PROGRAMME and BOOK of ABSTRACTS

JOCLAD 2025

3 - 5 APRIL

PORTO, PORTUGAL

P.PORTO
ISCAP

CLAD
Associação Portuguesa de Classificação e Análise de Dados

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

Programme and Book of Abstracts

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

3–5 April 2025

Porto, Portugal

https://sites.google.com/view/joclad2025

Sponsors

Axis Porto Business & SPA Hotel
Banco de Portugal
Câmara Municipal de Matosinhos
CEOS.PP – Centro de Estudos Organizacionais e Sociais do Politécnico do Porto
Escola de Mergulho do Norte
Instituto Nacional de Estatística/Statistics Portugal
Instituto Superior de Contabilidade e Administração do Porto
Leya
PSE — Produtos e Serviços de Estatística

Organisers

Associação Portuguesa de Classificação e Análise de Dados (CLAD)
Instituto Superior de Contabilidade e Administração do Porto

Programme and Book of Abstracts

XXXII Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2025)

Editors: M. Rosário Oliveira, Bruno de Sousa, Jorge Cadima, Maria Eduarda Silva, Nuno Moniz, Paula Brito, Susana Faria

Publisher: CLAD

Printed: Statistics Portugal

ISBN 978-989-35097-2-2

Depósito legal: 544678/25

Number of copies: 125

Preface

Since its founding, the Portuguese Association for Classification and Data Analysis (CLAD) has played a vital role in fostering collaboration, innovation, and scientific exchange in Data Science. As a member of the International Federation of Classification Societies (IFCS), CLAD promotes research and knowledge dissemination across the fields of classification, data analysis, and statistics, building bridges between academia, public institutions, and industry. JOCLAD has established itself as an essential annual forum where data science professionals, researchers, and students come together to exchange knowledge, foster collaboration, and drive innovation in the field.

The 32nd edition of JOCLAD takes place at the Instituto Superior de Contabilidade e Administração do Porto (ISCAP), a distinguished school within the Polytechnic Institute of Porto. Renowned for its strong academic programs in accounting, auditing, and management, ISCAP fosters a dynamic learning environment with strong industry connections, making it a prime venue for JOCLAD 2025.

The scientific program of JOCLAD 2025 focuses on a broad spectrum of topics, including clustering, time series analysis, supervised learning, generalized linear models, and practical applications in data analysis, among others, with 25 oral contributed talks and 28 poster presentations. A specialized short course is offered by Mia Hubert and Peter Rousseeuw on methods for dealing with outliers in multivariate statistics, providing participants with valuable tools for robust data analysis. In addition, three keynote plenary lectures by distinguished international experts address the latest advances in robust dimension reduction (Mia Hubert and Peter Rousseeuw), generative artificial intelligence (António Branco), and high-dimensional regression (Angela Montanari), highlighting key developments in these evolving fields.

The thematic sessions at JOCLAD 2025 reflect the ongoing collaboration between CLAD and key institutions that contribute to the advancement of data science and its applications. We are pleased to feature a session from Banco de Portugal, which explores the insights gained from in-depth data analysis. Statistics Portugal addresses the evolving role of administrative data in the production of official statistics and evidence-based policy. The Portuguese Statistical Society leads a session focusing on methodological advances in analyzing complex spatial and temporal data.

In addition to these institutional contributions, the thematic sessions also feature perspectives from the corporate sector. Feedzai and NOS present innovative applications of machine learning and data analytics, highlighting both the challenges and opportunities faced by the industry in leveraging data-driven solutions.

CLAD's dedication to nurturing young talent continues with the award of six scholarships for students presenting high-quality work at JOCLAD 2025. These scholarships, granted to promising master's and doctoral candidates, reflect our commitment to supporting the next generation of researchers in the data science community. The jury for the CLAD Scholarships 2025 comprises M. Rosário Oliveira (University of Lisbon, Chair), A. Manuela Gonçalves (University of Minho), and João Gama (University of Porto).

In recognition of Professor Fernando Nicolau's role in founding the CLAD, the Fernando Nicolau Award was established to honor his legacy. Now in its third edition, this award recognizes outstanding contributions to the field of data classification and analysis, with the winner being publicly announced during this event. The jury for the award consists of Helena Bacelar Nicolau (University of Lisbon), Gilbert Saporta (CNAM-CEDRIC, Paris), Mário Figueiredo (University of Lisbon), and Paulo Gomes (NOVA IMS Information Management School).

Each abstract published in this volume has been evaluated by the Scientific Committee, composed by M. Rosário Oliveira, M. Eduarda Silva, Susana Faria, Jorge Cadima, and Nuno Moniz, whose work significantly contributed to the quality of the JOCLAD 2025 programme. We thank all the authors who submitted an abstract, as well as the session chairs. We express our gratitude to all sponsors for their generous contribution to JOCLAD 2025, with a special mention to our long-lasting partners, Statistics Portugal and Banco de Portugal. Finally, we convey our thanks to the members of the Local Organising Committee, Cristina Lopes, Cristina Torres, Isabel Vieira, Lurdes Babo, and João Cordeiro, whose enthusiasm and tireless work have made JOCLAD 2025 come true.

In addition to the scientific program, JOCLAD 2025 offers a rich and engaging social program. Participants have the opportunity to visit the historic Livraria Lello and savor the renowned excellence of Port wine. The social dinner, offered in a cosy space of Matosinhos city, near the Atlantic Ocean, feature a delightful selection of traditional Portuguese delicacies, providing a relaxing atmosphere for networking and cultural exchange. These moments of cultural immersion and social interaction are essential to the JOCLAD experience, strengthening the sense of community and collaboration beyond the academic sessions.

As we come together in Porto, JOCLAD 2025 reaffirms its mission to promote scientific exchange, collaboration, and the dissemination of knowledge in data science, classification, and statistical analysis. We warmly welcome all participants and look forward to three days of insightful discussions, knowledge sharing, and networking in this beautiful and inspiring city.

Porto, April 2025

Chair of the Scientific Programme Committee
M. Rosário Oliveira

Conference Chair
Cristina Lopes

President of CLAD
Paula Brito

Organisation

President of CLAD

Paula Brito

Chair of JOCLAD 2025

Cristina Lopes (Instituto Superior de Contabilidade e Administração do Porto, Instituto Politécnico do Porto)

Local Organising Committee

Cristina Lopes (ISCAP – IPorto & CEOS.PP)
Cristina Torres (ISCAP – IPorto & CEOS.PP)
Isabel Vieira (ISCAP – IPorto & CEOS.PP)
Lurdes Babo (ISCAP – IPorto & CEOS.PP)
João Cordeiro (UBI & LIAAD INESC-TEC)

Chair of the Scientific Programme Committee

M. Rosário Oliveira (IST – Universidade de Lisboa & CEMAT)

Scientific Programme Committee

M. Rosário Oliveira (IST – Universidade de Lisboa & CEMAT)
Bruno de Sousa (FPCE – Universidade de Coimbra & CINEICC)
Jorge Cadima (ISA – Universidade de Lisboa & CEAUL)
Maria Eduarda Silva (FEP – Universidade do Porto & LIAAD INESC-TEC)
Nuno Moniz (University of Notre Dame, USA)
Susana Faria (Universidade do Minho & CMAT)

Contents

Programme Overview **xi**

Programme **xv**

Abstracts **1**

Short Course **3**
Recent developments in robust multivariate statistics 5

Keynote Lectures **7**
Cellwise robust dimension reduction 9
Generative AI for the Portuguese language 11
Shrinking or screening: solutions for high dimensional regression 13

Thematic Session: Bank of Portugal **15**
Inequality and wealth distribution among Portuguese households: introducing the new Distributional Wealth Accounts 17
Enhancing data consistency: an NLP-based approach for harmonizing non-resident entities 19
Beyond distance: a data-driven assessment to cash accessibility in Portugal . . 21

Thematic Session: Statistics Portugal **23**
Estimation of free-riding in plastic package waste using placed-on-market and business turnover information 25
Improving the accuracy of administrative data on property transactions using a network algorithm 27
Clustering expenditure patterns: an application to the Portuguese Household Budget Survey 29

Thematic Session: CLAD 2025 Scholarships **31**
Are complex networks useful for generating synthetic time series? 33
Supervised statistical learning methods in the presence of spatial correlation . 35
Statistical analysis of event impacts: the case of the Feira São Mateus 37
Spatial imputation in rotating panel data: an analysis using stochastic partial differential equations (SPDE) with INLA 39
Analyzing topological gait descriptors for improved classification of parkinsonism 41

Thematic Session: CLAD Corporate **43**
Strategic integration of machine learning and artificial intelligence at NOS . . 45
Low-latency graph methods for fraud detection systems 47

Thematic Session: SPE **49**
Cross-correlation analysis to identify the drivers of phytoplankton biomass in Atlantic coastal bays 51
On the performance evaluation of algorithms for the identification of ARMA models 53
A hierarchical Bayesian geostatistical model for zero-inflated and extreme spatial data: analysing sardine egg density in Portugal 55

Contributed Sessions **57**
Detecting Airbnb host profiles with HiMC: the multilevel clustering methodology 59
Clustering for points of interest identification: insights from recent research . . 61
Real estate market dynamics in the Oporto municipality 63
Exploring voter turnout in Portuguese legislative elections through municipal profiling 65
Iterative GMM for bias reduction in state-space models 67
A model-based approach for clustering zero-inflated count time series 69
Clustering health data time series with the generalized affinity coefficient . . . 71
Optimizing energy use in agricutural irrigation systems: a data-driven approach for sustainable practices 73
A classification method based on a cloud of spheres 75
Randomly perturbed random forests 77
Hypothesis testing for goodness-of-fit in generalized partially linear models using projections 79
The effect of distress on the health and well-being of workers. PLSc SEM estimator 81
Digital transformation in Europe: insights from a multilevel multivariate probit regression model 83
Enhancing fuzzy forests with consensus clustering for unbiased and robust feature selection 85
ClustOfVar: global vs local standardization 87
Clustering density-valued data 89
A statistical comparison of external training load metrics during congested versus non-congested periods in football 91
Time series features and forecasting of community pharmacy sales 93
Towards a guide to include the social perspective in engineering programs: an international perspective 95
From data to stories: statistics and creativity in data journalism 97
Enhancing data quality in real-time environments: metrics and applications in digital industry 99
R&D in Portuguese companies 101
Classification of districts in Costa Rica using geospatial data 103

Applying the weighted aggregated sum product assessment method to the risk classification of sectors for greenhouse gas emission 105
Supervised machine learning methodologies for longitudinal data 107

Poster Sessions **109**
Predicting undergraduate dropout at the Polytechnic University of Coimbra . 111
Open government data: a global perspective with a focus on Portugal 113
Selection of variables influencing math scores in PISA data using LASSO and elastic net 115
Bitcoin price prediction with statistical and neural networks forecast models . 117
Forecasting Ibovespa: statistical *vs* neural models 119
The perceived value of cooperative membership in organic cocoa production: PLSc-SEM approach 121
Analysis of consumer perceived value in food products using PLSc-SEM 123
The role of social media and influencers in restaurant decision-making 125
Organisational climate and employee well-being 127
A support vector machine model for stock risk classification 129
Lab-grown diamond price estimates: an interpretation 131
Variable selection in low and high-dimensional mixtures of linear regression models 133
Trends in mean sea level and extreme events: particular case of Leixões 135
Titanium prostheses in middle ear reconstruction: a statistical analysis of audiometric outcomes 137
Statistical approach to establishing geochemical baselines for metal concentrations in marine sediments: a case study of Madeira Island 139
Production of papayas in an aquaponics system 141
Domestic violence and economic and social vulnerability 143
The influence of socio-economic conditions on public health 145
Weekend and holiday work by young people 147
Influence of physical readiness on fatigue variation during infantry officer training course - a case study 149
Comparative robustness of machine learning methods for genomic prediction . 151
Exploring dynamic neural field self-organizing maps for dimensionality reduction, visualization and classification 153
Assessing risk: an extremal inference methodology 155
Analyzing a method for estimating the tail index 157
Classification rules for a folded directional distribution 159
A comparative analysis of residential water consumption models and forecasting approaches 161
PAGEC: advancing attributed network analysis with joint embedding and clustering 163
Clustering, time series and risk analysis in surface water quality monitoring . . 165

Author Index **167**

Programme Overview

Thursday, 3 April – ISCAP Building

08:30	Registration	Hall of Grande Auditório
09:00	**Short Course – M. Hubert, P. J. Rousseeuw**	Room 13.2
10:30	Coffee Break	Restaurant of ISCAP
11:00	**Short Course** (cont.)	Room 13.2
12:30	Lunch Time	Restaurant of ISCAP
13:30	Registration	Hall of Grande Auditório
14:00	**Opening Session**	Grande Auditório
14:30	**Keynote Lecture I – M. Hubert, P. J. Rousseeuw**	Grande Auditório
15:30	Coffee Break	Restaurant of ISCAP
16:00	**Parallel Sessions I**	Room 13.1 & Room 13.2
19:00	**Visit to Livraria Lello**	Porto
19:30	**Porto de Honra**	Pólo Zero – Porto

Friday, 4 April – ISCAP Building

08:30	Registration	Hall of Grande Auditório
09:00	**Parallel Sessions II**	Room 13.1 & Room 13.2
10:10	**Parallel Sessions III**	Room 13.1 & Room 13.2
11:10	Coffee Break	Restaurant of ISCAP
11:30	**Poster Session I**	Hall of Grande Auditório
12:00	**Keynote Lecture II – A. Branco**	Grande Auditório
13:00	Lunch Time	Restaurant of ISCAP
14:30	**Thematic Session I – Bank of Portugal**	Grande Auditório
15:30	**Thematic Session II – Statistics Portugal**	Grande Auditório
16:30	Coffee Break	Restaurant of ISCAP
16:50	**Thematic Session III – CLAD 2025 Scholarships**	Grande Auditório
18:30	General Assembly of CLAD	Grande Auditório
20:30	**Social Dinner**	Casa do Ribeirinho – Matosinhos

Saturday, 5 April – ISCAP Building

09:00	**Parallel Sessions IV**	Room 13.1 & Room 13.2
10:10	**Thematic Session IV – CLAD Corporate**	Grande Auditório
11:10	Coffee Break	Restaurant of ESTG
11:30	**Poster Session II**	Hall of Grande Auditório
12:00	**Fernando Nicolau Award**	Grande Auditório
12:30	**ISCAP Hackathon Awards**	Grande Auditório
13:00	Lunch Time	Restaurant of ESTG
14:30	**Thematic Session V – SPE**	Auditório 1
15:30	**Keynote Lecture III – A. Montanari**	Auditório 1
16:30	**Closing Session**	Auditório 1

Programme

Thursday, 3 April

08:30 Registration – Hall of Grande Auditório

09:00 **Short Course** – Room 13.2
Recent developments in robust multivariate statistics
Mia Hubert and Peter J. Rousseeuw, p. 5
Chair: Adelaide Freitas

10:30 **Coffee Break**

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

12:30 **Lunch Time**

13:30 Registration – Hall of Grande Auditório

14:00 **Opening Session** – Grande Auditório

14:30 **Keynote Session I** – Grande Auditório
Cellwise robust dimension reduction
Mia Hubert and Peter J. Rousseeuw, p. 9
Chair: M. Rosário Oliveira

15:30 **Coffee Break**

16:00 **Parallel Sessions I**

	Room 13.1 **Clustering I** Chair: José G. Dias	Room 13.2 **Time series analysis** Chair: Maria Eduarda Silva
16:00	**Detecting Airbnb host profiles with HiMC: the multilevel clustering methodology**, Maria Gonçalves, Pedro Campos, p. 59	**Iterative GMM for bias reduction in state-space models**, Marco Costa, Magda Monteiro, p. 67
16:20	**Clustering for points of interest identification: insights from recent research**, Flora Ferreira, p. 61	**A model-based approach for clustering zero-inflated count time series**, Luís Sousa, Isabel Pereira, Magda Monteiro, p. 69
16:40	**Real estate market dynamics in the Oporto municipality**, Francisco Sousa Matos, Pedro Duarte Silva, p. 63	**Clustering health data time series with the generalized affinity coefficient**, Ana Paula Nascimento, Mónica Vieira, Brígida Mónica Faria, Alexandra Oliveira, Cristina Prudêncio, Helena Bacelar-Nicolau, p. 71
17:00	**Exploring voter turnout in Portuguese legislative elections through municipal profiling**, Fábio Coutinho, Joana Leite, p. 65	

19:00 **Visit to Livraria Lello** – Porto

19:30 **Porto de Honra** – Pólo Zero

Friday, 4 April

08:30 Registration – Hall of Grande Auditório

09:00 **Parallel Sessions II**

	Room 13.1 **Optimization and Statistical learning** Chair: Pedro Duarte Silva	Room 13.2 **Latent variables and Generalized linear models** Chair: Susana Faria
9:00	**Optimizing energy use in agricutural irrigation systems: a data-driven approach for sustainable practices**, José Brito, Conceição Rocha, Renato Fernandes, Pedro Guimarães, Filipe Silva, p. 73	**Hypothesis testing for goodness-of-fit in generalized partially linear models using projections**, Rui Costa-Miranda, Rita Gaio, Christian Heumann, Wenceslao González-Manteiga, p. 79
9:20	**A classification method based on a cloud of spheres**, Tiago Dias, Paula Amaral, p. 75	**The effect of distress on the health and well-being of workers. PLSc SEM estimator**, Luís M. Grilo, Helena L. Grilo, p. 81
9:40	**Randomly perturbed random forests**, Laura Anderlucci, Angela Montanari, p. 77	**Digital transformation in Europe: insights from a multilevel multivariate probit regression model**, José G. Dias, Lucas de Souza, p. 83

10:10 **Parallel Sessions III**

	Room 13.1 **Clustering II** Chair: Margarida G. M. S. Cardoso	Room 13.2 **Data analysis applications I** Chair: Cristina Lopes
10:10	**Enhancing fuzzy forests with consensus clustering for unbiased and robust feature selection**, Mouhamadou Lamine Ndao, Ndèye Niang, Genane Youness, Gilbert Saporta , p. 85	**A statistical comparison of external training load metrics during congested versus non-congested periods in football**, Paulo Barreira, Luísa Novais, Francisco Tavares, João Pedro Araújo, p. 91
10:30	**ClustOfVar: global vs local standardization**, Adelaide Freitas, Juliana Castanheira, Ana Aida Sá, p. 87	**Time series features and forecasting of community pharmacy sales**, Maria Inês Vicente, Joana Leite, p. 93
10:50	**Clustering density-valued data**, Rui Nunes, Paula Brito, Sónia Dias, p. 89	**Towards a guide to include the social perspective in engineering programs: an international perspective**, Teresa Barros, Alexandra Albuquerque, Inês Braga, Paula Carvalho, p. 95

11:10 **Coffee Break**

11:30 **Poster Session I** – Hall of Grande Auditório

Chair: Isabel Vieira

Predicting undergraduate dropout at the Polytechnic University of Coimbra
Marta Simões, Joana Leite, António Paulino, Isabel Pedrosa, p. 111

Open government data: a global perspective with a focus on Portugal
Inês Rocha, Clara Viseu, Manuela Larguinho, p. 113

Selection of variables influencing math scores in PISA data using LASSO and elastic net
Beatriz Silva, Susana Faria, p. 115

Bitcoin price prediction with statistical and neural networks forecast models
João Peixoto, Carlos Grilo, José Martins, p. 117

Forecasting Ibovespa: statistical *vs* neural models
Elysiario Santos, Carlos Grilo, José Martins, p. 119

The perceived value of cooperative membership in organic cocoa production: PLSc-SEM approach
Ibrahim Prazeres, Maria Raquel Lucas, Ana Marta-Costa, Pedro Damião Henriques, Luís M. Grilo, p. 121

Analysis of consumer perceived value in food products using PLSc-SEM
Eunice Venâncio, Maria Raquel Lucas, Ana Marta-Costa, Pedro Damião Henriques, Luís M. Grilo, p. 123

The role of social media and influencers in restaurant decision-making
Carla Henriques, Suzanne Amaro, Ana Desiderati, p. 125

Organisational climate and employee well-being
Vera Valente, Cláudia Amanajás, Hugo Carvalho, Cristina Lopes, Isabel Vieira, Lurdes Babo, Cristina Torres, p. 127

A support vector machine model for stock risk classification
Faustino Sachimuco, Gaspar J. Machado, Irene Brito, p. 129

Lab-grown diamond price estimates: an interpretation
Margarida G. M. S. Cardoso, Luís Chambel, p. 131

Variable selection in low and high-dimensional mixtures of linear regression models
Ana Moreira, Susana Faria, p. 133

12:00 **Keynote Lecture II** – Grande Auditório
Generative AI for the Portuguese language
António Branco, p. 11

Chair: João Cordeiro

13:00 **Lunch Time**

14:30 **Thematic Session I – Bank of Portugal** – Grande Auditório
What can we discover when we drill-down data further?

Chair: Luís Teles

14:30 **Inequality and wealth distribution among Portuguese households: introducing the new Distributional Wealth Accounts**, Gonçalo Amado, p. 17
14:50 **Enhancing data consistency: an NLP-based approach for harmonizing non-resident entities**, Carolina Costa, Tiago Pinho Pereira, p. 19
15:10 **Beyond distance: a data-driven assessment to cash accessibility in Portugal**, André Costa, Fábio Gomes, Hugo Cipriano, Pedro Cruz, Vânia Lopes, p. 21

15:30 **Thematic Session II – Statistics Portugal** – Grande Auditório
From data collection to decision making: the growing use of administrative data in Official Statistics

Chair: Pedro Campos

15:30 **Estimation of free-riding in plastic package waste using placed-on-market and business turnover information**, João S. Lopes, Filipa Chambel, Nuno Romão, p. 25

15:50 **Improving the accuracy of administrative data on property transactions using a network algorithm**, Alexandre Cunha, João Poças, Sofia Rodrigues, Paulo Saraiva, p. 27

16:10 **Clustering expenditure patterns: an application on the Portuguese Household Budget Survey**, Eduarda Góis, Maria Manuel Pinho, Cristina Gonçalves, Carla Afonso, p. 29

16:30 **Coffee Break**

16:50 **Thematic Session III – CLAD 2025 Scholarships** – Grande Auditório

Chair: M. Rosário Oliveira

16:50 **Are complex networks useful for generating synthetic time series?**, Jaime Vale, Vanessa Freitas Silva, Maria Eduarda Silva, Fernando Silva, p. 33

17:10 **Supervised statistical learning methods in the presence of spatial correlation**, Beatriz Ferreira, Raquel Menezes, p. 35

17:30 **Statistical analysis of event impacts: the case of the Feira São Mateus**, A. Catarina Gonçalves, Ana I. Melo, Marco Costa, p. 37

17:50 **Spatial imputation in rotating panel data: an analysis using stochastic partial differential equations (SPDE) with INLA**, Antonio Loría-García, Lígia Henriques-Rodrigues, Pedro Campos, p. 39

18:10 **Analyzing topological gait descriptors for improved classification of parkinsonism**, Jhonathan Barrios, Wolfram Erlhagen, Miguel Gago, Estela Bicho, Flora Ferreira, p. 41

18:30 General Assembly of CLAD – Grande Auditório

20:30 **Social Dinner** – Casa do Ribeirinho – Matosinhos

Saturday, 5 April

9:00 **Parallel Sessions IV**

	Room 13.1 **Data analysis applications II** Chair: Bruno de Sousa	Room 13.2 **Supervised Learning** Chair: Sónia Dias
9:00	**From data to stories: statistics and creativity in data journalism**, Cláudia Silvestre, Helena Figueiredo Pina, p. 97	**Classification of districts in Costa Rica using geospatial data**, Luis Eduardo Amaya Briceño, Erick Alfredo Vásquez Murillo, p. 103
9:20	**Enhancing data quality in real-time environments: metrics and applications in digital industry**, Eliana Costa e Silva, Óscar Oliveira, Bruno Oliveira, p. 99	**Applying the weighted aggregated sum product assessment method to the risk classification of sectors for greenhouse gas emission**, Irene Brito, p. 105
9:40	**R&D in Portuguese companies**, Lídia Maria Galvão Rodrigues Praça, p. 101	**Supervised machine learning methodologies for longitudinal data**, Elsa Soares, Inês Sousa, p. 107

10:10 **Thematic Session IV – CLAD Corporate: Feedzai, NOS** – Grande Auditório

Chair: Carlos Ferreira

10:10 **Strategic integration of machine learning and artificial intelligence at NOS**, Diogo Santos, p. 45

10:40 **Low-latency graph methods for fraud detection systems**, Jacopo Bono, p. 47

11:10 **Coffee Break**

11:30 **Poster Session II** – Hall of Grande Auditório

Chair: Lurdes Babo

Trends in mean sea level and extreme events: particular case of Leixões
Dora Carinhas, Pedro Rodrigues, Miguel Picoto, Paulo Infante, p. 135

Titanium prostheses in middle ear reconstruction: a statistical analysis of audiometric outcomes
Ana Matos, José Marques Santos, Javier Gavilan, p. 137

Statistical approach to establishing geochemical baselines for metal concentrations in marine sediments: a case study of Madeira Island
Dora Carinhas, Sandra Moreira, Anabela Oliveira, Carla Palma, Aurora Rodrigues, p. 139

Production of papayas in an aquaponics system
Fernando Sebastião, Judite Vieira, Luís Cotrim, Ounísia Santos, Maria Rodrigues, Daniela Vaz, Vânia Ribeiro, Raul Bernardino, p. 141

Domestic violence and economic and social vulnerability
Ana Ribeiro, Ana Martins, Daniela Fernandes, Juliana Barbosa, Lurdes Babo, Cristina Torres, Isabel Vieira, Cristina Lopes, p. 143

The influence of socio-economic conditions on public health
Patrícia Pinto, Ivanise Gomes, Jéssica Martins, Lurdes Babo, Cristina Torres, Isabel Vieira, Cristina Lopes, p. 145

Weekend and holiday work by young people
Jéssica Nadais, Catarina Ferreira, Isabel Gomes, Cristina Lopes, Lurdes Babo, Cristina Torres, Isabel Vieira, p. 147

Influence of physical readiness on fatigue variation during infantry officer training course – a case study
João Fonseca, Rui Lucena, André Fonseca, Nuno Almeida, Paula Simões, p. 149

Comparative robustness of machine learning methods for genomic prediction
Vanda M. Lourenço, Joseph O. Ogutu, Hans-Peter Piepho, p. 151

Exploring dynamic neural field self-organizing maps for dimensionality reduction, visualization and classification
Paulo Barbosa, Flora Ferreira, Estela Bicho, Wolfram Erlhagen, p. 153

Assessing risk: an extremal inference methodology
Marta Ferreira, Elisa Moreira, p. 155

Analyzing a method for estimating the tail index
Marta Ferreira, Liliana Monteiro, p. 157

Classification rules for a folded directional distribution
Adelaide Figueiredo, Fernanda Figueiredo, p. 159

A comparative analysis of residential water consumption models and forecasting approaches
Eliana Costa e Silva, Tatiana Cunha, Flora Ferreira, p. 161

PAGEC: advancing attributed network analysis with joint embedding and clustering
Lazhar Labiod, Mohamed Nadif, p. 163

Clustering, time series and risk analysis in surface water quality monitoring
A. Manuela Gonçalves, Irene Brito, Ana Pedra, p. 165

12:00 **Fernando Nicolau Award** – Grande Auditório

Chair: Paulo Gomes

12:30 **ISCAP Hackathon Awards** – Grande Auditório

Chair: Rita Gaio

13:00 **Lunch Time**

14:30 **Thematic Session V – SPE** – Auditório 1
Temporal and spatial modelling

Chair: A. Manuela Gonçalves

14:30 **Cross-correlation analysis to identify the drivers of phytoplankton biomass in Atlantic coastal bays**, Helena Mouriño, p. 51

14:50 **On the performance evaluation of algorithms for the identification of ARMA models**, Sónia Gouveia, p. 53

15:10 **A hierarchical Bayesian geostatistical model for zero-inflated and extreme spatial data: analysing sardine egg density in Portugal**, Soraia Pereira, Raquel Menezes, Maria Manuel Angélico, Tiago Marques, p. 55

15:30 **Keynote Lecture III** – Auditório 1
Shrinking or screening: solutions for high dimensional regression
Angela Montanari p. 13

Chair: Paula Brito

16:30 **Closing Session** – Auditório 1

Abstracts

Short Course

3 April, 9:00 - 10:30, 11:00 - 12:30, Room 13.2

Recent developments in robust multivariate statistics

Mia Hubert[1], **Peter J. Rousseeuw**[1]
[1] KU Leuven, Department of Mathematics, Belgium, mia.hubert@kuleuven.be, peter@rousseeuw.net

---

The topic of this short course is dealing with outliers in multivariate statistics. The first part is about discriminant analysis in the presence of outlying cases, with visualizations. The second part gives an introduction to cellwise outliers and how to detect them.

