ivone Figueiredo**[3], **Andreia Monteiro**[4], **Isabel Natário**[5]
[1]NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, and Military Academy Research Center - Military University Institute (CINAMIL), Portugal, pc.simoes@campus.fct.unl.pt
[2] Centre of Statistics and its Applications (CEAUL), Faculty of Sciences of the University of Lisbon, Portugal, mlucilia.carvalho@gmail.com
[3] Portuguese Institute for Sea and Atmosphere (IPMA), Portugal, ifigueiredo@ipma.pt
[4] NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, Portugal, andreiaforte50@gmail.com
[5] NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, and Department of Mathematics, NOVA University of Lisbon, Portugal,icn@fct.unl.pt

---

Black Scabbardfish (BSF) captures are modelled using a geostatistical analysis combined with a preferential sampling technique which enables to better capture the variability of the BSF captures providing a more realistic pattern of BSF distribution. This approach allows a better knowledge os BSF spatial distribution assuming that the selection of the sampling locations depends on the values of the observed variable of interest. In order to construct a survey design to improve the BSF abundance estimates, in Portuguese waters, geostatistical sampling design strategies for preferentially sampled data are investigated.

Keywords: geostatistics, preferential sampling, sampling design, inla, point process

---

Black Scabbardfish (BSF) is a deep-water species that occurs in continental waters at depths greater than 800 m. On the portuguese coast BSF constitutes an important commercial resource. In the absence of dedicated deep-water research surveys in this area, the spatial distribution of its abundance is mainly inferred from commercial deep-water longline fishery that operates along the continental slope.
The Portuguese Institute of Sea and Atmosphere (IPMA) provided georeferenced data on the location of the fishing hauls and the corresponding captures for a number of differently sized vessels belonging to the BSF fishing fleet. It is intended to use this information, understood as preferentially sampled data, combined with environmental covariates, to predict where the species is likely to exist, also in unsampled locations, for management

and conservation purposes, in order to ensure the sustainability of commercial fisheries and protect the biodiversity of species that are of high interest for consumption.
The objective of this study is two-folded. In one hand, to perform species distribution modelling of the BSF data by using a geostatistical model-based method that takes preferentiability into account. Under a Bayesian approach and resorting to INLA methodology, by taking the stochastic partial differential equations (SPDE) for the spatial effects of the geostatistical model for the captures and the Log-Cox point process model (LGCP) for the locations, BSF captures can be analysed considering several different covariates and random effects [2, 3]. On the other hand, considering the present modelling of the BSF captures, predictions can be made at several potential sampling locations (unobserved locations) in order to construct a survey design to improve the BSF abundance estimates in Portuguese waters. Here different design classes are investigated, namely random, inhibitory and adaptative geostatistical sampling designs [4, 1].

Acknowledgements This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects PREFERENTIAL, PTDC/MAT-STA/28243/2017, UIDP/00297/2020 (Center for Mathematics and Applications) and UIDB/00006/2020 (Centre of Statistics and its Applications).

References

[1] M. Chipeta, D. Terlouw, K. Phiri, and P. Diggle. Adaptive geostatistical design and analysis for prevalence surveys. *Spatial Statistics*, 15:70–84, 2016.

[2] P. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232, 2010.

[3] M. Pennino, I. Paradinas, J. Illian, F. Muñoz, A. López-Quílez, and D. Conesa. Accounting for preferential sampling in species distribution models. *Ecology and evolution*, 9(1):653–663, 2019.

[4] P. Simões, M. L. Carvalho, I. Figueiredo, A. Monteiro, and I. Natário. Geostatistical sampling designs under preferential sampling for black scabbardfish. In R. Bispo, L. Henriques-Rodrigues, R. Alpizar-Jara, and M. de Carvalho, editors, *Recent Developments in Statistics and Data Science*, pages 137–151, Cham, 2022. Springer International Publishing.

22 April, 10:40 - 11:00, Auditorium Francisco Sampaio

An approach to estimate residential real estate prices with scarce information

Marco Marto[1], **Catarina Duarte Ribeiro**[2], **João Barrias**[3], **Paulo Batista**[4]
[1] Research Unit in Governance, Competitiveness and Public Policies (GOVCOPP), Universidade de Aveiro, marcovmarto@ua.pt
[2] Departamento de Ciências Sociais, Políticas e do Território, Universidade de Aveiro, catarinaduarteribeiro@ua.pt
[3] Departamento de Ciências Sociais, Políticas e do Território, Universidade de Aveiro, joao.barrias@ua.pt
[4] Departamento de Ciências Sociais, Políticas e do Território, Universidade de Aveiro, pauloricardolb@ua.pt

---

The Portuguese population has been decreasing according to the last census data. In addition, there is a trend to abandon some regions in the interior of the Portuguese mainland territory due to the economic attractiveness, more job opportunities, and better quality of life in regions nearer the coast and cities. The real estate sector reflects these social, economic, and demographic dynamics. This study aims to apply machine learning and data science's predictive models to estimate the price (per sq. m.) of residential real estate which has more limited accessibility in terms of pedestrian networks and at the same time a low representation in the recent history of online sales. The objective is to estimate the prices of residential real estate through methodologies of information and knowledge transfer. The results suggest as the best predictive models consider: regressions based on k-nearest neighbors (KNN), random forests (RF), support vector regression (SVR), and ensembles that include combinations of the previous models' estimations. Besides this work being applied to the municipalities of Aveiro and Ílhavo, the approach can easily be adapted to other territories.

Keywords: regression predictive models, residential real estate, knowledge transfer, limited data scenario, machine learning

---

The real estate sector plays a relevant role in the social, economic, and demographic dynamics in the territories because supports the attractiveness and population fixation together with the general interest services and job opportunities offered. The prices of residential real estate are a driver in this context and can aid to have positive migration rates in territories and influence the economy in general, as demonstrated recently by the economic crises which started with the real estate sector in U.S. in 2008.

One of the methods most used in the real estate sector is the comparative method which is based on finding the value of the real estate we want through recent previously known market transaction values' of similar or substitute samples of real estate. The question is how to estimate satisfactory residential real estate values when there are only a few or no samples of similar residential real estate recently transitioned in the market. This challenging question can be answered with regression predictive econometric and machine learning models ([1]; [2]) which can give us satisfactory estimations of residential real estate prices. Moreover, the imputation technique used to estimate the prices were based on knowledge and information transfer from similar residential real estate in other locations in the same municipalities, considering the values of real states that belong to the same cluster of characteristics.
Limited accessible residential real estate is defined as residential real estate with less than 5 neighbors at a pedestrian-accessible distance lesser than 500 meters with the aid of the shortest path Dijkstra algorithm in networks and DBSCAN clustering algorithm. Figure 1 shows all the DBSCAN-colored (not grey) clusters, our sample of unclustered residential real estate represented with the grey color is the 111 residential real estate with limited information for which we want to estimate the prices.

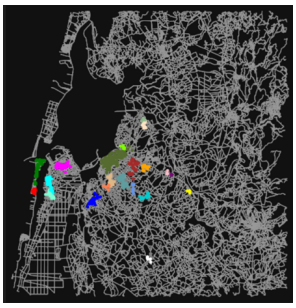

Figure 1: clusters' map (22) of accessibility in the network (colored) of residential real estate developed with the DBSCAN algorithm (OpenStreetMap)

Acknowledgements The Foundation for Science and Technology, I.P provided the funding for the fellowship of one of the authors (Marco Marto) which allowed him to develop this work. The company Prime Yield, SA provided the datasets.

References

[1] Y. Chen, X. Liu, X. Li, Y. Liu, and X. Xu. Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied Geography*, 75:200–212, 2016.

[2] J. L. Marques, P. Batista, E. A. Castro, and A. Bhattacharjee. Spatial Automated Valuation Model (sAVM)–from the notion of space to the design of an evaluation tool. *In International Conference on Computational Science and Its Applications. Springer, Cham*, pages 75–90, 2021.

22 April, 10:00 - 10:20, Room A1.2

Compositional data vectors: how useful they can be?

Adelaide Freitas[1], **Marta Maltez**[2], **Marco Costa**[3]
[1] Department of Mathematics & CIDMA, University of Aveiro, adelaide@ua.pt
[2] CIDMA, University of Aveiro, martamaltez@ua.pt
[3] Águeda School of Technology and Management & CIDMA, University of Aveiro, marco@ua.pt

---

In this work, we explore two examples of compositional vectors from the point of view of interpretation. The main propose is to highlight the potential of data analyses when this type of structure can be presented in the data set. In particular, the interpretation of Principal Component Analysis when applied on compositional vectors is illustrated and discussed.

Keywords: principal component analysis, compositional vector

Compositional data are multivariate observations providing quantitative descriptions of the parts of a whole. In mathematical notation, a D-multivariate vector is a D-compositional observation whenever all these D components are positive numbers and only relative information about the D parts of a whole (e.g. proportions, percentages) is relevant. This implies that the important information between components is given by ratios rather than differences. This concept can be extended to a observation defined by a composition of p D-compositional variables (i.e., p variables each one with D-part compositional components). This type of multivariate observation is referred to as a (p-dimensional) compositional data vector. The p-dimensional compositional data vectors are defined in a vectorial space (Simplex space) with two basic algebraic operations: the perturbation operator and the power transformation. These operators are deduced from component-wise operations of compositional data and correspond to the addition of two vectors and the multiplication of a vector by a real number on the Simplex space, respectively [2].
In [5], it was developed the principal component analysis (PCA) for modelling compositional data vectors. For such, the definition of sample mean, sample variance, sample covariance and sample correlation coefficient for compositional data vectors and the eigen-decomposition of the variance-covariance matrix were considered in the Simplex space. In the present study, to demonstrate the usefulness and interpretation of PCA for compositional data vector, this statistical technique is used on two real data sets to reduce dimension while explaining most total variability of the data with a small number of compositional vectors.
The first data set concerns the evolution of diet patterns of various European countries over a period of six decades. It is a extension of a previous work ([3]) with the data extracted freely from the Our World in Data ([4]). The focus is on a temporal composition of the

consumption in the diet of the four macronutrients: animal protein, vegetal protein, fat, and carbohydrate.
The second data set has been extracted from the Portuguese Institute of Statistics (INE) [1]. It summarizes the number of deaths by cause of death, age group and region in 2020. In this data set context, Portugal was divided into 25 regions, corresponding to the the third level of the Nomenclature of Territorial Units for Statistics (NUTS), NUTS III classification. The causes of death considered consist of a set of cause groups based on the International Statistical Classification of Diseases (ICD), divided by age groups at 20-year intervals. One of the challenges of this dataset is the existence of zeros. It is known that compositions should always be treated in terms of logratios which implies that zero components cannot be directly dealt with, as logarithms of zero values are undefined [2]. For this reason, the replacement strategy was applied.
Graphical methods will be applied to illustrate patterns resulting from the first principal components, they also compositional vectors.

Acknowledgements This work was supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

References

[1] Instituto Nacional de Estatística (2023). Óbitos (n.º) por local de residência (nuts - 2013), grupo etário e tipologia de áreas urbanas; anual, 2020. Extracted: 2022-11-16.

[2] Peter Filzmoser, Karel Hron, and Matthias Templ. *Applied Compositional Data Analysis. With Worked Examples in R.* 11 2018.

[3] Mariana Pinto, Marco Costa, and Adelaide Freitas. Evolution of diet patterns over time in europe from 1963 to 2013: an exploratory analysis using pca for compositional data vectors. *pecial Issue - Statistics on Health Decision Making: Real World Data. Journal on Statistics in Health Making Decision (extended abstract)*, 2022.

[4] Hannah Ritchie, Pablo Rosado, and Max Roser. Diet compositions. *Our World in Data*, 2017. https://ourworldindata.org/diet-compositions.

[5] Huiwen Wang, Liying Shangguan, Rong Guan, and Lynne Billard. Principal component analysis for compositional data vectors. *Computational Statistics*, 30:1079–1096, 2015.

22 April, 10:20 - 10:40, Room A1.2

A supervised clustering algorithm for preventing fraud in edge attributed network components

Pedro Campos[1], **Célia Carvalho**[2]
[1] Faculdade de Economia do Porto, LIAAD INESCTEC, & INE, pcampos@fep.up.pt
[2] Universidade do Minho, Escola de Ciências, celia-c-carvalho@sapo.pt

We develop a fraud detection supervised algorithm for edge attributed network components, EdgeX, to identify suspicions of fraudulent acts in a financial transactions' network. We use symbolic data analysis to characterize each component subgraph extracted from the original network, by aggregating data into symbolic objects, where each object corresponds to a different component, allowing us to find the determinants of fraudulent transactions through a supervised clustering algorithm.

Keywords: supervised clustering, symbolic data analysis, network components, edge-attributed network, Paysim

In a competitive environment, fraud can be a critical problem when the prevention procedures are not robust enough. Traditional practices of tax evasion through illegal transactions have been improved in the last years with the use of computers and mobile communications. Fraud is domain-specific, and there is no one-solution-fits-all method among fraud detection techniques [3]. We use Symbolic Data Analysis (SDA) [2] to aggregate data containing a summary of financial transactions, and apply a supervised learning approach. Chacón and Rodríguez [1] explore symbolic data analysis with supervised approaches, including classical linear regression models, tree-based regression models, K-nearest neighbors regression, support vector machines regression, and regression using neural networks, using the center method and the center and range methods in the context of each regression models considered.

In this work we developed a fraud detection supervised algorithm, EdgeX, to identify suspicions of fraudulent acts. Several financial transaction systems work in a basis of P2P (peer-to-peer) payments, corresponding to electronic transactions from one person or business to another. Using P2P links from an edge list, we build network components (a particular case of subgraphs), which are portions of the network that are disconnected from each other. In each component several peers are connected forming a closed group. Each component contains several edges with associated information (like in edge attribute networks), namely the label of the existence of fraudulent transactions, the total amount of transfers, and the type of transfer. Then, we use Symbolic Data Analysis at the level of the components for fraud classification, allowing us to find the determinants of fraudulent transactions. A symbolic supervised clustering algorithm on edges' attributes is then

applied. The goal is to discover interesting associations among different variables with respect to a property of interest. Santos and Campos [5] used SUWAN to assemble elements of a graph, based on their structural and compositional characteristics, while it provides class-uniform clusters, based on a predefined target variable. From a supervised clustering perspective, some authors argue that classical techniques of clustering do not guarantee that objects of the same class are grouped together, but some solutions exist that can deal with this limitation by improving a measure of clusters purity. An illustrative application of the method is made with a variant of PaySim [4], a Synthetic Financial Dataset for Fraud Detection, containing more than 6 million of transactions (the edges) between more than 2 million users (the nodes).

References

[1] J.E. Chacón and O. Rodríguez. Regression models for symbolic interval-valued variables. *Entropy*, 23:429–, 2021.

[2] E. Diday. Introduction à l'approche symbolique en analyse des données. *Premières Journées Symbolique-Numérique*, Université Paris IX Dauphine, Paris, France, 1987.

[3] B.K. Jha, G.G. Sivasankari, and K.R. Venugopal. Fraud detection and prevention by using big data analytics. *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India:267–274, 2020.

[4] E. Lopez-Rojas, A. Elmir, and S. Axelsson. Paysim: A financial mobile money simulator for fraud detection. *arXiv preprint*, arXiv:1606.04158, 2016.

[5] B. Santos and P. Campos. Suwan – a supervised clustering algorithm with attributed networks. *Intelligent Data Analysis*, to appear, 2023.