Keywords: cellwise outliers, classmaps, discriminant analysis, minimum covariance determinant, silhouette plot

---

In this short course we discuss two important topics in multivariate statistics: robust classification and cellwise outliers.
In the first part we study classification in the presence of outliers and mislabeled observations. Here, outlier refers to a case, that is, a row of the data matrix that behaves differently from the overall pattern. We study discriminant analysis based on robust estimates of location and scatter, such as the Minimum Covariance Determinant estimator (MCD). Next we introduce class maps and silhouette plots as graphical diagnostic plots for visualizing various aspects of classification results.
The second part addresses cellwise outliers. These are suspicious cells (entries) that can occur anywhere in the data matrix, and might not reveal themselves in the individual variables separately. We first describe the cellwise paradigm and address the detection of outlying cells. Then we look at a cellwise robust version of the MCD that can also deal with casewise outliers and missing values.

References

[1] J. Raymaekers and P. J. Rousseeuw. The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, 119:2610–2621, 2024.

[2] J. Raymaekers, P. J. Rousseeuw, and M. Hubert. Class maps for visualizing classification results. *Technometrics*, 64:151–165, 2022.

[3] P. J. Rousseeuw and W. Van den Bossche. Detecting deviating data cells. *Technometrics*, 60:135–145, 2018.

Keynote Lectures

3 April, 14:30 - 15:30, Grande Auditório

Cellwise robust dimension reduction

Mia Hubert[1], **Peter J. Rousseeuw**[1]
[1] KU Leuven, Department of Mathematics, Belgium, mia.hubert@kuleuven.be, peter@rousseeuw.net

---

We present two robust dimension reduction methods. The first is a principal component analysis (PCA) method that can simultaneously deal with casewise outliers, cellwise outliers, and missing cells. The second is an extension to multiway data.

Keywords: cellwise outliers, iteratively reweighted least squares, missing values, principal components, tensor data

---

The prevalence of ever larger datasets poses substantial challenges for statistical analysis. A common issue is the presence of outliers and missing data, caused by a variety of factors such as measurement errors or rare and unexpected events. Multivariate data are typically represented by a rectangular matrix in which the rows are the cases (objects) and the columns are the variables (measurements). Depending on the situation, outliers may be undesirable errors which can adversely affect the data analysis, or valuable nuggets of unexpected information. Either way we want to detect them. To date research was mainly focused on *casewise outliers*, also called rowwise outliers, which are cases not generated by the same mechanism as the clean data. Many robust methods have been developed for this framework. These casewise robust methods work by downweighting or deleting outlying cases, and require that at least 50% of the cases are clean.

In recent years also *cellwise outliers* have come in the spotlight. These are deviating entries (cells) that can occur anywhere in the data matrix. They are especially common in high dimensional data, because then there are many variables in which something can go wrong. A typical example is process control data recorded by many sensors, some of which are defective some of the time, whereas we want to extract the information in the clean cells. This resembles the missing data setting, but it is harder because we do not know which cells are outlying, especially because they need not reveal themselves when looking at the individual columns separately.

Even a relatively small proportion of outlying cells can contaminate over half the rows, which may cause casewise robust methods to fail. Therefore, other approaches are being developed that can deal with outlying cells. For a recent review of cellwise methods and their properties, see [5]. Most of these methods were developed for relatively low-dimensional settings. However, we also need methods that can deal with high dimensions. Classical PCA is strongly affected by both rowwise and cellwise outliers because it is a least squares method. In contrast, the MacroPCA method [2] was the first to address the issues

of rowwise outliers, cellwise outliers, and missing values simultaneously. However, it is a combination of elements from earlier methods and lacks a unifying underlying principle. Today we present the cellPCA method [1] which also deals with all three issues, and is the first method to do so by minimizing a single objective function, which makes it possible to study its properties. It combines two robust losses that effectively mitigate the effect of rowwise and cellwise outliers, while skipping missing cells. The method is computed by an iteratively reweighted least squares (IRLS) algorithm. We provide the casewise and cellwise influence functions of the estimator, as well as its asymptotic distribution. We also present enhanced graphical displays of outliers, and illustrate the method on real data.
Next we show how this methodology can be extended towards multiway or tensor data. Multilinear Principal Component Analysis (MPCA) is a new formulation for tensor decomposition, proposed in [4]. It extends classical PCA by working directly with the original tensor data and performs dimensionality reduction across all tensor modes, by identifying bases in each mode that enable the projected tensors to capture the maximum variation present in the data. Similar to PCA, MPCA is sensitive to both casewise and cellwise outliers as it relies on an alternating least squares algorithm. To alleviate this sensitivity, robust adaptations of MPCA have been proposed [3] but they can either deal with casewise outliers or with cellwise outliers, but not both. There does not yet exist an MPCA method that can simultaneously handle both types of outliers as well as missing values. We show how our new cellMPCA method addresses all these issues simultaneously.

References

[1] F. Centofanti, M. Hubert, and P. J. Rousseeuw. Robust principal components by casewise and cellwise weighting. *arXiv preprint arXiv:2408.13596*, 2024.

[2] M. Hubert, P. J. Rousseeuw, and W. Van Den Bossche. MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, 61(4):459–473, 2019.

[3] K. Inoue, K. Hara, and K. Urahama. Robust multilinear principal component analysis. In *2009 IEEE 12th International Conference on Computer Vision*, pages 591–597, 2009.

[4] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19(1):18–39, 2008.

[5] J. Raymaekers and P. J. Rousseeuw. Challenges of cellwise outliers. *Econometrics and Statistics*, to appear, https://doi.org/10.1016/j.ecosta.2024.02.002 , 2025.

4 April, 12:00 - 13:00, Grande Auditório

Generative AI for the Portuguese language

António Branco
Faculdade de Ciências da Universidade de Lisboa, Portugal,
ambranco@ciencias.ulisboa.pt

This talk will provide an updated overview of the current development of Large Language Models (LLMs) of generative Artificial Intelligence (gen AI) for the Portuguese language that are widely open, i.e. that are open source, and openly distributed for free under a most open license, and furthermore are scientifically validated.

The crucial importance of open AI for the democratization of AI will be discussed, including its key societal and strategic impacts, as well as the actual relevance of the open LLMs, which tend to be unfairly played down usually on the basis of equivocation about their smaller size and respective capacities and usefulness.

Keywords: generative AI, Large Language Models (LLMs), Portuguese language, democratization of AI

This lecture aims to provide an updated overview of the development of Large Language Models (LLMs) for Generative AI, specifically for the Portuguese language. It will focus on open models, meaning open-source AI models that are freely distributed under open licenses, with the goal of making artificial intelligence accessible to all. Furthermore, the lecture will discuss the scientific validation of these models, which is essential to ensure their accuracy, efficiency, and applicability across various fields.

One of the central themes of the lecture will be the growing importance of open AI and how it plays a crucial role in the democratization of artificial intelligence. Unlike proprietary AI models, which are often controlled by large corporations with restricted access, open models provide equal opportunities for individuals, researchers, developers, and organizations of all sizes. This accessibility facilitates the creation of innovative solutions, the advancement of scientific research, and the ethical and inclusive application of AI.

The lecture will also explore the social and strategic impacts of open LLMs, discussing how their free distribution can transform various sectors of society, such as education, healthcare, science, business, and the creative industries. Open models help reduce barriers to access to knowledge and technology, enabling more people to use and develop AI-based applications, regardless of their location or socioeconomic context.

Another key point will be the analysis of public perception regarding open LLMs, which are often underestimated or undervalued compared to larger proprietary models with greater visibility. A significant part of this undervaluation is based on the misconception that

smaller models with fewer parameters are less effective or useful than their larger counterparts. This stigma is often driven by a lack of understanding of the real capabilities of these models, which, despite being smaller in size, can be highly effective, efficient, and highly applicable in various contexts, particularly in the Portuguese language, where access to large-scale specialized models is still limited.

The lecture will also address the fundamental role that open LLMs play in strengthening communities of developers and researchers, creating a collaborative ecosystem where everyone can contribute to the advancement and improvement of the technology. Unlike closed solutions, which hinder collaboration, open models encourage collective innovation and allow for the adaptation and customization of models to meet specific local, cultural, and linguistic needs, such as those of the Portuguese language.

5 April, 15:30 - 16:30, Auditório 1

Shrinking or screening: solutions for high dimensional regression

Angela Montanari
University of Bologna, Italy, angela.montanari@unibo.it

It is well known that, in multiple linear regression, the least squares estimates of the regression coefficients are not unique when p (the number of variables) is larger than n (the number of units) and even when, for p slightly smaller than n, a unique solution to the least squares problem exists, the estimates can be really unstable making inference completely unreliable. In this talk we will review different solutions that have been proposed in the statistical literature to tackle these issues and propose some new ones. This presentation is a joint work with Laura Anderlucci, Matteo Farnè and Giuliano Galimberti from the University of Bologna.

Keywords: linear regression, random projection, variable screening, regularized covariance estimation

---

It is well known that, when dealing with high dimensional data, most of the classical multivariate methods cannot be applied or give unreliable results and it is known as well that, when the number of observed variables p is large, the relevant information may be contained in an lower-dimensional subset of the observed variables.

In the context of multiple linear regression this means that, for high dimensional data, the least squares estimates may be non unique and thus regularization approaches or sparsity assumptions on the vector of coefficients need to be adopted, in order to obtain reliable estimates of the model parameters.

Starting from the well known ridge regression approach, many regularized solutions have been proposed. We review a few of them and focus on a new proposal [5] that explored the possibility to replace the sample covariance matrix of the predictors by a regularized one, obtained under the assumption that the true covariance matrix of the predictors follows a low rank plus sparse structure [4].

From a different perspective, high-dimensionality can be dealt with by resorting to variable selection. In this framework the ordinary approach based on stepwise methods has turned out to produce very unstable results and new alternative solutions have recently appeared in the literature. The problem, for instance, has been addressed by either directly applying l_1 norm regularization to the original data [7] or by screening the variables to identify the most relevant ones and then applying an l_1 penalty to the selected subset [3]. The reasons

for this two-step approach lie in the high computational load inherent in the penalized approach.

Here we describe a new method [1] for variable screening and selection in multiple linear regression which is based on random projections. The use of random projections to reduce the dimensionality of a data set is becoming increasingly popular in the multivariate statistical literature. The common trait of the most effective solutions consists in randomly combining the p columns of the data matrix X, thus mapping the data onto a random d-dimensional (with $d \ll p$) subspace on which classical analyses can be performed. The results obtained on different random projections are then summarized by ensemble methods in order to obtain the final estimates.

We present some novel results exploiting axis aligned random projections and compare our solution with the one in [6] which is still based on random projections but differs in the way the members of the ensemble are selected.

References

[1] L. Anderlucci, M. Farnè, G. Galimberti, and A. Montanari. Sparse linear regression via random projections ensembles. *Preprint*, 2025.

[2] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680, 2013.

[3] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

[4] M. Farnè and A. Montanari. A large covariance matrix estimator under intermediate spikiness regimes. *Journal of Multivariate Analysis*, 176:104577, 2020.

[5] M. Farnè and A. Montanari. High-dimensional regression coefficient estimation by nuclear norm plus l_1 norm penalization. *Stat*, 12(1):e548, 2023.

[6] Y. Tian and Y. Feng. Rase: A variable screening framework via random subspace ensembles. *Journal of the American Statistical Association*, 118(541):457–468, 2023.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Thematic Session
Bank of Portugal

4 April, 14:30 - 14:50, Grande Auditório

Inequality and wealth distribution among Portuguese households: introducing the new Distributional Wealth Accounts

Gonçalo Amado
Banco de Portugal, gfamado@bportugal.pt

This article examines Distributional Wealth Accounts (DWA) for Portugal, which combine national accounts and survey data to provide internationally comparable insights into household wealth distribution. In last years, Portuguese households experienced a rise in net wealth accompanied by decreases in inequality, though absolute changes were driven mainly by the top 10%. Debt reduction boosted net wealth for the bottom half, while asset appreciation benefited richer households. On average, the wealthiest 10% hold over nine times the wealth of the poorer half, reflecting stark differences in balance sheet composition.

Keywords: inequalities, wealth distribution, national accounts, Portuguese households

This article presents the experimental Distributional Wealth Accounts (DWA) published by the ECB, a novel high-frequency dataset on household wealth consistent with National Accounts statistics. By linking Quarterly Sector Accounts with household survey data, DWA provide new insights on the growth of household wealth on a quarterly basis. The article provides details on the DWA for Portugal, discussing the evolution of household wealth and inequality between 2010Q2 and 2024Q2.

The data indicate that the overall increase in the net wealth of Portuguese households over the past decade has been accompanied by a reduction in inequality. This can be seen by the reduction in the Gini coefficient. The wealth share of the poorest 50% of the population has been rising slightly since 2016 (from 5% to 8% in 2024Q2). Also, the ratio of net wealth between the top 10% and the bottom 50% has been decreasing in the recent decade. One of the main drivers of the reduction in this ratio, and therefore in inequality, is the continued reduction in the mortgage debt of the poorest half of the population.

However, for the period under analysis, this ratio had an average value of 9.76. This means that, on average, the wealthiest 10% are more than nine times richer than the combined poorer half of the population. DWA data highlights that the wealthiest 10% of the population holds more than half (57%) of total net wealth in Portugal at the end of the reference period.

Another meaningful result is the overall increase in absolute value of the net wealth being driven mainly by the 10% richest decile. Figure 1 shows that the quarterly changes in the total net wealth are guided by the top 10% (yellow columns).

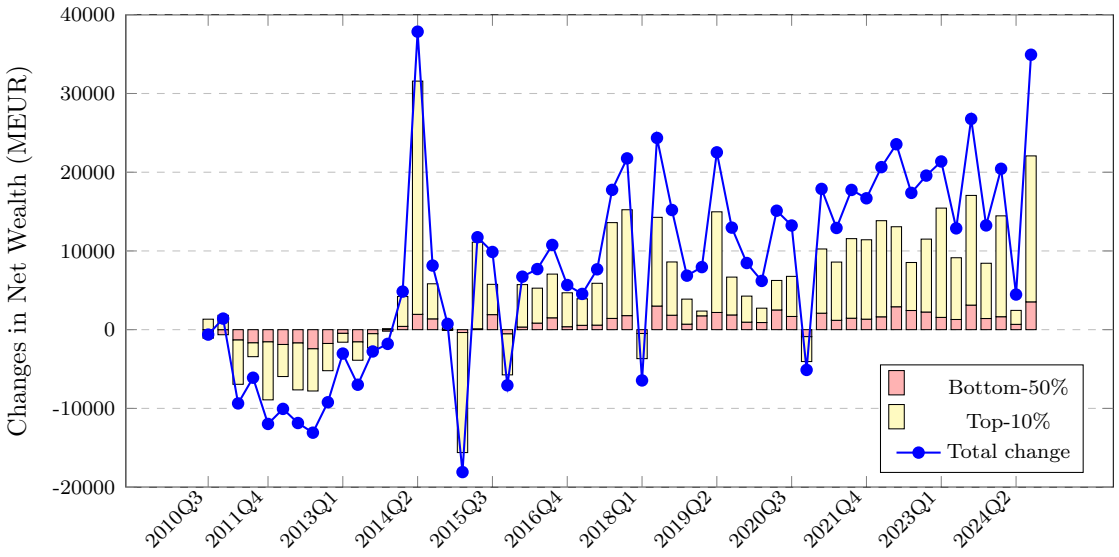

Figure 1: Quarterly Changes in Net Wealth, Portuguese households (2010Q3 - 2024Q2)

Looking only to top 10%, changes in net wealth were due to financial assets increases (mainly equity and deposits) and, more recently, by housing wealth positive changes. The total assets of the richest have risen steadily over the years, while their liabilities have remained relatively stable. Therefore, the debt-to-assets ratio (another indicator available in the database) decreased continuously.
While the net wealth of the bottom 50% of the population has increased, the quarterly changes, in absolute terms, are less significant compared to those of the top 10%. However, in relative terms, the wealth growth for the former group remains more substantial. The quarterly variations in the net wealth of the poorest half of the population primarily stem from reductions in mortgage debt, with a smaller contribution from increases in housing wealth. The overall rise in real estate prices in Portugal is reflected across all population deciles, contributing to a general increase in housing assets.
Over the last ten years, the difference between the relative growth in wealth of the poorest half and the richest top 10% in Portugal has been positive and has only been greater in Ireland, compared to other European countries. In this two countries, this difference increased mainly due to the reduction in mortgage debts of the poorest 50%. The same happened, in a less significant level, to the other countries with available data. This reduction in mortgage debt could be due either to debt amortization or to households moving to higher wealth deciles as a result of the general increase in housing prices.

References

[1] Marco Moreno. *Inequality and wealth distribution among Irish households: introducing new Distributional Wealth Accounts*. Central Bank of Ireland, 2024.

4 April, 14:50 - 15:10, Grande Auditório

Enhancing data consistency: an NLP-based approach for harmonizing non-resident entities

Carolina Costa[1], **Tiago Pinho Pereira**[1]
[1] Banco de Portugal, cgcosta@bportugal.pt, tppereira@bportugal.pt

This paper presents a tool designed to detect and correct errors and inconsistencies in the identification of non-resident entities, ensuring harmonization and temporal consistency. Tested across seven countries, the method achieves over 90% accuracy and can be adapted for other applications, such as standardizing categorical variables like addresses.

Keywords: firm identifiers, NLP models, IES database, RIAD, iBACH database

Every year, non-financial firms report their business group structure through Informação Empresarial Simplificada (IES). From this information, the Central Balance-Sheet Database of Banco de Portugal develops a business group database which contains resident and non-resident entities and the equity participation between them. Regarding non-resident entities, evidence shows that declarant firms not always report their name and tax number in the same way which leads to inconsistencies in their identification and generates duplicates for the same non-resident entity.

Business group structure data constitutes a relevant source of information used for several purposes, including risk analysis, integration of the data into a reference database, and supporting economic analysis from other departments and researchers. However, institutions responsible for processing and maintaining these databases face considerable challenges in ensuring its quality and consistency, particularly in the context of categorical data and human language.

To address these challenges, comparison with other databases such as Register of Institutions and Affiliates Database (RIAD) and iBACH database as well as Natural Language Preprocessing (NLP) models are employed to assess text similarity, thus facilitating the identification of identical entities even when represented differently across the dataset.

This paper outlines the ongoing development of a tool designed to detect and correct errors and inconsistencies in the non-resident entities database, ensuring its harmonization and temporal consistency. This tool is implemented using Python and SQL languages and operates through two main steps:

- **Treating errors and inconsistencies**
  In this stage, we apply data cleaning and transformation techniques to deal with issues such as syntax errors in the names of entities (e.g., incorrect capitalization, special characters, articles, and legal forms). The process also includes the detection of anomalous values and duplicates.

- **Defining a unique name and identifier**
  Due to the linguistic particularities of each country, this part of the analysis is performed on a country-by-country basis rather than on the entire database. However, the target is to expand this analysis to all countries represented in the data. This step focuses on diagnosing and treating five key inconsistencies: (1) Identifier according to the Value Added Tax (VAT) number format; (2) Invalid identifier; (3) Duplicates; (4) Identifiers with more than one associated name; (5) Names with more than one associated identifier. To handle these inconsistencies, we follow an 11-step procedure centered on an NLP algorithm called RapidFuzz, which evaluates string similarity using a ratio based on the Levenshtein Distance. Additional control measures are also implemented to minimize errors and enhance accuracy.

This method has been tested and validated across datasets from seven countries (Spain, France, Italy, Slovakia, Poland, Netherlands and Norway), yielding positive results. Overall, the success rates are above 90%, demonstrating the tools' effectiveness.

Table 1: Before and after of a fictional company in the database

	Identifier	**Name**
Before	FR20332710563	Saint Lucas Recording
Before	FR2710563	Saint Lucas Recording
Before	FR20332710563	SAINT LUCAS RECORDING
Before	FR20332710563	Saint Lucas Recording, S.A.
After	FR332710563	SAINT LUCAS RECORDING

When applying this method, one should keep in mind that its computational demands are high, especially when processing substantial volumes of data. Additionally, certain cases may require detailed analysis, and linguistic variations across countries can introduce further complexity. Despite these challenges, the method is highly versatile and valuable for information processing. It can also be easily adapted to other circumstances, such as the standardization and treatment of other categorical variables like addresses.

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

4 April, 15:10 - 15:30, Grande Auditório

Beyond distance: a data-driven assessment to cash accessibility in Portugal

André Costa[1], **Fábio Gomes**[1], **Hugo Cipriano**[1], **Pedro Cruz**[1], **Vânia Lopes**[1]
[1] Banco de Portugal, anfcosta@bportugal.pt, fegomes@bportugal.pt, hcipriano@bportugal.pt, pfcruz@bportugal.pt, vilopes@bportugal.pt

This study focuses on the spatial distribution of cash access points across Portugal, analysing its adequacy relative to population needs through a two-dimensional assessment. It integrates sociodemographic factors — including tourism intensity and point of sale usage — with geospatial data measuring distances between people and cash access points. Using coordinate mapping and statistical analysis, we developed an interactive dashboard that illustrates these relationships, revealing significant disparities across regions, ultimately providing valuable insights for the management of the cash access point network regarding coverage and efficiency.

Keywords: cash access, spatial analysis, sociodemographic analysis, data visualization

Maintaining the supply of cash is one of Banco de Portugal's critical functions. Despite growing digital alternatives, recent 2024 data shows that cash remains a fundamental payment method in the Euro area where 52% of point of sale transactions are conducted in cash, with Portugal exhibiting a slightly higher rate of 54% [1]. Furthermore, cash accessibility remains a crucial aspect of financial inclusion [2].

This study extends traditional distance-based analyses of cash access points across Portugal by examining their adequacy relative to the population needs. By adopting a two-dimensional analytical approach, we integrated geographic distances with sociodemographic variables as proxies of cash demand to assess the adequacy of the infrastructure [3].

Our methodology leveraged postal codes as population proxies to comprehensively analyse cash accessibility patterns. For each postal code, using euclidian distances, we identified the three nearest cash access points for two groups: ATMs (generally available 24/7) and banks and payshops (time-restricted services). Furthermore, we enhanced our analysis using road network calculations to determine actual driving distances and times since terrain features such as mountains, rivers, and coastal areas can significantly increase travel time.

Additionally, we developed demand indicators for cash using Statistics Portugal data. These indicators aimed to profile areas by cash preference levels (based on age and education), pressure demand on access points (including tourism effects) and point of sale eletronic payment availability [4].

This multidimensional demographic profile enables a deeper understanding of each region's specific cash access requirements beyond pure geographical coverage.
This analysis culminated in an interactive dashboard, showcased by the figures below, enabling stakeholders to dynamically explore both geographic and demographic dimensions simultaneously, highlighting critical areas thus providing valuable feedback on network adequacy for decision-makers.

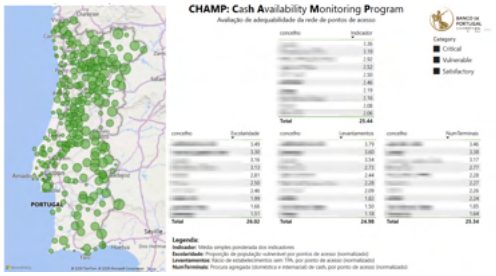

Figure 1: Sociodemographic overview

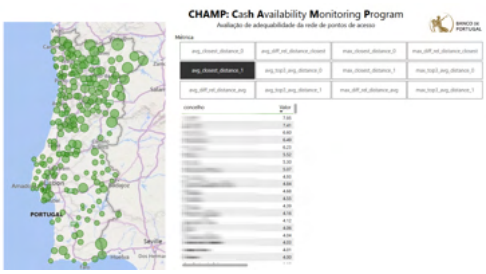

Figure 2: Distance overview

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.

Acknowledgements We express our gratitude to the organisers of the 2nd Banco de Portugal Datathon for challenging us into this compelling project, and to the Issue and Treasury Department for providing invaluable data access and enriching business insights.

References

[1] European Central Bank. Study on the payment attitudes of consumers in the euro area (SPACE), 2024.

[2] D. Posada Restrepo. Cash infrastructure and cash access vulnerability in Spain. *Banco de Espana Article*, 23:21, 2021.

[3] A. Zamora-Pérez. Guaranteeing freedom of payment choice: access to cash in the euro area. *Economic Bulletin Articles*, 5, 2022.

[4] L. Hernández and H. Esselink. The use of cash by households in the euro area. *ECB Occasional Paper Series*, 2017.

Thematic Session
Statistics Portugal

4 April, 15:30 - 15:50, Grande Auditório

Estimating free-riding in plastic package waste using placed-on-market and business turnover information

João S. Lopes[1], **Filipa Chambel**[1], **Nuno Romão**[1]
[1] Statistics Portugal, joao.lopes@ine.pt, filipa.chambel@ine.pt, nuno.romao@ine.pt

We present a study to calculate the annual amount of undeclared plastic packages placed-on-market (i.e. free-riding). The methodology uses information on annual business turnover and reported placed-on-market values. Using this information, we estimate an average amount of packaging placed-on-market per turnover. The businesses are clustered by similar characteristics like activity field. Then, for each cluster, a hypothetical waste generation can be calculated.

Keywords: plastic packages, placed-on-market, free-riding, business turnover

Since January 1 2021, a new source of revenue for the European Union (EU) budget has been created based on waste generated from non-recycled plastic packaging (plastic own resource).
The system of own resources, including the plastic own resource, was established by the EU for the period 2021-2027 (Decision 2020/2053) [2]. EU Council Regulation 2021/770 of April 30 2021 [3] established the basis for the application of the plastic own resource, which is calculated in accordance with Article 6a of Directive 94/62/EC [4] and the methodology set out in Decision 2005/270/EC [1], in particular Article 6 thereof. This revised methodology must be applied by Member States from reference year 2020 onwards.
Like most EU member states, Portugal calculates the generation of plastic packaging waste based on the placed-on-the-market (PoM) approach. Since 2018, reporting the PoM of all packaging, including plastic packaging, on the information platform of the Portuguese Environment Agency (APA) has been mandatory.
Due to the potential for non-declaration of plastic packaging PoM (free-riding), a methodology was developed to estimate free-riding based on the turnover (TO) of businesses based in Portugal. The information on the TO of businesses based in Portugal comes from the Integrated Company Accounts System calculated by Statistics Portugal (INE). This system integrates statistical information from businesses and administrative data, namely from the Simplified Business Information (IES), complemented with data on self-employed workers received through protocols with the Tax Authority (AT).
In this study, we present the methodology for estimating free-riding, which is based on calculating the ratio between the PoM and the TO of the reporting companies. Firstly, businesses are grouped by similar characteristics (i.e. Classification of Economic Activity, CAE; and Number of People Employed, NPS). Secondly, for each grouping, the average amount of PoM per TO of the reporting companies is estimated. Thirdly, the hypothetical

non-declared PoM is calculated by multiplying this average amount by the TO of the non-declarant companies. Fourth, the hypothetical total PoM can be calculated by adding the declared PoM and the estimated undeclared PoM. Fifth, the proportion of undeclared PoM in the hypothetical total can be indicative of the presence of free-riding within the specific grouping.