Poster Sessions

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Linear regression for symbolic density-valued data

Rui Nunes[1], **Paula Brito**[2], **Sónia Dias**[3]
[1] Faculdade de Ciências da Universidade do Porto & LIAAD-INESC TEC, Portugal, up201400313@up.pt
[2] Faculdade de Economia da Universidade do Porto & LIAAD-INESC TEC, Portugal, mpbrito@fep.up.pt
[3] Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal, sdias@estg.ipvc.pt

---

Symbolic data has gained importance due to data nature, both in dimension and in the type of analysis that we are now interested in performing. In this context we are interested in analyzing not the individuals but the groups they belong to. Symbolic Data Analysis provides a methodology that allows performing the analysis of such kind of units. In this work, we follow a continuous approach, going from histograms to densities. With this new variable type, we are now able to apply Functional Data Analysis approaches to symbolic data.

Keywords: symbolic data, functional data analysis, linear regression

---

The data that we encounter in our problems is becoming increasingly large and intricate. Additionally, our focus has shifted from individual behavior to group behavior, necessitating some form of data aggregation. Previous aggregation techniques relied on a single value, typically a central descriptive measure, to describe a group of individuals. However, this approach results in a loss of information inherent to the data, namely the variability that it comprises. To address this issue, Symbolic Data Analysis (SDA)[2] was introduced, which accounts for data with inherent variability. It is important to note that this variability may stem from either the recorded data itself or result from their aggregation. SDA considers two types of aggregation: *Contemporary*, where data are recorded at the same point in time and first-level units are not analyzed, and *Temporal*, where data is recorded at different points in time for the same individual, but time is not a concern. A symbolic variable Y is defined by a mapping

$$\begin{array}{rll} Y\colon & E & \to \mathcal{B} \\ & i & \mapsto Y(i) = \xi_i \end{array} \tag{1}$$

defined on a set E of statistical units. For a Histogram-valued variable Y, to each unit i corresponds an histogram $Y(i)$, that can be represented by the classical representation of a Histogram $H_{Y(i)}$ or its corresponding quantile function $\Psi_{Y(i)}(t)$ with $t \in [0, 1]$. The Mallows distance is a suitable metric for assessing the similarity between distributions. However, the quantile function representation of a Histogram-valued variable has certain

limitations. Although it offers some algebraic benefits, using quantile functions means working in a semi-vector space because $\lambda\Psi_X(t)$ with $\lambda < 0$ is no longer a quantile function. To address this issue in linear regression modeling, the Symmetric Distribution Model (DSD)[1] was developed. The DSD uses, together with the observed quantile function $\Psi_X(t)$, the symmetric distribution $-\Psi_X(1-t)$ to overcome the constraints of working in the space of the quantile functions.
In this study, we focus on a density-valued variable instead of a histogram-valued variable, where each unit i is represented by its density or by the corresponding quantile function. Working with continuous functions, we can employ Functional Data Analysis (FDA) approaches. Functional Linear Models (FLM)[3] are an extension of classical linear models and can be categorized into three types: *Scalar-on-function* regression (where the response is a scalar but the regressors are curves), *Function-on-scalar* regression (where the response is a curve but the regressors are scalars), and *Function-on-function* regression (where both the response and regressors are curves). As our problem involves a function-on-function model, and we are using quantile functions to represent variables, we have a special case of the model known as the Concurrent Model [3]. However, FLM cannot guarantee that the functional response is a quantile function. Therefore, in this work, we consider an extension of the DSD model to continuous variables using a quantile function as the functional response. We write the Linear Regression for Symbolic Density-valued Data as:

$$\Psi_{\hat{Y}_{(i)}}(t) = v + \sum_{j=1}^{p} \left(a_j \Psi_{X_{(i)j}}(t) - b_j \Psi_{X_{(i)j}}(1-t) \right) \tag{2}$$

with $v \in \mathbf{R}$, $a_k, b_k \geq 0$ and $t \in [0,1]$. We may obtain the regression coefficients by minimizing the Sum of Squared Errors (SSE):

$$\begin{aligned} \min \quad & SSE = \sum_{i=1}^{n} D_M^2 \left(\Psi_{Y_{(i)}}(t), \Psi_{\hat{Y}_{(i)}}(t) \right) = \sum_{i=1}^{n} \int_0^1 \left(\Psi_{Y_{(i)}}(t) - \Psi_{\hat{Y}_{(i)}}(t) \right)^2 dt \\ \text{s.t.} \quad & a_k, b_k \geq 0,\ k = 1, \cdots, p \end{aligned} \tag{3}$$

The model was applied to a dataset of 31 European countries' GDP between 1995 and 2022. The study focused on analyzing the behavior of the "Import Goods" component in relation to the other four components. The previous findings showed a good fit. The next step involves studying a penalized model and validating it through cross-validation.

References

[1] Sónia Dias and Paula Brito. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(2):75–113, 2015.

[2] Edwin Diday. The symbolic approach in clustering and related methods of data analysis. *Proceedings of IFCS, Classification and Related Methods of Data Analysis, 1987*, pages 673–384, 1988.

[3] Piotr Kokoszka and Matthew Reimherr. *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, 2017.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Analyzing compositional data using distributions defined on the hypersphere

Adelaide Figueiredo[1]
[1] Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC Porto,adelaide@fep.up.pt

In this study we transform the compositional data into directional data using the square-root transformation. For clustering the compositional data, we apply to the directional data the identification of a mixture of distributions on the hypersphere and for the classification of compositional data into predefined groups, we apply Bayes rules to the directional data. We apply these methods for analyzing several sets of compositional data and we compare our results with those obtained using log-ratio transformations.

Keywords: compositional data, directional data, hypersphere

Compositional data are vectors whose components are non-negative values and constrained to a constant sum, for example vectors of proportions that sum one. This type of data arise in many areas, including Agriculture, Economics, Environment, Geology, Medicine and Psychology.
The statistical analysis of compositional data was introduced by Aitchison [1] and has received much attention lately. These data need to be transformed, before applying the standard statistical techniques designed for the Euclidean space. The methods based on the log-ratios of the components are the natural ways of analysing compositional data and this type of transformations has been introduced in the literature for handling compositional data ([1]). These transformations cannot be applied with zero components, without using certain strategies, as for example those suggested in [1].
Alternatively, the square-root transformation can be used to transform compositional data into directional data (unit vectors on the surface of the hypersphere), and then modeled using distributions defined on the hypersphere. This transformation has the advantage of allowing zero components to be analysed.
The statistics of directional data has also developed a lot in the last years (see for example, Mardia and Jupp [4]) and many applications of directional data have arisen recently in the areas of Machine Learning, Text Analysis, Bioinformatics, among others.
In this study for analyzing compositional data, we apply the square-root transformation to the compositional data to obtain directional data. Then, we model the obtained data using a distribution defined on the hypersphere, such as the Watson distribution defined on the hypersphere or the von Mises-Fisher distribution (see [4]).
For clustering compositional data we identify a mixture of the distributions defined on the

hypersphere and for the classification of compositional data into predefined groups, we use the Bayes classification rules based on the distribution defined on the hypersphere.
We apply these methods to the compositional data sets already analysed by Korhonová *et al.* [3], Filzmoser *et al.* [2] and Tsagris *et al.* [5] using other transformations. The results obtained for these data sets with the square-root transformation are very satisfactory and are similar to those obtained with other transformations for handling compositional data. As the data sets analysed do not contain zero components, it would be also interesting to consider data sets with zero components and compare the results obtained with the several transformations.

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the project LA/P/0063/2020.

References

[1] J. Aitchison. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44(2):139–177, 1982.

[2] Hron K. Filzmoser, P. and M. Templ. Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 33(27):585–604, 2012.

[3] Hronb K. Klimcíková D. Muller L. Bednár P. Korhonová, M. and P. Barták. Coffee aroma-statistical analysis of compositional data. *Talanta*, 80:710–715, 2009.

[4] K..V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley and Sons, Chichester, London, 2000.

[5] Preston S. Tsagris, M. and A.T.A. Wood. Improved classification for compositional data using the α-transformation. *Journal of Classification*, 3(1):243–261, 2016.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Survival forests in lifetime analysis

Cecília Castro[1], **Ana Paula Amorim**[2]
[1] Centre of Mathematics, University of Minho, cecilia@math.uminho.pt
[2] Centre of Mathematics, University of Minho, apamorim@math.uminho.pt

The Random Survival Forest method is a non-parametric approach that can be used for modeling time-to-event data in presence of censored data, by using a censoring-specific splitting criterion providing an accurate estimate of the survival function even when not all individuals have a known survival time. This is one of the key strengths of the RSF algorithm, and sets it apart from other machine learning algorithms that are not specifically designed for survival analysis. Unlike Cox proportional hazards regression, Random Survival Forest, RSF, does not assume a particular distribution for the lifetimes or a linear-exponential form for the treatment effects, allowing for a more flexible modeling approach. However, Cox proportional hazards regression is generally considered to be more interpretable compared to RSF, as the latter involves building a large number of decision trees and combining their results to make predictions.
The comparison of RSF and Cox proportional hazards regression can provide insights into the strengths and limitations of each method and inform the choice of the most appropriate method for a specific problem.
This work aims to present the Survival Forests methodology and apply it to public domain data, and compare the results obtained with those resulting from the Cox proportional hazards model, by using performance metrics as prediction error curve, PEC, and the concordance index, c-index.

Keywords: lifetime, log-rank test, performance metrics, random survival forest

Cox proportional hazards regression model is a statistical technique for analyzing time-to-event data. It was developed by David Cox in 1972 [2] and is widely used in various fields. The model is based on the proportional hazards assumption, which states that the hazard ratio (the risk of an event occurring at a given time) between two individuals remains constant over time.
The application of Random Forests to censored lifetime data is a relatively recent development in the field of survival analysis. Random Forests is a machine learning algorithm for classification and regression that was introduced in 2001 by Leo Breiman [1]. In the context of survival analysis, the algorithm Random Survival Forest, RSF, was introduced by Ishwaran and Kogalur in 2008 [3] and it was adapted to handle censored data.
However interpretability is a key consideration in choosing a statistical method for survival analysis, as it affects the ability to understand and communicate the results of the analysis.

In this regard, Cox proportional hazards regression is generally considered to be more interpretable, compared to RSF, providing a straightforward interpretation of the effect of each predictor variable on the hazard of the event of interest. This makes it easy to understand the relationship between each predictor variable and the risk of experiencing the event. RSF, on the other hand, is a more complex and less interpretable method, as it involves building a large number of decision trees and combining their results to make predictions. It can be challenging to understand the relationship between the predictor variables and the event of interest in an RSF model, as the relationship is likely to be more complex and even non-linear.

To compare the survival times of two or more groups of individuals and determine whether the survival times of the groups are significantly different or not, is commonly used the log-rank test. The test works by dividing the total follow-up time into a series of intervals, and counting the number of events (such as death) that occur in each interval. The observed number of events in each interval is then compared to the expected number of events, which is based on the overall event rate across all groups.

In the context of the RSF algorithm, the log-rank test is used as a censoring-specific splitting criterion to determine the best way to split the data into subgroups during the tree-growing process. The goal of the splitting is to create subgroups with as much difference in survival times as possible, as determined by the log-rank test statistic.

In this work we use two commonly used measures to evaluate the performance of a Cox proportional hazards model – prediction error curve and the concordance index.

A prediction error curve, PEC, shows the relationship between the predicted and observed event times. The concordance index, c-index, is a measure of the model's ability to rank individuals by their risk of experiencing the event, such that those who experience the event earlier are ranked higher. A value of 0.5 indicates that the model is not better than random, while a value of 1.0 indicates perfect concordance.

RSF provided lower prediction error compared to Cox proportional hazards regression and also provided higher c-index values. However, one must be caution, and more simulation studies have to be done with several different conditions on covariates, including non-linear relationships and interactions between variables, because the performance of RSF and Cox proportional hazards regression may vary depending on the data.

Acknowledgements This work was supported by Portuguese funds through the CMAT within projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[3] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *Ann. Appl. Statist.*, 2(3):841–860, 2008.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

How much time do we spend on the sofa?

António Balau[1], Fábio Rodrigues[1], Sofia Ribeiro[1], Cristina Lopes[2], Cristina Torres[2] Lurdes Babo[2], Isabel Vieira[2]
[1] ISCAP, Polytechnic University of Porto, csofiamarquesr@gmail.com, j.balau@gmail.com, fabiosereno.rs@gmail.com
[2] CEOS.PP, ISCAP, Polytechnic University of Porto, cristinalopes@iscap.ipp.pt, ctorres@iscap.ipp.pt, lbabo@iscap.ipp.pt, mivieira@iscap.ipp.pt

This study seeks to quantitatively and qualitatively analyse individual use of a common type of furniture present in our homes and life: the sofa (or couch). A questionnaire was developed to gather data, and the statistical analysis addressed the relationship between the use of the sofa and the professional situation of each individual, the time of usage, and the days of greatest use.

Keywords: sofa, leisure time, hypothesis tests, correlation, factor analysis

The sofa assumes a relevant role in terms of rest, leisure and social life, and its importance for the individual changes according to the demographic, social and family nature of each user[1]. The aim of this study is to understand the habits of using the sofa. As research methodology, the first phase was the elaboration of a questionnaire to collect data on the use of the sofa. It contained 21 questions, qualitative, quantitative and in Likert scale.
Data collection was carried out using the Google Forms tool, obtaining 258 responses, of which 256 were considered valid. The next phase consisted of applying the most appropriate quantitative methods to analyse the data, with the statistical software IBM SPSS.
The research hypothesis were:
Hypothesis 1: *There are significant differences in the number of hours per day we spend, on average, on the couch, depending on marital status.*
Hypothesis 2: *There are significant differences in the number of hours per day we spend, on average, on the couch, depending on the professional situation.*
Hypothesis 3: *There are significant differences in the number of hours per day we spend, on average, on the couch, depending on age.*
The majority (96%) of respondents reported to have a sofa. In the sample, 64% were female and 35.6% were male, aged between 16 and 78 years. The majority (57.8%) were married, followed by single people (32.8%). Monday tends to be the day with the least use of the sofa (24 responses) and Sunday is the day of greatest use (205 responses). 86.59% of the respondents use the sofa essentially between 20:00 and 24:00. On average, respondents spend 1.64 hours per day on the couch (1h38m). Through ANOVA and t-tests, it was concluded, with a significance level of 5%, that there are no significant differences between the mean number of daily hours spent on the sofa regarding the Marital Status categories. Therefore, hypothesis 1 is not confirmed.

Hypothesis 2 was validated through 95% confidence intervals and hypothesis tests. It was concluded that retired people and unemployed people spend more hours on the couch than full time active people, which may make sense considering that they have more time available to be on the couch.
Hypothesis 3 was confirmed with regression: $Y_i = 1.530 - 0.266X_{1_i} + 0.014X_{2_i} + \epsilon_i$, where Y is the *hours spent on the sofa*, X_1 is *how many people share the sofa with*, X_2 is *age* and ϵ the residuals. It was concluded that, for the same age, when the number of people sharing the sofa increases by one, the time spent on the sofa decreases by approximately 16 minutes (p=0.001) and that, for a person who is 1 year older sharing the sofa with the same number of people, the time spent on the couch increases by approx. 1 minute (p=0.022).
The reliability of the questions about the uses of the sofa was measured by Cronbach's alpha ($\alpha = 0.675$). A factor analysis was performed to form 2 groups with strongly related variables. After orthogonal rotation (Fig.1), Factor 1 represented the *recreational uses* (using the cell phone, watching TV, sleeping and resting) and Factor 2 represented the *biological and intellectual uses* (reading and eating).