Table 1: Estimation of total plastic packages PoM for the years 2018-2022

	2018	2019	2020	2021	2022
Reported	225.5	232.8	226.3	229.3	237.3
Free-ride	38.9 (14.7%)	35.2 (13.2%)	44.6 (16.5%)	47.1 (17.0%)	24.5 (9.3%)
Total	264.4	268.0	270.9	276.4	261.8

Following the estimation of annual free-riding for specific groups of businesses from 2018 to 2022, we calculated the total plastic packages PoM by considering both reported and unreported values (Table 1). Our results are consistent with estimations across EU (personal communication), but there is space for improvement in future work. Namely, in the choice of grouping of businesses with potential of free-riding, and in the decision of presence/absence of free-riding within specific groupings.

References

[1] European Commission. Commission decision of 22 march 2005 establishing the formats relating to the database system pursuant to directive 94/62/ec of the european parliament and of the council on packaging and packaging waste. *Off J EU*, L 86/5.4:325–334, 2005.

[2] Council EU. Council decision (eu, euratom) 2020/2053 of 14 december 2020 on the system of own resources of the european union and repealing decision 2014/335/eu, euratom. *Off J EU*, L 424:1–10, 2020.

[3] Council EU. Council regulation (eu, euratom) 2021/770 of 30 april 2021 on the calculation of the own resource based on plastic packaging waste that is not recycled, on the methods and procedure for making available that own resource, on the measures to meet cash requirements, and on certain aspects of the own resource based on gross national income. *Off J EU*, L 165/11.5, 2021.

[4] European Parliament. European parliament and council directive 94/62/ec of 20 december 1994 on packaging and packaging waste. *Off J EU*, L 365/31.12:10–23, 1994.

4 April, 15:50 - 16:10, Grande Auditório

Improving the accuracy of administrative data on property transactions using a network algorithm

Alexandre Cunha[1], **João Poças**[1], **Sofia Rodrigues**[1], **Paulo Saraiva**[1]
[1] Statistics Portugal, alexandre.cunha@ine.pt, joao.pocas@ine.pt, sofia.rodrigues@ine.pt, paulo.saraiva@ine.pt

---

Real estate administrative data offers great possibilities for the production of official statistics but also many challenges such as incomplete or inconsistent records. An edge-attributed network algorithm was developed to identify which records represent real transactions, which are incomplete or inconsistent data, and to assign an unique identifier linking records to a single property transaction. This approach can help improve the accuracy of property transaction identification, supports better data integration, and facilitates more efficient tracking and management of property ownership records.

Keywords: edge attributed networks, weight-constrained networks, administrative data, data consistency

---

The integration of administrative data into official statistics offers substantial benefits, including resource optimization, reduced processing time, and enriched datasets with broader coverage. However, these datasets often pose challenges due to their design, which prioritizes administrative functions over statistical rigor. The real estate datasets provided to Statistics Portugal exhibit quality issues, frequently containing provisional or incomplete records that complicate the accurate aggregation and reconciliation of information at the property level [1].

To address these limitations, each property is represented as a graph, with owners as nodes and transaction attributes as edges, where the percentage of the transaction defines the edge weight. Household ownership administrative data and historical transaction records are incorporated into the model, creating more robust graphs. Initial outlier detection is applied to remove records likely to generate invalid edges.

Constraint rules are based on network relationships and edge weights. Nodes are prioritized by ownership status and their role within the graph; for example, if a node represents a future seller, its edges are given higher priority. An edge is considered valid if the cumulative sum of the its node and all its neighboring nodes does not exceed 100 percent. Nodes' edges are ordered by transaction date, with more recent transactions assigned higher priority and, older edges of the same node, consider invalid.

Nodes that fail predefined validation criteria, such as network strength, degree, or transaction consistency, undergo a more complex analysis. This involves generating all possible

cumulative sum combinations within the graph to identify inconsistencies and ensure coherence in property transactions. This step aims to consider as valid older data over more recent data if it better satisfies the model's constraint rules.
The output of the algorithm determines whether each edge is classified as a valid or invalid component of a property transaction. Additionally, an unique identifier is assigned to each property transaction, allowing for clear differentiation and traceability of transactions within the dataset. The process of transaction identification leverages graph theory principles by calculating the shortest path from designated starting nodes to every other node within the graph. The path length is used as a distinguishing factor to separate and classify different transactions associated with the same property. This ensures that the transaction structure is properly captured and aligned with the underlying network relationships.
The results are then systematically compared to those derived from current production methods, enabling an assessment of the algorithm's accuracy and precision. By evaluating the model against established benchmarks, the approach ensures that the new method not only aligns with existing standards but takes advantage of the principles of graph theory to enhance the reliability and interpretability of real estate transaction data in official statistics.

References

[1] R. Evangelista and Â. Teixeira. Using different administrative data sources to develop house price indexes for Portugal. pages 3–6, 2014.

4 April, 16:10 - 16:30, Grande Auditório

Clustering expenditure patterns: an application to the Portuguese Household Budget Survey

Eduarda Góis[1], **Maria Manuel Pinho**[1], **Cristina Gonçalves**[1], **Carla Afonso**[1]
[1] Statistics Portugal, eduarda.gois@ine.pt, mmanuel.pinho@ine.pt, cristina.goncalves@ine.pt, carla.afonso@ine.pt

We apply the clustering methodology to the Portuguese HBS 2022/2023 data to enable the characterization of households' expenditure profiles. We conclude that high-income households are in a small number and better represented by large size families with children, owning their home, allocating a significant share of their expenditure to education. As income decreases, the larger the homogeneous groups of income become, allocating a higher share of their expenditure to food, housing and transport.

Keywords: household, expenditure, cluster analysis, household budget survey

The national Household Budget Survey (HBS) is part of the European HBS project and is in line with the recommendations agreed between Eurostat and the Member States. Its main goal is the five-year calculation of the structure of household expenditure on goods and services according to COICOP (Classification of Individual Consumption by Objective), contributing to the updating of the weights of the Consumer Price Index, to the estimation of the National Accounts private consumption and to approximating the population diet using the quantities of food purchased. It was launched at the national level in the late 1960s, and since then there have been 10 rounds, being one of Statistics Portugal's most consolidated large-scale statistical, targeting a representative sample of the population living in the national territory, stratified regionally (NUTS 2 level).
In addition to the survey results available at Statistics Portugal and open to the public, the anonymised data files, available for research purposes, can be used for detailed analyses on the allocation of Portuguese households' expenditure. Our strategy is to, first, apply the clustering methodology to the 2022/2023 data (the most recent HBS wave) to enable the characterization of the current expenditure profiles and then, using an exploratory descriptive analysis, compare the 2022/2023 data with the two previous rounds (2010/2011 and 2015/2016) in order to identify possible significant changes in the expenditure pattern of Portuguese households.
The main aim of our work is to create homogeneous groups of households according to their expenditure pattern, using a cluster analysis able to produce, as Cormack [1] suggests, internally cohesive and externally isolated groups, which requires homogeneity within clusters and heterogeneity between clusters. The grouping variables are the previously standardised monetary expenditures (in euros) for the COICOP first level (13 Divisions), with the

application of the survey weighting. Excluding non-monetary expenditure from the grouping variables implies disregarding self-consumption, income in kind on employment and owner-occupied imputed housing rents.
We decided to exclude Division 13, Personal care, social protection and miscellaneous goods, since it has a residual component and does not bring any discriminatory power. We also decided to aggregate Division 09, Recreation, sport and culture, and Division 11, Restaurants and accommodation services, since they capture to some extent similar consumption purposes. In line with the classification of housing acquisition as an investment expenditure, we also decided to remove from the grouping variables the purchase of vehicles (Group 071) and of recreative durable goods (Class 0912).
Given the high number of HBS 2022/2023 cases (11,701 households in the sample), a non-hierarchical classification technique is used, the K-means method, consisting of transferring an individual to the cluster whose centroid is at the shortest distance. According to Steinley [2], the K-means method is the most popular non-hierarchical clustering and is particularly useful for large datasets as hierarchical techniques are far more computationally demanding. To achieve the best differentiation and interpretative interest regarding the expenditure profile of the households, our choice is to produce four clusters (K=4).
According to the final results of the 2022/2023 HBS round for Portugal, the mean annual household expenditure in 2022/2023 was €23,900, about 2/3 of which were concentrated in housing (39.3%), food (12.9%) and transport (12.1%). The results suggest that households with dependent children spend annually on average €9,731 more than households without dependent children, which leads to a mean monthly expenditure €811 higher. This difference extends to all COICOP divisions. Households in the top income quantile (20% of households with the highest incomes) spent more than twice as much as households in the bottom income quantile (20% of households with lower incomes).
Based on the K-means results, we conclude that high-income households are few in number and are better represented by larger families with children, who own their homes and allocate a significant portion of their expenditure to education. As income decreases, the income groups become larger and more homogeneous, with a higher share of their expenditures allocated to food, housing, and transportation.

References

[1] R. M. Cormack. A review of classification. *Journal of the Royal Statistical Society, Series A (General)*, 134:321–367, 1971.

[2] D. Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, pages 1–34, 2006.

Thematic Session
CLAD 2025 Scholarships

4 April, 16:50 - 17:10, Grande Auditório

Are complex networks useful for generating synthetic time series?

Jaime Vale[1], Vanessa Freitas Silva[2], Maria Eduarda Silva[3], Fernando Silva[2]
[1] Faculdade de Ciências, Universidade do Porto, up201700191@edu.fc.up.pt
[2] CRACS-INESC TEC, Faculdade de Ciências, Universidade do Porto, vanessa.silva@fc.up.pt, fmsilva@fc.up.pt
[3] LIAAD-INESC TEC, Faculdade de Economia, Universidade do Porto, mesilva@fep.up.pt

Time series data are vital for multidisciplinary applications but often cannot be shared due to privacy concerns. The generation of synthetic data has been appointed as a solution to share data yet preserving its privacy, under a utility privacy trade-off. Leveraging on complex networks approach to represent time series, this study proposes a framework to generate synthetic time series. The utility of the synthetic data is assessed by statistical and structural time series metrics, while the privacy is assessed by usual privacy metrics for time series.

Keywords: synthetic time series generation, complex networks, time series mappings

A time series is a sequence of observations in chronological order and a key part of today's data ecosystem. However, time series often contain sensitive information and thus cannot be shared easily even for research purposes. The generation of synthetic data has been appointed as a solution to address this critical challenge [4]. Time series analysis via complex networks offers an alternative approach for analyzing time series data. There are several mapping methods to transform time series into complex networks, enabling time series analysis within the network domain [3]. In particular *Quantile Graph (QG)* is a transition-based mapping method that transforms time series data into a network, where each data sample quantile is represented by a node in the network and the weighted and directed links indicate the probability of transitions between consecutive data points across quantiles. Additionally, this approach supports reverse mapping, that is, from the generated quantile graph it is possible to obtain a time series that is statistically similar to the original [1], offering a method to synthesize time series data. This approach can help increase data volumes, which is crucial for the accuracy of modern machine learning and artificial intelligence methods [2], and enable the sharing of sensitive time series data securely through synthetic data [4]. This work aims to (1) develop a framework to synthesize data using complex networks, (2) assess the utility of the generated synthetic data through statistical and structural metrics, and (3) assess data privacy using established metrics such as Mutual Information and Conditional Entropy.

In this work, we proposed a framework which uses the time series-to-complex network mapping *Quantile Graph (QG)* and its reverse version for synthetic time series data generation. We utilize the reverse method, which we termed *Inverse Quantile Graph (InvQG)*, as a way to synthesize time series data that statistically resembles the original. We perform a set of empirical evaluations conducted on time series from established statistical models, to analyze data utility through statistical features using *tsfeatures* package. Additionally, we compute structural metrics of complex networks using the *NetF* framework to assess similarity between original and synthetic data. Subsequently, we assessed the privacy by computing Mutual Information and Conditional Entropy on the original and synthetic data sets generated by *InvQG*, as well as on the original time series perturbed with added noise. The synthetic time series generated using *InvQG* exhibited high utility, with some models showing statistical and structural properties closely mirroring those of the original data. Others models show increased variability and reduced temporal correlations, mainly in time series with specific characteristics like trends, long memory effects and state-dependency. Principal component analysis demonstrated consistent patterns, validating the preservation of key data characteristics in synthetic data. Privacy analysis showed that synthetic data outperformed the noise perturbation technique, with lower Mutual Information and higher Conditional Entropy, indicating better privacy preservation. Additionally, synthetic data preserves more key properties of the original data than noisy data, making it more useful. *InvQG* method offers a scalable solution for generating synthetic data, effectively addressing challenges in analyzing sensitive data. By balancing privacy and utility, it can supports data-driven research and applications, specially in fields with strict privacy regulations.

Acknowledgements National Funds partially fund this work through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within project 2023.13039.PEX. DOI 10.54499/2023.13039.PEX — `https://doi.org/10.54499/2023.13039.PEX` The authors also would like to acknowledge the project LA/P/0063/2020— `https://doi.org/10.54499/LA/P/0063/2020`

References

[1] A. S. L. O. Campanharo, M. I. Sirer, R. D. Malmgren, F. M. Ramos, and L. A. Nunes Amaral. Duality between time series and networks. *PloS one*, 6(8):e23378, 2011.

[2] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar. Using GANs for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*, pages 464–483, 2020.

[3] V. F. Silva, M. E. Silva, P. Ribeiro, and F. Silva. Novel features for time series analysis: a complex networks approach. *Data Mining and Knowledge Discovery*, 36(3):1062–1101, 2022.

[4] J. Yoon, D. Jarrett, and M. Van der Schaar. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

4 April, 17:10 - 17:30, Grande Auditório

Supervised statistical learning methods in the presence of spatial correlation

Beatriz Ferreira[1], **Raquel Menezes**[2]
[1] Centre of Mathematics (CMAT), University of Minho, Braga, Portugal, pg52211@alunos.uminho.pt
[2] Centre of Mathematics (CMAT), University of Minho, Guimarães, Portugal, rmenezes@math.uminho.pt

Abstract
This project applies machine learning methodologies to data analysis with spatial correlation, focusing on comparing techniques such as Random Forest, Random Forest Regression Kriging, and Boosting. The study evaluates their predictive performance, variable selection processes, and computational efficiency, highlighting their advantages and limitations. The findings aim to provide insights for selecting appropriate methods in spatial data contexts.

Keywords: geostatistics, random forest, boosting

The analysis of data exhibiting spatial correlation presents unique challenges that standard statistical and machine learning methods often fail to address adequately. This study aims to evaluate and compare the performance of three supervised learning techniques: Random Forest (RF), Random Forest Regression Kriging (RFRK), and Extreme Gradient Boosting (XGBoost), in the context of spatial data.
The motivation for this work lies in the growing need to apply robust methodologies capable of handling the complexities inherent in spatial datasets. These datasets often exhibit non-stationarity, heterogeneity, and dependencies that can compromise the assumptions underlying traditional modeling approaches. By leveraging machine learning techniques, we aim to explore their capacity to address these challenges while providing accurate predictions and insights.
The methods were applied to the dataset **soil250** from the `geoR` package [2]. This dataset includes several soil chemistry properties measured on a regular grid of 10×25 points spaced by 5 meters. The target variable for this analysis was **CTC (cation exchange capacity)**. The analysis focused on evaluating the predictive performance, variable importance, and computational efficiency of each method.

Methodology

1. **RF:** A learning method that constructs multiple decision trees for a given problem by using bootstrap random samples of the data. RF is robust to overfitting and handles non-linear relationships well.

2. **RFRK:** An extension of RF that incorporates spatial autocorrelation by combining the RF predictions with kriging, a geostatistical method. This approach accounts for spatial dependencies in the residuals of the RF model [1].

3. **XGBoost:** A boosting algorithm that combines multiple simple models to create a more robust and accurate predictive model. It trains base models iteratively, where each new model corrects the errors of the previous ones to minimize a risk function through stage-wise optimization in the function space [3].

Results and Discussion: The results indicate that XGBoost outperforms RF and RFRK in terms of predictive accuracy, as measured by metrics such as RMSE, MAE and R^2. The iterative optimization and ability to reduce residual errors at each step likely contribute to its superior performance. However, RFRK showed significant advantages in capturing spatial autocorrelation, evidenced by improved residual diagnostics and spatial prediction maps.
Variable importance analysis revealed that predictors such as *coordinates*, *elevation* and *enxofrar content* were consistently ranked as the most influential across all models. This finding underscores the importance of spatial information in driving the observed patterns.
Despite its strengths, XGBoost requires careful hyperparameter tuning and is computationally more demanding than RF. In contrast, RF offers a simpler implementation and faster runtime but lacks the ability to explicitly account for spatial dependencies. RFRK bridges this gap but at the cost of increased computational complexity.
Conclusion: This study highlights the strengths and limitations of RF, RFRK, and XGBoost in analyzing spatial data. XGBoost is recommended for scenarios prioritizing predictive accuracy, while RFRK is better suited for applications where spatial autocorrelation is critical. RF remains a viable option for exploratory analyses and scenarios with limited computational resources.
Future work will involve applying these methods to additional real-world datasets and exploring hybrid approaches that combine the strengths of each technique. By advancing the methodological toolkit for spatial data analysis, this research contributes to more accurate and actionable insights in fields such as environmental science, soil chemistry, and geostatistics.

Acknowledgements This study was supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within Individual Scholarship UMINHO/BIM/2024/130 and project DOI 10.54499/UIDB/00013/2020.

References

[1] M. Dumelle, M. Higham, and J. M. Ver Hoef. spmodel: Spatial statistical modeling and prediction in R. *PLOS ONE*, 18(3):1–32, 2023.

[2] P. Ribeiro Jr and P. Diggle. geor: A package for geostatistical analysis. *R-News*, 1(2), 2001.

[3] F. Sigrist. Gradient and Newton boosting for classification and regression. *Expert Systems with Applications*, 167:114080, 2021.

4 April, 17:30 - 17:50, Grande Auditório

Statistical analysis of event impacts: the case of the Feira São Mateus

A. Catarina Gonçalves[1], **Ana I. Melo**[2], **Marco Costa**[3]
[1] ESTGA - Universidade de Aveiro, a.goncalves@ua.pt
[2] ESTGA & CIPES - Universidade de Aveiro, ana.melo@ua.pt
[3] ESTGA & CIDMA - Universidade de Aveiro, marco@ua.pt

Abstract
This study analyzes the impacts of Feira de São Mateus in Viseu using a mixed methods approach. Based on visitor questionnaires and semi-structured interviews with merchants and organizers, it examines social, economic, and environmental effects. Quantitative methods include hypothesis testing, ANOVA, and Spearman's correlation, while content analysis is used for interviews. Results highlight positive economic and social impacts, though environmental aspects need improvement.

Keywords: data analysis, events, impact of events, mixed methods

Over the past two decades, the number of events designed to attract tourists worldwide has grown exponentially. As a result, cities have increasingly incorporated such events into their strategies to draw both domestic and international tourists [1]. In fact, in recent years, events have become instrumental in helping cities to achieve their objectives, which include infrastructure development, job creation, investment attraction, regional promotion, and enhancing their public image.

According to [3], events can be defined as gatherings of people at a specific time and place with a purpose of celebration, commemoration, communication, education, or leisure. Additionally, [2] classifies events according to their purpose, planning, size, form, content, and format.

This study adopted a mixed-methods approach, combining both quantitative and qualitative techniques to provide a more comprehensive understanding of the event under analysis. The combination of surveys and semi-structured interviews, accounting for the various stakeholders involved, enabled a more comprehensive analysis.

For visitors, in-person surveys were conducted during the 2023 edition of the Feira de São Mateus, as they were leaving the event. The study defined event visitors as the target population and employed a non-probabilistic quota sampling method, stratified according to data from the national population of NUTS III Viseu Dão Lafões. A total of 353 responses were collected, based on the established quotas, and the data was subsequently processed and analyzed using SPSS.

For merchants, 50 semi-structured interviews were conducted during the event, using random sampling to represent various economic sectors. Additionally, to capture the organizational perspective, semi-structured interviews were conducted post-event with a representative from Viseu Marca and a city councilor from the Municipality of Viseu. This combination of complementary methods facilitated data triangulation, ensuring a robust and multidimensional analysis.

The Feira de São Mateus attracts a diverse audience, including young people, families, and emigrants, with a significant proportion of visitors coming from outside the municipality of Viseu, comprising approximately one-third of the sample. The event is particularly popular among individuals with higher education levels, and the majority of visitors have a strong tradition of attending the event annually (75.72%). Additionally, 93.8% of respondents expressed interest in returning, and 97.7% stated they would recommend the event, indicating high levels of visitor satisfaction. Organizers highlighted the event's increased visitor numbers and revenue, while merchants mentioned a decline in the quality of the event and organizational shortcomings.

In terms of impact, the event yielded social and economic positive effects and environmental negative effects. Economically, it contributed to job creation and generated business opportunities for local companies. Socially, the event promoted the city as a tourist destination and increased Viseu's recognition. However, negative impacts were reported, including traffic congestion, increased noise, and inconveniences for residents. Environmental issues, such as waste generation and pollution, were also highlighted.

The study concludes that despite its challenges, the Feira de São Mateus' strengths outweigh its weaknesses, underscoring the event's enduring success over the years.

References

[1] H. M. B. Alves, A. M. C. Cerro, and A. V. F. Martins. Impacts of small tourism events on rural places. *Journal of Place Management and Development*, 3(1), 22–37, 2010. https://doi.org/10.1108/17538331011030257

[2] G. A. J. Bowdin, I. McDonnell, R. Harris, W. O'Toole, J. Allen, and L. Jago. *Events Management* (4th ed.), Routledge, 2023. https://doi.org/10.4324/9781003044963

[3] P. Yürük, A. Akyol, and G. G. Şimşek. Analyzing the effects of social impacts of events on satisfaction and loyalty. *Tourism Management, 60*, 367–378, 2017. https://doi.org/10.1016/j.tourman.2016.12.016

4 April, 17:50 - 18:10, Grande Auditório

Spatial imputation in rotating panel data: an analysis using stochastic partial differential equations (SPDE) with INLA

Antonio Loría-García[1], **Lígia Henriques-Rodrigues**[1], **Pedro Campos**[2]
[1] Centro de Investigação em Matemática e Aplicações, Universidade de Évora, antonio.loria@ucr.ac.cr, ligiahr@uevora.pt
[2] LIAAD - INESC TEC, School of Economics and Management, University of Porto, and Statistics Portugal, pcampos@fep.up.pt

Abstract
Attrition is a challenge in surveys of the labor force with rotating panel designs. This study uses spatial imputation with stochastic partial differential equations (SPDE) using data from the Portuguese Labor Survey, from 2023-2024. Imputation methods were used to treat missing values for net income along covariates such as age, sex, academic degree, and months of employment, aimed at improving estimations done by Portugal´s National Statistics Institute (INE), as a contribution to the development of imputation methods for spatially structured rotating panel data for complex survey designs.

Keywords: rotating panel data, attrition, spatial imputation, INLA, SPDE

The central problem in longitudinal surveys is attrition and rotating panel data surveys are not an exception. A comprehensive definition of a rotating panel design, described by [4], characterizes it as a combination of repeated cross-sectional and panel designs. We have special interest in individuals or households that, for many reasons, are dropped out of the labor force surveys, causing attrition, and how this missing information can be "filled" with imputation. In national statistics with rotating panel data designs, it becomes fundamental to understand the attrition rate to determine whether respondents are participating enough to obtain accurate inferences. Depending on how and why these missing values exist, the validity of inferences may be compromised.
The objective of this research relies on using a spatial imputation approach for net income, with stochastic partial differential equations (SPDE) method using integrated nested Laplace approximation (INLA) [3], comparing the results with classic INLA and multiple imputation. We used trimestral data from the Portuguese Labor Force Survey between 2023 and 2024. According to [3], INLA process is computed in three steps: 1) approximating the posterior marginal of certain parameter θ by using the Laplace approximation, 2) computing the Laplace approximation, or the simplified Laplace approximation for selected values of θ, to improve on the Gaussian approximation, and 3) combining the previous two

steps using numerical integration. The SPDE models that use INLA are greatly explained by [1] based on the wide studies of [2] to obtain an SPDE solution, using random Gaussian fields (RGF) and the Mátern covariance.
The results were developed using the programming language R and the IDE RStudio, applying a methodology that involved regressing net income as the response variable against one or more covariates—such as age, sex, academic degree, and months employed—using the available SPDE functions in the INLA-R library.
The models were used to predict the missing values in the net income variable of the first wave, simulating the full sample. This "complete" net income variable on the first wave is used as a lagged covariate for the second wave, and the process will be repeated subsequently for the next waves. If SPDE modeling is useful for imputation, the lagged covariates should improve the results of the next wave. The results of multiple imputation and classic INLA models without spatial data were included, to determine the potential benefits (if exist) of using geographic data and lagged covariates with SPDE and INLA.
A limitation was the inaccessibility of specific individual locations. The proposed solution to this problem was to randomly determine the location of the household according to the residence district based on NUTS-II, as the unique geographical reference available. The lack of specific geographic data for individuals or households opens the door to further research on the topic, especially to obtain better grids for SPDE with INLA analysis.

Acknowledgements This research was done with the support of Centro de Investigação em Matemática e Aplicações (CIMA), Universidade de Évora, Portugal, under reference CIMA/BD1/2023, projeto UIDP/04674/2020, Financiamento Plurianual 2020-2023 do CIMA, Fundação para a Ciência e a Tecnologia (FCT/MCTES) and the program of mobility for postgraduate studies of Universidad de Costa Rica (UCR), Costa Rica.

References

[1] E. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC, 2018.

[2] F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, 2011.

[3] H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

[4] R. Tourangeau, M. Zimowski, R. Ghadialy, and S. Pedlow. Introduction to panel surveys in transportation studies. Technical Report DOT-T-98-3, National Opinion Research Center, 1997.

4 April, 18:10 - 18:30, Grande Auditório

Analyzing topological gait descriptors for improved classification of parkinsonism

Jhonathan Barrios[1], **Wolfram Erlhagen**[1], **Miguel Gago**[2,3], **Estela Bicho**[4], **Flora Ferreira**[1]
[1] Centre of Mathematics, School of Sciences, University of Minho,
id10605@uminho.pt, wolfram.erlhagen, fjferreira@math.uminho.pt
[2] Neurology Department, Hospital da Senhora da Oliveira,
miguelgago@hospitaldeguimaraes.min-saude.pt
[3] School of Medicine, Life and Health Sciences Research Institute (ICVS), University of Minho,
[4] Algoritmi Centre, School of Engineering, University of Minho,
estela.bicho@dei.uminho.pt

Abstract
This study examines Topological Data Analysis (TDA) to improve parkinsonism classification based on gait. We applied persistent homology to extract topological descriptors—Betti Curves, Persistence Landscapes, and Silhouettes —to capture nonlinear gait dynamics. Combined with a Random Forest model, these descriptors distinguished between Control and Parkinson's, as well as idiopathic Parkinson's disease and vascular Parkinsonism. Betti Curves outperformed others, indicating TDA's potential in identifying disease-specific gait changes.

Keywords: gait analysis, machine learning classification, Parkinson's disease, persistent homology, topological data analysis

The differential diagnosis between Idiopathic Parkinson's Disease (IPD) and Vascular Parkinsonism (VaP) is essential for effective clinical management and tailored treatment strategies. However, the significant phenotypic overlap between IPD and VaP presents challenges for accurate differentiation using conventional approaches [1, 2]. This study investigates the application of Topological Data Analysis (TDA) to improve the classification of Parkinsonian subgroups by revealing nonlinear features in gait dynamics that are often undetectable by traditional methods.
Using gait time series data from 15 IPD patients, 15 VaP patients, and 36 healthy controls (CO), persistent homology was employed to extract topological descriptors, including Betti Curves (BC), Persistence Landscapes (PL), and Silhouettes (SL). These descriptors capture multiscale topological features from persistence diagrams and were used as input to a Random Forest classifier. Parameters such as embedding dimension,

time delay, and the number of bins were optimized to ensure robust feature extraction and classification performance [3].
The classification between CO and PD patients demonstrates the application of topological descriptors as features. The results demonstrate that BC outperformed PL and SL across multiple gait variables, achieving high classification accuracy and Area Under the Curve (AUC). For IPD vs VaP classification, BC achieved an AUC of 0.80 and an accuracy of 83%, particularly excelling in features like Strike Angle and Peak Swing. The superior performance of BC highlights its ability to summarize persistent topological features effectively, as opposed to PL and SL, which exhibited greater sensitivity to noise and dimensionality [4].
This work underscores the potential of TDA to advance gait analysis for neurodegenerative disease diagnostics by integrating topological descriptors with machine learning models. These findings align with previous studies demonstrating TDA's capacity to enhance the understanding of complex gait dynamics. Future efforts will focus on refining topological parameters and extending this methodology to larger datasets, contributing to more robust and scalable diagnostic tools for clinical use.