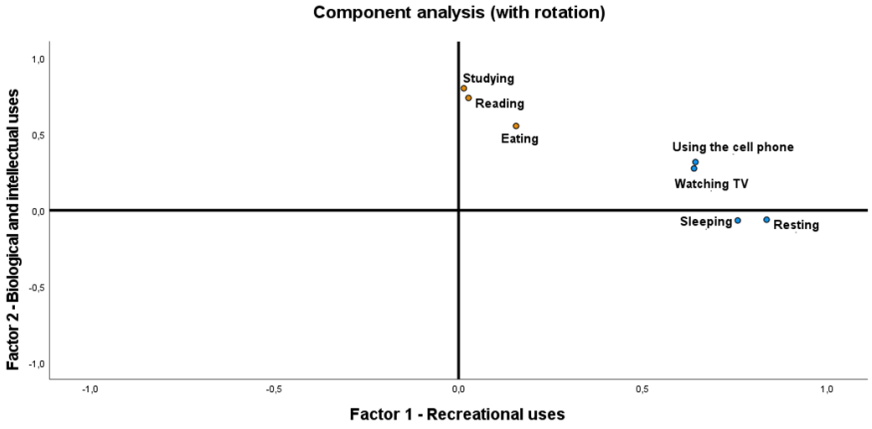

Figure 1: Factor analysis of the purposes for which the sofa is used

We can conclude that the heterogeneity is large and that people use their sofa for various purposes; mostly to watch TV, use the cell phone and rest; on the contrary, fewer people use the sofa to play, eat and study. The study allowed to understand that the time spent on the couch differs according to job situation, age, and the number of people that the sofa is shared with.
Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] Charles N. Wagner. *Glued to the Sofa: Exploring Guilt and Television Binge-Watching Behaviors*. LAP Lambert Academic Publishing, 2016.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Identification of potential causes in the number of beneficiaries of social disability pension in small municipalities in the northern region of Portugal

Cristina Torres[1], **Lurdes Babo**[1], **Isabel Vieira**[1], **Isabel Cristina Lopes**[1], **Rui Monteiro**[2], **Carla Ferreira**[2], **Inês Bem-Haja**[2]
[1] CEOS.PP, ISCAP, Polytechnic University of Porto, ctorres@iscap.ipp.pt, lbabo@iscap.ipp.pt, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt
[2] ISCAP, Polytechnic University of Porto, ruimanuelmonteiro@gmail.com, 2939577@iscap.ipp.pt, inesbemhaja@gmail.com

The relationship between the number of road accidents with casualties, the level of economic activity, and the number of beneficiaries of the social disability pension is studied using cluster analysis, applied to a sample of 46 municipalities in the northern region of Portugal. The results allow the grouping of municipalities into two clusters using the between-groups linkage method with squared Euclidean distance. Cluster 1 municipalities are predominantly located in the region's hinterland and have a lower number of companies, road accidents with casualties and disability pensioners compared to Cluster 2 municipalities. This study aims to give a contribution to local governments' decision-making process in order to enhance the economic activity, reduce the region's human desertification, and improve municipal road conditions.

Keywords: clusters analysis, hierarchical method, social disability pension, small municipalities, road accidents

The desertification of the hinterland of Portugal increased in the last decade. Living in the interior of the country presents many challenges, one of which is the lack of jobs in quantity and quality. Also, depending on where people live, the absence or reduced frequency of transport is a problem. There is a need to have your own vehicle for almost all activities that require travel, since the low population density has a negative impact on the availability of public transport. On the other hand, the more remote the location is, the worse the road networks tend to be. This work focuses essentially on the real problem of the beneficiaries of social disability pensions, that are unable to benefit from their work earnings, in result of accidents or even illness, and so they are also deprived of better living conditions. In the analysis of the case under study, the data was retrieved from the Statistical Yearbook of the Northern Region 2018 [1]. The goal was to analyze only the small municipalities (with a population of less than 20,000 inhabitants), which resulted in a total of 46 municipalities. In this study, multivariate analysis (cluster analysis) was used in

order to characterize the small municipalities of the northern region of Portugal regarding the number of existing companies (as a proxy of the economic activity), the number of road accidents with casualties and the number of beneficiaries of social disability pension. Combining the different variables through the various agglomerative clustering methods, it was concluded that the between-groups linkage method with squared Euclidean distance was the one that obtained a dendrogram that better distinguished the formed clusters. Two clusters were generated, and the geographic location of each cluster can be seen in Fig.1 (Cluster 1 in yellow and Cluster 2 in green).

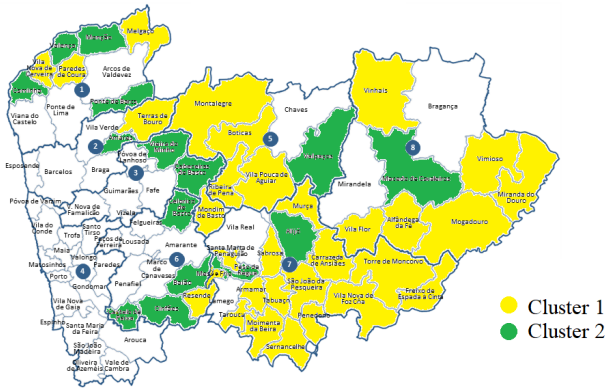

Figure 1: Cluster 1 and Cluster 2 Municipalities.

It was found that the municipalities that are part of Cluster 1 comprises the innermost municipalities, with few road accidents with casualties, not many disability pensioners and with a small number of companies when compared with Cluster 2. This study allowed to notice that for the municipalities in Cluster 1 it is important to boost the economic activity by increasing the number of companies and, consequently, augmenting job opportunities. On the other hand, regarding Cluster 2, although the number of companies is higher, in order to reduce the number of beneficiaries due to disability, it is important to improve road conditions and thus cut down road accidents. This study points out some indicators that local governments should monitor to improve decision-making.

Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] Instituto Nacional de Estatística: Anuário Estatístico da Região Norte - 2018. `https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=410498016&PUBLICACOESmodo=2`, 2019. Accessed: 2022-01-10.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Evolution of mean sea level: particular case of the port of Viana de Castelo

Dora Carinhas[1], **Miguel Picoto**[2], **Paulo Infante**[3]
[1] Instituto Hidrográfico; IIFA/Universidade de Évora, dora.carinhas@hidrografico.pt
[2] Marinha Portuguesa, goncalves.picoto@marinha.pt
[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt

---

Climate change and its consequences are one of the dominant themes today. One of the most obvious consequences of global warming is the tendency for the rise of the sea. In this work, it is intended to extend studies already carried out at the Hydrographic Institute on the mean sea level rise using time series analysis to tide records from the port of Viana do Castelo which shows a rising trend with no evidence of any acceleration.

Keywords: time series models, mean sea level, tide gauge records, seasonality, sea level trend

---

The rising trend of the mean sea level (MSL) has been studied for several decades at a global level, with studies identifying upward trends in the MSL oscillating between 1.5-3.2 mm/year during the 20th and early 19th centuries. In recent years we have heard about floods, landslides, as well as increased levels of coastal erosion. There is currently great concern about the rise of the MSL because this phenomenon has, on a global scale, a great impact on humanity, namely on economic and social activities [3]. This is particularly true for Portugal; in 2008, the distribution of the population living in coastal regions compared to the national population was 83% [4].

Sea level rise is primarily caused by two factors associated with global warming: water added by melting polar ice caps and the expansion of sea water as it warms [2]. This work addresses sea level change from tide gauge data instaled in Viana do Castelo. The long serie of data allowed us to determine that the mean sea level is rising, in Viana do Castelo, the increase was 123 millimeters in 30 years of analysis, which corresponds to a trend of 2.86 ± 0.89 mm/ year [1]. The tendency for the average level to rise was deduced through linear regression and the result was further compared with that obtained through the autoregressive neural network model. The characteristic monthly average levels were also calculated, and the existence of seasonality throughout the year was verified (Figure 1).

Significant annual and semi-annual variations in mean sea level have been observed due to changes in atmospheric pressure, water density and ocean circulation. In summary, during the summer months variations in water density tend to predominate and in the winter months variations of meteorological origin [5].

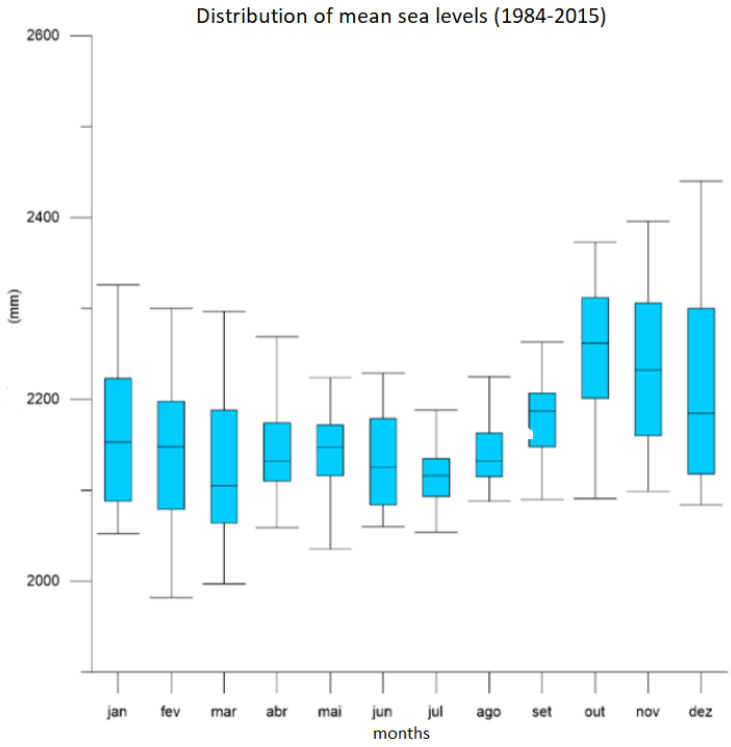

Figure 1: Behavior of mean sea levels throughout the year in Viana do Castelo(1984 to 2015).

References

[1] V. Mendes S. Barbosa D. Carinhas. Vertical land motion in the iberian atlantic coast and its implications for sea level change evaluation. *Journal of Applied Geodesy*, 14:361–378, 2020.

[2] N.J. White A. Church. A 20th century acceleration in global sea-level rise. *JGeophysical Research Letters*, 33, 2006.

[3] A. Cazenave G. Le Cozannet. Sea level rise and its coastal impacts. *Earth's Future 2*, pages 15–34, 2013.

[4] EUROSTAT. *Eurostat regional yearbook 2011.* Publications Office of the European Union, European Union, 2011.

[5] IOC. *Manual de Medição e Interpretação do Nível Médio do Mar.* Manuais e Guias No. 14, Reino Unido, 1985.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Statistical analysis of humanoids' arm movements

Eliana Costa e Silva[1], **Gianpaolo Gulletta**[2], **Estela Bicho**[2], **Wolfram Erlhagen**[3]

[1] CIICESI, ESTG, Politécnico do Porto, Portugaleos@estg.ipp.pt
[2] Centre Algoritmi/Dept. of Industrial Electronics, University of Minho, Portugal d6468@dei.uminho.pt, estela.bicho@dei.uminho.pt
[3] Centre of Mathematics/Dept. of Mathematics and Applications, University of Minho, Portugal wolfram.erlhagen@math.uminho.pt

---

In the paradigm of "Industry 5.0", contributing to more human-centric and resilient industries, human-like motion is a key feature for intuitive human–robot interactions. Here, simulated reaching movements of a humanoid robot are analyzed in order to evaluate its accordance with a well-known regularity of human arms' movements – the *one-sixth power law* (1/6-PL). Specifically, non-linear regression models are considered for obtaining the slope in the log-space of these movements. Results suggest that the movements reflect the 1/6-PL.

Keywords: non-linear regressions, ANOVA, Kruskal-Wallis test, humanoid robots, reaching movements

---

The *one-sixth power law* (1/6-PL) is a regularity between the velocity and the geometry of a hand path found in studies on human arm motor control. This law was introduced by Pollick et al. in [3] and states that the tangential velocity of the hand, v, is inversely related to the one-sixth power of the square of the curvature, k, multiplied by the torsion, τ, i.e. $v = \alpha \left(k^2|\tau|\right)^{-1/6}$.

Here, 600 unconstrained reaching movements of a humanoid robot, generated using the Human-like Upper-limb Motion Planner (HUMP), proposed in [1], are statistically analyzed. Specifically, the kinematics of a robotic hand is studied to identify the similarities with the 1/6-PL, demonstrating that the HUMP algorithm is capable of consistently planning human-like hand trajectories that obey the 1/6-PL. Assuming that the exponent of the 1/6-PL is unknown, the constant α and the exponential γ are obtained by applying the logarithm transformation and the approximately least-squares regression values can be found by linear regression in the log-space: $\log(v) = \log(\alpha) - \gamma \log(k^2|\tau|)$. This expression expresses a proportional relationship between the logarithm of the tangential hand velocity, $\log(v)$, and the logarithm of the square curvature multiplied by the torsion, $\log(k^2|\tau|)$. This equation also constrains the relation between curvature and torsion of the movements. Six sessions of 100 movements each, were generated by the HUMP planner, with the position of the target's hand randomly selected from a $50\,cm \times 60\,cm \times 60\,cm$ paralelepipedal in front of the robot. From these, the points with curvature, k, very close to zero were

excluded to avoid numerical problems in the calculus of the logarithm, and points with absolute torsion, τ, less than 2 m^{-1} were excluded in order to avoid unreal cups raised when the torsion changes sign [3]. The linear regression models were validated and residual analysis was performed, for a significant level of $\alpha = 1\%$, using RStudio (version 1.4.1106) and R Statistical Software (version 4.0.5) [4]. A total of 367 movements (56 and 57 for sessions 3 and 5; 63 for sessions 4 and 6; and 64 for sessions 1 and 2) yield valid models, therefore only these movements were analyzed.
Coefficients of determination, R^2, between 0.8713 and 0.9939 were obtained. This expresses the high explainability of the obtained models. In fact, 87.13% to 99.39% of the variability of the logarithm of the hand velocity, $\log(v)$, is explained by $\log(k^2|\tau|)$.
For both the parametric ANOVA test (p-value=0.131) and the non-parametric Kruskal–Wallis test (p-value= 0.160), no significant differences between the slopes of the regressions estimated for the six sessions were found. A similar analysis was performed to infer if there were significant differences between the bias for all planned movements. No significant differences ($\alpha = 1\%$) were found for the mean (p-value=0.202$>$ 0.01) and the median (p-value=0.288$>$ 0.01). Thus, the 367 movements of the six sessions were analyzed together. The estimated slopes, i.e. γ, are all negative and mostly concentrated between -0.21 and -0.195. The maximum slope was approximately $-0.1788 < -1/6$, only three of the regressions present $\gamma \geq -0.18$ and 1% have slopes ≥ -0.1839. Further, the average values of the slope, γ, were approximately -0.202. Note that, the 99% confidence interval was $[-2.02, -0.201]$, which does not include -1/6. However, similar results were also observed in human experiments. Specifically, in [3] the slopes range between -0.21 to -0.15 for different human subjects.

Acknowledgements Eliana Costa e Silva has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through project UIDB/04728/2020. This work has been partially supported by project "I-CATER - Intelligent Robotic Coworker Assistant for Industrial Tasks with an Ergonomics Rationale" (Ref PTDC/EEI-ROB/3488/2021), financed by FCT with national funds through the state budget.

References

[1] G. Gulletta, E. Costa e Silva, W. Erlhagen, R. Meulenbroek, M. F. Costa, and E. Bicho. A human-like upper-limb motion planner: Generating naturalistic movements for humanoid robots. *International Journal of Advanced Robotic Systems*, 18, 2021.

[2] J. Maindonald and J. Braun. *Data Analysis and Graphics Using R: An Example-Based Approach*. Cambridge University Press, 3rd edition, 2010.

[3] F. E. Pollick, U. Maoz, A. A. Handzel, P. J. Giblin, G. Sapiro, and T. Flash. Three-dimensional arm movements at constant equi-affine speed. *Cortex*, 45:325–339, 2009.