Acknowledgements We thank the Foundation for Science and Technology (FCT) for the financial support provided through the doctoral scholarship with reference 2023.02242.BDANA and the support of Portuguese funds through the Center of Mathematics and FCT, within the projects UIDB/00013/2020 and UIDP/00013/2020. In adition, The J. Barrios thanks Associação Portuguesa de Classificação e Análise de Dados (CLAD) for the CLAD2025 grant.

References

[1] J. C. M. Zijlmans, S. E. Daniel, A. J. Hughes, T. Révész, and A. J. Lees. Clinicopathological investigation of vascular parkinsonism, including clinical criteria for diagnosis. *Movement Disorders*, 19(6):630–640, 2004.

[2] J. B. Lehosit and L. J. Cloud. Early Parkinsonism: Distinguishing idiopathic Parkinson's disease from other syndromes. *Journal of Clinical Outcomes Management*, 21(2):150–153, 2015.

[3] F. Chazal and B. Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021.

[4] Y. Yan, O. M. Omisore, Y.-C. Xue, H.-H. Li, Q.-H. Liu, Z.-D. Nie, J. Fan, and L. Wang. Classification of neurodegenerative diseases via topological motion analysis—a comparison study for multiple gait fluctuations. *IEEE Access*, 8:96363–96377, 2020.

Thematic Session
CLAD Corporate

5 April, 10:10 - 10:40, Grande Auditório

Strategic integration of machine learning and artificial intelligence at NOS

Diogo Santos
NOS, SGPS, Lisbon, Portugal, Diogo.AnSantos@nos.pt

Abstract
NOS leverages extensive data and advanced Machine Learning/Artificial Intelligence to enhance customer experience and operational efficiency. This talk covers predictive models like decision trees and regression, real-world case studies, and measurable outcomes. It also addresses scaling Machine Learning from experiments to structured, agile workflows, emphasizing standardized processes and automation. Key lessons, challenges, and NOS's future innovation strategies are highlighted.

Keywords: machine learning, artificial intelligence, operational efficiency, predictive modeling, digital transformation

NOS, as a telecommunications provider, strategically leverages extensive datasets generated through continuous and diverse customer interactions, enabling numerous practical applications of advanced Machine Learning (ML) and Artificial Intelligence (AI). These applications focus primarily on customer experience enhancement and operational efficiency. This talk will highlight specific methodologies and real-world case studies where predictive ML models, including decision trees and regression analyses, were successfully applied. Each case study encompasses detailed analyses from problem contextualization and methodological approaches through to quantifiable business outcomes.
Further, we address organizational challenges associated with scaling ML initiatives, notably transitioning from isolated experimentation and ad-hoc prototyping towards a structured, agile, and collaborative workflow. Institutionalizing standardized processes, deploying robust data engineering practices, and automating model management in production have significantly boosted productivity and innovation capacity.
We conclude by discussing the key lessons learned, ongoing challenges, and future strategic directions, reaffirming NOS's commitment to sustained technological innovation and intelligent data utilization.

5 April, 10:40 - 11:10, Grande Auditório

Low-latency graph methods for fraud detection systems

Jacopo Bono
Feedzai, jacopo.bono@feedzai.com

Fraud detection systems must assess events with millisecond latency. Due to the increasing prevalence of scams, where victims lose funds that are then rapidly dispersed through multiple mule accounts, there is a need to integrate network-based information into detection processes. However, performing graph-based computations within stringent latency constraints poses significant challenges, often forcing compromises between the timeliness of the data and the complexity of the graph operations. In this paper, we present an overview of our recent research into encoding real-time graph information specifically tailored for low-latency production environments.

Keywords: dynamic graphs, low latency, fraud detection

Financial transactions form dynamic graphs linking entities like accounts, merchants, and devices. Fraud typically manifests as deviations from standard entity behaviors and requires detection within a tight window between transaction initiation and authorization. Recent years have seen an increase in scam activities, where fraudsters use social engineering to move funds into mule accounts. These funds often pass through multiple accounts to obscure origins, making a network-based detection approach essential.
Efficiently encoding graph information within millisecond latencies is challenging. Traditional graph-based methods involve sampling graph neighborhoods and extracting features, but this process is computationally intensive, prompting trade-offs between data timeliness and context size, either by heavily restricting the neighborhood or performing graph operations asynchronously.
Our work aims to avoid neighborhood sampling by using stateful models and message-passing triggered by interactions. Graph recurrent neural networks (GRNNs) [1], similar to traditional RNNs, update multiple entity states and exchange information upon each interaction, naturally incorporating temporal dynamics and neighborhood information. Nevertheless, training GRNNs using standard backpropagation methods is impractical [1]. We have therefore investigated two variations to overcome using backpropagation. The first method involves fixing recurrent updates, leading to a graph feature engineering system that can deliver low-latency graph information effectively [2]. The second method proposes learnable but linear recurrent dynamics, therefore allowing real-time recurrent learning instead of backpropagation through time. This simplifies the training process, iterating through interactions chronologically without needing to keep past interactions in memory

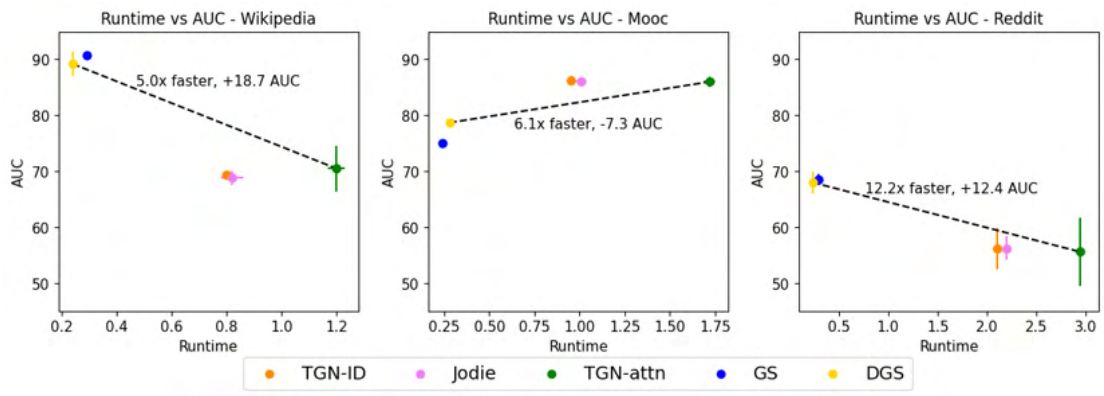

Figure 1: Trade-off between test ROC-AUC and inference runtime for three dynamic graph datasets (Wikipedia, Mooc, Reddit). Figure reproduced from [3]

as is required by backpropagation through time [3]. We evaluated our methods against state-of-the-art techniques (Jodie [4], TGN [5]), achieving comparable performance while significantly reducing inference latency, as shown in Figure 1.

References

[1] J. Bravo, J. Bono, H. Ferreira, P. Saleiro, and P. Bizarro. Mind the truncation gap: challenges of learning on dynamic graphs with recurrent architectures. *Transactions on Machine Learning Research*, 2024.

[2] A. N. Eddin, J. Bono, D. O. Aparício, H. Ferreira, J. T. Ascensão, P. Ribeiro, and P. Bizarro. From random-walks to graph-sprints: a low-latency node embedding framework on continuous-time dynamic graphs. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 176–184, New York, NY, USA, 2023. Association for Computing Machinery.

[3] A. N. Eddin, J. Bono, D. O. Aparício, H. Ferreira, P. M. P. Ribeiro, and P. Bizarro. Deep-graph-sprints: Accelerated representation learning in continuous-time dynamic graphs. *Transactions on Machine Learning Research*, 2024.

[4] S. Kumar, X. Zhang, and J. Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2019.

[5] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on Graph Representation Learning*, 2020.

Thematic Session
SPE

5 April, 14:30 - 14:50, Auditório 1

Cross-correlation analysis to identify the drivers of phytoplankton biomass in Atlantic coastal bays

Helena Mouriño
CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, mhnunes@fc.ul.pt

---

In this presentation, we will mainly apply a modified cross-correlation technique to identify the drivers of phytoplankton biomass in two Portuguese coastal bays: Lisbon Bay and Lagos Bay. We will analyse a nine-year chlorophyll-a time series, as chlorophyll-a is considered a proxy for phytoplankton concentration. The analysis will proceed as follows: First, the periodic behaviour of chlorophyll-a concentration in each bay will be modelled as a combination of the periodicities found statistically significant by the Hartley test applied to the periodogram's peaks and the one-week lag of chlorophyll-a concentration. Then, the chlorophyll-a time series will be cross-correlated with various meteorological and oceanographic variables at different time lags.

Keywords: time series, cross-correlation, simultaneous confidence intervals, spectral analysis, phytoplankton

---

Phytoplankton are microscopic, primarily unicellular algae that live in the ocean and constitute the foundation of the marine food chain. Like terrestrial plants, phytoplankton contain the pigment chlorophyll-a, which gives them a greenish colour. Based on this characteristic, chlorophyll-a can be used as a proxy for phytoplankton biomass in the visible light region of the ocean. Monitoring chlorophyll-a levels provides a simple and cost-effective method for tracking phytoplankton biomass.

In upwelling regions such as the Western Iberian Coast, phytoplankton biomass patterns are primarily influenced by water column stratification, nutrient availability and the intensity and persistence of upwelling conditions. Several studies have highlighted the role of specific environmental drivers in shaping phytoplankton variability across space and time. The analysis of chlorophyll-a time series and correlated environmental variables can lead to identifying the relative importance of phytoplankton variability drivers [1]. In this study, we will examine the influence of selected meteorological and oceanographic (MetOc) variables on chlorophyll-a levels.

The present study focuses on nine years (2008–2016) of chlorophyll-a data from two coastal bays, Lisbon Bay and Lagos Bay. The data under analysis were collected weekly. We began by assessing the extent to which small latitudinal differences and/or coastline orientations (Lisbon *versus* Lagos) contribute to distinct periodic behaviours of chlorophyll-a concentration. To this end, we estimated the periodic behaviour of chlorophyll-a concentration

in each bay by combining the results of the Hartley test applied to the ordinates of each periodogram with the partial-adjustment model [2, 3].
Next, a modified cross-correlation technique was applied to quantify the relationship between chlorophyll-a concentration and the following MetOc variables at different time lags: sea surface temperature, surface photosynthetically available radiation, upwelling index, mixed layer depth and precipitation. The approximate variance and covariances of the sample cross-correlations - used to assess their significances - were derived from Bartlett's theorem [4]. Simultaneous confidence intervals for different lags were also computed.

Acknowledgements This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

References

[1] J. E. Cloern, P. C. Abreu, J. Carstensen, L. Chauvaud, R. Elmgren, J. Grall, H Greening, and J. O. Roger. Human activities and climate variability drive fast-paced change across the world's estuarine – coastal ecosystems. *Global Change Biology*, 22:513–529, 2016.

[2] G. G. Judge, W. E. Griffiths, R. Carter-Hill, H. Lütkepohl, and T.-C. Lee. *The Theory and Practice of Econometrics, 2nd Edition*. John Wiley & Sons, New Jersey, 1985.

[3] H. Mouriño and Barão M. I. A comparison between the Linear Regression Model with autocorrelated errors and the Partial Adjustment Model. *Stochastic Environmental Research and Risk Assessment*, 24:499–511, 2010.

[4] R. S. Tsay. *Analysis of Financial Time Series, 3rd Edition*. John Wiley & Sons, New Jersey, 2010.

5 April, 14:50 - 15:10, Auditório 1

On the performance evaluation of algorithms for the identification of ARMA models

Sónia Gouveia
University of Aveiro, Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), sonia.gouveia@ua.pt

Identifying the order of an ARMA model is crucial in time series analysis. Traditional methods use ACF and PACF, while automated approaches like the Hyndman-Khandakar (HK) algorithm rely on Information Criteria to improve computational efficiency, but lack thorough validation. This paper assesses the HK performance through simulations, comparing the selected orders with those of the generating processes. The results highlight strengths and limitations of the algorithm, and offer insights for enhancing the task of model selection.

Keywords: ARMA models, order identification, Box-Jenkins, Hyndman-Khandakar

The analysis of univariate time series data using Autoregressive and Moving Average (ARMA) models constitutes a cornerstone of data analysis, yet selecting the appropriate model order (p, q) remains a pivotal challenge.

The publication of Box and Jenkins' seminal book in 1970 brought significant advancements in the field of ARMA model order selection [3]. The book provided a practical and unified approach to model selection based on an iterative cycle of stages, including (1) the identification of a candidate model to be tentatively entertained, (2) the estimation of its parameters and (3) a diagnostic checking aiming to answer the key question "is the model adequate?". Initial reactions to the approach were mixed, with some literature questioning "Is Box-Jenkins a waste of time?" [1, 2], but it has since become a foundational guideline in time series analysis. The following editions of the book underwent substantial revisions and expansions, incorporating contributions from new authors [4], including Greta Ljung (1941 - 2024) who collaborated with George Box (1919 - 2013) in the development of the Ljung-Box statistical test, widely used to assess the presence of autocorrelation in residuals, helping to determine whether a model adequately captures the underlying temporal structure observed in the data [6].

Nowadays, the process of building an ARMA model typically involves examining the empirical Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) and evaluating numerous candidate models using an Information Criteria (IC). Automated methodologies have been developed to streamline this task, typically involving batch analysis of a predefined set of orders or systematically exploring all feasible combinations within

specified ranges (similar to a grid search procedure). While systematic, these methodologies can be computationally expensive.
Sequential search strategies for ARMA model order selection provide a more efficient alternative to grid search methods. For example, the Hyndman-Khandakar (HK) algorithm explores a range of models with different (p, q) values, evaluating each candidate combination using an IC [5]. The algorithm starts with a set of initial models, selects the one with the lowest criterion value and iteratively adjusts parameters to explore neighboring models until no further improvement is detected. Despite its appeal, this algorithm has not been rigorously evaluated, raising concerns about its reliability in identifying the correct model order.
This study addresses the question "Is Hyndman-Khandakar a waste of time?" by evaluating the effectiveness of the HK algorithm in determining the order of an ARMA model. Specifically, the performance of the HK algorithm is evaluated through a systematic simulation study, comparing the selected orders over realizations with those of the generating processes, considering different factors such as the ARMA structure, the values of the parameters and the length of the time series.

Acknowledgements This work was partially supported by FCT (Fundação para a Ciência e a Tecnologia, I.P., https://www.fct.pt/) under the scope of the research unit 00127-IEETA (https://www.ieeta.pt/).

References

[1] O. D. Anderson. An appraisal of the Box-Jenkins approach to univariate time series analysis. *Metrika*, 24:187–194, 1977.

[2] O. D. Anderson. Is Box-Jenkins a waste of time? *De Economist*, 125(2):254–263, 1977.

[3] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Series in Time-Series Analysis and Digital Processing. Holden-Day, First edition, 1970.

[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, New Jersey, 5 edition, 2015.

[5] R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26(3), 2008.

[6] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

5 April, 15:10 - 15:30, Auditório 1

A hierarchical Bayesian geostatistical model for zero-inflated and extreme spatial data: analysing sardine egg density in Portugal

Soraia Pereira[1], **Raquel Menezes**[1], **Maria Manuel Angélico**[2], **Tiago Marques**[3]
[1] CEAUL, University of Lisbon; CMAT, University of Minho,
soraia.pereira@math.uminho.pt
[2] IPMA, Portugal, mmangelico@ipma.pt
[3] CREEM, University of St Andrews; CEAUL, University of Lisbon,
tiago.marques@st-andrews.ac.uk

Modeling spatial datasets characterized by an excess of zeros and extreme values poses significant challenges in statistical analysis. Such data structures are common in ecological studies, particularly in marine biology, where the presence or absence of a species and rare high-density occurrences can impact population management strategies. This study introduces a hierarchical Bayesian geostatistical model designed to effectively handle these complexities in the context of sardine egg density along the Portuguese coast.

Keywords: geostatistical modeling, zero-inflated data, extremes, hierarchical Bayesian models, sardine egg density

The ability to predict where, and when, a species will be, and at what densities it will occur when present, is fundamental knowledge to effective management and conservation of wild species. Here in particular we intend to understand the spatial distribution of the sardine eggs density in the coast of Portugal and Gulf of Cadiz, using data from the spring acoustics-trawl survey conducted anually by the Portuguese National Institute for the Sea and Atmosphere (IPMA). We will look at this as a geostatistical problem, where the main objective is the prediction of a variable of interest over a domain, based on the values observed at a limited number of points. Kriging is one of the most classical approaches to spatial prediction in the point-referenced data setting. However, inference on such models is not straightforward due to the sparse covariance matrices. That problem is well known in the literature as big n problem. To overcome the computational costs, Lindgren et al (2011) [1] proposed a new approach based on stochastic partial differential equation (SPDE) models. The idea is to approximate the Gaussian field by a Gaussian markov random field, a discretized version. This approximation can be easily implemented using the integrated nested Laplace approximation (INLA) methodology ([3]). Here, we will adopt these frameworks to support the implementation of suggested model.
Since the data presents a high proportion of zeros and overdispersed data compared to one of the most common distributions to model sardine eggs densities, such as a Gamma

distribution, we propose a geostatistical Gamma-GP hurdle model, where GP represents the Generalized Pareto distribution.
The model we propose here has some points of contact with the model of Opitz et al (2018) [2]. The authors developed a Bayesian generalized additive modeling framework tailored to estimate complex trends in marginal extremes observed over space and time, where the latent random effects were modeled using Gaussian process priors. Here, we will extend this model incorporating a binomial distribution to model the presence/absence, and that will be applied to the ecological problem here presented.
The integrated modeling strategy is broadly applicable to similar ecological problems and other contexts involving zero-inflated or heavy-tailed spatial data.

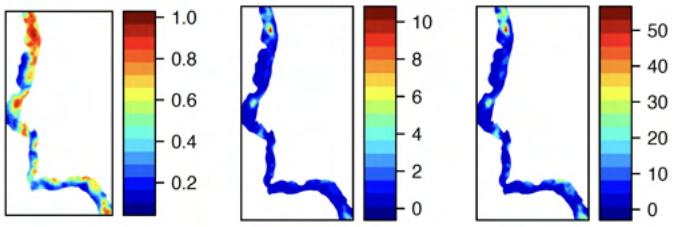

Figure 1: Posterior mean of: probability of presence (left), right truncated positive sardine eggs density conditional to the presence (middle), exceedances (right)

Acknowledgements This work is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025, UIDB/00006/2020, UIDB/04050/2020, PTDC/MAT-STA/28243/2017 and PTDC/MAT-STA/28649/2017. The survey data analysed was collected under the framework programme PNAB: Portuguese Marine Surveying Programme - P03M02 (EU Data Collection Framework EUDCF, FEAMP), and the current work was developed within the scope of project SARDINHA2020 - Ecosystem approach towards a sustainable sardine fishery exploitation (Mar2020-MAR-01.04.02-FEAMP-0009).

References

[1] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[2] T. Opitz, R. Huser, H. Bakka, and H. Rue. Inla goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*, 21(3):441–462, 2018.

[3] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Contributed Sessions

3 April, 16:00 - 16:20, Room 13.1

Detecting Airbnb host profiles with HiMC: the multilevel clustering methodology

Maria Gonçalves[1], **Pedro Campos**[2]
[1] School of Economics and Management, University of Porto, m.joao2000@hotmail.com
[2] LIAAD - INESC TEC, School of Economics and Management, University of Porto, and Statistics Portugal, pcampos@fep.up.pt

---

We provide a new methodology by applying multilevel segmentation to Airbnb hosts, focusing on hierarchically structured and interconnected data. This work compares traditional single-level clustering with HiMC, Hierarchical Multilevel Clustering. The study highlights the advantages of HiMC in capturing complex interactions between reviews, properties, and hosts, offering actionable insights for platform managers and hosts.

Keywords: multilevel clustering, Kmeans clustering, clustering cohesion and comparison, Airbnb

---

The traditional clustering approach consists of grouping different types of variables while ignoring the inherent complexity of the relationships between different layers. By not accounting for the interrelationships between different levels of data, it can lead to a loss of important nuances and a superficial understanding of the observed patterns [1]. In contrast, multilevel clustering adopts a more robust approach by recognizing and incorporating the hierarchy within the data. Studies such as [2] show that the application of these models can significantly improve quality indicators in data analysis, particularly in environments that require the processing of large data volumes. The present study introduces an innovative methodology (HiMC - Hierarchical Multilevel Clustering), by applying multilevel segmentation to Airbnb hosts, focusing on hierarchically structured and interconnected data.
The HiMC Framework is structured across three levels:

1. **Level 1: Clustering Reviews within Properties**

   - For each property i (where $i = 1, 2, \ldots, N$), we have a set of reviews denoted as R_{ij}, where $j = 1, 2, \ldots, M_i$, with M_i being the number of reviews for property i. Each review R_{ij} is represented by a vector of features, such as cleanliness, communication, and location:

$$R_{ij} = [x_{ij1}, x_{ij2}, \ldots, x_{ijm}] \tag{1}$$

Then, for each property i, clustering is performed on the reviews $\{R_{i1}, R_{i2}, \ldots, R_{iM_i}\}$ to obtain a set of $K_i^{(1)}$ clusters:

$$C_i^{(1)} = \{C_{i1}^{(1)}, C_{i2}^{(1)}, \ldots, C_{iK_i^{(1)}}^{(1)}\} \tag{2}$$

where $C_{ik}^{(1)}$ represents the k-th cluster of reviews for property i.

2. **Level 2: Clustering Properties within Hosts**

   - At this level, the input data are the clusters $C_i^{(1)}$ obtained in Level 1. Each host h (where $h = 1, 2, \ldots, H$) manages a set of properties P_{hi} (where $i = 1, 2, \ldots, N_h$ for host h, with N_h being the number of properties managed by host h). For each host h, clustering is performed on the properties $\{C_1^{(1)}, C_2^{(1)}, \ldots, C_{N_h}^{(1)}\}$, resulting in a set of $K_h^{(2)}$ clusters:

$$C_h^{(2)} = \{C_{h1}^{(2)}, C_{h2}^{(2)}, \ldots, C_{hK_h^{(2)}}^{(2)}\} \tag{3}$$

   where $C_{hk}^{(2)}$ represents the k-th cluster of properties for host h.

3. **Level 3: Clustering Hosts**

   - At this level, the input data are the clusters $C_h^{(2)}$ obtained in Level 2. Here, hosts are clustered based on the properties they manage.Clustering is performed on all hosts $\{C_1^{(2)}, C_2^{(2)}, \ldots, C_H^{(2)}\}$, resulting in a set of $K^{(3)}$ clusters:

$$C^{(3)} = \{C_1^{(3)}, C_2^{(3)}, \ldots, C_{K^{(3)}}^{(3)}\} \tag{4}$$

   where $C_k^{(3)}$ represents the k-th cluster of hosts.

We have also applied a single-level clustering to serve as a benchmark and compared the two approaches using the Silhouette Index, Davies-Bouldin Index, ARI, NMI, and the contingency matrix. We concluded that HiMC outperforms single-level clustering in several key areas. HiMC higher cohesion and separation between clusters, leading to more accurate and meaningful segmentation of Airbnb hosts. While both models share some common patterns, the moderate agreement shown by the ARI and NMI metrics highlights that HiMC uncovers additional details and nuances, offering a more refined segmentation of hosts.

References

[1] J. K. Holodinsky, P. C. Austin, and T. S. Williamson. An introduction to clustered data and multilevel analyses. *Family Practice*, 37(5):719–722, 2020.

[2] I. S. Lebedev and M. E. Sukhoparov. Improving the quality indicators of multilevel data sampling processing models based on unsupervised clustering. *Emerging Science Journal*, 8(1):355–371, 2024.

3 April, 16:20 - 16:40, Room 13.1

Clustering for points of interest identification: insights from recent research

Flora Ferreira
Centre of Mathematics, School of Sciences, University of Minho,
fjferreira@math.uminho.pt

Identifying stop locations in GPS trajectories is vital for understanding travel patterns. This study examines various data mining methods for stop identification, focusing on clustering algorithms like K-means and HDBSCAN, as well as neural network approaches using Dynamic Neural Fields (DNF). Recent advancements effectively capture shorter stops with higher visit frequencies that conventional methods often overlook. The presentation highlights the practical applications of these methodologies and their potential.

Keywords: stop locations, GPS trajectories, clustering algorithms

Identifying stop locations on GPS trajectories is essential to understand travel patterns and extract meaningful travel information. Accurate identification of these locations facilitates effective navigation, improves transportation planning and the user experience in smart urban environments.

Various data mining methods have been employed to extract stop locations from GPS trajectories, with clustering algorithms being the most prevalent. For instance, the K-means clustering algorithm is commonly used. However, density-based clustering algorithms such as DBSCAN and its derivatives have gained popularity due to their ability to detect clusters of varying shapes and densities, providing advantages over K-means in certain scenarios [4]. Hierarchical clustering algorithms, such as agglomerative hierarchical clustering, have also been used to identify different visited locations. Hierarchical clustering methods, such as agglomerative clustering, offer intuitive interpretations of clusters and have also been utilized for location identification. More recently, HDBSCAN, which combines the strengths of both DBSCAN and hierarchical clustering, has shown effectiveness in identifying regions of interest [2].

Recent advances include the use of recurrent neural network models, such as Dynamic Neural Fields (DNF), which allow the continuous integration of spatial and temporal information from GPS data [3]. The DNF approach has shown superior performance compared to traditional clustering algorithms such as K-means and HDBSCAN, particularly in capturing shorter stops with higher visit frequencies that are often neglected by conventional methods. This methodology is particularly advantageous when both the duration and frequency of location visits need to be considered. When the temporal integration of GPS data at a specific location surpasses the bump formation threshold, the location

is recorded. In contrast, places visited briefly and infrequently are automatically ignored. This joint representation of spatial and temporal information about stop locations provides a new perspective for driver assistant systems. Furthermore, an innovative solution for the identification of the stop location using K-means enhanced by a user interaction correction mechanism is presented in [1]. This integration addresses the challenges posed by dynamic mobility patterns and GPS coordinate ambiguities by allowing users to refine the clustering of stop points into Points of Interest (POIs).
This presentation will explore these recent approaches and discuss their advantages in depth. In addition, it will highlight practical applications and prospective research directions, illustrating how these methodologies can transform transportation systems and effectively capture daily user routines.

Acknowledgements 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] P. Dias, F. Ferreira, P. M. F. Guimarães, W. Wojtak, W. Erlhagen, S. Monteiro, E. Sousa, and E. Bicho. A machine learning approach for points of interest extraction and event classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 69–82. Springer, 2024.

[2] C. Fernandes, F. Ferreira, W. Erlhagen, S. Monteiro, and E. Bicho. A deep learning approach for intelligent cockpits: learning drivers routines. In *Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimarães, Portugal, November 4–6, 2020, Proceedings, Part II 21*, pages 173–183. Springer, 2020.

[3] F. Ferreira, W. Wojtak, C. Fernandes, P. Guimarães, S. Monteiro, E. Bicho, and W. Erlhagen. Dynamic identification of stop locations from gps trajectories based on their temporal and spatial characteristics. In *International Conference on Artificial Neural Networks*, pages 347–359. Springer, 2021.