[4] R Core Team. *Computational Many-Particle Physics*, volume 739 of *Lecture Notes in Physics*. Vienna, Austria, 2017.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Application of data reduction methods in the creation of SoResilere – Social resilience index applied to flood affected municipalities

Rita Jacinto[1], **Fernando Sebastião**[2], **João Ferrão**[3], **Eusébio Reis**[4]
[1] Centro de Estudos Geográficos, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa and Laboratório Associado Terra, Lisboa, Portugal, ritajacinto@edu.ulisboa.pt
[2] LSRE-LCM - Laboratory of Separation and Reaction Engineering – Laboratory of Catalysis and Materials, Polytechnic of Leiria and ALiCE - Associate Laboratory in Chemical Engineering, Faculty of Engineering, University of Porto and School of Technology and Management, Polytechnic of Leiria, Portugal, fsebast@ipleiria.pt
[3] Instituto de Ciências Sociais, Universidade de Lisboa (ICS-UL), Lisboa, Portugal, joao.ferrao@ics.ulisboa.pt
[4] Centro de Estudos Geográficos, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa and Laboratório Associado Terra, Lisboa, Portugal, eusebioreis@edu.ulisboa.pt

---

SoResilere is the first Social Resilience index applied to Portuguese flood disaster affected municipalities. Official statistics were not sufficient for the data needs and therefore categorical data were created. Principal Component Analysis and Categorical Principal Component Analysis were applied to the quantitative and categorical datasets respectively, which resulted into two subindexes. SoResilere index is the combination of the subindexes. SoResilere spatial distribution was mapped with and without component weighting.

Keywords: social resilience, resilience index, floods, municipalities resilience assessment, principal component analysis

---

Climate change effects on the frequency and intensity of climate-related events, such as floods, which have been very destructive, are expected to worsen in Portugal [3]. Social resilence is an emergent scientific field, wich aims to support disaster risk governance. A relevant percentage of assessments (40%) are qualitative, as referred by studies which revised almost two hundred articles on this topic [1]. There are several studies on the social aspects of floods, but few have been developed at municipality level. In Portugal, SoResilere, the first flood social resilience index, was built and applied to portuguese flood disaster affected municipalities. Figure 1 presents SoResilere index methodology as well as the workflow to its creation. Based on Jacinto *et al.* (2020) [2], six flood resilience dimensions were considered: (1) Individuals, (2) Society, (3) Governance, (4) Built Environment, (5) Natural Environment, and (6) Disaster. Two datasets were obtained and

data dimension reduction methodologies such as Principal Component Analysis (PCA) and Categorical Principal Component Analysis (CATPCA) were applied, resulting into two subindexes.

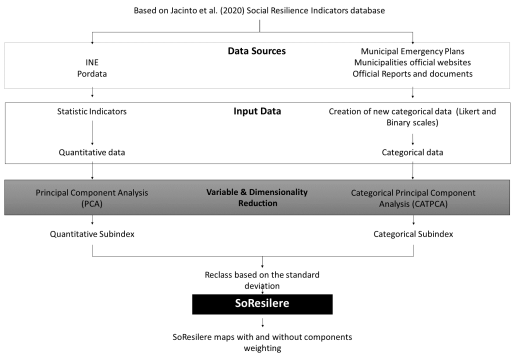

Figure 1: SoResilere methodological scheme

The results with PCA involved six retained principal components, explaining 67.197% of the total variance and the results with CATPCA included five retained principal components, explaining 78.975% of the total variance. In order to better explore the SoResilere results, the subindexes were calculated and mapped with and without component weighting. Component weighting did not affect the results of 55.5% of the municipalities. Although SoResilere status improved in 22.22% of the case studies with component weighting, no evidence was found on component weighting benefits since there were no spatial associations between the status changes and the case studies.

Acknowledgements This research was funded by the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology), I.P. [grant number SFRH/BD/119228/2016 to Rita Jacinto; UIDB/00295/ 2020 + UIDP/00295/2020]. Eusébio Reis was funded by the Portuguese Foundation for Science and Technology (FCT I.P.), under the MIT-Portugal project "Multi-risk Interactions Towards Resilient and Sustainable Cities (MIT-RSC)" (MIT-EXPL/CS/0018/2019.

References

[1] H. Cai, N.S.N. Lam, Y. Qiang, L. Zou, R.M. Correll, and V. Mihunov. A synthesis of disaster resilience measurement methods and indices. *International Journal of Disaster Risk Reduction*, 31:844–855, 2018.

[2] R. Jacinto, E. Reis, and J. Ferrão. Indicators for the assessment of social resilience in flood-affected communities - a text mining-based methodology. *Science of The Total Environment*, 744:140973, 2020.

[3] I. Kotzee and B. Reyers. Piloting a social-ecological index for measuring flood resilience: A composite index approach. *Ecological Indicators*, 60:45–53, 2016.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Alto Minho regional performance through SDG11: a cluster analysis

Helena Sofia Rodrigues[1], **Ângela Silva**[2], **Jorge Esparteiro Garcia**[3]
[1] Instituto Politécnico de Viana do Castelo and CIDMA - Center for Research & Development in Mathematics and Applications, sofiarodrigues@esce.ipvc.pt
[2] Instituto Politécnico de Viana do Castelo and ADiT-LAB, angela.a@esce.ipvc.pt
[3] Instituto Politécnico de Viana do Castelo and ADiT-LAB and INESC TEC, jorgegarcia@esce.ipvc.pt

---

The Sustainable Development Goals (SDG) for 2030 is a project from United Nations, for all countries, regions and institutions, for peace and prosperity for people and the planet. The focus of this study is SDG11 – Sustainable Cities and Communities – and it is centered on the municipalities in the region of Alto Minho. Based on the similarities and differences between municipalities, a performance analysis was carried out, as a tool to inform how much effort is need to achieve the goal.

Keywords: SDG11, Alto Minho, regional performance, cluster analysis

---

Sustainable Development Goal 11 (SDG11), also known as "Sustainable Cities and Communities" aims to make cities and human settlements inclusive, safe, resilient, and sustainable. This goal is important because cities are where most of the world's population lives and will continue to grow, with over two-thirds of the global population expected to live in urban areas by 2050. This means that cities have a significant impact on people's quality of life and the environment.

Inclusive cities provide access to basic services and opportunities to all, regardless of income, race, or other factors. Safe cities reduce crime and violence and promote peace and security. Resilient cities are better able to withstand and recover from natural disasters and other shocks, such as pandemics. Sustainable cities reduce greenhouse gas emissions, promote sustainable transport and urban planning, and protect the natural environment.

Achieving SDG11 will require collaboration between different levels of government, the private sector, and civil society. This includes improving urban planning and management, promoting green and public spaces, and improving access to affordable housing, transport, and basic services. By making cities more sustainable, we can improve the lifes of billions of people and reduce the environmental impact of urbanization.

Through the database provided by ODSLocal platform [**?**], there were selected 9 indicators for SDG11, for the year of 2020. The choice of these indicators is related to the availability of data, discretized until the municipality level.

Due to have different indicators unity to measurement, some data is in different orders of magnitude. Example given, the Municipalities' expenditure on biodiversity and landscape protection per inhabitant has a scale associated with €/hab. Therefore it was necessary to normalize the data, becoming each indicator score between 0 and 100 [**?**].
A cluster analysis was carried out, using the Hierarchical Cluster Analysis method [**?**]. This is a general approach to cluster analysis, with the main goal to group together objects (in this case, regions) that are close to another in term of indicators. In this case, it was selected the single linkage, where the distance between two clusters is the minimum distance between members of the two clusters. The results were obtained through IBM SPSS v27.

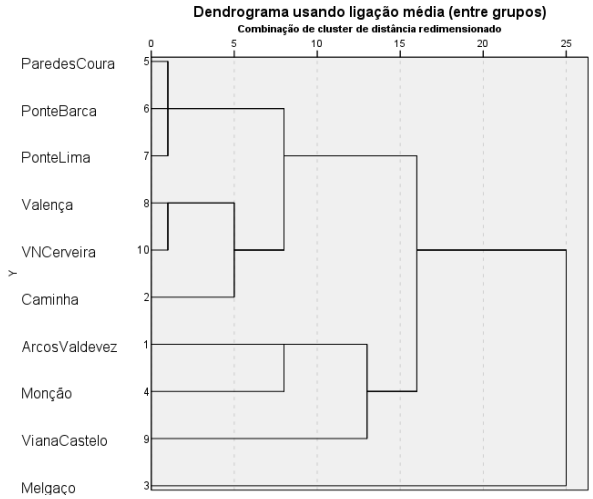

Figure 1: Dendogram of municipalities

Figure 1 shows the dendogram of the municipalities, using the data from SDG11. Melgaço is a municipality that stands out, because it is the region that is near to achieve the majority of the 9 indicators expected to 2030, while Paredes de Coura, Ponte da Barca e Ponte de Lima are the ones that have to make additional efforts to trace the path.

Acknowledgements This work is supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020 (Rodrigues).

References

[1] J.P. Cling, S. Eghbal-Teherani, M. Orzoni, and C. Plateau. The interlinkages between the sdg indicators and the differentiation between eu countries: It is (mainly) the economy. *Statistical Journal of the IAOS*, 36:455–470, 2020.

[2] ODSLocal. `https://odslocal.pt` (last accessed january 2023).

[3] C. Zhang, Z. Sun, Q. Xing, J. Sun, T. Xia, and H. Yu. Localizing indicators of SDG11 for an integrated assessment of urban sustainability—a case study of Hainan Province. *Sustainability*, 13:11092, 2021.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Gender equality in wages in Portugal, between 1994 and 2020

Ana Freitas[1], Elenice Santos[1], Marileide Silva[1], Vanessa Lima[1], Isabel Vieira[2], Cristina Lopes[2], Cristina Torres[2], Lurdes Babo[2]
[1] Instituto Federal de Educação, Ciência e Tecnologia do Mato Grosso, tais.goes@ifmt.edu.br, elenice.santos@ifmt.edu.br, marileide.silva@ifmt.edu.br, vanessa.alves@ifmt.edu.br
[2] CEOS.PP, ISCAP, Polytechnic of Porto, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt, ctorres@iscap.ipp.pt, lbabo@iscap.ipp.pt

The present study aims to analyse gender equality in wages in Portugal using data available in Pordata, for the period 1994-2020. The idea is to verify whether the average weekly working time is decisive for defining the value of the salary, if there is a correlation between the value of the base salary and the number of people with a higher education degree and, finally, whether more men and women with a higher education degree influences their income. Various statistics technics were performed using IBM SPSS Statistics, version 28.0. The results obtained enable to state that the world of work in Portugal is unfair to women with higher education.

Keywords: gender equality, wages, data analysis, higher education

After decades of work and dozens of equal payment laws, in many professions and positions where men and women do the same job, women still continue to earn less than men. Despite considerable progress in combating gender inequality, with support from various initiatives, in practice the data are still unfavorable, considering the average earnings of women compared to men. At the United Nations World Economic Forum held in 2020, the gender pay gap was estimated to reach 23%. Was also revealed that, considering the current situation, it would only be possible to have equal pay between genders in 257 years [3]. In an attempt to reverse this, the UN inserts Gender equality among the 17 Sustainable Developement Goals: "Goal 5 - Achieve gender equality and empower all women and girls" [3]. The European Union also dealt with equality of remuneration in the Treaty of Rome, which came into effect in 1958 [2]. The European Institute for Gender Equality was created in 2010 with the aim of strengthening the promotion of gender equality in the EU, as a tool for making policies and measuring the progress through the Gender Equality Index. This measurement is performed every year with a score of 1 to 100 for all Member States, where a score of 100 means that equality between men and women has been achieved [1]. According to this index, Portugal currently occupies the 15th position, with a score of 62.8 out of 100. With this score, Portugal is 5.8 points below the EU average.

The data for this study was collected in the Pordata database [4], for the period between 1994 and 2020. The nine variables used are: Higher Education Graduates, Number of people enrolled in higher education, Monthly Average Base Remuneration, Average Monthly Earning, Male Ratio of the Employed Population, The number of men per 100 women, Employed Population, Higher Education Level Unemployed Population, Average Weekly Duration of Work, Active Population.
The results of the exploratory descriptive analysis indicate that, in average, for each graduated man, there are 1.6 women who are also graduates of higher education, which means that in Portugal there are more women completing higher education than men. A similar proportion also occurs among those enrolled in higher education for every man enrolled, there are 1.2 women also enrolled in higher education. Although women are better qualified, the average monthly base salary, i.e. the amount that men are entitled to receive for a month of work is 22.55% higher than that of women. The difference is even greater when analysing the average earnings, that is, the amount paid by the employer to the workers, adding overtime, vacations, bonuses, among others.
Through ANOVA and t-tests, it was concluded, with a significance level of 1%, that there are significant differences between the Average Base Remuneration, Average Earning and Average Weekly Duration of Work for men and women. Discriminant analysis was also used to understand the effect that these variables had on the difference between genders, with all variables having proved to be statistically significant ($p < 0.001$) to discriminate between genders over the years under study.
This study confirmed that, in Portugal, between 1994 and 2020, there are more women with higher education degrees than men, but their wages are much lower.

Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] European institute for gender equality. `https://eige.europa.eu/pt/in-brief`. Accessed: 2022-12-18.

[2] European union, aims and values. `https://european-union.europa.eu/principles-countries-history/principles-and-values/aims-and-values_en`. Accessed: 2022-12-19.

[3] Organização das Nações Unidas - ONU: A agenda 2030 - objetivos de desenvolvimento sustentável. `https://unric.org/pt/Objetivos-de-Desenvolvimento-Sustentavel`. Accessed: 2022-12-19.

[4] PORDATA - Estatísticas sobre Portugal e Europa. `https://www.pordata.pt/Portugal`. Accessed: 2023-01-02.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Characterization of mobbing in Portuguese accounting professionals using Leymann inventory of psychological terror scale items

Irene Oliveira[1], **António Dias**[2], **Margarida Simões**[3], **Ana Paula Monteiro**[4], **Rui Silva**[5]
[1] Department of Mathematics, University of Trás-os-Montes e Alto Douro and CEMAT - Center for Computational and Stochastic Mathematics, ioliveir@utad.pt
[2] Department of Economy, Sociology and Management, University of Trás-os-Montes e Alto Douro, acgdias@utad.pt
[3] Department of Education and Psychology, University of Trás-os-Montes e Alto Douro, margaridas@utad.pt
[4] Department of Education and Psychology, University of Trás-os-Montes e Alto Douro, apmonteiro@utad.pt
[5] Department of Economy, Sociology and Management, University of Trás-os-Montes e Alto Douro, rui.silva@utad.pt

We examined the phenomenon of mobbing among Portuguese Accounting Professionals, by measuring mobbing behaviour using items from the Leymann Inventory of the Psychological Terror Scale (LIPT). In this study, nonparametric tests were used to assess the hypothesis of association between LIPT scale items and socio-demographic characteristics such as Gender, Years in the profession, Academic Degree and Institutional Bound. This characterization allowed us to verify that, in general, professional accountants do not experience mobbing in their workplace, however, some items were identified as of greatest concern.

Keywords: mobbing, LIPT scale, accounting professionals

The term mobbing was initially introduced in Sweden by Leymann, [1], and it means the act of harassing or psychologically terrorizing other people in the workplace. Leymann Inventory of Psychological Terror (LIPT) Scale was validated for Portuguese Accounting Professionals in 2021, [2]. The LIPT scale has five main dimensions, namely: 11 items of Self Expression Effects (SEE); 7 items Occupational Situation Effects Quality of Life (OSEQL); 5 items of Self Contacts Effects (SCE); 15 Social Reputation Effects (SRE)items; and 7 Health Effects (HE) items. A sample of 419 Portuguese Accounting Professionals was studied, and Chi-square tests were performed, to determine which scale LIPT items were

associated with Gender (G), Accounting years (Y), Academic degree (D) and Institutional Bound (I). Mann-Whitney U and Kruskal-Wallis tests, with Dunn's multiple comparisons, were applied to compare medians between two or more than two independent groups, respectively. The significance threshold was set at 0.05.