[4] R. A. Hamid and M. S. Croock. A developed gps trajectories data management system for predicting tourists' poi. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(1):124–132, 2020.

3 April, 16:40 - 17:00, Room 13.1

Real estate market dynamics in the Oporto municipality

Francisco Sousa Matos[1], **Pedro Duarte Silva**[1]
[1] Universidade Católica Portuguesa, Católica Porto Business School, Portugal, fsousamatos@gmail.com, psilva@ucp.pt

Real estate markets reflect similarities of structural and location characteristics, leading to distinct housing submarkets. This paper studies the existence, and evolution, of housing submarkets in the municipality of Oporto in 2019 and 2022. Those submarkets were identified with the recent methodology of hierarchical cluster analysis with contiguity restrictions. Results identified three relatively stable submarkets, whose delimitation, and characterization gives valuable information for homeowners, municipalities, lenders, and real estate investors.

Keywords: real estate markets, market segmentation; spatially restricted cluster analysis

In recent decades, the availability of affordable housing in Portugal has been a major concern for the Portuguese government and society. However, this market is far from homogeneous showing strong differences over time and between different areas. This work aims is to find segments the real estate market in Oporto municipality, and its evolution, in the years 2019 and 2022.

The housing market is characterized by being segmented and structured according to a complex pattern that takes into account various elements and not just following a homogeneous process of spatial organization [4]. In fact, residential housing has various and diverse characteristics. These include physical characteristics, neighborhood attributes and location factors [4]. The presentation of multiple attributes means that the different social groups that participate in the market and have different preferences and economic capacities organize themselves into clusters both in territorial as well as in social and economic terms [4]. Thus, there is spatial heterogeneity in the housing market and, as a result, there are several reasons why the analysis of the segmentation of this market. The development and analysis of submarkets makes it possible to understand the particularities of each submarket, improving the ability of investors and lenders to assess the risk associated with housing investments. At the same time, housing consumers themselves acquire information on how submarket boundaries are defined [1]. From a strategic perspective, the definition of submarkets contributes to a better understanding of possible problems in specific submarket areas. This recognition is beneficial for fiscal assessment and community development, providing an effective source for planners and policymakers to investigate dynamic change in the housing system [3].

The issue of housing is a current and emerging concern in Portugal, providing a motivation for studying the issue of defining submarkets in the municipality of Oporto. The lack of similar studies in this part of the country, makes this study particularly useful for many government officials and economic agents. In this work we identified submarkets by hierarchical cluster analysis with contiguity restrictions [2]. Although this methodology is arguably the most appropriate for any clustering problems with spatial constraints, it is also relatively new, and we are not aware of its application to any other housing segmentation study.

Our results identified four clusters in 2019 and 2022, with three of them sharing most members and characteristics for both years. The three common submarkets may be characterized in the following way: Submarket 1 - Mostly small apartments located predominantly around the Asprela university area. Submarket 2 - Historic properties in Oporto center with a low presence of high-rise apartments. Submarket 3 - High-value properties in the noble area of Douro's Foz. Furthermore, in 2019 our analysis suggested a fourth cluster comprised mostly by old single-story houses in Oporto suburbs, while in 2022 there seemed to be a small independent submarket in the area between Paranhos and Campanhã.

References

[1] A. C. Goodman and T. G. Thibodeau. The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics*, 35(2):209–232, 2007.

[2] G. Guénard and P. Legendre. Hierarchical clustering with contiguity constraint in R. *Journal of Statistical Software*, 103:1–26, 2022.

[3] B. Keskin and C. Watkins. Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies*, 54(6):1446–1462, 2017.

[4] J. L. Marques, E. Castro, A. Bhattacharjee, and P. Batista. Spatial heterogeneity across housing sub-markets in an urban area of portugal. In *Proceedings from ERSA 2012 Congress - Regions in Motion - Breaking the Path*, pages 1–21, 2012.

3 April, 17:00 - 17:20, Room 13.1

Exploring voter turnout in Portuguese legislative elections through municipal profiling

Fábio Coutinho[1], **Joana Leite**[1,2,3]
[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal, iscac12825@alumni.iscac.pt, jleite@iscac.pt
[2] CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Portugal
[3] Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal

Electoral participation is a cornerstone of democracy, yet rising abstention and voter disengagement in Portugal pose significant challenges. This study examines electoral abstention, political entropy, and a sociodemographic index across 308 municipalities to define municipal profiles using cluster analysis. Four unique profiles emerge, revealing distinct regional patterns that are analyzed for the 2002, 2011 and 2022 legislative elections, enriching the understanding of electoral dynamics in Portugal.

Keywords: elections, abstention, entropy, cluster analysis, municipalities

The growing abstention in many countries and, in particular, in Portugal threatens the legitimacy of democratic institutions and reflects a widespread disillusionment with the political system. This study addresses the phenomenon of electoral participation in the context of Portuguese democracy, focusing on three particularly significant electoral moments in Portugal's recent history: the legislative elections of 2002, 2011, and 2022. These years were selected for their regular intervals, enabling the observation of changes over two decades, and also their coincidence with census years, allowing for the incorporation of a more time-aligned sociodemographic information.

The primary goal of the research is to understand electoral participation patterns at the municipal level in Portugal, employing cluster analysis to identify municipal profiles. Three variables are used to form the clusters: electoral abstention, political entropy, and a sociodemographic index. Electoral abstention refers to the percentage of eligible voters who did not participate in the election. Political entropy, as suggested in [2], is used as an indicator of electoral uncertainty. It is calculated based on the share of votes obtained by each party, with the higher values indicating greater uncertainty in the election outcome. The sociodemographic index is adapted from the Municipal Human Development Index in [3], considering purchasing power, education level, and longevity, where higher values reflect greater development. This index provides a portrait of the characteristics of Portugal's 308 municipalities. All data sources to construct these variables are public, namely

the Ministry of Internal Administration and the Pordata database. Following [1], the clustering technique used is k-means, which is widely recognized for its efficiency in pattern identification and clarity in data segmentation, with the number of clusters determined using the silhouette index.
The results of the analysis reveal that municipalities can be grouped into four meaningful clusters that reflect unique combinations of sociodemographic characteristics and electoral patterns, which allow for the definition of four profiles. These profiles highlight significant regional differences that evolve over the years. For example, in 2002, there are almost no municipalities with high abstention, low entropy, and a low development index. However, in 2011, the number of municipalities with this profile increases significantly in the northeast of mainland Portugal and also in the Azores, before seeing another modest increase again in 2022. This trend demonstrates the growing regional disconnection from democratic institutions in these areas, offering valuable insights into the regional disparities in electoral participation.

References

[1] A. Akarca and C. Başlevent. Persistence in regional voting patterns in turkey during a period of major political realignment. *European Urban and Regional Studies*, 18:184–202, 2011.

[2] J. Gill. An entropy measure of uncertainty in vote choice. *Electoral Studies*, 24:371–392, 2005.

[3] P. R. Nietto, M. C. Nicoletti, and N. C. Sacco. Analyzing electoral data using partitional and hierarchical clustering algorithms. In A. Abraham, S. Pllana, G. Casalino, K. Ma, and A. Bajaj, editors, *Intelligent Systems Design and Applications (ISDA 2022)*, volume 646, pages 53–64. Springer, 2023.

3 April, 16:00 - 16:20, Room 13.2

Iterative GMM for bias reduction in state-space models

Marco Costa[1], **Magda Monteiro**[1]
[1] ESTGA - Águeda School of Technology and Management, CIDMA - Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt, msvm@ua.pt

State-space models (SSM) offer a flexible framework for modeling dynamic systems, where latent variables evolve according to a stochastic process, and noisy observations are linked through a linear relationship. This study introduces a novel estimation method for SSM parameters, leveraging the double-iterated Generalized Method of Moments. This approach aims to address biases in parameter estimation, particularly in small samples and high-variance scenarios, outperforming traditional methods such as maximum likelihood.

Keywords: state-space models, GMM, parameter estimation, Kalman filter, simulation

A general linear state-space model represents a dynamic system in which the state evolves over time following a linear process, while observations are linked to the hidden state through a linear relationship. The state equation is given by:

$$\beta_t = \Phi_t \beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon) \tag{1}$$

where: $\beta_t \in \mathbb{R}^s$ is the state vector at time t, $\Phi_t \in \mathbb{R}^{s \times s}$ is the state transition matrix that governs how the state evolves from $t-1$ to t, and, ε_t is the process noise, assumed to be normally distributed with zero mean and covariance matrix $\Sigma_\varepsilon \in \mathbb{R}^{s \times s}$, i.e., $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$. The observation equation is given by:

$$Y_t = H_t \beta_t + e_t, \quad e_t \sim \mathcal{N}(0, \Sigma_e) \tag{2}$$

where $Y_t \in \mathbb{R}^m$ is the observation vector at time t, $H_t \in \mathbb{R}^{m \times s}$ is the observation matrix that maps the state vector to the observation space in time t, e_t is the measurement noise, assumed to be normally distributed with zero mean and covariance matrix $\Sigma_\varepsilon \in \mathbb{R}^{m \times m}$, i.e., $e_t \sim \mathcal{N}(0, \Sigma_e)$. The process noise e_t and the measurement noise ε_t are assumed to be uncorrelated across time steps, i.e., $e_t \perp \varepsilon_s$ for $\forall t, s$. The model includes a set of parameters, denoted as θ. The vector θ may encompass various parameters depending on the model's specification, but it typically includes the covariance matrices Σ_e and Σ_ε and the state transition matrix Φ, so that $\theta = \{\Phi, \Sigma_e, \Sigma_\varepsilon\}$. Growing interest in adopting time series modeling in fields is evident from the increasing number of review articles and guides dedicated to applying these models in such contexts. These resources often focus on

computational methods and practical considerations essential to effectively implement time series approaches, [1, 2, 3], namely the parameters estimation. The most widely considered method for estimating unknown parameters is the maximum likelihood method, where the log-likelihood function for the state model, based on the observed series $\{Y_t\}_{t=1}^n$ and the normality of errors, is given by:

$$\ell(\mu, \Phi, \Sigma_\varepsilon, \Sigma_e) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left[\log|\Sigma_t| + \eta_t'\Sigma_t^{-1}\eta_t\right], \tag{3}$$

where η_t is the one-step-ahead prediction error, and $\Sigma_t = H_t P_t H_t' + \Sigma_\varepsilon$ is its covariance. The maximization of the log-likelihood function ℓ is typically performed numerically, as closed-form solutions are not available for this class of models. Standard optimization algorithms such as quasi-Newton methods (e.g., BFGS) or the Expectation-Maximization (EM) are commonly used to obtain the parameter estimates. The proposed GMM2i approach iteratively refines parameter estimates by combining the stochastic properties of SSMs with Kalman filter predictions. The method uses first-order Taylor expansions to propagate biases and employs the generalized method of moments to minimize prediction errors. A simulation study was conducted using a univariate local level model with an AR(1) state process. The parameters included the autoregressive coefficient (ϕ), and variances of state and observation errors (σ_ϵ^2, σ_e^2). The study compared GMM2i, ML, and a hybrid method across varying sample sizes and noise levels. Results demonstrated that GMM2i outperformed ML in small samples (e.g., $n = 25$), particularly under high variance. For larger samples ($n = 200$), all methods showed comparable accuracy, though GMM2i retained an edge in estimating variance components.

Acknowledgements This work is supported by CIDMA under the FCT (Portuguese Foundation for Science and Technology) Multi-Annual Financing Program for R&D Units.

References

[1] M. Auger-Méthé, K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. Mills Flemming, A. Nielsen, G. Petris, and L. Thomas (2021), A guide to state–space modeling of ecological time series. Ecological Monographs, 91(4):e01470. https://doi.org/10.1002/ecm.1470

[2] P. Poncela, E. Ruiz, and K. Miranda (2021), Factor extraction using Kalman filter and smoothing: This is not just another survey, International Journal of Forecasting, 37:4, 1399-1425, https://doi.org/10.1016/j.ijforecast.2021.01.027

[3] K. Newman, R. King, V. Elvira, P. de Valpine, R. S. McCrea, and B. J. T. Morgan(2023), State-space models for ecological time-series data: Practical model-fitting. Methods in Ecology and Evolution, 14, 26–42. https://doi.org/10.1111/2041-210X.13833

3 April, 16:20 - 16:40, Room 13.2

A model-based approach for clustering zero-inflated count time series

Luís Sousa[1,2], **Isabel Pereira**[1,2], **Magda Monteiro**[2,3]
[1] Department of Mathematics - University of Aveiro, luissousa2@ua.pt
[2] CIDMA - University of Aveiro, isabel.pereira@ua.pt
[3] ESTGA - University of Aveiro, msvm@ua.pt

This work adopts a model-based clustering approach using finite mixture models. It includes a method to determine the optimal orders of ZINAR processes to enter the finite-mixture model, as well as an expectation-maximization algorithm for parameter estimation. A simulation study was conducted across three increasingly challenging scenarios. This aimed to evaluate the accuracy of the proposed clustering methodology under varying degrees of complexity, with closer parameter proximity presenting greater difficulty in distinguishing between the groups.

Keywords: time series, clustering, model-based, zero inflation, ZINAR

The methodology employed in this study aims to cluster data consisting of n time series of counts, each with T observations, using a model-based clustering technique. It's assumed that each cluster of time series comes from a ZINAR(p*) process, such that [1]

$$X_t = \alpha_p \circ X_{t-p} + \epsilon_t \tag{1}$$

where $\alpha_p \in [0, 1]$, the innovations component, ϵ_t, is a sequence of i.i.d. zero-inflated Poisson distributions [2] with parameters ρ (the inflation parameter) and λ and $\circ$ is the binomial thinning operator. The data can be seen as a finite-mixture model, such that

$$f(X|\Theta) = \sum_{g=1}^{G} \pi_g f_g(X|\theta_g) \tag{2}$$

where π_g refers to the mixing proportion of each process and $\theta_g = (\alpha_g, \rho_g, \lambda_g)$. The methodology used was based on the works of Roick et al. (2020) [3]
The first step of the algorithm is to choose the orders of the ZINAR(p*) processes to be included in the mixture model, which arise from an analysis of the empirical partial autocorrelation function, Figure 1 and Figure 2. Then, k-means clustering is used to obtain initial values for the parameters $\theta_g = (\alpha_g, \rho_g, \lambda_g)$ and mixing proportions π_g. The main goal of this algorithm is to predict the group membership of each time series, z_{ig}, which takes value 1 if a certain time series object, $(x_i)_t$ belongs to group g and is 0 otherwise. In order to predict the value of the membership variable z_{ig}, as well as the the parameters

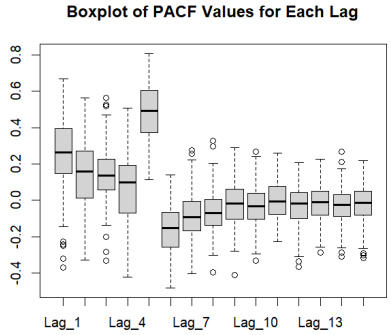

Figure 1: Partial autocorrelation boxplots for one replica of the easy scenario simulated time series

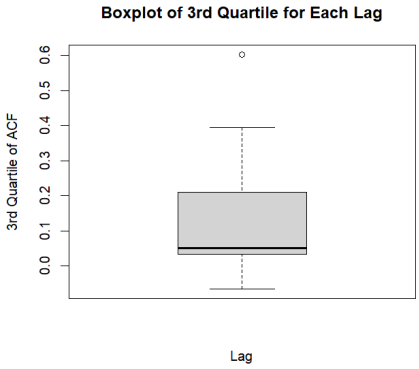

Figure 2: Third quartiles of the lags' PACF boxplot

for each of the processes, θ_g and its mixing proportions π_g, an EM algorithm is employed. In each E-step, the membership of the series is updated, and in the M-step, the mixing proportions are updated, and a weighted-likelihood function for each group is optimized in order to obtain updated parameter estimates. Finally, the last step of the algorithm is to select the combination of ZINAR processes which are most fit for the data. Then, when applying the EM algorithm to all of the combinations, one would choose the combination which had the lowest BIC.

Four cases of simulations are presented: in the first case, 100 replications of 200 series were simulated, each with 50 observations that came from a mixture of 2 INAR(5*); in the second case, the time series were simulated from a mixture of 2 ZINAR(5*); in the third case, a mixture of 1 INAR(2*) and 1 INAR(3*) and in the last case, a mixture of 1 ZINAR(2*) and 1 ZINAR(3*). Each of these cases were tested in 3 scenarios with increasing difficulty, based on how distinguishable the values of the parameters were for both processes. The results are presented in tables which contain the parameter estimations, the BIC of the combination selected and the accuracy for each of the scenarios.

Acknowledgements This work is supported by CIDMA under the FCT (Portuguese Foundation for Science and Technology) Multi-Annual Financing Program for *R&D* Units.

References

[1] M. A. Al-Osh and A. A. Alzaid. First-order integer-valued autoregressive (inar(1)) process. *Journal of Time Series Analysis*, 8:261–275, 1987.

[2] M. A. Jazi, G. Jones, and C. Lai. First-order integer valued ar processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, 33(6):954–963, 2012.

[3] T. Roick, D. Karlis, and P. McNicholas. Clustering discrete-valued time series. *Advances in Data Analysis and Classification*, 15:209–229, 2020.

3 April, 16:40 - 17:00, Room 13.2

Clustering health data time series with the generalized affinity coefficient

Ana Paula Nascimento[1], **Mónica Vieira**[1], **Brígida Mónica Faria**[2], **Alexandra Oliveira**[2], **Cristina Prudêncio**[1], **Helena Bacelar-Nicolau**[3]
[1] RISE-Health, Center for Translational Health and Medical Biotechnology Research (TBIO), E2S-Polytechnic of Porto, Porto, Portugal, ananascimento@ess.ipp.pt, mav@ess.ipp.pt, cprudencio@ess.ipp.pt
[2] Artificial Intelligence and Computer Science Laboratory (LIACC member of LASI), University of Porto, Porto, Portugal, E2S-Polytechnic of Porto, Porto, Portugal, monica.faria@ess.ipp.pt, aao@ess.ipp.pt
[3] Faculty of Psychology, University of Lisbon (FPUL), Lisboa, Portugal, Institute of Environmental Health, Faculty of Medicine, University of Lisbon (ISAMB-FMUL), Lisboa, Portugal, hbacelar@psicologia.ulisboa.pt

---

Diabetes *mellitus* (DM) is one of the non-communicable diseases. The behavior of Disability-Adjusted Life Years (DALYs) over time, in the context of DM, for each country was modeled by ARIMA models [3]. Agglomerative hierarchical cluster analysis was employed using the associated distance of the generalized Affinity coefficient to measure the dissimilarity between ARIMA models [2]. The hierarchy shows three country clusters based on DALY trends over the years.

Keywords: diabetes, DALYs, ARIMA models, affinity coefficient, distance

---

Public Health focuses on studying and preventing diseases, promoting longer lifespans, and improving quality of life through coordinated efforts and informed decision-making. The primary health metric used by the Global Burden of Disease (GBD) study is the disability-adjusted life year (DALY) [1]. This metric serves as a comprehensive tool to evaluate the effects of diseases, injuries, and risk factors. It merges two key components: years of life lost (YLL) due to early mortality and years lived with disability (YLD), presenting a complete picture of the overall health impact of a specific condition. By synthesizing mortality and morbidity data, DALYs offer deeper insights into the actual burden a disease places on a population [3]. When analyzing the health data for the European Region it is clear to see that non-communicable diseases play a big role in the overall burden of disease [3]. Diabetes mellitus (DM) is one of the most severe non-communicable diseases and it can be prevented with a healthy diet and physical exercise, since obesity and overweight are factors that contribute to the increase in diabetes' prevalence. Analyzing data regarding DM is essential in order to define public and health policies. The behavior of DALYs over time, in the context of DM across several countries constitutes time series data. This time series

data can be represented by Autoregressive Integrated Moving Average (ARIMA) models. The ARIMA models that best fit each time series were estimated [3]. Analyzing these time series data is essential for identifying geographic and demographic patterns, offering deeper insights into health disparities across populations [3]. Therefore, agglomerative hierarchical cluster analysis was employed to explore patterns and relationships among various countries in the context of DM using the estimated ARIMA models. To measure the dissimilarity among different countries in the context of DM the associated distance of the generalized Affinity coefficient between ARIMA models is used [2]. The resulting hierarchy reveals three main clusters of countries, as shown in Figure 1.

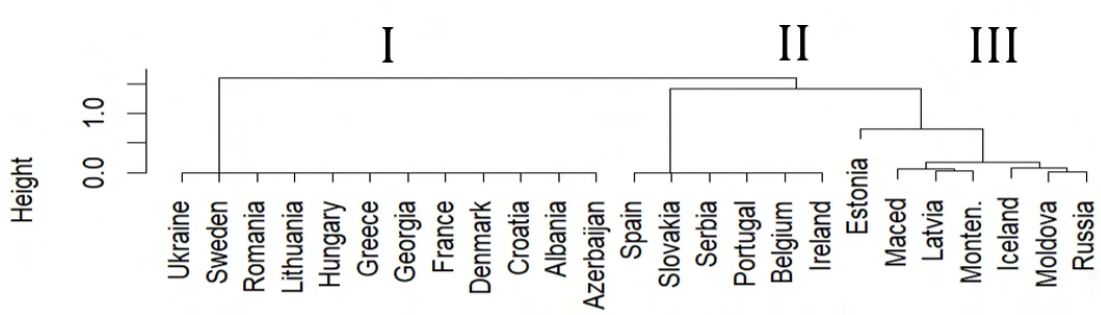

Figure 1: Dendrogram based on the square root of 2(1-Affinity) distance and average linkage aggregation criterion between countries in DM context

I - The values in this cluster show an upward trend over the years with fluctuations.
II - The values in this cluster show an upward trend over the years with a small fluctuation.
III - The values in this cluster show an upward trend over the years without fluctuations.
These clusters present the similarities between DALYs rate values from 1990 to 2019 per country and it can be seen that for all groups, there is an increase in DALYs values over the years, indicating a decrease in life-span. However, in Group I, there are periods where these values decrease, suggesting phases of increased life-span, while in Group III, DALYs values consistently increased over the years. These results may help health authorities in the communication between countries and implementation of health policies in order to promote better and longer life-span, as well as economic benefits to different countries.

References

[1] C. P. Benziger, G. A. Roth, and A. E. Moran. The global burden of disease study and the preventable burden of ncd. *Global Heart*, 11(4):393, Dec 2016.

[2] A. P. Nascimento, A. Oliveira, B. M. Faria, R. Pimenta, M. Vieira, C. Prudêncio, and H. Bacelar-Nicolau. Affinity coefficient for clustering autoregressive moving average models. *Comp. and Mathematical Methods*, 2024:5540143, 2024.

[3] C. Vinhal. *Comparing Time Series Forecasting Models for Health Indicators: A Clustering Analysis Approach.* E2S, Polytechnic of Porto, RECIPP, Porto, 2024, to appear.

4 April, 9:00 - 9:20, Room 13.1

Optimizing energy use in agricultural irrigation systems: a data-driven approach for sustainable practices

José Brito[1], **Conceição Rocha**[1], **Renato Fernandes**[1], **Pedro Guimarães**[1], **Filipe Silva**[2]
[1] INESC-TEC, jose.brito@inesctec.pt, conceicao.n.rocha@inesctec.pt, renato.s.fernandes@inesctec.pt, pedro.souzagui021@gmail.com
[2] Herdade do Esporão, filipe.silva@esporao.com

This study addresses the optimization of energy use in irrigation systems focusing on three objectives: minimizing energy consumption, maximizing the use of renewable energy with the use of solar production forecasts and reducing energy costs by strategically purchasing electricity during lower prices periods based on the Portugal tetra-horary price schedule. These objectives are achieved using a Mixed Integer Linear Programming (*MILP*) model to provide optimized irrigation schedules.

Keywords: *MILP*, renewable energy, irrigation systems

Efficient resource management is the key to sustainable agriculture, particularly in regions where water demand exceeds natural availability. This study leverages a *MILP* model, [2], to optimize irrigation practices in vineyards and olive groves, ensuring energy efficiency, environmental sustainability and cost effectiveness.
The *MILP* model is designed around three main goals:

1. Minimizing energy consumption: The model seeks to reduce the total energy required for irrigation without compromising water delivery.
2. Maximizing renewable energy use: Weekly forecasts of solar energy production, [1], are integrated into the model to prioritize the use of sustainable energy sources for irrigation.
3. Minimizing energy costs: The model assigns weight factors to the four periods defined by Portugal's tetra-horary pricing schedule[1], aligning energy use with economic efficiency.

The application of the theoretical model requires knowledge of physical parameters of the system, which are typically unknown and change over time. To address this challenge, this study explores the use of real data provided by the Herdade do Esporão team[2], including energy consumption and the area to be irrigated over time. Based on this historical data, a

[1] `https://poupaenergia.pt/tarifas-e-ciclos-horarios/`.
[2] `https://esporao.com/en`.

predictive model for energy consumption as a function of the irrigated area was developed. The analysis focused on data consistency and the behavior of consumption patterns over time for different irrigation periods.
It was observed that energy consumption remains nearly constant during the irrigation period, with significant increases or decreases occurring before or after this period. To enhance the model's robustness, the median energy consumption during the irrigation period was used for each sector in the dataset. The median consumption values were found to vary linearly with the irrigated areas, allowing the use of a linear model, derived from the collected data, as an approximation to the physical model for the implemented irrigation system. This model will then be incorporated into the *MILP* model.
The *MILP* model integrates irrigation demand, solar energy availability, and energy market dynamics to optimize weekly irrigation schedules, aligning with the main objectives. Solar energy forecasts prioritize irrigation during high renewable energy periods, while weight factors from the tetra-horary pricing schedule, guide grid energy usage to lower-cost periods, enhancing cost efficiency.
Simulations based on the *MILP* model demonstrated its effectiveness in optimizing irrigation practices. This highlights the robustness of the framework, allowing it to better account for the unique characteristics and variability of each irrigation system. By incorporating renewable energy forecasts, weighted cost prioritization for the pricing schedule, and advanced data preprocessing techniques, the model balances efficiency, sustainable resource management, and economic viability.
Although this is an ongoing work, its ability to generate consistent irrigation schedules demonstrates robustness. This research sets a benchmark for future innovations in agricultural energy management, contributing to sustainable practices.

Acknowledgements This article is co-financed by Component 5-Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021–2026, within project Vine and Wine Portugal - Driving Sustainable Growth Through Smart Innovation, with internal project reference number 67.

References

[1] J. R. Andrade and R. J. Bessa. Improving renewable energy forecasting with a grid of numerical weather predictions. *IEEE Transactions on Sustainable Energy*, 8(4):1571–1580, 2017. Accessed: 2025-01-07.

[2] O. Koné, C. Artigues, P. Lopez, and M. Mongeau. Event-based milp models for resource-constrained project scheduling problems. *Computers & Operations Research*, 38(1):3–13, 2011. Project Management and Scheduling.

4 April, 9:20 - 9:40, Room 13.1

A classification method based on a cloud of spheres

Tiago Dias[1], **Paula Amaral**[2]
[1] NOVA Math, tme.dias@campus.fct.unl.pt
[2] FCT UNL and NOVA Math, paca@fct.unl.pt

In [1], we propose a binary classification model to distinguish a specific class that corresponds to a characteristic that we intend to identify (e.g. fraud, spam, disease). The classification model is based on a connected cloud of spheres that circumscribes the points of such target class. To solve the Connected Cloud of Spheres Problem, a quadratic model with continuous and binary variables (MINLP) is proposed with the minimization of the number of spheres. This classification model is effective when the structure of the class to be identified is highly non-linear and non-convex, also adapting to the case of linear separation. Unlike neural networks, the classification model is transparent, with the structure perfectly identified. Finding the global optima for large instances is quite challenging. To address this, a heuristic approach is proposed.