The results from significant nonparametric tests showed that the following items were of main concern and were associated with the demographic characteristics (G, Y, D, I):
Your superior restricts the opportunity for you to express yourself (SEE1-I); You are constantly interrupted (SEE2-I); Colleagues restrict your opportunity to express yourself (SEE3-Y); Written threats are sent (SEE9-I); Contact is denied through looks or gestures (SEE10-I,Y); Contact is denied through innuendo (SEE11-I); People do not speak with you anymore (SCE1-I,G); You cannot talk to anyone; Access to others is denied (SCE2-I,G); You are treated as if you are invisible (SCE5-I); People talk badly about you behind your back (SRE1-I); Unfounded rumors about you are circulated (SRE2-I); You are ridiculed (SRE3-I); Supervisors take away assignments so that you cannot create new tasks to do (OSEQL2-I); You are given meaningless jobs to carry out (OSEQL3-D,I); You are given jobs that are below your qualifications (OSEQL4-D,I); You are continually given new tasks (OSEQL5-D,I); Threats of physical violence are made (HE2-I); Damaging your workplace or home (HE6-D).

Although mobbing in accountancy profession is scarce, mobbing behaviors at workplace need attention. The assessment of the various factors of mobbing in the workplace allowed to verify which are the main concerns of professionals for a future reflection about mobbing in the workplace and how that influences the psychological climate and social life of workers. In addition, this work provided global information and allowed to perceive intensity of the set of behaviors and strategies of mobbing suffered by individuals.

Acknowledgements The work is supported by national funds, through the FCT Portuguese Foundation for Science and Technology under the projects UIDB/04011/2020, UIDB/04630/2020 and UID/MULTI/04621/2019.

References

[1] H. Leymann. Mobbing and psychological terror at workplaces. *Violence and Victims*, 5:119–126, 1990.

[2] R. Silva, M. Simões, A.P. Monteiro, and A. Dias. Leymann inventory of psychological terror scale: Development and validation for portuguese accounting professionals. *Economies*, 9(3):1–15, 2021.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Birth rate in Portugal

Maria Guerra[1], **Ana Pessoa**[1], **Ana Rodrigues**[1], **Maria Mesquita**[1], **Lurdes Babo**[2], **Isabel Vieira**[2], **Cristina Lopes**[2], **Cristina Torres**[2]
[1] ISCAP, Polytechnic of Porto, 2220284@iscap.ipp.pt, 2220262@iscap.ipp.pt, 2220263@iscap.ipp.pt, 2221759@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Polytechnic of Porto, lbabo@iscap.ipp.pt, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt, ctorres@iscap.ipp.pt

This paper presents a study on the birth rate in Portugal. The quantitative data analysis of 308 municipalities concludes that the variables "live births of mothers residing in Portugal with Portuguese and foreign nationality", "fertility rate of the various age groups", "live births of mothers residing in Portugal according to their level of education" and "births by gender" are strongly correlated, presenting common factors that significantly influence the birth rate.

Keywords: birth rate, correlation, factor analysis

A study carried out by the Francisco Manuel dos Santos Foundation [3] concluded that around 25% of individuals expect to have only one child. Those with a higher level of education, generally become parents later. However, women with higher levels of education show a greater intention to have a larger number of children, despite actually having fewer. The Portuguese population has been decreasing, and since 2011, Portugal has lost around 196,000 inhabitants. Exceptions to this trend occurred only in the years 2019 and 2020. In 2019, the Portuguese population increased compared to 2018 by about 19,300 inhabitants and 75,700 inhabitants when the years 2020 and 2019 are compared. However, this growth is not due to the increase in the birth rate among the Portuguese population, but mainly, to the positive migration balance [2].
The data in this work were collected from the PORDATA database [1] considering 308 municipalities in Portugal. The study considered 17 variables related to live births of mothers residing in Portugal, with Portuguese and foreign nationality and according to their level of education; fertility rate for various age groups and, births by gender.
ANOVA and t-tests allowed to conclude, with a significance level of 1%, that there are significant differences between the average live births of Portuguese and foreign mothers living in Portugal. Obviously, the average number of births of Portuguese mothers is higher than that of foreign mothers. Furthermore, significant differences were observed between the average fertility rate by age, with the highest value in the 30-34 age group. Concerning the level of mothers' education, significant differences were found where mothers with higher education exhibiting a high average score. Regarding the number of births by gender, the study concluded that, on average, more boys were born than girls and this

difference is statistically significant (p-value < 0.001).
A factor analysis with VARIMAX rotation was performed producing 4 factors. The scree plot (Figure 1) and the Kaiser's criterion were used to determine the number of factors to be retained. These factors represented the mother's education; the fertility rate at

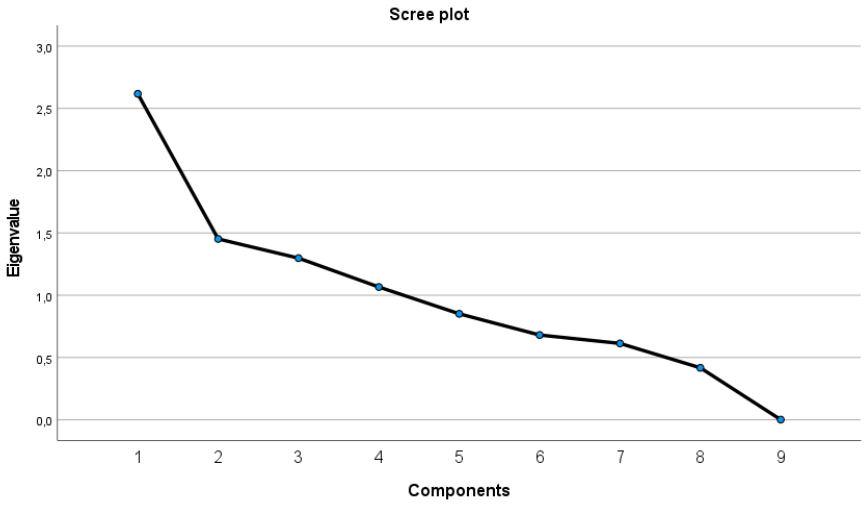

Figure 1: Factor analysis scree plot

extreme ages; the fertility rate in the 25-29 age group and the fertility rate in the 20-24 age group. However, it should be highlighted that there are factors with a very small number of variables, namely the 3rd and the 4th factors.

Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] Pordata - estatísticas sobre portugal e europa. `https://www.pordata.pt/Portugal`. Accessed: 2022-12-01.

[2] Carlos Neves. Portugal vive crise demográfica, só atenuada pelos imigrantes. `https://www.dn.pt/sociedade/pais-vive-crise-demografica-so-atenuada-pelos-imigrantes-15505591.html`, 2022. Accessed: 2022-12-27.

[3] Maria João Rosa and Paulo Ferreira. Cada vez menos? `https://www.ffms.pt/pt-pt/estudos/cada-vez-menos`, 2021. Accessed: 2022-12-10.

21 April, 10:40 - 11:00, Hall of Library Barbosa Romero

Assessment of exhaustion, cognitive weariness and physical fatigue of security services workers: PLSc-SEM approach

Luís M. Grilo[1,2,4,5,6], **Tiago F. Braz**[3], **Helena L. Grilo**[2,3]
[1] Universidade de Évora, Portugal, luis.grilo@uevora.pt
[2] Universidade Aberta, Portugal
[3] Instituto Politécnico de Tomar (IPT), Portugal
[4] NOVA Math (Centro de Matemática e Aplicações), FCT NOVA, Universidade Nova de Lisboa, Portugal
[5] Ci2 (Centro de Investigação em Cidades Inteligentes), IPT, Portugal
[6] CIICESI, ESTG, P.Porto, Portugal

To assess the facets of burnout to which security services workers are exposed, a survey was carried out in a Portuguese public institution using the Shirom Melamed Burnout Measure (SMBM). A reflective theoretical Structural Equation Model (SEM) was proposed, and the consistent Partial Least Squares (PLSc) estimator was applied to mimic common factor model results. The exogenous latent construct 'emotional exhaustion' has a direct positive effect on constructs 'cognitive weariness' and 'physical fatigue', but it also has an indirect effect on the last through the mediator construct 'cognitive weariness'. The estimated path model shows high predictive quality of the human perceptions, after applying the PLSpredict technique.

Keywords: latent variables, ordinal manifest variables, psychosocial risks, survey, well-being

In a Portuguese public institution, the security service workers are exposed to a lot of human contact and experience stressful moments during their workday. This situation can lead to 'mental exhaustion', which is the core component of burnout, and it is related with 'cognitive weariness' and 'physical fatigue'. A survey was conducted using a Portuguese version of the internationally validated Shirom Melamed Burnout Measure (SMBM) [1, 2], with 14 variables expressed in an ordinal scale of seven categories. These observed/manifest variables operationalize the burnout multidimensional construct, which consists of the 'emotional exhaustion', 'cognitive weariness' and 'physical fatigue'. We proposed a hypothetical structural path model, consistent with the specialized literature and expressing a priori perceptions about the causal relationships between those latent constructs, where 'physical fatigue' is the target latent variable. A sample (primary data with no missing values) of 115 workers was obtained and to estimate the reflective (i.e.,

with relationships from the latent constructs to the manifest variables) Structural Equation Model (SEM) we used the consistent Partial Least Squares (PLSc) estimator, which applies a correction for attenuation to consistently estimate SEM with common factors [4, 3]. The nonparametric PLSc-SEM method maximizes the explained variance of endogenous constructs, does not consider distribution assumptions and works well with small sample sizes. The main results reveal that the latent constructs 'emotional exhaustion' and 'cognitive weariness' have statistical significant effects on 'physical fatigue'. From the application of the novel PLSpredict technique, we conclude that the model has high predictive power, since none of the manifest variables of the target latent construct 'physical fatigue' in the PLSc-SEM analysis yields higher prediction errors compared to the naïve LM (linear model) benchmark, in terms of RMSE (root mean squared error) and MAE (mean absolute error). The frequent monitoring of psychosocial risks in the workplace is recommended in order to ensure the well-being and satisfaction of workers, so that they can have a high-quality professional performance. We hope that the results achieved in this study can contribute to a better understanding of the state of burnout in security service workers.

Acknowledgements This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications).

References

[1] G. Armon, A. Shirom, and S. Melamed. The big five personality factors as predictors of changes across time in burnout and its facets. *Journal of Personality*, 80(2):403–427, 2012.

[2] A. R. Gomes. *Medida de "Burnout" de Shirom-Melamed (MBSM).* Relatório técnico não publicado. Braga: Escola de Psicologia, Universidade do Minho, 2012.

[3] L. M. Grilo, E. J. Pereira, J. P. Maidana, and M. Stehlík. On stochastic aspects of impact modeling of the innovation incentive system and business internationalization: evidence from portuguese smes. *Stochastic Analysis and Applications*, DOI: 10.1080/07362994.2023.2166532, 2023.

[4] J. F. Hair, G. T. M. Hult, C. M. Ringle, M. Sarstedt, N.P. Danks, and S. Ray. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R.* A Workbook, Springer, 2022.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

The use of the EM, CEM and SEM algorithms for fitting finite mixtures of linear mixed models: a simulation study

Luísa Novais[1], **Susana Faria**[2]
[1] Department of Mathematics, University of Minho, Portugal,
luisa_novais92@hotmail.com
[2] Department of Mathematics, University of Minho, Portugal, sfaria@math.uminho.pt

---

In this work we compare the performance of the Expectation-Maximization (EM), Classification Expectation-Maximization (CEM) and Stochastic Expectation-Maximization (SEM) algorithms in the estimation of the parameters for finite mixtures of linear mixed models. In order to evaluate their performance, we carry out a simulation study. The results show that the CEM algorithm is the least computationally demanding algorithm, although the three algorithms provide similar maximum likelihood estimates for the parameters.

Keywords: mixture models, maximum likelihood estimation, iterative algorithms, simulation study

---

Finite mixture models are a well-known method for modelling data that arise from a heterogeneous population. In regression analysis, it has been a popular practice for modelling unobserved population heterogeneity through finite mixtures of regression models. Within the family of mixtures of regression models, finite mixtures of linear mixed models have also been applied in different areas of application since, besides taking into account the heterogeneity in the population, they also allow to take into account the correlation between observations from the same individual (see McLachlan and Peel [3]).
One of the main issues in mixture models regards the estimation of the parameters. The maximization of the log-likelihood function in mixture models is complex, producing in many cases infinite solutions whereby the maximum likelihood estimator may not exist, at least globally. In order to solve the problem, it is common to resort to iterative methods in the estimation of the parameters, in particular to the Expectation-Maximization (EM) algorithm (Dempster et al. [1]).
In the attempt to overcome the slow convergence and the selection of initial values by the EM algorithm, several modified versions of this algorithm were developed over the years, the most relevant being the Classification Expectation-Maximization (CEM) and the Stochastic Expectation-Maximization (SEM) algorithms (see Novais and Faria [4]). The CEM algorithm incorporates a classification step, C-step, between the E- and M-steps, which consists of assigning each observation to one of the components of the mixture

model, the one that corresponds to the maximum posterior probability. On the other hand, the SEM algorithm incorporates a stochastic step, S-step, between the E- and M-steps, which consists of simulating a realization of the unobserved indicator for each individual, by choosing it randomly from its conditional distribution.
In this work we compare the performance of the three algorithms in the estimation of the parameters of mixtures of linear mixed models through a simulation study under different configurations. These configurations concern the variation of the number of components, the sample size, the number of fixed-effects and the error distributions. For this, we analyse the computational effort of each algorithm by studying the mean number of iterations necessary to achieve convergence, we analyse two statistical properties of the estimators (the bias and the mean square error (MSE)) and we also analyse goodness of fit by computing the root mean-squared error of prediction (RMSEP) through 10-fold cross-validation.
Based on the mean number of iterations for convergence, we conclude that the CEM algorithm always converge in fewer iterations than the EM algorithm, whereas the slow convergence of the SEM algorithm can be a drawback to its use. On the other hand, the bias and MSE of the parameter estimates for the three algorithms show that the three algorithms present approximately the same behaviour, with all the estimates having small bias and MSE.
Finally, for the study of the goodness of fit, the values of the RMSEP show that, although once again the three algorithms perform in a similar way, the SEM algorithm performs generally better whenever the error variance increases and, on the other hand, it seems that both the EM and the CEM algorithms perform better for smaller error variances.
In conclusion, in our simulation study it can be seen that the three algorithms provide similar maximum likelihood estimates for the parameters, both in the sense of lower bias and MSE and in the sense of goodness of fit. However, the CEM algorithm is the less computationally demanding algorithm out of the three algorithms, that is, it is always the one converging in fewer iterations, so we reccommend its use in every situation.

Acknowledgements The research of L. Novais was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference SFRH/BD/139121/2018.

References

[1] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1):1–38, 1977.

[2] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[3] G. McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[4] Luísa Novais and Susana Faria. Comparison of the EM, CEM and SEM algorithms in the estimation of finite mixtures of linear mixed models: a simulation study. *Computational Statistics*, 36(4):2507–2533, 2021.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Tree-based classification methods for customer NPS analysis

Inês Carvalho[1], **Susana Faria**[2], **Ana Freitas**[3]
[1] Universidade do Minho, pg42544@alunos.uminho.pt
[2] Centro de Matemática e Departamento de Matemática, Universidade do Minho, sfaria@math.uminho.pt
[3] Sonae MC, ancfreitas@mc.pt

Companies are becoming more focused on understanding their customers to improve customer satisfaction and consequently customer loyalty and thus increase sales. The Net Promoter Score (NPS) is a widely adopted measure of customer loyalty and predictor of sales growth.