Keywords: classification, MINLP, non-linear separation, interpretable machine learning

Machine learning models have been widely applied across various domains, often without critically examining the underlying mechanisms. Black-box models, such as Deep Neural Networks, pose significant challenges in terms of counterfactual analysis, interpretability and explainability. In situations where understanding the rationale behind a model's predictions is essential, exploring more transparent machine learning techniques becomes highly advantageous.

In this presentation, we introduce a novel binary classification model called a Connected Cloud of Spheres [1]. The model is formulated as a Mixed-Integer Non-linear Programming (MINLP) problem that seeks to minimize the number of spheres required to classify data points. This approach is particularly well-suited for scenarios where the structure of the target class is highly non-linear and non-convex, but it can also adapt to cases with linear separability.

The aim is, given a target class that is separated from another class by a non-convex surface, to approximate this surface by a cloud of connected spheres, as in Fig. 1.

Unlike neural networks, this classification model retains data in its original feature space, eliminating the need for kernel functions or extensive hyperparameter tuning. Only one parameter may be required if the objective is to maximize the separation margin, similar to Support Vector Machines (SVMs). This simplicity enhances the model's transparency and interpretability, avoiding the complexities of black-box approaches.

One of the primary challenges of this approach is finding global optima for large datasets. To address this, we propose a heuristic solution that performs well on commonly tested

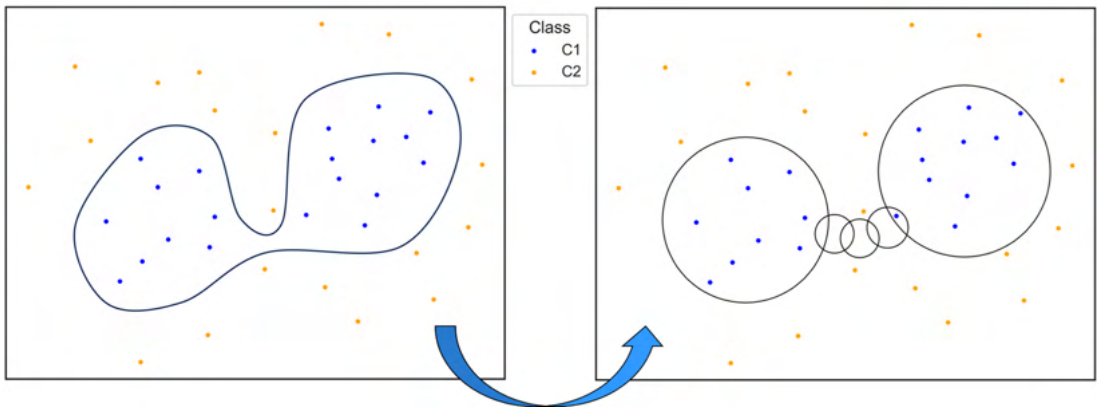

Figure 1: Approximation of a separation surface by a connected cloud of spheres

real-world problems - when compared to state-of-the-art algorithms, the heuristic delivers competitive results, highlighting the model's effectiveness and practical applicability.

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 project UIDB/00297/2020 (Center for Mathematics and Applications).

References

[1] T. Dias and P. Amaral. A classification method based on a cloud of spheres. *EURO Journal on Computational Optimization*, 11:100077, 2023.

4 April, 9:40 - 10:00, Room 13.1

Randomly perturbed random forests

Laura Anderlucci[1], **Angela Montanari**[1]
[1] University of Bologna, Italy, laura.anderlucci@unibo.it, angela.montanari@unibo.it

In supervised classification, a change in the distribution of a single feature, a combination of features, or the class boundaries, may be observed between the training and the test set. This situation is known as dataset shift. As a result, in real data applications, the common assumption that the training and testing data follow the same distribution is often violated. In order to address dataset shift we propose to randomly introduce more variability in the training set by sketching the input data matrix resorting to random projections of units. We then modify the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data. Results on real data show that perturbing the training data via matrix sketching improves the prediction accuracy of test units that have a different distribution in terms of variance structure.

Keywords: classification, dataset shift, data perturbation

Dataset shift occurs when the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries ([2]), i.e., when the training and test joint distributions are different:

$$P_{tr}(\mathbf{X}|\mathbf{y}) \neq P_{te}(\mathbf{X}|\mathbf{y}),$$

where $P_{tr}(\mathbf{X}|\mathbf{y})$ indicate the features' distribution according to $\mathbf{y}$ in the training set, while $P_{te}(\mathbf{X}|\mathbf{y})$ in the test set.

An example of dataset shift can be found in the classification of metastatic cancer. Specifically, when cancer spreads, the secondary cancer cells usually look like abnormal versions of the primary cancer cells (in the tissue where the cancer began). For example, if breast cancer spreads to the lungs, the metastatic tumor in the lung is made up of cancerous breast cells (not lung cells) and is then described as metastatic breast cancer (not lung cancer).
A tool that can be employed to identify the tissue of origin of the metastasis is the modelling of Micro-RNA (mi-RNA) expression profiles. Recently, [3] proposed a miRNA-based tissue classifier to identify the tissue origin of metastatic tumors and they tested on a dataset including 205 primary tumors and 131 metastatic tumors, representing 22 different tumor origins. Exploratory data analysis carried on these data highlighted three main aspects:

- The cancer classes are highly imbalanced;

- The primary and the metastatic cells have similar expression mean values;
- The primary and the metastatic cells have different variability.

Such features suggest that the problem can indeed be cast within the dataset shift framework.
In this work we propose a novel method that accurately classifies the secondary from the primary cancer cells, allowing to identify the most likely tissue of origin of metastatic samples. In order to deal with dataset shift, data perturbation via Random Projections (RP) is proposed.
In [1], it was shown that using random projections to generate new 'compressed' points would allow not only to fix the imbalance problem, but also to explore information outside the convex hull of the original data, so as to avoid the risk of overfitting. The perturbation induced by matrix sketching modified the variability of the training set while preserving the mean values.
Based on such result, we propose a novel classifier that modifies the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data. Results on real data show that perturbing the training data via matrix sketching improves the prediction accuracy of test units that have a different distribution in terms of variance structure.

References

[1] R. Falcone, L. Anderlucci, and A. Montanari. Matrix sketching for supervised classification with imbalanced classes. *Data Mining and Knowledge Discovery*, 36(1):174–208, 2022.

[2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

[3] N. Rosenfeld, R. Aharonov, E. Meiri, S. Rosenwald, Y. Spector, M. Zepeniuk, H. Benjamin, N. Shabes, S.t Tabak, A. Levy, et al. Micrornas accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4):462–469, 2008.

4 April, 9:00 - 9:20, Room 13.2

Hypothesis testing for goodness-of-fit in generalized partially linear models using projections

Rui Costa-Miranda[1], **Rita Gaio**[1], **Christian Heumann**[2], **Wenceslao González-Manteiga**[3]

[1] Faculty of Sciences of the University of Porto and Centre of Mathematics of the University of Porto, rui.miranda@med.up.pt, argaio@fc.up.pt

[2] Faculty of Mathematics, Computer Science and Statistics of the Ludwig Maximilian University of Munich, chris@stat.uni-muenchen.de

[3] Faculty of Mathematics of the University of Santiago de Compostela, wenceslao.gonzalez@usc.es

---

Building on recent advancements in testing the goodness-of-fit of Generalized Partially Linear Models (GPLM), we develop a test based on a residual-marked empirical process considering residuals derived from a kernel-based local polynomial estimation. A simulation study using data generated from Poisson models was designed to compare the size and power of the test in different scenarios. Our results show that it can outperform an existent methodology.

Keywords: generalized partially linear models, goodness-of-fit, hypothesis testing, local polynomial kernel estimation

---

Generalized Partially Linear Models (GPLM) extend Generalized Linear Models (GLM) by allowing a combination of linear predictors and nonparametric components. An example is given by the equation:

$$g(\mu_i) = X_i^T \beta + \gamma(Z_i) \tag{1}$$

where X_i is a vector of covariates for subject $i = 1, \ldots, n$, Z_i is a scalar, β is a finite dimensional parameter and $\gamma(\cdot)$ is a smooth function. Also, $g(\cdot)$ is a link function relating the expected value of the (conditional) response with the partially linear predictor. As usual, it is assumed that $Y_i|X_i, Z_i$ follows a distribution belonging to the exponential family of distributions. The model is estimated by extending local polynomial fitting to generalized linear models. More precisely, we start by estimating the nonparametric function locally by maximum likelihood, where, at each value of Z, the function γ is approximated by a first-degree polynomial using Taylor series expansion [1]. This assumes a local GLM, which can be estimated by kernel iteratively reweighted least squares algorithm (IRLS). Given $\widehat{\gamma(Z)}$, the parameters β can be estimated as part of a global GLM.

The starting points for our specification test are those developed by Escanciano (2006)[2] and Li et al (2024)[3]. The first is foundational, and was designed for parametric linear regression models, projecting residuals along one-dimensional directions, which simplifies the analysis in high-dimensional spaces and does not require the choice of bandwidths.

The corresponding test statistic employs a Cramér–von Mises-type measure in a residual-marked empirical processs and a wild-bootstrap procedure for estimating the critical values through the empirical distributions of the test statistic. More recently, Li et al (2024)[3] extended Escanciano's test to deal with nonlinear link functions and nonparametric structures that can be present in GPLM.
In the test we are proposing, we consider the heteroscedastic linearized model formulation that is obtained from the IRLS algorithm. Then, residuals obtained from the working response variable and the estimated partially linear predictor are considered in Escanciano's test statistic. This approach contrasts with that of Li and Liang's [3], which includes the original GPLM residuals. Table 1 presents the results of a simulation study designed to evaluate the finite sample performance of the two tests, named newHT$_n$ [2] and LiHT$_n$ [3]. In order to consider different data generating processes, we chose $a = 0$ (situation corresponding to a GPLM) and $a = 2$ from the following Poisson model family:

$$log(\mu_i) = 2 + X_i + aX_iZ_i + \frac{1}{2}cos(2\pi Z_i)$$

The results were then based on 500 Monte Carlo replications, where, first, a sample of size n, $X \sim Ber(0.4), Z \sim U(0,1)$ and $Y \sim Pois(\mu_i)$ is simulated, a Poisson partially linear model (1) is fitted, and then the described hypothesis tests are applied. A wild-bootstrap process was repeated 200 times, and nominal levels were set at 0.01, 0.05 and 0.10. Two different sample sizes, $n = 50$ and 100, were considered. The estimated rejection probabilities seem to be in favor of the hereby Escanciano-type proposed test.

Table 1: Empirical size and power of tests

		n=50			n=100			n=200		
	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
a=0	newHT$_n$	0.04	0.09	0.16	0.04	0.09	0.16	0.05	0.11	0.16
	LiHT$_n$	0.06	0.14	0.20	0.06	0.16	0.22	0.13	0.23	0.32
a=2	newHT$_n$	0.01	0.04	0.10	0.17	0.54	0.81	0.91	1	1
	LiHT$_n$	0	0	0.02	0.05	0.33	0.69	0.84	0.99	1

Acknowledgements Rui Costa-Miranda was granted a doctoral research fellowship financed by national funds through FCT, under the reference 2024.03100.BD. Rita Gaio was partially supported by CMUP, under the project with reference UIDB/00144/2020.

References

[1] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

[2] J. C. Escanciano. A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051, 2006.

[3] X. Li, H. Liang, W. Härdle, and H. Liang. Model checking for generalized partially linear models. *Test*, 33(2):361–378, 2024.

4 April, 9:20 - 9:40, Room 13.2

The effect of distress on the health and well-being of workers. PLSc-SEM estimator

Luís M. Grilo[1,3,4], **Helena L. Grilo**[2]
[1] Departamento de Matemática, Universidade de Évora, Portugal, luis.grilo@uevora.pt
[2] Secção de Matemática, Universidade Aberta, Lisboa, Portugal, helenagrilo56@gmail.com
[3] CIMA (Centro de Investigação em Matemática e Aplicações), Universidade de Évora, Évora, Portugal
[4] NOVA Math (Centro de Matemática e Aplicações), Universidade NOVA de Lisboa, Portugal

Occupational stress has been recognized as a potential cause of mental disorders, affecting thinking, feelings, mood and behavior. We proposed a Structural Equation Modeling with "stress" as the latent exogenous construct while the endogenous constructs are "burnout", "sleeping troubles" and "Depressive Symptoms" (target construct). The Copenhagen Psychosocial Questionnaire was applied to workers from a Portuguese company. The consistent Partial Least Squares estimator was used to estimate a model, providing information on workers' health and well-being, namely that "stress" has a strong direct effect on both "depressive symptoms" and "burnout".

Keywords: bootstrap, non-normal data, structural equation model, survey

The harmful effects of distress (negative stress) on mental health and well-being of workers in different sectors of activity are now recognized, while associated disorders appear to become a true epidemic. The specialized literature usually considers that excessive and continuous occupational "stress" wears down the immune system, often leading to mental and physical illnesses, including "burnout", "sleeping troubles" and "depressive symptoms" ([4, 1], among others). As regards "burnout", the World Health Organization considers that it results from chronic stress in the workplace that has not been successfully managed. The Portuguese version of the Copenhagen Psychosocial Questionnaire (COPSOQ) – containing observed variables (measured on a Likert-type scale, with five categories) that operationalize those latent constructs – was applied to workers from a Portuguese company. The study sample comprised 256 valid questionnaires, representing approximately 50% of the company's workforce. Of these, 75% were completed by women. Furthermore, 55.1% of the sample comprised workers aged 20–39 years, and 45.3% possessed more than nine years of formal education. According to both the literature on this scientific topic and the empirical experience of those responsible for the company in the area of Medicine and Occupational Safety, research hypotheses were formulated and a theoretical reflexive

Structural Equation Modeling (SEM) was proposed, where "stress" is the latent exogenous construct and "depressive symptoms" is the target endogenous construct (such as [4, 1]). The consistent Partial Least Squares (PLSc) estimator (from the variance-based-SEM family) was developed after modifications of the original PLS estimator to mimic the common factors model of the Covariance-Based-SEM family of estimators [3]. It retains all the strengths of traditional PLS, such as the ability to handle complex models while it performs well with small sample sizes. Since PLSc-SEM does not require data coming from a multivariate normal distribution [3], it was used to estimate a model which was evaluated considering the usual criteria related to the measurement and structural submodels. The bootstrap resampling method was also used [3, 2] and all parameters are statistically significant with $R^2 = 81.7\%$ for the target latent construct "depressive symptoms". Additionally, the PLSpredict algorithm (allowing the use of a holdout sample to determine the predictive capability of the model on new unseen data [3]), was also applied. Furthermore, the Root Mean Square Error (RMSE) was used to compare the performance of the estimated PLSc-SEM model with a simple Linear Model (LM), in terms of prediction. No "depressive symptoms" indicators have prediction errors greater than the naive LM benchmark, allowing to consider a good predictive power of the estimated model.

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 project UIDB/04674/2020, DOI 10.54499/UIDB/04674/2020 (https://doi.org/10.54499/UIDB/04674/2020).

References

[1] L. T. de Beer, J. Pienaar, and S. Rothmann. Job burnout's relationship with sleep difficulties in the presence of control variables: a self-report study. *South African Journal of Psychology*, 44(4):454–466, 2014.

[2] L. M. Grilo, T. F. Braz, J. P. Maidana H. L. Grilo, and M. Stehlík. Modeling the facets of burnout in Lisbon airport border officers using plsc-sem estimator. *Stochastic Analysis and Applications*, 2023.

[3] 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.* Cham: Springer, Switzerland, 2022.

[4] H.-F. Hsieh, Y. Liu, H.-T. Hsu, S.-C. Ma, H.-H. Wang, and C.-H. Ko. Relations between stress and depressive symptoms in psychiatric nurses: The mediating effects of sleep quality and occupational burnout. *Int. J. Environ. Res. Public Health*, 18(7327), 2021.

4 April, 9:40 - 10:00, Room 13.2

Digital transformation in Europe: insights from a multilevel multivariate probit regression model

José G. Dias[1], **Lucas de Souza**[2]
[1] Iscte – Instituto Universitário de Lisboa, jose.dias@iscte-iul.pt
[2] UNIFOR – Universidade de Fortaleza, lucaslfsouza@unifor.br

The digital transformation is a phenomenon that has become very relevant in recent years. This paper analyzes the progress of the digital transformation in the European business sector using a European survey. We use a Bayesian multilevel multivariate probit regression model to investigate whether the probability of adopting digital technologies by the EU businesses can be explain by their characteristics. Results show that countries are at different stages of progress in the EU.

Keywords: multivariate probit analysis, multilevel analysis, Bayesian analysis, European Union, digital transformation

Business digital transformation involves the integration of digital technologies into all business areas and leads to substantial challenges in the way firms operate. It involves not only the introduction of new technologies, but also a cultural change that requires adaptation to new market realities.

Digital transformation in Europe is driven by a number of factors, including the growing demand for digital services, the need to improve operational efficiency and the pressure to internationalize the businesses. It involves technologies such as artificial intelligence, the internet of things (IoT), big data or cloud computing that allow optimizing processes and new business models. The introduction of new technologies often demands restructuring of business processes, which create resistance to change due to the lack of employee's digital skills and often demand investments in technological infrastructure. Companies that successfully implement digital strategies and data-driven decision-making benefit from increased corporate performance.

To understand and characterize the level of achievements, the European Commission launched a survey covering companies in all EU Member States to assess the level of progress of the digital transformation. Therefore, this study aims to analyze the factors that influence organizations to adopting digital technologies. To this end, managers were asked whether they implemented seven digital technologies: 1 - Artificial intelligence, 2 - Cloud computing, 3 - Robotics, 4 - Smart devices, 5 - Big data analytics, 6 - High speed infrastructure, and 7 - Blockchain. To model these joint binary choices, we implemented a multilevel multivariate probit regression.

Let the binary variable y_{ijm} denote the response of individual $i \in \{1, \ldots, n_j\}$ in country $j \in \{1, \ldots, J\}$ to digital technology $m \in \{1, \ldots, M\}$. The multivariate probit model models multiple binary outcomes simultaneously and m indexes each of these outcomes. Let y^*_{ijm} be the latent variable or choice propensity on a continuous scale. Then, we have

$$y_{ijm} = \begin{cases} 1, & \text{if } y^*_{ijm} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The latent variable is defined by the linear component $y^*_{ijm} = \beta_{0m} + \beta_{1m} x_{ij1} + \ldots, + \beta_{Km} x_{ijK} + u_j + \epsilon_{ijm}$, where β_{0m} is the intercept (fixed effect) for the m-th technology; β_{km} is the coefficient for the k-th predictor for the m-th outcome; u_j is the random effect for the j-th country; ϵ_{ijm} is the residual error term for the i-th manager in the j-th country for the m-th outcome. As fixed effect factors, we consider variables such as size of the company, turnover, turnover growth, age, and sector. Random effects were included to account for the lack of statistical independence at country level. That is, it measures the overall effect of the digital transformation at country level, which indicates the stage of the country regarding digital transformation. We assume that the distribution of the random effects is $u_j \sim N(0, \sigma^2_u)$. Residual errors ϵ_{ijm} are assumed to be multivariate normal with a variance-covariance matrix $\mathbf{V}$, where $\mathbf{V}$ has values of 1 on the diagonal and correlations $\rho_{mm'}$ as off-diagonal elements. This residual structure allows dependence between multiple binary outcomes. Residuals and random effects are assumed to be independent.

The Bayesian specification of the model regarding prior distributions is set as default (non informative), given the domain of the parameters. The Bayesian estimation used the MCMC algorithms, in this case the Hamiltonian Monte Carlo (HMC). We ran 4 chains for 8000 iterations with 4000 as burn in and 4000 for sampling from the posterior distribution (retaining every fourth sample). Convergence was assessed by $\hat{R}$ and *n_eff* statistics. The 4000 samples were used to characterize the posterior distribution. All the analyses were conducted using Rstudio and Stan. We assessed the effect of the fixed effects on the multivariate binary data based on whether the 95% credible interval overlapped zero.

This paper identified the important triggers that should become priority to leverage the growth and innovation in the digital age. In addition, knowing which technologies are being implemented and what factors lead to the adoption of these technologies provides a meaningful overview of where organizations are investing, what knowledge they consider most important, and which technologies are more difficult to adopt. In this way, these results can help inform government policies to provide knowledge to help organizations adopt new technologies.

Acknowledgements This work was financially supported by Fundação para a Ciência e Tecnologia (UIDB/00315/2020).

References

[1] S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85:347–361, 1998.

4 April, 10:10 - 10:30, Room 13.1

Enhancing fuzzy forests with consensus clustering for unbiased and robust feature selection

Mouhamadou Lamine Ndao[1,2], **Ndèye Niang**[2], **Genane Youness**[1,2], **Gilbert Saporta**[2]
[1] LINEACT CESI, IDF, Nanterre, France, {mlndao, gyouness}@cesi.fr
[2] CNAM-CEDRIC, Paris, France, {gilbert.saporta, ndeye.niang_keita}@cnam.fr

---

This study presents the Fuzzy Forests algorithm, which uses consensus clustering to improve feature selection in high-dimension data and address multicollinearity issues. While Fuzzy Forests mitigates feature selection biases, its effectiveness relies on the clustering method used. Our proposed consensus clustering framework enhances robustness and reduces variability in results, demonstrating better feature independence through extensive simulations.

Keywords: feature clustering, feature selection, high-dimension, ensemble clustering, ensemble learning

---

High-dimensional data is becoming increasingly common across fields such as genomics, sensometrics, and biomedical sciences, presenting challenges like the curse of dimensionality, multicollinearity, and redundancy, which compromise model interpretability and performance [4]. Traditional dimensionality reduction methods, such as Principal Component Analysis (PCA) and feature clustering, address these issues by simplifying data structure but often sacrifice interpretability or fail to consider the target feature in supervised contexts.

Feature selection methods such as Random Forest [1] offer feature importance measures (VIMs) but are biased when some features are highly correlated and independent features are often ignored in the feature selection process. The Fuzzy Forests algorithm [2] extends Random Forest by incorporating feature clustering and recursive feature elimination to reduce multicollinearity effects and provide unbiased feature selection based on VIMs. However, the Fuzzy Forests's results heavily depends on the choice of the feature clustering algorithm, as different methods yield varying feature partitions, impacting the final results. To address this limitation, we propose a robust extension of Fuzzy Forests that integrates consensus clustering, combining multiple partitions into a stable and reliable consensus [6, 5]. This approach leverages the strengths of diverse clustering algorithms, improves partition stability, and reduces noise sensitivity. Extensive simulations, varying data structures and the relationship between predictors (X) and target features (Y) demonstrate the superior performance of our consensus-based Fuzzy Forests in correcting VIM biases and selecting relevant features. The proposed algorithm outperforms all Fuzzy and Random Forest algorithms in selecting relevant features, even if they are in an independent cluster.

Applications to real-world datasets, including the Liver_Expr gene expression dataset and the SECOM [3] semiconductor manufacturing dataset, further validate its effectiveness in handling high-dimension data, improving interpretability, and addressing challenges like multicollinearity and class imbalance. These results promote the robustness and utility of the proposed method for feature selection and predictive modeling in complex, high-dimension contexts.

References

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

[2] D. Conn, T. Ngun, G. Li, and C. M. Ramirez. Fuzzy forests: Extending random forest feature selection for correlated, high-dimensional data. *Journal of Statistical Software*, 91:1–25, 2019.

[3] M. McCann and A. Johnston. SECOM. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C54305.

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[5] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.

[6] A. Topchy, A. K. Jain, and W. Punch. A mixture model for clustering ensembles. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 379–390. SIAM, 2004.

4 April, 10:30 - 10:50, Room 13.1

ClustOfVar: global vs local standardization

Adelaide Freitas[1,2], **Juliana Castanheira**[1], **Ana Aida Sá**[1]
[1] Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal, adelaide@ua.pt, juliana.ac@ua.pt, anaaida@ua.pt
[2] Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

ClustOfVar is a technique for clustering variables in mixed data. It is already implemented in R software with both hierarchical and partitioning clustering algorithms. Both use normalized data, given by removing the mean and dividing by the standard deviation of each numerical variable. We propose a new standardization conditioned on the categories of some predefined qualitative variable in the dataset. The effect of this new type of normalization in ClustOfVar will be studied using several real datasets.

Keywords: standardization, clustering, mixed data

In many real situations, individuals in the dataset are described by both quantitative and qualitative variables. For example, in customer segmentation for a retail store, customers can be described by numerical variables such as "age" and "monthly spending" and categorical variables such as "most frequent payment method" (categories: cash, credit card) and "membership tier" (categories: standard, premium, VIP). Moreover, categorical variables can be used to typify individual groups (e.g., customer groups by payment method).
ClustOfVar is a technique of clustering variables for mixed (quantitative and qualitative) data that was introduced by[1] and involves a hierarchical clustering algorithm and a k-means type partitioning algorithm. Both clustering algorithms focus on maximizing the homogeneity of the clusters. This homogeneity provides information about the quality of the association between each variable in the cluster and a new numerical synthetic variable, which corresponds to the first principal component of PCAMIX applied to all variables within the cluster [1, 2]. The ClustOfVar technique is implemented in the R package `ClustOfVar` [1].
In many clustering procedures, data preprocessing is required. In `ClustOfVar`, the algorithm is executed on normalized data. For quantitative variables, this involves standardization by removing the mean and dividing by the standard deviation of each variable. For qualitative variables, it is achieved through the standardization of the corresponding indicator matrix. This standardization step is essential to ensure all variables contribute equally to the analysis, preventing variables with larger scales from dominating the results and minimizing biases from discrepant scales.
When we have categorical variables, they can be understood as an aid for classifying or grouping individuals. Moreover, two standardized numerical variables can be correlated

because both are associated with a given categorical variable, potentially biasing the identification of variable groups in ClustOfVar.
We intend to analyze the effect of this type of (local) standardization versus the (global) standardization for the quantitative variables, as already implemented in the ClustOfVar technique. Given the difference between these two types of standardization, it was necessary to make several changes to the functions of the R package `ClustOfVar`.
Preliminary results with R datasets show that the clusters formed can differ depending on the standardization procedure used, with local standardization outperforming the global one in several situations. For instance, for the well-known `mtcar` dataset, cophenetic coefficient values are consistently higher when local standardization is applied.
In this project, a summarization of the behavior of the ClustOfVar technique under global and local standardization will be presented, based on several real datasets and using different evaluation measures of the quality of the clustering.

Acknowledgements This work was partially supported by CIDMA (Center for Research and Development in Mathematics and Applications, University of Aveiro) under the Fundação para a Ciência e a Tecnologia (FCT) Multi-Annual Financing Program for R&D Units.

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] J. Saracco and M. Chavent. Clustering of variables for mixed data. In D. Fraix-Burnet and S. Girard (Eds.), editors, *Statistics for Astrophysics: Clustering and Classification*, volume 77, pages 121–169. EAS Publications Series, 2016.

4 April, 10:50 - 11:10, Room 13.1

Clustering 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 Analysis offers a framework for analyzing complex data with intrinsic variability. This study focuses on clustering data described by density-valued variables, employing Kernel Density Estimation for representation. Hierarchical methods, specifically agglomerative clustering with complete linkage, are employed using distance measures such as Bhattacharyya, Hellinger, and Jeffreys divergence. The methodology is applied to U.S. Communities and Crime data, highlighting insights into state-level patterns across five numerical features.

Keywords: clustering, symbolic data, distances, density-valued data, kde

---

The data encountered in contemporary research problems are increasingly large and complex. Furthermore, there has been a paradigm shift from focusing on individual behavior to analyzing group behavior, which requires the aggregation of data. Traditional aggregation methods relied on summarizing data into a single central descriptive measure, such as the mean or median, to represent a group of individuals. However, this approach inherently reduces the richness of the data by ignoring its variability. To address this limitation, Symbolic Data Analysis (SDA)[1] was introduced in the 80's of the last century. SDA is designed to accommodate data with intrinsic variability, which may arise from either the original recorded data or as a result of aggregation processes. SDA distinguishes between two primary types of aggregation: *Contemporary*: Data are recorded at a single point in time, and the analysis focuses on higher-level entities (e.g., groups of first-level units) rather than the first-level units themselves. Temporal Aggregation: Data are collected at multiple time points for the same individual, and time is not a primary concern; in this case, the analysis focuses on the original first-level units. A symbolic variable Y is defined by a mapping on a set E of statistical units.

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

For a *Density-valued variable* Y, to each unit, i corresponds a density function estimated from the micro data applying a Kernel Density Estimator (KDE).

Clustering aims at identifying groups of similar units within a dataset. This study focuses on hierarchical clustering, with particular emphasis on agglomerative approaches. Distance measures are pivotal for density-based clustering. Table 1 lists the distances used.

Table 1: Distances between density functions

Bhattacharyya	$D_B(f_1(x), f_2(x)) = -ln(BC(f_1(x), f_2(x)))$
Hellinger	$D_H(f_1(x), f_2(x)) = \sqrt{1 - BC(f_1(x), f_2(x))}$
Mallows	$D_M(\Psi_1(t), \Psi_2(t)) = \sqrt{\int_0^1 (\Psi_1(t) - \Psi_2(t))^2 dt}$
Total Variation	$D_{TV}(f_1(x), f_2(x)) = \frac{1}{2}\int_{\mathcal{X}} \lvert f_1(x) - f_2(x)\rvert dt$
Jeffreys divergence	$D_J(f_1(x), f_2(x)) = D_{KL}(f_1(x), f_2(x)) + D_{KL}(f_2(x), f_1(x))$

In distances defined in Table 1, $BC(f_1(x), f_2(x)) = \int_{\mathcal{X}} \sqrt{f_1(x) f_2(x)} dx$; $D_{KL}(f_1(x), f_2(x)) = \int_{\mathcal{X}} f_1(x) ln(f_1(x)/f_2(x)) dt$ is the Kullback–Leibler divergence; $f_1(x), f_2(x)$ are density functions where we assume the same domain $\mathcal{X}$ and $\Psi_1(t), \Psi_2(t)$ are the quantile functions of density $f_1(x)$ and $f_2(x)$ respectively, with $t \in [0, 1]$. In hierarchical clustering methods, the choice of the linkage criteria plays a crucial role in defining the clustering structure. For this study, the Complete Linkage method is employed, which is mathematically defined as: $D^C(A, B) = \max_{a \in A, b \in B} D^U(a, b)$ where A and B represent two clusters, and $D^U(a, b)$ denotes the distance between units, considering that each unit is described by p variables, then $D^U(a, b) = \sqrt{\sum_{j=1}^{p} D(a_j, b_j)^2}$, where distance D is defined in Table 1.
This methodology is applied to the U.S. Communities and Crime dataset, focusing on the following four numerical variables: % of people 25 and over with less than a 9th grade education; % of people 16 and over who are employed; % of population who are divorced; % of *immigrants* who immigrated within last 10 years; Total number of violent crimes per 100K population. We aggregate the micro data by state. To decide on the number of clusters, we use the well-known Silhouette coefficient [2]. The clustering allowed putting in evidence groups of states with similar distributions of the considered variables.

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020.

References

[1] E. 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.

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

4 April, 10:10 - 10:30, Room 13.2

A statistical comparison of external training load metrics during congested versus non-congested periods in football

Paulo Barreira[1], **Luísa Novais**[2], **Francisco Tavares**[3], **João Pedro Araújo**[3]
[1] First Team Physical Performance Coach, Manchester United FC, United Kingdom, paulo.barreira@manutd.co.uk
[2] Department of Mathematics, University of Minho, Portugal, luisa_novais92@hotmail.com
[3] Medical and Performance Department, Sporting Clube de Portugal SAD, Portugal, fstavares@sporting.pt, jparaujo@sporting.pt

In football, periods when official matches are interspersed by 5 days or less have become a present reality. These congested fixture periods can occur frequently throughout a football season, with players competing in both domestic and international competitions. The purpose of this study is to analyse and compare the performance of external training load metrics over a series of back-to-back matches during congested and non-congested periods, on players with consecutive rates of match exposure above 75% of total match time.

Keywords: data analysis, mechanical load, physical performance, football

Football calendars and season fixtures are becoming increasingly demanding for football players, both from a physical and mental health perspective, with congested periods of competition occurring frequently throughout a season. These periods of matches being played with less than five days of interval (back-to-back matches) may compromise the ability to fully recover physically, as this process can last up to three to five days. Entities such as the Fédération Internationale des Associations de Footballeurs Professionnels (FIFPRO) have been creating awareness of the detrimental effects for players' health, and for football as a sport, of the increasing number of matches and congested periods.
Regardless of the rationale associated with performance decrements during congested periods in football, research has been contradictory in this matter, with distance performances at several velocity thresholds seeming to be unaffected by the congested calendar. Despite the existent research on physical performance during congested fixtures, research is scarce in analysing the effect of these periods on players with a high consecutive rate of match participation.
In this study, we statistically compare the performance of external training load metrics over a series of back-to-back matches during congested (CP) and non-congested (NCP) periods. External load metrics include velocity-related distances, for instance, we analysed

high-speed running (HSR), sprint and higher velocity zones such as above 80% (V80) and above 90% (V90) of maximal speed. These variables, represented in meters, were presented as an intensity measure (meter per minute), because there were differences in minutes played by each player.
Matches from two seasons (2022-2023 and 2023-2024) from a Portuguese male professional football team were analysed. In total, congested and non-congested sequences of matches from 22 football players were included. A minimum and maximum number of two and five consecutive matches, respectively, with a minimum participation of 70 minutes from the players, from CP and NCP periods, were included for analysis. A match was considered a back-to-back match (i.e. a CP match) if between the final whistle of a match and the next kick-off there was a period of less than five days. For NCP sequences, a match was considered a consecutive match within a NCP sequence if a period of more than five days and of less than eight days existed from the previous match, as that would be a regular one game per week schedule.
Comparisons of total match load, first and second halves, were performed for consecutive matches within each period and for the same matches between the two periods (CP versus NCP). Additionally, a whole team analysis and a positional analysis was also performed, where four positions were considered: central-defenders (CD), wingbacks (WB), central-midfielders (CM) and forwards (F).
Future studies should further investigate how different external load metrics, such as accelerations and decelerations, either in the form of volume or intensity, could potentially continue to clarify the effect of consecutive matches during both periods on performance.

References

[1] P. Barreira, J. R. Vaz, R. Ferreira, J. P. Araújo, and F. Tavares. External training loads and soft-tissue injury occurrence during congested versus noncongested periods in football. *International Journal of Sports Physiology and Performance*, 19(10):1068–1075, 2024.

[2] A. Dellal, C. Lago-Peñas, E. Rey, K. Chamari, and E. Orhant. The effects of a congested fixture period on physical performance, technical activity and injury rate during matches in a professional soccer team. *British Journal of Sports Medicine*, 49(6):390–394, 2015.

[3] R. Julian, R. M. Page, and L. D. Harper. The effect of fixture congestion on performance during professional male soccer match-play: a systematic critical review with meta-analysis. *Sports Medicine*, 51:255–273, 2021.

4 April, 10:30 - 10:50, Room 13.2

Time series features and forecasting of community pharmacy sales

Maria Inês Vicente[1], **Joana Leite**[2]
[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal, iscac16490@alumni.iscac.pt
[2] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal, jleite@iscac.pt

The pharmaceutical retail sector is rapidly evolving, leveraging its ability to digitally process data to enable forecasting that optimizes inventory management and meets consumer demand. This study examines and forecasts twelve time series of medication sales from a community pharmacy, grouped by pharmacotherapeutic classifications, using ARIMA and ETS models. The forecasting method selected for each series is then evaluated against the identified time series features, providing insights for easier implementation in this context.

Keywords: pharmaceutical retail, sales forecast, STL decomposition, ARIMA, ETS

Health promotion and care provision are becoming increasingly relevant in progressively aging societies, as is the case in Portugal. The pharmaceutical sector is undergoing significant advancements, driven by the ability to digitally collect and store information. As highlighted in [1], technological evolution enables sales forecasting in retail, offering numerous benefits for pharmacy management and also enhancing the customer experience through a better understanding of purchasing patterns. Despite these technological opportunities, community pharmacies often face challenges in implementing these tools and fully integrating them into their operations.

This study aims to contribute by addressing the challenges in forecasting for pharmaceutical retail. Specifically, based on the survey in [1], it focuses on widely accepted forecasting methods in this sector, assessing their effectiveness for forecasting, ease of implementation, and alignment with time series characteristics to provide more precise and actionable recommendations for practical application. To achieve this, historical monthly sales data from a community pharmacy, covering the period between January 2013 and March 2023, is analyzed. Given the wide variety of products typically offered in community pharmacies, the focus is on respiratory system medicines. As emphasized in [4], these medicines are particularly significant due to their frequent purchase patterns, especially during periods of seasonal demand, and their relevance during the COVID-19 pandemic. Sales data is aggregated by pharmacotherapeutic classifications, resulting in twelve distinct time series

to be examined, with aggregation necessary due to the discontinuation of certain medicines over time.
The analysis begins with Seasonal-Trend decomposition using LOESS (STL decomposition), which enables visualization of the time series components, measurement of the strength of its trend and seasonality, and detection of outliers. This is followed by statistical hypothesis testing for stationarity and normality. Several forecasting methods are then tested for one-step-ahead forecasts, including Naïve and Seasonal Naïve, Autoregressive Integrated Moving Average (ARIMA), and state-space Exponential Smoothing, known as ETS (Error, Trend, Seasonal). The automated versions of ARIMA and ETS modeling, based on the algorithms described in [3], are used for their efficiency and ease of use, as they are better suited for this context. Following best practices recommended in [2], the evaluation is performed using time series cross-validation, with classical accuracy measures reported.
The results indicate that pharmacotherapeutic classifications with high sales volumes typically exhibit time series with strong seasonality, which tend to favor ARIMA models. In contrast, for series with weaker seasonality, ETS models are generally preferred. However, for these highly seasonal series, the accuracy measures suggest that historical sales data alone cannot capture all relevant information, highlighting the need for future work to integrate additional sources, such as influenza cases reported by the Portuguese National Health Service.

References

[1] A. Burinskiene. Forecasting model: The case of the pharmaceutical retail. *Frontiers in Medicine*, 9, 2022.

[2] H. Hewamalage, K. Ackermann, and C. Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37:788–832, 2023.

[3] R. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 3rd edition, 2021.

[4] S. Romano, H. Galante, D. Figueira, Z. Mendes, and A. T. Rodrigues. Time-trend analysis of medicine sales and shortages during covid-19 outbreak: Data from community pharmacies. *Research in Social and Administrative Pharmacy*, 17:1876–1881, 2021.

4 April, 10:50 - 11:10, Room 13.2

Towards a guide to include the social perspective in engineering programs: an international perspective

Teresa Barros[1], **Alexandra Albuquerque**[1], **Inês Braga**[1], **Paula Carvalho**[1]
[1] CEOS.P, ISCAP, Polytechnic of Porto, mtmtb@iscap.ipp.pt, alexalb@iscap.ipp.pt, inesbraga@iscap.ipp.pt, paulacarvalho@iscap.ipp.pt

Awareness of environmental hazards and resource depletion from electronics is rising yet integrating social perspectives in engineering education lacks consensus. A survey of lecturers and students explored this issue. Findings reveal age minimally impacts sustainability views, with older individuals prioritizing social impacts. Engineering education moderately addresses social, economic, and political factors, though effectiveness varies by country. Social science students show greater concern for sustainability and social development than engineering students.

Keywords: social perspective, engineering programs, survey, guide, technical/social competencies

There is increasing awareness and concern about the environmental hazards and use of limited Earth resources in producing goods and services. Reviewing literature, we found that there are different strategies but the most applied are specific projects, conferences, courses, and end-of-studies projects. Other authors mention that the most frequently used methodologies to integrate transversal competencies in engineering are case studies and questionnaires, while the least frequent are interviews and literature reviews. Some studies [2] emphasize that sustainability should be included in all teaching programs and throughout all courses and point out that several universities worldwide, such as Kaunas University of Technology, teach sustainability programs in all knowledge areas to raise awareness. Yet, there are constraints arising from the curricula adaptation. Some research [1] concluded that difficulties stemming from the curriculum structure and planning, where most professors felt that the teaching load was already too high to incorporate additional content, and challenges observed during teaching practice, primarily involving the difficulty of integrating relevant transversal competencies with technical content.
With the purpose of developing a guide to the inclusion of the social perspective in engineering programs (bachelor's and master's degrees) it was decided to collect information on the experience, preferences and opinions of the lecturers and students as a first approach to the design of the modules that will form the mentioned guide. A comprehensive survey by means of a questionnaire was conducted (between September and October 2024) to assess the current state of integration and identify areas for improvement, garnering 203

valid responses from international students and lecturers collected from the HEIs that are involved in a transnational project with the authors of this research.
The survey findings revealed a strong consensus on the importance of societal impacts in engineering practices. Respondents rated the need to consider these aspects at an impressive 6.10 on a 7-point scale, with sustainability emerging as the top priority, achieving the highest average score of 6.25 from the following items: Infrastructures and Industry, Natural resources and raw materials, social development, social policies and legal regulations. This strong agreement between the respondents regarding importance of societal impacts in engineering practices indicates widespread recognition of the need to align technological advancements with societal well-being. The average response (4.64), although above the midpoint (3.5), is noticeably lower than other results. This suggests that there is still room for improvement in the way education addresses these broader societal aspects. For example, when the respondents were asked whether education should incorporate these considerations, the average response is much higher (5.78). The difference between satisfaction with the level of inclusion of these aspects and the importance of their inclusion in the educational process may indirectly mean that people don't feel totally satisfied with the way the necessary aspects are addressed in the current situation, or that they see barriers to incorporating these aspects. This discrepancy underscores the urgent need for reforms to better integrate social perspectives into engineering education.
Cultural differences also emerged prominently in the data highlighting the influence of cultural and institutional contexts on educational priorities. Furthermore, disciplinary distinctions were evident, as social science students displayed significantly more concern for sustainability and social development than their engineering peers. Findings highlight the pressing need to align engineering education with broader sustainability goals and societal expectations. By addressing the identified gaps, fostering interdisciplinary collaboration, and tailoring educational practices to regional and cultural contexts, the project aims to create a comprehensive guide to embedding social perspectives in engineering programs. This initiative aspires to cultivate a new generation of engineers who possess not only technical expertise but also a deep commitment to social responsibility and sustainable development.

Acknowledgements This work is financed by Portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UID/05422: Centre for organisational and Social Studies of Polytechnic of Porto.

References

[1] A. G. Abo-Khalil. Integrating sustainability into higher education challenges and opportunities for universities worldwide. *Heliyon*, 10(9):e29946, 2024.

[2] C. Pacher, M. Woschank, B. M. Zunk, and E. Gruber. Engineering education 5.0: a systematic literature review on competence-based education in the industrial engineering and management discipline. *Production and Manufacturing Research*, 12(1), 2024.

5 April, 9:00 - 9:20, Room 13.1

From data to stories: statistics and creativity in data journalism

Cláudia Silvestre[1], **Helena Figueiredo Pina**[2],
[1] Escola Superior de Comunicação Social, LIACOM, CEAUL, csilvestre@escs.ipl.pt
[2] Escola Superior de Comunicação Social, LIACOM, hpina@escs.ipl.pt

Data journalism has become increasingly prominent, expanding its influence within newsrooms and playing a vital role in society. It not only leverages the growing availability of data enabled by new technologies, but also fosters greater public understanding of data and visual information. This work underscores the importance of collaboration between journalism and data analytics, highlighting the critical role of statistics throughout the process, from data collection and analysis to effectively and clearly communicating insights.

Keywords: data journalism, statistics, creativity, visualization, communication

In recent years, concerns about the credibility of news media have grown, with many questioning the accuracy and impartiality of reporting. Data journalism offers a promising path to rebuild public trust and uphold journalism's societal role by promoting transparency and openness. By sharing data sources and methodologies, news organizations can demonstrate their commitment to accountability [3]. Achieving this requires collaboration among professionals with diverse yet complementary skills.

Data journalism occupies a dynamic space at the intersection of statistics, creativity, and storytelling, merging analytical rigor with artistic innovation to produce impactful narratives. Statistical analysis lies at the heart of this discipline, serving as a foundational tool for extracting meaningful insights, ensuring data accuracy, and establishing credibility [4]. Robust statistical methodologies enable data journalists to assess data quality, uncover patterns and trends, and draw evidence-based conclusions, setting the stage for storytelling that is both compelling and trustworthy.

Creativity is equally indispensable in data journalism, transforming raw numbers into vibrant, engaging narratives that resonate with diverse audiences. Through the thoughtful design of data visualizations and storytelling techniques, creativity bridges the gap between complex datasets and audience understanding. These visual and narrative elements are not merely aesthetic; they are integral to enhancing comprehension, fostering emotional connections, and encouraging engagement [2]. Data visualization emerges as the pivotal medium in this process, translating abstract numerical insights into accessible and emotionally compelling forms that drive the narrative forward.

This work emphasizes the essential synergy between statistical rigor and creativity in data journalism. While creativity serves as a powerful tool to captivate and inform, its application must remain faithful to the underlying data. Ethical responsibility is central to

this balance, ensuring that data visualizations and narratives amplify truth rather than distort it. Statistical expertise plays a crucial role in maintaining this integrity, offering a safeguard against misrepresentation while enabling journalists to craft stories that are not only visually striking but also grounded in factual accuracy [1].

Acknowledgements This work is financed by the Polytechnic University of Lisbon under the IDI&CA program with the reference: IPL/IDI&CA2023/MOOC-JorDt ESCS.

References

[1] H. Bhaskaran, G. Kashyap, and H. Mishra. Teaching data journalism: A systematic review. *Journalism Practice*, 18(3):722–743, 2022.

[2] A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar. Designing with pictographs: Envision topics without sacrificing understanding. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4515–4530, 2021.

[3] M. F. de Lima-Santos. The entanglements between data journalism, collaboration and business models: A systematic literature reviewp. *Scientific American*, 12(2):256–281, 2023.

[4] D. Spiegelhalter. *The Art of Statistics: Learning from Data.* Penguin, UK, 2019.

5 April, 9:20 - 9:40, Room 13.1

Enhancing data quality in real-time environments: metrics and applications in digital industry

Eliana Costa e Silva[1], **Óscar Oliveira**[1], **Bruno Oliveira**[1]
[1] CIICESI, ESTG, Politécnico do Porto, Portugal,
eos@estg.ipp.pt,oao@estg.ipp.pt,bmo@estg.ipp.pt

Real-time data environments face challenges in maintaining high data quality due to the continuous influx of large, heterogeneous data sets. Addressing issues, such as incompleteness, inconsistency, and inaccuracy, is critical for ensuring reliable analytics and decision-making. This work synthesizes insights from previous studies to propose and validate data quality metrics. These metrics enable monitoring and assessment of data quality in streaming pipelines, ensuring robust foundations for analytics in Digital Industry contexts. The results show the applicability of the proposed metrics on data quality assessment. Further, it was observed that longer time-windows on the longitudinal score yielded less reactive scores but increased computational effort, while shorter ones enhanced reactivity. Also, adjusting the decay rate controls score behavior without altering the computational cost.

Keywords: Data quality, monitoring, real-time manufacturing, digital industry

In the era of Digital Industry, the proliferation of Internet of Things (IoT) devices and sensors has revolutionized real-time monitoring. However, the continuous and dynamic nature of data streams poses significant challenges. In fact, data quality issues, such as, missing values, outliers, and inconsistencies, can lead to erroneous insights and operational inefficiencies. To effectively tackle these challenges, it is imperative to develop robust and specialized data quality metrics specifically tailored to the unique demands of streaming scenarios.
The following metrics have been designed to address real-time data quality challenges:

- **Weighted Quality Score (WQS):** Provides a snapshot of data quality at a specific time by evaluating adherence to predefined quality rules.
- **Longitudinal Weighted Quality Score (LWQS):** Measures historical data quality, assigning greater importance to recent data while accounting for past trends.
- **Quality Score Delta (QSD):** Quantifies the difference between WQS and LWQS, offering insights into the evolution of data quality over time.

These metrics rely on a weighted sum method, therefore allowing stakeholders to assign importance to specific quality dimensions and rules based on context and relevance.
A service-oriented data ingestion pipeline was implemented in a manufacturing environment. The pipeline integrated tasks such as raw data ingestion, time-series processing, and quality assessment. This pipeline, built around Apache Kafka, demonstrated the effectiveness of the proposed metrics, in terms of: **Enhanced Monitoring:** Real-time alerts for data quality issues enabled immediate corrective actions; **Improved Reliability:** Continuous quality checks ensured the integrity of data used for analytics; **Operational Insights:** Historical analysis of metrics highlighted recurring issues and their root causes.
The Metrics were applied to evaluate data streams involving sensor readings of temperature, pressure, and speed, illustrating their practical utility in maintaining data integrity.
The application of WQS, LWQS, and QSD in real-time scenarios highlights their potential to transform data quality management. However, future work should focus on:

- **Dynamic Weighting Schemes:** Refining the weighting of dimensions and rules based on evolving data characteristics.
- **Automated Rule Generation:** Leveraging machine learning to identify and adapt quality rules in dynamic environments.
- **Broader Applications:** Extending these metrics to other domains, such as healthcare and finance, where real-time data is critical.

Maintaining high data quality in real-time environments is a cornerstone of effective analytics and decision-making. The proposed metrics provide a robust framework for assessing and enhancing data quality in streaming pipelines. By addressing key challenges in data integrity, these metrics contribute to operational efficiency and data-driven innovation.

Acknowledgements This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through projects UIDB/04728/2020 and UIDP/04728/2020.

References

[1] C. Batini and M. Scannapieco. *Data and Information Quality: Dimensions, Principles, and Techniques*. 2016.

[2] E. Costa e Silva, Ó. Oliveira, and B. Oliveira. Enhancing real-time analytics: Streaming data quality metrics for continuous monitoring. In *ICoMS 2024 Proceedings*, 2024.

[3] A. Goknil, P. Nguyen, S. Sen, D. Politaki, H. Niavis, K. J. Pedersen, A. Suyuthi, A. Anand, and A. Ziegenbein. A systematic review of data quality in cps and iot for industry 4.0. *ACM Computing Surveys*, 55(14s):1–38, 2023.

[4] Ó. Oliveira and B. Oliveira. An extensible framework for data reliability assessment. In *International Conference on Enterprise Information Systems (ICEIS) - Proceedings*, 2022.

5 April, 9:40 - 10:00, Room 13.1

R&D in Portuguese companies

Lídia Maria Galvão Rodrigues Praça
Polytechnic Institute of Bragança, lpraca@ipb.pt

The purpose of this study is to characterize and evaluate the Research and Development (R&D) activities of Portuguese companies, from 2016 to 2022. This is a descriptive statistical analysis of R&D activities, from an internal and external perspective of the companies and by sector of activity, based on the Community Innovation Survey. It was possible to observe an increase in R&D activities, especially intramural activities, as well as in activities carried out internally, on a continuous basis.

Keywords: companies, extramural, intramural, R&D activity, sector

Research and Development (R&D) means all creative work carried out in a systematic way, with a view to expanding the body of knowledge, including knowledge of man, culture and society, as well as the use of this body of knowledge in new applications [4]. It can be carried out intramurally (R&D carried out within the company) and/or extramurally (external acquisition of R&D). Intramural R&D concerns activities carried out by the company to create new knowledge or to solve scientific or technical problems (it may include software development within the company when it falls within this scope). Extramural R&D concerns the contracting or acquisition of R&D services from other companies or public or private research organizations and may have to do with: (i) Acquisition of machinery, equipment, software and buildings; (ii) Acquisition of existing knowledge (know-how) in other companies or organizations; (iii) Training for innovation activities; (iv) Introduction of innovations to the market, (v) Design, among others. Still regarding intramural R&D, it is important to know whether this occurs continuously, which presupposes that the company has permanent personnel carrying out R&D activities within the company, or whether it only occurs occasionally, when necessary. It is in this context that the present study is intended to be carried out, applied to Portuguese companies and using a descriptive statistical analysis as its methodology, based on the results obtained through the Community Innovation Survey - CIS, for the period between 2016 and 2022. These surveys are biennial and regulated by the European Union/Eurostat, in accordance with the Oslo Manual, being carried out obligatorily since 1993, in all member states and whose results are published with a two-year lag. Therefore, this study is based on three surveys: CIS 2018 [1], CIS 2020 [2] and CIS 2022 [3]. The CIS 2018 survey collected information from 2016 to 2018 and the results were published in 2020; the CIS 2020 survey collected information from 2018 to 2020 and the respective results were made public in 2022; and the CIS 2022 survey collected information from 2020 to 2022, the results of which are

the most recently published, in 2024. From the analysis of the aforementioned results, it was possible to conclude that: In the 2016-2018 three-year period, 7,5% had intramural R&D activities, 4,3% developed these activities on an ongoing basis and 4,0% contracted R&D services to other companies or public or private research organizations (extramural R&D). In the 2018-2020 three-year period, 9,3% had intramural R&D activities, 4,7% developed these activities continuously and 3,6% contracted R&D services to other companies or public or private research organizations (extramural R&D). In the 2020-2022 three-year period, 10,2% had intramural R&D activities, 4,9% developed these activities on an ongoing basis and 4,2% contracted R&D services to other companies or public or private research organizations (extramural R&D). Over the period under analysis, there was a significant percentage increase in R&D activities, especially intramural, which was accompanied by an identical increase in activities carried out internally, on a continuous basis. It should be noted that this percentage fluctuation in activities carried out on a continuous basis presupposes that Portuguese companies have valued the human resources dedicated to R&D to the detriment of meeting this need on an occasional or extramural basis. We can also see that in the last three years under analysis, there was an increase in R&D activities as a whole, as well as in each of the categories (internal and external). By sector of activity, industry or services, none is clearly demarcated throughout the period under analysis, with both being R&D activity. In conclusion, the study allows us to foresee positive results regarding the role of Portuguese companies in the scope of R&D activities, which need, however, to be proven with a more complete analysis to be carried out in the future and with the inclusion of size, economic activity (CAE), geographic region (NUTS), among other aspects.

References

[1] DGEEC. Principais resultados: CIS 2018-inquérito comunitário à educação. `https://www.dgeec.mec.pt`, 2020.

[2] DGEEC. Principais resultados: CIS 2020-inquérito comunitário à educação. `https://www.dgeec.mec.pt`, 2022.

[3] DGEEC. Principais resultados: CIS 2022-inquérito comunitário à educação. `https://www.dgeec.mec.pt`, 2024.

[4] OCDE Eurostat. *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data. The Measurement of Scientific and Technological Activities*. OECD Publishing, 2005.

5 April, 9:00 - 9:20, Room 13.2

Classification of districts in Costa Rica using geospatial data

Luis Eduardo Amaya-Briceño[1], **Erick Alfredo Vásquez Murillo**[2]
[1] University of Costa Rica - Guanacaste Campus, luis.amaya@ucr.ac.cr
[2] Yaipan S.A, contact@erickvasm.com

In recent decades, the collection and use of geospatial data has generated interest in various fields, including economics, biology, and health.

In our paper, we present our experience collecting, cleaning, and transforming such data. We then used it to classify Costa Rica's districts and cantons. The results were compared with other indices developed by government agencies in the country.

Keywords: geospatial data, google cloud, human development index, principal components, classification

---

Every day, within the analysis of data, the use of geospatial data becomes more relevant, these allow us to show a more complete image of the events [4] in our environment. They are not always easily accessible, either due to the cost or access of them or the computational management.

Using geospatial data, we share our experience in the characterization of the districts and cantons of Costa Rica. This with the aim of being able to contrast our segmentation with existing groupings at the cantonal level, such as the human development index [3] or the consumer confidence index prepared by the University of Costa Rica [2]. To do this, we collect information related to social variables (schools, hospitals), economic variables (bars, restaurants, malls, banks)and religious variables (churches), this information was provided by Google Maps [1], we used a Google Cloud API, which returns the values found for a variable indicated in that area based on a center and radius. In our process, we made a circle overlay of the entire country. The process generated millions of records, which were cleaned and transformed by us, to build a summary table of information that served as input for subsequent statistical analysis.