This study had the goal of classifying the customer's Net Promoter Score (NPS) class of a given retail company. Namely, predict the customer as a Promoter, Passive or Detractor, through the use of Data Mining (DM) techniques, as well as determine the most important variables and thus understand which effects can impact the customer's classification.

Keywords: net promoter score, data mining, classification, decision trees, random forests

---

Reichheld [4] proposed the concept of Net Promoter Score (NPS) as a metric that measures the likelihood of an existing customer recommend a company to another prospective customer. The author claimed that the willingness of a customer to make a recommendation to friends or colleagues would determine customer loyalty and consequently the company's growth. NPS is derived from a single question: "How likely is that you would recommend company X to a friend or colleague?", on a scale of 0 to 10, where 0 means "Not at all Likely" and 10 means "Extremely Likely". The customers are then categorized into three groups of survey responders depending on the range of their NPS: Promoters (9–10), Passives (7–8) and Detractors (0–6).

This study's project was developed in a major Portuguese food-based retailer for one of their non-food stores. The data analysed is a combination of survey and loyalty card data, which combines customers' promoter scores, their characteristics and their purchasing habits.

We propose a Data Mining approach to the NPS multi-class classification problem. Initially, due to the discrepancy of the customer feedback which resulted into an extremely imbalanced dataset, several resampling techniques, namely Random Oversampling, Random Undersampling and Synthetic Minority Oversampling Technique (SMOTE) [2], are applied to handle class imbalance. Two different machine learning algorithms are used:

a baseline algorithm of Decision Tree (DT), and an ensemble learning of Random Forest (RF) [1]. Overall, the results from both methods indicate a poor predictive performance, with low AUC values.
Through the RF model, the importance of each variable in the prediction can be analysed [3]. The most relevant variables are variables related to the customer account longevity, the customer's age and variables related to the store traffic and customer's transactions for different periods of the day, during the week and the weekend.
In order to understand why the RF model is behaving more erroneously, an error analysis is performed. The results show that the classifier has difficulty distinguishing the minority classes, namely, the Detractors and Passives, but it has good performance in predicting the class Promoters. We are also able to determine that the Detractors seem to be unsatisfied customers, with bad shopping experiences.
In a business sense, this methodology can be leveraged to distinguish the Promoters from the rest of the consumers, since the Promoters are more likely to provide good value in long term and can benefit the company by spreading the word for attracting new customers.

Acknowledgements
This work was also supported by the Portuguese Foundation for Science and Technology (FCT) in the framework of the Strategic Funding UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 2002.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, New York, NY, 1 edition, 2013.

[4] F. F. Reichheld. The one number you need to grow. *Harvard Business Review*, 81:46–55, 2003.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Prediction of bankruptcy one-to-three-year-ahead

Vera Rabaça[1], **Mário Basto**[2], **José M. Pereira**[3]
[1] Polytechnic Institute of Cávado and Ave, Portugal, veracatarinasousa14@gmail.com
[2] Polytechnic Institute of Cávado and Ave, Portugal, mbasto@ipca.pt
[3] Polytechnic Institute of Cávado and Ave, Portugal, jpereira@ipca.pt

Predicting corporate insolvency is vital to the economies of all nations. Numerous statistical models for predicting insolvency have been introduced and evaluated in a variety of scenarios. Using data from 2017, 2018, and 2019, this study applies the logistic regression model and its ridge and lasso variants to portuguese SMEs in the textile industry and examines their ability to predict the companies' viability in 2020. Some results were unexpected.

Keywords: business bankruptcy, logistic regression, ridge regression, lasso

Insolvency occurs when a company cannot pay its debts or has insufficient assets to fulfill its obligations. Financial and societal costs are enormous. Consequently, it is essential to have very accurate forecasting models for the future status of businesses so that the required measures can be undertaken. Most of the research employs statistical or artificial intelligence methods to predict if organizations would go bankrupt. This study uses logistic regression to predict SME (Small and Medium-Sized Enterprises) insolvency in the textile industry. The ridge and lasso models, two forms of logistic regression, were also used to build two extra models to see if they could generate a more precise forecast than logistic regression. Financial data was gathered from the SABI database that includes insolvent and healthy textile SMEs.

The logistic regression model is governed by the dependent variable's success probability, recorded as one or zero. Ridge [1] and lasso [3] regressions are better at handling multicollinearity and overfitting. Ridge and lasso regressions require a nonnegative fit parameter λ present in logistic regression's log-likelihood function.

In ridge, as λ increases, estimated coefficients approach zero but never reach it. All independent variables are included in the model [2].

In contrast, lasso regression is an alternative to conventional logistic regression that provides a reduction in the number of independent variables by setting some coefficients to exactly zero [2].

Table 1 depicts the overall accuracy of the three models.

Logistic regression and its ridge and lasso variants achieved similar global accuracy rates in 2019 (Table 1). Two and three years in advance (years 2018 and 2017), the global accuracy rates differed somewhat between the three models. In 2017 and 2018, the ridge and lasso regressions achieved a greater global overall accuracy rate than the classic logistic regression, with an unexpected increase in 2017 (Table 1).

Table 1: Three-year overall accuracy for the three statistical approaches.

	Logist regression	**Ridge**	**Lasso**
Year 2017	Accuracy = 64.55%	Accuracy = 77.72%	Accuracy = 80.83%
Year 2018	Accuracy = 69.64%	Accuracy = 72.79%	Accuracy = 73.69%
Year 2019	Accuracy = 74.65%	Accuracy = 73.53%	Accuracy = 74.30%

Similarly, the discriminant ability of the ROC curves to distinguish between active and bankrupt enterprises was generated for the three models. The area under the ROC curve was calculated as a measure of the model's quality.
Using data from 2019, the data indicate that the conventional logistic regression has a modest benefit over the other two models in terms of discriminant ability (ROC curve) to distinguish between operating and insolvent businesses (area 0.8107 versus 0.7781 for ridge and 0.7802 for lasso), however there are no significant differences when using data from 2018 (area 0.7846 versus 0.7922 for ridge and 0.7916 for lasso). Using 2017 data, ridge and lasso regressions perform significantly better (area 0.7311 versus 0.8231 for ridge and 0.8461 for lasso). In contrast to the standard logistic regression, the discriminant capacity of ridge and lasso regressions improve as the distance to the year 2020 increases.
In short, the global overall accuracy rate is comparable for all three models in 2019, but in 2017 and 2018, the ridge and lasso regressions achieved a greater global overall accuracy rate than the classic logistic regression, with an unexpected increase in 2017. The results obtained using the ROC curve were very similar. The results of ridge and lasso regressions were noticeably improved with the 2017 data.

References

[1] S. Cessie and J.C. Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201, 1992.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, London, 2013.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Monitoring and prediction of the air quality towards sustainable work environments

Jorge Siopa[1*], **Bruno Gonçalves**[1,2], **Luís Aires**[1] **e Marcelo Gaspar**[1,3]
[1] Higher School of Technology and Management Polytechnic Institute of Leiria, Leiria, Portugal
[2] Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal
[3] ADAI –LAETA - Industrial Aerodynamics Development Association, Coimbra University, Coimbra, Portugal
jorge.siopa@ipleiria.pt, bruno.gonçalves@ipleiria.pt, luis.aires@ipleiria.pt, marcelo.gaspar@ipleiria.pt *corresponding author

---

Health, safety and well-being are often related to the air quality of urban areas, as well as with occupational safety and health in many workplaces [4]. Hence, it is crucial to not only monitor the air quality of such locations, but also to be able to predict potential changes in air quality to avoid such conditions. This will help to promote better working conditions and increase the social sustainability of existing and future workplaces. Thus, it is urgent to measure and monitor the quality of the air we breathe, because it is with this assessment of air quality that the health impacts, caused by air pollution, can be known and solved [1] [2].

Keywords: air quality prediction, air pollution, public health, sustainable campus network, nustainable development goals.

The present work aims at carrying out a study of air quality measurement and monitoring integrated in an initiative of the Sustainable Campus Network [3] implemented at the Polytechnic Institute of Leiria. In this study, in addition to measuring the concentration of pollutants in the ambient air at Campus 2 of this Higher Education Institution during two consecutive weeks, a comparative analysis of the measured parameters with the meteorological conditions experienced at that location during the same time period was carried out.
Thus, in addition to the air quality parameters, nitrogen oxides, sulfur dioxide, ozone, carbon monoxide, hydrocarbons (BTX) and PM10, the parameters wind speed and direction, temperature and relative humidity, global solar radiation and precipitation of the ambient meteorological factors were also evaluated. Besides allowing the monitoring and discussion of these parameters in the quality of the air breathed in the ecosystem of this polytechnic, this work is intended to contribute not only to the promotion of improved sustainability on the campus of the Polytechnic of Leiria, as well as the other campuses of Portuguese higher education institutions integrated in this Network Sustainable Campus.

Based on the pilot study the air quality at Campus 2 of the ESTG of the IP Leiria was measured. After applying the quadratic multilinear regression models (equation 1) and analysing the results we conclude that the correlation for SO_2 is non-existent (R=0.122), for NO and NO_X is also very weak ($0.465 < R < 0.575$), for CO, NO_2 and $PM10$ is medium ($0.733 < R < 0.780$) and for $PM25$ and O_3 is very high ($0.842 < R < 0.910$). This allows predicting with high accuracy the concentration of these pollutants from the meteorological and traffic intensity parameters.

$$f(x_i, \bar{\beta}, \bar{\alpha}, \omega_1, \omega_2) = \beta_0 + \sum_{j=1}^{m-1} \left(\beta_j x_{ij} + \alpha_j x_{ij}^2\right) + \omega_1 \cos(x_{im}) + \omega_2 \sin(x_{im}) \quad (1)$$

where

x_{ij} - Measure in the time period i (15min) for the parameter j
x_{im} - Entry wind direction angle.

Table 1: Regression Parameters

Continuous Parameters		Binary Parameters
$\alpha\beta5$	Precipitation (mmH2O)	$\beta0$ Mean Value
$\alpha\beta6$	R. Global (W/m^2)	$\beta1$ Saturday
$\alpha\beta7$	Pressure (KPa)	$\beta2$ Sunday
$\alpha\beta8$	Wind Speed (m/s)	$\beta3$ Rush Hour
$\alpha\beta9$	Temperature (°C)	$\beta4$ Carnival
$\alpha\beta10$	Rel. Humidity	-
$\omega1\omega2$	Wind Heading (°)	-

For example, the $PM10$ concentration equation is

$$f_{\text{PM10}}(x_i) = 20.34 + 3.63x_{i2} - 3.64x_{i3} + 5.37x_{i4} + 0.55x_{i5} - 8.83x_{i6} - 2.49x_{i6}^2 - 1.12x_{i7} + 1.27x_{i7}^2 + 7.41x_{i8} - 0.93x_{i8}^2 + 13.85x_{i9} + 1.64x_{i9}^2 + 2.50x_{i10} - 6.581\cos(x_{i11}). \quad (2)$$

References

[1] Pantavou K. Rizos E. C. Sindosi O. A. Tagkas C. Seyfried M. Saldanha I. J. Hatzianastassiou N. Nikolopoulos G. K. Ntzani E. Markozannes, G. Outdoor air quality and human health: An overview of reviews of observational studies. environmental pollution. The bootstrap. *Environmental Pollution*, 306, 2022.

[2] Jayabal S. Ramesh Kumar A. Vinoth V. Prabhakaran, J. Air quality assessment in indoor and outdoor environments: A review. The bootstrap. *Materials Today: Proceedings*, 2022.

[3] Rede Campus Sustentável. http://www.redecampussustentavel.pt, 2022.

[4] Leidelmeijer K. Marsman G. de Hollander A. van Kamp, I. Urban environmental quality and human well-being. The bootstrap. *Landscape and Urban Planning*, 65(1–2):5–18, 2003.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Clustering on the unit hypersphere using non-negative matrix factorization

Lazhar Labiod[1], **Mohamed Nadif**[1]
[1] Centre Borelli UMR 9010 Université Paris Cité, lazhar.labiod@u-paris.fr
[1] Centre Borelli UMR 9010 Université Paris Cité, mohamed.nadif@u-paris.fr

We propose a new non-negative matrix factorization (NMF) framework inspired by the spherical *k*means objective function. We derive a new criterion and to optimize it, we develop an algorithm based on only one multiplicative update rule. In particular, we show that the spherical *k*-means is approximatively equivalent to the algebraic problem of NMF under some suitable constraints. Simplicity and efficiency are two major characteristics of our approach and numerical experiments on real document×term datasets demonstrate its interest.

Keywords: clustering, NMF, unit hypersphere

Clustering has received a significant amount of attention as an important problem with many applications, and a number of different methods have emerged over the years. In general, sparsity and high dimensionality are the problems encountered by the different existing clustering algorithms. It is the case of document×term matrices where each cell represents the frequency of a word in a document. Hence, the choice of an appropriated distortion measure can be crucial to the performance of a document/term clustering algorithm. Thus, in document clustering, considering the data as directional is a good alternative. Indeed, for high-dimensional data sparse or not, cosine similarity has been shown to be a superior measure to Euclidean distance [1]; the direction of a document vector is more important than its magnitude. This leads to a unit-vector representation, i.e., each document vector is normalized to be of unit length. Therefore it is shown that the spherical *k*means algorithm (SPKM) which is a k-means whose the criterion is based on a cosine similarity and where the centers are normalized to be of unit length, is one of the most effective clustering algorithm in this case [6].
However, despite the advantages of SPKM, in [2] the authors shown that the criterion of SPKM is associated to underlying restricted von Mises-Fisher distributions mixture where proportions of components (clusters) are assumed to be equal and the directional variance (or dispersion) of each cluster is the same for all clusters [4, 2]. Hence SPKM presents some drawbacks where the clusters are not well separated while the data are far from the model. Then, the authors proposed to use vMF mixture models with less constraints and for the estimation of the parameters and clustering they performed hard and soft clustering algorithms derived from the EM algorithm [3]. However, it should be noted that, the proposed approximation used, for estimation of cluster concentrations, suffers of the high

dimensionality. This difficulty is due to the concentration parameter which is non-trivial in high dimensions due to the functional inversion of the ratios of the Bessel functions. In their conclusion, the authors emphasized that investigations about the tradeoff between model complexity and sample complexity deserve to be studied in the directional data context. In the sequel, we will not consider the mixture approach but, we propose to consider another approach simple and efficient allowing to overcome the limits of SPKM. Even if bringing a solution to the clustering problem is not the main objective of non-negative factorization matrix (NMF) [5], this approach has appealed many authors for data clustering and particularly for document clustering. Different authors [7] emphasized that the NMF approach outperforms $k$means on most datasets since NMF seems to model varying distributions due to the flexibility of matrix factorization, as compared to the rigid spherical $k$means with equal proportions that $k$means objective function attempts to capture. For SPKM, note that the hard version of EM proposed in [2] is in fact a classification EM algorithm optimizing the classification log-likelihood of the restricted v-MF distributions mixture. It is natural to think that an appropriated NMF is susceptible to model varying distributions as compared to the underlying model that SPKM function attempts to capture. For this reason, in this paper, we chose to consider the objectives of SPKM in a NMF framework.

References

[1] Melissa Ailem, François Role, and Mohamed Nadif. Sparse poisson latent block model for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1563–1576, 2017.

[2] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[4] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175, 2001.

[5] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[6] Aghiles Salah and Mohamed Nadif. Directional co-clustering. *Advances in Data Analysis and Classification*, 13:591–620, 2019.

[7] Dingding Wang, Tao Li, and Chris Ding. Weighted feature subset non-negative matrix factorization and its applications to document understanding. In *2010 IEEE International Conference on Data Mining*, pages 541–550. IEEE, 2010.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

p-value or Bayes factor: three brief illustrations

Mário Basto[1], **Teresa Abreu**[2], **Ricardo Gonçalves**[3], **José M. Pereira**[4]
[1] Polytechnic Institute of Cávado and Ave, Portugal, mbasto@ipca.pt
[2] Polytechnic Institute of Cávado and Ave, Portugal, tabreu@ipca.pt
[3] Polytechnic Institute of Cávado and Ave, Portugal, rgoncalves@ipca.pt
[4] Polytechnic Institute of Cávado and Ave, Portugal, jpereira@ipca.pt

---

Hypothesis testing is used in the process of making decisions. The *p-value* is typically used. A value below 0.05 is generally taken as evidence against the null hypothesis H_0. Another approach is Bayesian analysis, which relies on the Bayes factor. However, any of these options is not without controversy. Extreme views are held by many. Nevertheless, both approaches can be valuable. Three short examples are given.

Keywords: hypothesis testing, p-value, Bayes factor

A low *p-value* reveals nothing about the evidence against H_1, hence it can inflate the evidence against H_0. In addition, the *p-value* does not answer the question of how plausible is the research hypothesis H_1 given the data. In Bayesian analysis, prior parameters are modified for any quantity of new data to yield posterior values. Bayes factor contrasts the data under two hypotheses, offering evidence for and against either. However, it can be very sensitive to prior values.

Example 1. Consider the following hypotheses pertaining to a certain disease:
$H_0 : Mary\ does\ not\ have\ the\ disease$ versus $H_1 : Mary\ has\ the\ disease$

Imagine a worthless diagnostic test that, regardless of whether a person is ill or not, produces 99 percent negative and 1 percent positive results.
For a positive test, $p\text{-}value = P(test\ positive|Mary\ does\ not\ have\ the\ disease) = 0.01$.
For a level of 1%, this is a significant result, there is strong evidence against H_0. Since the test reveals no information about the disease, it would be ludicrous to conclude that Mary is most certainly infected.
In contrast, the Bayes factor is one, meaning neither hypothesis is more likely than the other:

$$BF_{01} = \frac{P(test\ positive|Mary\ does\ not\ have\ the\ disease)}{P(test\ positive|Mary\ has\ the\ disease)} = \frac{0.01}{0.01} = 1.$$

Example 2. Consider the hypotheses testing on a binomial variable's success rate π:
$H_0 : \pi = 0.5$ versus $H_1 : \pi \neq 0.5$

A sample of $n = 200$ trials was collected to choose amongst the hypotheses. There were 80 successes and 120 failures. Under H_0, $\pi = 0.5$, the *p-value* is given by:
$p\text{-}value = 2 \times P(\bar{X} \leq 80|\pi = 0.5) = 0.005685$.
Given this very low *p-value*, H_0 is rejected, and the evidence is very strong against H_0.
Using the Bayesian binomial test implemented in the software JASP with a uniform prior distribution, $BF_{01} = 0.206$, the data are 4.853 (1/0.206) times more probable under H_1 than H_0. The evidence against H_0 is moderate.
Using a symmetric beta prior with parameters $\alpha = \beta = 7$, $BF_{01} = 0.0909$, indicates that the result is approximately 11 times more likely under H_1 than under H_0, which constitutes strong evidence against H_0.
Using a beta prior with values $\alpha = 10$ and $\beta = 50$ (in this scenario, there is a prior greater probability of getting much fewer successes than failures), $BF_{01} = 20.7046$, a complete turnaround in the evidence. Now, strong evidence supports H_0's claim. This occurs because H_1 is a composite hypothesis, and the Bayes factor computes a weighted average, based on the prior, of the likelihood of the observed data under all the alternative effect sizes. Bayesian priors affect the evidence. Prior choice can be crucial.
For the same findings and beta prior with values $\alpha = 10$ and $\beta = 50$, but with a sample size three times larger, $BF_{01} = 0.0284$ ($BF_{10} = 35.2547$). There is very strong evidence supporting H_1 revealing another reversal in the evidence, due to the decrease in the posterior distribution dispersion.

Example 3. Consider the hypotheses for the mean of a normal population with known standard deviation of 10:
$H_0 : \mu = 50$ versus $H_1 : \mu = 60$

The sample size for the study was one hundred. The sample mean was $\bar{x} = 55$.
The *p-value* is given by $P(\bar{X} \geq 55|\mu = 50) = 0.00000029$.
Since the *p-value* reveals very strong evidence against H_0, $\mu = 60$ must represent the actual population mean. Because the z statistic's distribution is symmetrical and the sample mean falls between the two hypothetical values, swapping the hypotheses will have no effect on the *p-value*, meaning that $\mu = 50$ must now be the population mean. This occurs due to the fact that the data are never compared to H_1.
The Bayes factor is one for both the original and permuted hypotheses, since $f(55|\mu = 50) = f(55|\mu = 60)$, indicating no conclusive evidence for either hypotheses.

In short, both the p-value and the Bayes factor are effective tools, but they must be used critically in order to make better decisions.

References

[1] M.J. Bayarri and J.O. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19:58–80, 2004.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Does the qualifications index influence the gross value added in Portuguese municipalities?

Marco Marto[1], **João Lourenço Marques**[2], **Mara Madaleno**[3]
[1] Research Unit in Governance, Competitiveness and Public Policies (GOVCOPP), Universidade de Aveiro, marcovmarto@ua.pt
[2] Departamento de Ciências Sociais, Políticas e do Território, Universidade de Aveiro, jjmarques@ua.pt
[3] Departamento de Economia, Gestão, Engenharia Industrial e Turismo, Universidade de Aveiro, maramadaleno@ua.pt

Portuguese territories' distribution of wealth and people with higher qualifications appears to be related. On the one hand, more concentrated in certain municipalities, and on the other hand, each time more unbalanced between municipalities. The Portuguese population with higher qualifications tends to be more concentrated close to big cities and in the littoral. Therefore, this work studies the effect of the qualifications on the concentration of higher values (and lower values) for the gross value added in Portuguese municipalities. It is used a DEA (data envelopment analysis) to explain the effect of municipalities' qualifications on the values of gross value added comparatively for the years: 2001, 2011, and 2021. The asymmetric distribution of the gross value added among municipalities is explained as an output with DEA, while the index of qualifications is the input variable. Some social and political measures need to be developed in order to make the distribution of gross value added and qualified people more balanced, maybe starting by reducing (average) qualification inequalities values among municipalities.

Keywords: qualifications index, gross value added, DEA optimization, Portuguese municipalities, spatial analysis

The territorial distribution of more qualified people by the Portuguese municipalities has a greater concentration close to the main universities, in their metropolitan areas, and littoral in general. Despite this asymmetric distribution of qualified labor, in some municipalities is not only the qualified labor that determines the value of gross value added. The gross value added is influenced by the outliers. One of the aims of this work is to identify the municipalities which perform better than expected (or worse than expected) considering in DEA ([3]) optimization the qualifications' index as input and the measure of wealth as output. Related work was developed by Marto et al. ([2]), considering EU NUTS 2.
The qualifications index was defined by Marques et al. ([1]) for a study about accessibility to primary schools and social inequalities. Their definition is used in this work for the

purpose of calculation of the municipality's qualification index which is used as the input variable for DEA optimization. The output variable, as previously mentioned is a measure of wealth. Concerning the development of DEA optimization, it is important to start to analyze whether the relation between the qualification index and wealth is linear or non-linear. In other others, it is better described in DEA optimization by constant or variable returns to scale.
The results can identify various groups of municipalities with different behaviors, considering the input and output variables. In some (groups of) municipalities and for various reasons the results are better or worse than expected and part of this work is dedicated to understanding why. Another output of this work is the evolution and relation between variables for the years 2001, 2011, and 2021.

Acknowledgements The Foundation for Science and Technology, I.P provided the funding for the fellowship of one of the authors (Marco Marto) which allowed him to develop this work.

References

[1] João Lourenço Marques, Jan Wolf, and Fillipe Feitosa. Accessibility to primary schools in portugal: a case of spatial inequity? *Regional Science Policy & Practice*, 13(3):693–707, 2021.

[2] Marco Marto, João Lourenço Marques, and Mara Madaleno. An evaluation of the efficiency of tertiary education in the explanation of the performance of gdp per capita applying data envelopment analysis (dea). *Sustainability*, 14(23):15524, 2022.

[3] Subhash C Ray. *Data envelopment analysis: theory and techniques for economics and operations research*. Cambridge university press, 2004.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Estimation of the dispersion parameter in count models

Rui Miranda[1], **Rita Gaio**[2]
[1] Faculdade de Ciências da Universidade do Porto, up201804962@edu.fc.up.pt
[2] Faculdade de Ciências da Universidade do Porto e Centro de Matemática da Universidade do Porto, argaio@fc.up.pt

---

The Poisson regression model is generally the first choice for fitting count data in the context of generalized linear models. However, the equality condition between the mean and the variance in the conditional response is sensitive and difficult to evaluate, limiting its use in real-world problems. The suitability of existing methods for assessing over-dispersion in Poisson regression models will be discussed.

Keywords: poisson regression, dispersion parameter, pearson estimator, mean deviance estimator

---

The theory of Generalized Linear Models states[1] that the response variable conditional on the explanatory variables, $Y|\boldsymbol{x}$, follows a distribution belonging to the exponential family of distributions, whose general expression for the probability (density) function is given by

$$\mathbb{P}_Y(y;\theta(\boldsymbol{x}),\phi) = a(y,\phi)\,\exp\Big(\frac{y\theta(\boldsymbol{x}) - k(\theta(\boldsymbol{x}))}{\phi}\Big)$$

with location parameter $\theta(\boldsymbol{x}) \in \mathbb{R}$ and dispersion parameter $\phi > 0$.

For a Poisson regression with conditional mean $\mu(\boldsymbol{x})$ we have $\theta(\boldsymbol{x}) = log(\mu(\boldsymbol{x}))$ and $\phi = 1$, due to the equality between mean and variance. The fact that the dispersion parameter is a constant allows us to raise the question of the estimation of ϕ.

The Mean Deviance and the Ratio of the Pearson X^2 statistic to its Degrees of Freedom are known estimators for ϕ. They are defined, respectively, by:

$$\hat{\phi}_D = \frac{D(y,\hat{\mu}(\boldsymbol{x}))}{n-p-1} \quad \text{and} \quad \hat{\phi}_P = \frac{X^2}{n-p-1} = \sum_i \frac{(y_i - \hat{\mu}_i(\boldsymbol{x}))^2}{(n-p-1)V(\hat{\mu}_i(\boldsymbol{x}))}$$

where D is the total deviance of the model and V is the variance function of $Y|\boldsymbol{x}$, evaluated at $\hat{\mu}_i(\boldsymbol{x})$. For the Poisson distribution, $V(\hat{\mu}_i(\boldsymbol{x})) = \hat{\mu}_i(\boldsymbol{x})$.

It is known that the statistics D and X^2 follow asymptotically a $\chi^2(n-p-1)$ distribution. Also, both estimators $\hat{\phi}_D$ and $\hat{\phi}_P$ are asymptotically unbiased and consistent.

In Poisson regression, the departure of $\hat{\phi}_P$, or $\hat{\phi}_D$, from $\phi = 1$, can be seen as a departure from the adequacy of the Poisson distribution to the data. In particular, $\phi > 1$ corresponds to over-dispersion, which is a problem often encountered in the analysis of count data that can lead to invalid inferences if not correctly addressed, such as biased parameter estimations corresponding to underestimated standard errors. Harrison indicated that over-dispersion can arise, for example, when models have been poorly specified, when there is clustering or an excess number of zeroes in the data, or just when the variance of the conditional response is truly greater than the mean[2].

The estimators $\hat{\phi}_D$ and $\hat{\phi}_P$ are advertised to keep their properties when applied in the framework of Generalized Linear Mixed Models, and their use is advocated, specially for $\hat{\phi}_P$[3]. However, the dispersion included by the random effects may interfere with the dispersion of the conditional response variable.

In this study, we evaluated a simulation procedure described by Zuur[3] to access the departure of the estimates of ϕ for a given model from the condition $\phi = 1$. More precisely, we generated data coming from a mixed-effects Poisson regression model, with $j \in \{1, 2, 3, 4\}$ repeated measurements per experimental unit:

- $Y_{ij}|b_{0i} \sim \mathcal{P}(\exp(\beta_0 + b_{0i}))$
- $Y_{i1}|b_{0i}, \ldots, Y_{i4}|b_{0i}$ are independent, $i = 1, \ldots, n$
- $Y_{i1}|b_{0i}$ independent from $Y_{i'1}|b_{0i'}$, $i \neq i'$
- $b_{0i} \sim \mathcal{N}(0, {\sigma_0}^2)$ are independent, $i = 1, \ldots, n$

with $n = 10, 50, 200$, $\beta_0 = 0.1, 0.5, 1$ and $\sigma^2 = 0.1, 1, 2$.

We were able to compare the performances of $\hat{\phi}_D$ and $\hat{\phi}_P$ in models with different values for $\mu_i(\boldsymbol{x})$, focusing on the cases where the mean was close to zero. In particular, $\hat{\phi}_P$ obtained better results than $\hat{\phi}_D$ as is known from the literature of Generalized Linear Models, stated by Dunn and Smith[1].

References

[1] Peter K. Dunn and Gordon K. Smyth. *Generalized Linear Models With Examples in R.* Springer Texts in Statistics. Springer, New York, 2018.

[2] Xavier A. Harrison. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2, 2014.

[3] Alain F. Zuur. *Mixed effects models and extensions in ecology with R.* Springer, 2009.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Perturbation methods: an application using R

Jorge Morais[1], **Rita Sousa**[2], **Susana Faria**[3]
[1] Universidade do Minho, pg42545@uminho.pt
[2] Banco de Portugal, Centro de Matemática e Aplicações, FCT-UNL, rcsousa@bportugal.pt
[3] Universidade do Minho, Centro de Matemática, sfaria@math.uminho.pt

A large part of the data from surveys or other sources cannot be published directly due to privacy and confidentiality concerns. Statistical Disclosure Control techniques suggest methods to modify data so that they can be published without revealing confidential information.

In this study, we describe and compare different perturbation methods. We also present several measures for data utility and disclosure risk to evaluate the method's performance. Finally, we illustrate an application of these methods using the R-Package *sdcMicro*.

Keywords: statistical disclosure control (SDC), disclosure risk, data utility, package *sdcMicro*

The demand for data access has been growing a lot in recent years. The compromise between the utility of the information provided and the protection of confidentiality is increasingly important.
Statistical Disclosure Control (SDC) techniques suggest methods to modify the statistical data without providing information that makes it possible to identify an observation from the perturbed dataset [4]. Applying SDC techniques to the original data may result in information loss and hence affect data utility. An optimal method of SDC is the one that minimizes the disclosure risk and maximizes the utility of data. In this study, we look for techniques that better balance these two perspectives.
Initially, we describe the SDC methods and we present several measures for data utility and disclosure risk to compare the method's performance.
According to the literature, the *Exact General Additive Data Perturbation* (EGADP) model provides the lowest disclosure risk and still allows all inferences made on the perturbed database to be the same as the original dataset. As for nonlinear models, the *Data Shuffling* model leads to perturbed values very similar to the original values and with low disclosure risk. Finnaly, the *Microaggregation* method offers very ineffective results comparing to the *Data Shuffling* and EGADP models [3].
We also apply the SDC methods to a real dataset which is made available by the *Banco de Portugal Microdata Research Laboratory* (BPLIM) for research purposes [1]. The SDC methods are applied using the R language, more specifically the *sdcMicro* package [2].

The conclusions from the literature do not exactly match with the results obtained from the application to the real dataset. The *Data Shuffling* model and the EGADP model do not perform well when compared to the other models presented. The EGADP model offers the lowest disclosure risk, however, it has higher information loss when compared to the noise models. The *Data Shuffling* model offers a lower disclosure risk, but also presents large values for the information loss. So we conclude that for the dataset under study the best models in the literature were over-performed by the *Noise Models*.
The choice of the most appropriate method always depends on the objective of the person responsible for the dataset, giving more emphasis to the risk of disclosure or the data utility.

Acknowledgements
This work was also supported by the Portuguese Foundation for Science and Technology (FCT) in the framework of the Strategic Funding UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

References

[1] Banco de Portugal Microdata Research Laboratory (BPLIM). Incentives systems data. Banco de Portugal, 2021.

[2] T. Matthias, M. Bernhard, and K. Alexander. Package *sdcmicro*. Technical report, 2021.

[3] C.R. Rao, R. Chakraborty, and P. K. Sen. *Handbook of Statistics Bioinformatics in Human Health and Heredity.* North Holland, 2012.

[4] M. Templ. *Statistical Disclosure Control for Microdata: Methods and Applications in R*, volume 1. Springer International Publishing, 2017.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Mortality and lethality rates in public health

Teresa Abreu[1], **Ricardo Gonçalves**[2], **José M. Pereira**[3], **Mário Basto**[4]
[1] Polytechnic Institute of Cávado and Ave, Portugal, tabreu@ipca.pt
[2] Polytechnic Institute of Cávado and Ave, Portugal, rgoncalves@ipca.pt
[3] Polytechnic Institute of Cávado and Ave, Portugal, jpereira@ipca.pt
[4] Polytechnic Institute of Cávado and Ave, Portugal, mbasto@ipca.pt

To make an informed decision, it is essential to comprehend statistical concepts that appear obvious but are not necessarily properly understood. Many people, including health professionals, routinely mistake healthcare numbers and statistics. One of the objectives of this study is to determine whether individuals understand the terms mortality rate and lethality rate. To accomplish this, an online questionnaire was created in october 2022.

Keywords: mortality rate, lethality rate, case fatality rate

Mathematical and statistical thinking permeate modern life to an unprecedented degree. Most often, they are utilized to back up an argument or add weight to a marketing text, statement, or idea. The risks and advantages of medical procedures are sometimes exaggerated or overlooked by physicians, nurses, and patients alike. The mortality rate and the lethality rate are two terms that, often, are interpreted in an incorrect manner.
The lethality rate, also known as the case fatality rate, is a measurement of the percentage of individuals who get a particular illness and die as a result. The lethality rate is determined by dividing the number of deaths attributable to the illness by the total number of cases. On the other hand, the mortality rate is calculated by dividing the number of deaths linked to a particular disease during a specified time period by the total population size. The key distinction lies in the denominator of the two ratios.
The lethality rate and the mortality rates are key indicators of the severity of a disease. They can assist policymakers and public health professionals in comprehending the impact of an illness on a population and guiding decisions on the distribution of resources. They can also be used to evaluate the efficacy of interventions or treatments.
From a public health perspective, the mortality rate is generally more relevant, as it reveals the impact of the disease in question on the population. A illness with a high lethality rate but low incidence may have a reduced impact on the population. On the other hand, a disease with low lethality rate but rapid population spread may have a larger effect on mortality. Nevertheless, from an individual standpoint, the lethality rate is crucial information for the physician who has a patient with a particular condition and must decide how to treat him. In this scenario, the mortality rate is less relevant.
This study focuses on a single question. It was asked:

Question: A new drug designed to prevent death from 'Navis' disease is intended to be launched on the market. The drug is taken as a preventative measure prior to contracting the disease. The results of the randomized, controlled, and double-blind clinical trial involving 2000 participants in each of the experimental and control groups are as follows:
500 individuals in the experimental group contracted 'Navis' disease. 10% of these individuals succumbed to the disease, or the lethality rate was 10%.
100 individuals in the control group contracted 'Navis' disease. 20% of these individuals succumbed to the disease, or the lethality rate was 20%.
In terms of public health, it can be concluded that:
A) The value of greatest interest in this trial is not the mortality rate in each group, but rather the lethality rate in each group and the results of the trial support its approval.
B) The value of greatest interest in this trial is not the lethality rate in each group, but rather the mortality rate in each group and the results of the trial support its approval.
C) The value of greatest interest in this trial is not the mortality rate in each group, but rather the lethality rate in each group, and the results of the trial do not support its approval.
D) The value of greatest interest in this trial is not the lethality rate in each group, but rather the mortality rate in each group, and the results of the trial do not support its approval.

The results in Table 1 are those obtained up until December 31, 2022.

Table 1: Distribution of responses.

Answers	**Frequency**	**Percent**
A	42	46.7%
B	14	15.6%
C	9	10.0%
D	25	27.8%

Options A and C give lethality rate the most importance in terms of public health. It was anticipated that individuals who place a larger emphasis on lethality would respond that the results of the study support its approval, that is, option A, given that lethality has been reduced with the treatment. The lethality rate is provided, so individuals who are unfamiliar with it can easily deduce its meaning.
In contrast, option B and D emphasize the mortality rate. Those who place a larger focus on the mortality rate were expected to respond that the results of the trial do not support its approval, that is, option D, as the treatment raised it. It is also possible that some of the option B respondents do not understand what the mortality rate actually means.
As anticipated, the response rates for alternatives A and D are higher. However, just 27.8% of respondents selected the best solution, option D. There is still a widespread lack of familiarity with numbers and statistics.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Modeling of hourly water consumption of residential clients in the north of Portugal

Tatiana Cunha[1], **Eliana Costa e Silva**[2], **Flora Ferreira**[3]
[1] ESTG, Politécnico do Porto, Portugal8180272@estg.ipp.pt
[2] CIICESI, ESTG, Politécnico do Porto, Portugaleos@estg.ipp.pt
[3] Centre of Mathematics/Dept. of Mathematics, University of Minho, Portugal
fjferreira@math.uminho.pt

Being able to model and forecast water consumption is essential, since this is a scarce resource that must be used rationally. Companies have implemented new smart water meters that allow collecting hourly and sub-hourly consumption data for each consumer. These data present different sources of seasonality that cannot be captured by commonly used models, such as the Autoregressive Integrated Moving Averages models. The objective of this work is the individualized study of hourly water consumption of residential clients of a company in the North of Portugal. TBATS model is applied to model the consumption of six residential clients, with the intent to forecast individual behavior.

Keywords: times series, TBATS, water consumption, forecasting

The TBATS model incorporates multiple and complex seasonalities and includes a trigonometric representation of the seasonal components based on Fourier series, Box-Cox transformation, ARMA errors, and seasonal patterns [1]:

$$y_t^{(w)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-1}^{(i)} + d_t \quad (1)$$

where $y_t^{(w)} = \frac{y_t^w - 1}{w}$ if $w \neq 0$ and $y_t^{(w)} = \log(y_t)$ if $w = 0$; $l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$ is the local level; $b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t$ is the short-run trend in period t; b is the long-run trend; $d_t = \sum_{i=1}^{p} \psi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$; denotes an ARMA($p$,$q$) process, and ε_t is a Gaussian white noise process with zero mean and constant variance σ^2. The ith seasonal component at time t is given as: $s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}$, with $s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t$ and $s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t$. Furthermore, α, β, $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$, $i = 1, .., T$ are the smoothing parameters, $\lambda_j^{(i)} = 2\pi j / m_i$ and k_i is the number of harmonics required for the ith seasonal component.

The dataset provided by the company has a total of 8,760 hourly observations of water consumption of six residential consumers (see Fig. 1). For the analysis, the statistical software R [3] was used. The outliers of each of the six time series were replaced by

linearly interpolated values using the neighboring observations, in a total of 5.89%, 9.18%, 3.69%, 3.14%, 12.72% and 5.64%, for consumers 1, 2, 3, 4, 5 and 6, respectively. Since the data includes non-positive values, the inverse hyperbolic sine transformation is used.
The observation of box-plots of water consumption at week-days and weekends suggests differentiable client patterns and the existence of daily and weekly seasonal patterns with length $m_1 = 24$ and $m_2 = 168$, respectively. These are found in the estimated TBATS models of the six clients presented in Table 1. For all, no Box-Cox transformation was required ($w = 0$).For all the other parameters of the TBAST models there are differences.

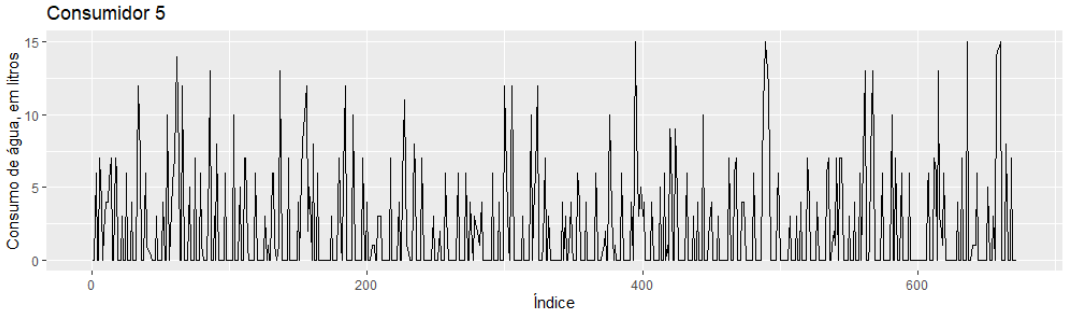

Figure 1: Hourly water consumption, in liters, of one of the clients.

Table 1: TBATS models for the six clients.

Client	p, q	m_1, k_1	m_2, k_2	α	$\gamma_1^{(1)}$	$\gamma_1^{(2)}$	$\gamma_2^{(1)}$	$\gamma_2^{(2)}$	σ	AIC
1	0, 0	(24,11)	(168,6)	0.017	0.0003	-0.0063	0.0052	-0.0009	0.94	3.183
2	0, 1	(24,10)	(168,6)	0.007	-0.0003	-0.0038	0.0042	-0.0015	0.88	3.150
3	0, 0	(24,7)	(168,6)	0.017	-0.0007	-0.0060	0.0033	0.0001	1.47	3.393
4	0, 0	(24,7)	(168,4)	0.001	-0.0019	-0.0049	0.0048	0.0024	1.19	3.279
5	0, 0	(24,5)	(168,6)	0.018	-0.0034	-0.0075	0.0037	0.0008	0.91	3.146
6	0, 0	(24,8)	(168,3)	0.043	-0.0022	-0.0049	0.0029	-0.0043	0.78	3.070

Acknowledgements THis work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through projects UIDP/04728/2020 and UIDB/04728/2020.

References

[1] A. M. de Livera, R.J. Hyndman, and R.D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106:1513–1527, 12 2011.

[2] E. Costa e Silva and F. Ferreira. Individualized monitoring and forecasting of water consumption: a preliminary study. In *AIP Conference Proceedings*. AIP Publishing LLC, 2022.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

22 April, 11:20 - 11:40, Hall of Library Barbosa Romero

Georeferenced analysis of vehicle-pedestrian collisions in Lisbon urban area from 2010 to 2019

Telma de Garção[1], **Nelson de Jesus**[2]
[1] Iscte-IUL, licenciatura em Ciência de Dados - 3.º ano telma_garcao@iscte-iul.pt
[2] Iscte-IUL, licenciatura em Ciência de Dados - 3.º ano nelson_carvalho_jesus@iscte-iul.pt

As the use of alternative travel modes increases, the analysis of the geographical distribution of traffic collisions with pedestrians becomes even more pertinent. With the use of a spatial approach, it becomes possible to assess contributing factors, such as location, creating room for more targeted measures that may lead to improvements in safety when considering vehicle-pedestrian interaction.

Keywords: traffic accidents, mobility challenge, geographic information, spatial analysis, Lisbon

This study was conducted as part of the curricular unit Applied Project in Data Science II, in the 3rd year of the Data Science Bachelor's Degree Course from ISCTE-IUL. The data used in this work was provided by the National Road Safety Authority (ANSR), the main source of information on road accidents in mainland Portugal. These data were gathered by security forces (GNR and PSP) and made available through the Road Traffic Statistical Report (BEAV). Road traffic security has always been a challenge in modern society. In the period between 2010 and 2019, there were 327 384 police-reported vehicle crashes, including 55 850 vehicle-pedestrian accidents, which resulted in 950 fatalities. After reading up on statistics produced by ANSR, a decision about the direction of this work was made. The importance of adding geographic information to data analysis became clearer. To this effect, accidents that took place in urban street settings were considered. The study's scope was reduced to the Lisbon metropolitan area and covered all its 24 municipalities. Three main goals were set:

1. Identification of sites, involving pedestrians, by municipality, from 2010 to 2019.
2. Identification of sites with higher rates of severe of injuries, also, from 2010 to 2019.
3. Finding locations with higher death tool, counting up to 30 days after the accident.

Latitude and longitude coordinates were used, and the spatial analysis was carried out by QGis® software. Figure 1 shows the geographical distribution of collisions, involving pedestrians, in the metropolitan area. Through a dataset analysis, municipalities as

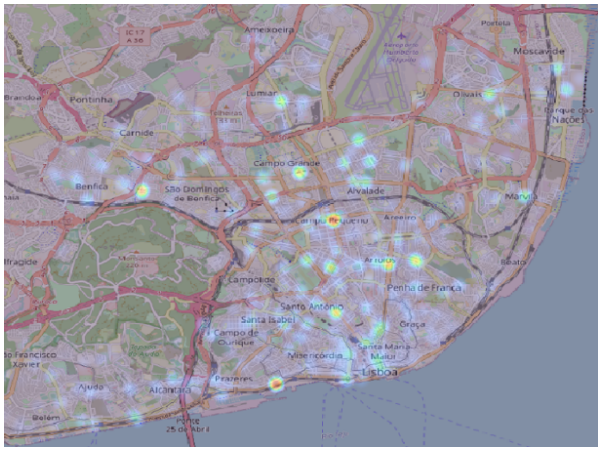

Figure 1: Distribution of traffic accidents in Lisbon urban area (2010-2019)

Misericórdia, Avenidas Novas, Arroios and São Domingos de Benfica scored high on the number of vehicle-pedestrian occurrences, from 2010 to 2019.
Locations where injuries of great severity were also identified. Among the 24 municipalities, the top-ranking hotspots were Olivais, Marvila, Alvalade, Avenidas Novas and Penha de França.
Upon integration with street lighting georeferenced data (https://lisboaaberta.cm-lisboa.pt/), the analysis time period was adjusted to 2016-2019. Overall, the northern urban area (including municipalities such as Marvila, Olivais, and also Penha de França) registered a higher number of fatalities, counting a period of 30 days after the occurrence.
In addition, the study allowed for spatial recognition of the most dangerous streets for vehicle-pedestrian interaction: Praça Paiva Couceiro and Avenida Infante D. Henrique.

Acknowledgments: to Professors Anabela Costa, Margarida Cardoso, and Fernando Batista, we thank their support, knowledge sharing, and valuable contributions.

References: QGis® software (https://www.qgis.org/en/docs/index.html).

Author Index

SPONSORS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


ipvc


ipvc estg


BANCO DE PORTUGAL
EUROSISTEMA


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


CÂMARA MUNICIPAL
VIANA DO CASTELO


CMAT
Centre of Mathematics
UNIVERSITY OF MINHO


FCT
Fundação
para a Ciência
e a Tecnologia


REPÚBLICA
PORTUGUESA


PSE
your data
specialists


CA
Crédito Agrícola