A principal component analysis and a classification using k-means were performed. The results found allow us to observe the formation of classes of districts that are comparable with what was observed or that can be validated with instruments such as the aforementioned indexes.

References

[1] Google Cloud. Google cloud platform, 2024. Accessed on 8 January 2024.

[2] Universidad de Costa Rica (UCR). Informe del Índice de confianza del consumidor (icc), febrero 2023, 2023. Accessed on 8 January 2024.

[3] Programa de las Naciones Unidas para el Desarrollo (PNUD). Atlas de desarrollo humano cantonal, 2024. Accessed on 8 January 2024.

[4] IBM. Geospatial data, 2024. Accessed on 8 January 2024.

5 April, 9:20 - 9:40, Room 13.2

Applying the weighted aggregated sum product assessment method to the risk classification of sectors for greenhouse gas emission

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

The weighted aggregated sum product assessment method is a multi-criteria decision making method that combines the weighted sum and the weighted product models for ranking a set of alternatives based on different criteria. The purpose of this work is to apply this method to the risk analysis of greenhouse gas emission by sectors, using different risk measures for the definition of the criteria and data corresponding to annual greenhouse gas concentrations measured in Portugal from 1990 to 2022.

Keywords: WASPAS, decision-making model, risk assessment, risk measures, greenhouse gas emission

Multi-criteria decision-making (MCDM) methods are decision support systems that permit making decisions under different criteria, by evaluating and ranking the alternatives with respect to the different criteria. The weighted aggregated sum product assessment (WASPAS) method is one of these methods, that combines the well known weighted sum model and weighted product model (see e.g. [4]).

In the present work, this method will be adapted to the risk analysis of greenhouse gas emissions by sectors. Since greenhouse gases trap heat and are responsible for raising the surface temperature of the earth, contributing to climate change, the reduction of greenhouse gas emissions is essential to slow the rate of global warming and mitigate its impact on environment and human health. In the European Union, the European Climate Law sets a target to reduce the emissions by at least 55% in 2030, compared to 1990 [3]. In order to fulfill these ambitions, risk measurement of greenhouse gases and the identification and risk classification of sectors responsible for the emissions is important for the implementation of appropriate solutions.

This study specifically analyses the emissions of the three most emitted greenhouse gases, CO_2 (carbon dioxide), CH_4 (methane), N_2O (nitrous oxide), in Portugal by different sectors: industrial processes, agriculture, forest and land use change, waste, energy, considering annual data from 1990 to 2022 (emissions in kt of CO_2 equivalent), extracted from the Portuguese Institute of Statistics (INE) [2]. The analysis will incorporate three different risk measures for establishing the assessment criteria [1], including mean, value at risk, and conditional tail expectation.

The method consists first in calculating the risk measures for each pollutant emitted by the five sectors (or alternatives: A_1 – industrial processes, A_2 – agriculture, A_3 – forest and land use change, A_4 – waste, A_5 – energy) based on the empirical distribution. Then, the WASPAS method will be applied to the risk decision matrix, whose entries are the risk values of the alternatives determined with the risk measures, in order to obtain a ranking of the alternatives. Table 1 contains the rankings obtained for the three greenhouse gases, where the order 1 corresponds to the lowest risk and the order 5 to the highest risk.

Table 1: Risk rank orders of the alternatives for each greenhouse gas

	CO_2	CH_4	N_2O
Industrial processes	5	1	1
Agriculture	2	4	5
Forest and land use change	4	3	4
Waste	1	5	3
Energy	3	2	2

Based on the proposed risk classification method, one can conclude that A_1 (industrial processes) is classified as the riskiest sector considering the emissions of CO_2, whereas it is classified as least risky sector with respect to the emission of the other two greenhouse gases. As for CH_4, the sector A_4 (waste) is the riskiest sector for the emission of this greenhouse gas, however this sector was classified as having the lowest risk for CO_2 emission. Considering N_2O, the sector A_2 (agriculture) represents the highest risk for the emission of this greenhouse gas.

Acknowledgements Irene Brito thanks support from FCT through the projects:
UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020),
UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020).

References

[1] H. Brachinger and M. Weber. Risk as a primitive: A survey of measures of perceived risk. *OR Spektrum*, 19:235–250, 1997.

[2] Instituto Nacional de Estatística. IP-Portugal. Emissão de principais gases de efeito de estufa por tipo de gás e setor de emissão, 2024. Accessed 27 December 2024.

[3] EEA. Total net greenhouse gas emission trends and projections in Europe, 2024. Accessed 7 January 2025.

[4] A. Mardani, M. Nilashi, N. Zakuan, N. Loganathan, S. Soheilirad, M. Saman, and O. Ibrahim. A systematic review and meta-analysis of SWARA and WASPAS methods: Theory and applications with recent fuzzy developments. *Applied Soft Computing*, 57:265–292, 2017.

5 April, 9:40 - 10:00, Room 13.2

Supervised machine learning methodologies for longitudinal data

Elsa Soares[1], **Inês Sousa**[1]
[1] Centre of Mathematics, School of Sciences, University of Minho, Braga, Portugal, id10725@uminho.pt, isousa@math.uminho.pt

Longitudinal data are characterized by repeated measurements over time. Traditional statistical models applied to longitudinal data have been widely used, but are notable limitations. Machine learning techniques are emerging as promising alternatives, especially for high-dimensional data. This study reviews these methodologies, comparing it with traditional models and discusses challenges such as correlation among repeated measures and missing data, highlighting their clinical relevance.

Keywords: dynamic prediction, longitudinal statistical models, mixed-effects machine learning, multi-task learning, supervised machine learning

Longitudinal data, characterized by repeated measurements of individuals over time, are a cornerstone of analysis in various fields, including biomedicine, social sciences, and economics. In biomedical contexts, longitudinal data are especially critical for understanding disease progression, predicting individual trajectories, and informing real-time decision making. Proper analysis of such data can yield substantial benefits, such as early diagnosis, identification of at-risk individuals, and the development of targeted preventive or therapeutic strategies [1].

Traditionally, longitudinal data analysis has been dominated by statistical methods, particularly Generalized Linear Mixed Models, introduced by Nelder and Wedderburn in 1972. Although effective in capturing basic trends and describe statistical associations between variables, these models face notable limitations in complex and high-dimensional settings. Specifically, they require the specification of parametric forms for variable relationships, which are often unknown a priori.

Machine learning has emerged as a powerful alternative for analyzing longitudinal data, proving to be significant promise in various clinical contexts [4]. For example, this approach has outperformed traditional algorithms in tasks such as long-term prediction of chronic heart disease [3]. Additionally, the integration of longitudinal big data into machine learning models has been demonstrated to improve predictive accuracy and facilitate personalized healthcare solutions [2].

The primary objective of this study is to provide a comprehensive review of supervised machine learning methodologies for longitudinal data, including Summary Features (SF), Longitudinal Features (LF), Multiple Instance Learning (MIL), Stacked vertically (SV),

Multi-Task Learning (MTL), and Mixed-Effects Machine Learning (MEML). Particular emphasis will be placed on MTL and MEML, as these methodologies offer significant advantages in modeling complex temporal dependencies and managing data heterogeneity. Furthermore, these approaches will be critically compared with traditional longitudinal models to evaluate their respective strengths and limitations.

Acknowledgements We thank the Fundação para a Ciência e a Tecnologia (FCT) for the financial support provided through the doctoral scholarship with reference UI/BD/154394/2023

References

[1] A. Cascarano, J. Mur-Petit, J. Hernández-González, and et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review*, 56(Suppl 2):1711–1771, 2023.

[2] S. Chen, E. Grant, T. T. Wu, and F. D. B. Bowman. Statistical learning methods for longitudinal high-dimensional data. *Review of Computational Statistics*, 6(1):10–18, 2014.

[3] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015.

[4] A. Sheetal, Z. Jiang, and L. Di Milia. Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology*, 72(3):1339–1364, 2023.

Poster Sessions

4 April, 11:30 - 12:00, Hall of Grande Auditório

Predicting undergraduate dropout at the Polytechnic University of Coimbra

Marta Simões[1], **Joana Leite**[1,2,3], **António Paulino**[1], **Isabel Pedrosa**[1,2]
[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal, iscac17289@alumni.iscac.pt, jleite@iscac.pt, antonio.paulino@estgoh.ipc.pt, ipedrosa@iscac.pt
[2] CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Portugal
[3] Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal

Student dropout is a current concern in Portuguese Higher Education Institutions due to its substantial implications, requiring the creation of evidence-based strategies. This study develops a predictive model for undergraduate dropout at the Polytechnic University of Coimbra, identifying key predictors from the institution's database. To facilitate adoption within the academic community, standard and interpretable machine learning algorithms are employed.

Keywords: dropout, higher education, prediction, machine learning, interpretability

Student dropout in higher education is a critical concern globally, particularly in Portugal, where recent statistics reveal that 11.1% of first-year students at public institutions dropped out of their courses during the 2021/22 academic year. This marks a noticeable increase of two percentage points compared to the 2019/20 academic year, according to the Directorate General of Education and Science Statistics (DGEEC) in the infocurso.pt portal, underscoring the urgent need for targeted interventions aimed at improving student retention. Dropout not only affects the academic and professional futures of students but also has far-reaching consequences for higher education institutions and society at large, including wasted resources, reduced workforce competitiveness, and social instability. In response to this challenge, this study aims to develop an interpretable predictive model for undergraduate dropout at the Polytechnic University of Coimbra (IPC), leveraging institutional data to identify its key predictors. The goal is to contribute to a more effective understanding of the factors leading to dropout and provide actionable insights for developing targeted retention strategies.

The significance of interpretability in machine learning models, particularly within educational settings, cannot be overstated. As defined in [1], interpretability refers to a system where users can not only see the outputs but also understand how inputs are mathematically mapped to those outputs. In the context of this study, interpretability enables stakeholders such as university administrators, faculty, and policymakers to comprehend

the reasoning behind predictions. This transparency enhances trust in the model and facilitates its adoption for informed decision-making.
The literature review plays a pivotal role in identifying key predictors of undergraduate dropout, drawing from studies that highlight a range of factors influencing student retention [2, 3]. These factors can be categorized into four dimensions: organizational (e.g., teaching methods, institutional conditions), personal life management (e.g., financial difficulties, family background), professional (e.g., employability, alignment with the course), and relational/social (e.g., social integration, peer relationships). The study uses these dimensions to guide the collection of data from the institution's database, which has data from the six organic units of IPC, for the 2020/21 and 2021/22 academic years. Data preprocessing techniques, including normalization and handling of missing values, ensure the quality and consistency of the data. Feature selection methods are then applied to focus on the most relevant predictors for dropout risk.
To ensure the results are accessible and actionable, two interpretable machine learning algorithms – k-nearest neighbors and decision trees – are used. The model's performance is assessed using accuracy, precision, and recall. Given the imbalanced nature of the data – dropout cases are less frequent than non-dropout cases – the precision metric is particularly important, as it indicates the proportion of predicted dropouts that are true dropouts. The results show interesting performance, with high accuracy and a promising level of precision, suggesting that the model can effectively identify at-risk students.

Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UID/05422: Centre for organisational and Social Studies of Polytechnic of Porto.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] P. Barroso, I. Oliveira, D. Noronha-Sousa, A. Noronha, C. Mateus, E. Vázquez-Justo, and C. Costa-Lobo. Dropout factors in higher education: A literature review. *Psicologia Escolar e Educacional*, 26, 2022.

[3] V. Realinho, J. Machado, L. Baptista, and M. V. Martins. Predicting student dropout and academic success. *Data*, 7:146, 10 2022.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Open government data: a global perspective with a focus on Portugal

Inês Rocha[1], **Clara Viseu**[1,2,3], **Manuela Larguinho**[1,2,3]
[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093, Coimbra, Portugal, a2022126914@alumni.iscac.pt, cviseu@iscac.pt, mlarguinho@iscac.pt
[2] CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Portugal
[3] Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal

This study provides an overview of Open Government Data (OGD), focusing on its components, principles, and perspectives. Explores the benefits and barriers associated with OGD and highlights its alignment with the Sustainable Development Goals. The study also examines the maturity of OGD in Portugal, with an emphasis on national and European evaluations. By addressing these topics, the study aims to foster a deeper understanding of the role of OGD in promoting transparency, innovation, and sustainable growth.

Keywords: open government data, Portugal, sustainable development goals, data accessibility, open data maturity

In today's increasingly technological world, vast amounts of data are being generated daily, which have become valuable for governments, businesses, and society as a whole. However, much of this data remains inaccessible due to storage and usage limitations [4]. Open Government Data (OGD) emerged as a solution to this problem, aiming to make public data freely available and reusable without restrictions [2]. Governments have implemented legislative measures and regulations that promote public access to the data they generate, thereby encouraging citizen participation and creating a global repository for data and information (Agência para a Modernização Administrativa, 2016). Then, the main goal of OGD is to create public value by ensuring that data is not only accessible but also effectively utilized by citizens. Governments, leveraging Information and Communication Technologies (ICT), play a crucial role in this process by supporting the opening of data and fostering public participation [3]. Through ICT, government data becomes available online in formats that are machine-readable and easily processed, enabling businesses and citizens to access, reuse, and generate new products or services. Actually, OGD brings multiple benefits, such as providing vital information on public services including healthcare, the environment, demographics, and education. Despite its many advantages, OGD adoption faces several barriers that prevent its full potential from being realized. These

barriers, identified by some authors, include institutional challenges, implementation difficulties, legal compliance issues, and concerns over data quality and technical access. Addressing these barriers is essential for overcoming the ambiguity that surrounds the widespread adoption of OGD. Moreover, OGD plays a crucial role in the achievement of the United Nations' Sustainable Development Goals by providing essential data on natural resources, government operations, public services, and demographic characteristics [1]. Open data allows countries to define priorities and develop strategies to address both national and global challenges. Turning to Portugal, the "dados.gov.pt" portal, launched in 2011, serves as the country's central platform for open data. It hosts a wide range of datasets provided by public organizations, which can be downloaded, modified, and reused by citizens, businesses, and other entities for diverse purposes (Agência para a Modernização Administrativa, 2016). Portugal's OGD initiatives are evaluated annually within the European context, with assessments focusing on key dimensions such as policy, portal infrastructure, impact, and data quality. These evaluations aim to measure how effectively Portugal manages and publishes open data, helping to identify areas for improvement and ensuring alignment with broader European standards. Through these evaluations, Portugal's progress in advancing open data is monitored, supporting ongoing improvements in data accessibility and usage.

References

[1] J. Gurin, L. Manley, and A. Ariss. Sustainable development goals and open data. *https://blogs.worldbank.org/en/digital-development/sustainable-development-goals-and-open-data*, 2015.

[2] M. A. Hossain, Y. K. Dwivedi, and N. P. Rana. State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2):14–40, 2016.

[3] M. Lněnička, A. Nikiforova, S. Saxena, and P. Singh. Investigation into the adoption of open government data among students: the behavioural intention-based comparative analysis of three countries. *Journal of Information Management*, 74(3):549–567, 2022.

[4] B. W. Wirtz, J. C. Weyerer, M. Becker, and W. M. Müller. Open government data: A systematic literature review of empirical research. *Electronic Markets*, 32(4):2381–2404, 2022.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Selection of variables influencing math scores in PISA data using LASSO and elastic net

Beatriz Silva[1], **Susana Faria**[1]
[1] Centre of Mathematics, Department of Mathematics, University of Minho,
a102385@alunos.uminho.pt, sfaria@math.uminho.pt

The importance of mathematical literacy is increasingly being recognized. However, various factors can impact students' mathematical literacy, resulting in diverse performance patterns. This study assessed the performance of LASSO and Elastic Net methods, along with the traditional forward stepwise regression approach, in selecting predictor variables associated with students' mathematics performance based on PISA 2022 data.

Keywords: mathematical literacy, PISA 2022, variable selection, LASSO, elastic net

Variable selection has become highly important in linear regression models, particularly in recent years, due to the growing complexity and dimensionality of most datasets, making it an essential step in any modeling process.
The development of new variable selection methods has become essential to address challenges related to computational complexity. Methods based on penalty functions have emerged as a solution, as they shrink a subset of coefficient estimates to zero. Among these, the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani [3] and Elastic Net (Enet) proposed by [4] are widely regarded as effective approaches, enabling both variable selection and parameter estimation in a model.
Penalized regression methods provide a powerful approach to analyzing the relationships within the extensive datasets collected in large-scale educational assessment studies. An example is the Programme for International Student Assessment (PISA)[2], an international evaluation that collects data from students, teachers, and parents to assess the skills and knowledge of 15-year-old students in mathematics, reading, and science.
This study applies both penalized regression techniques and the traditional forward stepwise regression approach to select predictor variables for Portuguese students' mathematics scores, using data from the PISA 2022. Specifically, it seeks to examine the impact of students' backgrounds, attitudes toward mathematics, home environment, parental involvement, and school-related factors on their mathematical literacy.
The dataset includes 6793 Portuguese students from 224 schools, drawn from the PISA 2022 assessment. Students' mathematics performance scores were used as the outcome variable, with 44 variables from the student questionnaires using as predictors.
Students' economic, social, and cultural status (ESCS), gender, repeat, self-efficacy toward formal and applied mathematics, preference for mathematics, students' access to ICT

at home, family support for self-directed learning were the most influential predictors associated with students' math performance.
Both LASSO regression and Enet produced similar findings, selecting fewer variables by shrinking insignificant ones to zero. This approach improves model interpretability and reduces the risk of overfitting.

Acknowledgements The research at CMAT 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.

References

[1] T. Hastie and R. Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35:579–592, 2029.

[2] OECD. Pisa 2022 technical report. OECD Publishing, Paris, 2023.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[4] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67:301–320, 2005.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Bitcoin price prediction with statistical and neural networks forecast models

João Peixoto[1], **Carlos Grilo**[2], **José Martins**[3]
[1] ESTG, Polytechnic of Leiria, 2220146@ipleiria.pt
[2] CIIC, ESTG, Polytechnic of Leiria, carlos.grilo@ipleiria.pt
[3] LIAAD-INESC TEC and ESTG, Polytechnic of Leiria, jmmartins@ipleiria.pt

Bitcoin's capped supply and price volatility have made it a popular speculative asset, drawing attention from investors seeking accurate forecasts to manage risk and seize opportunities. This study compares the performance of various forecasting models, from traditional statistical methods to advanced neural networks, providing insights into their ability to analyze price dynamics and predict Bitcoin's price movements.

Keywords: bitcoin, forecast, time-series analysis

The aim of this study is to predict bitcoin next day closing price [2]. Daily closing prices in US dollars for 10 complete years, from 2015 to 2024, were extracted from *yfinance*[1] using the ticker BTC-USD. Exogenous variables were also extracted from yfinance for the same time period. These included the US Dollar Index, Gold, S&P 500, Treasury Yield 5 Years, and Oil Prices. The price of Bitcoin during this period, illustrated in Figure 1, exhibited substantial volatility and growth, starting at $314.25 in January 2015 and climbing to an all-time high of $106,140.60 by December 2024. Data was divided into three subsets: training, validation, and testing. The training set covered 2015–2020, the validation set spanned 2021–2022, and the test set included 2023–2024.

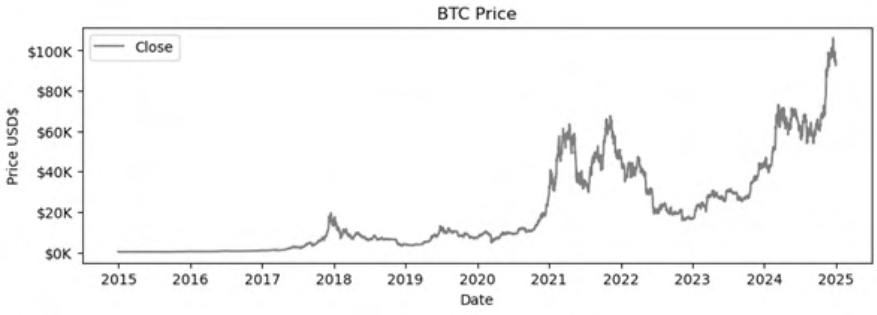

Figure 1: BTC Daily Closing Price 2015-2024

[1] https://pypi.org/project/yfinance/

Multiple models were tested, beginning with Auto Regressive Integrated Moving Average (ARIMA) and extending to neural network architectures, including Multilayer Perceptrons (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), DeepAR, Neural Basis Expansion Analysis Time Series Forecasting (N-BEATS), and Neural Hierarchical Interpolation for Time Series Forecasting (N-HITS). More recent models introduced in 2024, such as Patch Time Series Transformer (PatchTST), Time-series Dense Encoder (TiDE), and Kolmogorov-Arnold Networks (KAN) [1], were also evaluated. All models underwent automated hyperparameter fine-tuning using Mean Absolute Error (MAE) as the loss function. Two approaches were tested: one predicting the closing price directly and another forecasting Bitcoin's daily returns. Models were evaluated using Mean Absolute Percentage Error (MAPE) as performance metric, summarized in Table 1. To mitigate neural network randomness, results represent the mean of 30 runs with sequential random seeds and predicted returns were converted to price to standardize metrics for comparison.

Table 1: MAPE by model

Model	Price	Return	Price with Exogenous	Return with Exogenous
ARIMA	**1.777**	-	-	-
MLP	1.819	**1.780**	1.793	1.781
LSTM	3.549	1.796	4.052	**1.787**
GRU	6.082	**1.775**	4.490	1.981
DeepAR	3.146	**1.857**	-	-
NBEATSx	1.781	1.781	**1.781**	1.848
NHITS	**1.785**	1.791	1.787	1.859
PatchTST	**1.769**	1.853	-	-
Tide	1.782	1.784	**1.782**	1.786
KAN	1.780	1.781	1.785	**1.777**

The best MAPE for each model ranged from 1.769% to 1.857%, demonstrating that all models achieved high performance with minimal variation, effectively fitting the training data and generalizing to unseen data. However, when applied to a simple trading strategy—buying or holding Bitcoin if a price increase was predicted and selling if a decrease was forecast—only TiDE and N-BEATS marginally outperformed a basic buy-and-hold approach during the testing period. This suggests that for trading purposes, models may require optimization with alternative loss functions that minimize errors based on their impact on trading outcomes.

References

[1] Z. Liu et al. Kan: Kolmogorov-Arnold networks, 2024, https://arxiv.org/abs/2404.19756.

[2] S. Smyl, G. Dudek, and P. Pełka. Forecasting cryptocurrency prices using contextual es-adrnn with exogenous variables. In *Computational Science – ICCS 2023: Proceedings, Part I*, page 450–464. Springer-Verlag, 2023.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Forecasting Ibovespa: statistical *vs* neural models

Elysiario Santos[1], **Carlos Grilo**[2], **José Martins**[3]
[1] ESTG, Polytechinic of Leiria, 2220153@my.ipleiria.pt
[2] CIIC, ESTG, Polytechinic of Leiria, carlos.grilo@ipleiria.pt
[3] LIAAD-INESC TEC and ESTG, Polytechinic of Leiria, jmmartins@ipleiria.pt

This paper analyzes the predictive accuracy of statistical models and neural networks for forecasting the closing values of the Brazilian stock index (Ibovespa). By comparing traditional approaches, such as ARIMA, with modern machine learning models, including MLP, LSTM, TiDE, and KAN, we evaluate forecast errors using metrics like MAPE and RMSE. The results show that while recent algorithms perform well, traditional methods such as ARIMA and MLP achieve competitive accuracy, offering valuable insights to academia and investors.

Keywords: neural networks, ARIMA, Ibovespa, stocks index, forecast

The Brazilian financial market offers a diverse ecosystem, balancing stability and volatility through fixed and variable income assets. At the heart of this landscape is B3, the country's sole stock exchange, reflecting Brazil's economy and comprising over 400 listed companies in various sectors. This paper focuses on the Ibovespa, an index tracking the performance of B3's major stocks, aiming to forecast its closing values using traditional econometric models and advanced machine learning techniques.

The methodology involved selecting 10 years of historical data, from January 2013 to December 2023, normalizing them for consistency, and splitting the data set into training sets (75%), validation sets (10%) and tests (15%) [1]. Different configurations were used for ARIMA and neural networks. The AutoARIMA function and the Optuna library were used for hyperparameter optimization, and a sliding window approach was used so that a fair comparison could be made between ARIMA and neural networks [2]. Neural network predictions were denormalized before error calculation to ensure a correct comparison with real values. To ensure statistical significance, the results shown for neural networks are averages of 100 runs.

The results presented in Table 1 demonstrate consistent metrics, highlighting the reliability of the evaluations. We can see that all models achieve a MAPE value of less than 1%, except LSTM, GRU and DeepAR. ARIMA achieved the highest R^2 value (0.9700), indicating its superior ability to explain data variability, while also maintaining low error metrics (MAE of 1085.9085 and RMSE of 1370.4918). Among the neural network models, NHITS achieved the best overall performance, with the lowest MAE (1065.5834), RMSE (1357.6058), and MAPE (0.9573). In contrast, LSTM exhibited the highest error metrics, with a MAE of 1423.9116 and an RMSE of 1728.3206, despite achieving an R^2 of 0.9488.

Table 1: Results of Ibovespa Predictions

Model	MAE	RMSE	MAPE	R2
ARIMA	1085.9085	1370.4918	0.9671	**0.9700**
LSTM	1423.9116	1728.3206	1.2758	0.9488
GRU	1366.8244	1731.4238	1.2156	0.9478
MLP	1077.0929	1373.6700	0.9675	0.9675
NBEATS	1069.0547	1358.6425	0.9608	0.9683
NHITS	**1065.5834**	**1357.6058**	**0.9573**	0.9683
DeepAR	1373.1770	1685.2074	1.2350	0.9511
PatchTST	1069.8394	1366.6299	0.9615	0.9679
TiDE	1069.3313	1358.9113	0.9611	0.9682
KAN	1075.2727	1373.1950	0.9665	0.9676

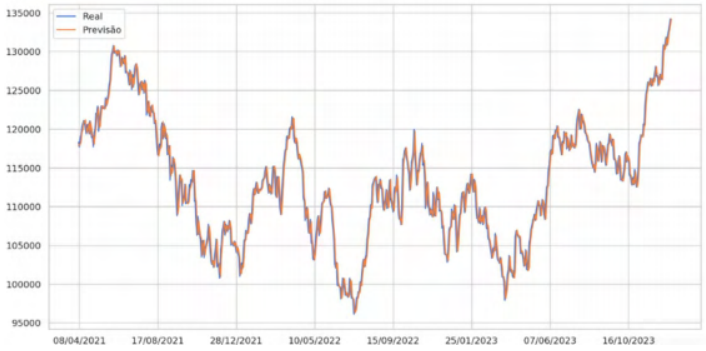

Figure 1: ARIMA Model, real *versus* predicted values

Figure 1 shows the values predicted by the ARIMA model compared to the real values. We can see that the predictions follow the real values very closely.
The results indicate that the models' performances are closely aligned, with slight variations in accuracy and error metrics. Overall, the models performed consistently well on the different evaluation criteria.

References

[1] F. Chollet. *Deep Learning with Python*. Manning Publications Co., 2018.

[2] R. M. Rapp. Grid Search Approach to Select and Calibrate Exponential Smoothing, SARIMA and LSTM Models for Demand Forecasting. Dissertação de mestrado, Escola Politécnica, Universidade de São Paulo, São Paulo, 2023.

4 April, 11:30 - 12:00, Hall of Grande Auditório

The perceived value of cooperative membership in organic cocoa production: PLSc-SEM approach

Ibrahim Prazeres[1], **Maria Raquel Lucas**[1,2], **Ana Marta-Costa**[3,4], **Pedro Damião Henriques**[1,5], **Luís M. Grilo**[6,7,8]
[1] MED (Mediterranean Institute for Agriculture, Environment and Development) & CHANGE (Global Change and Sustainability Institute), Portugal, ibrahim.prazeres@uevora.pt
[2] Department of Management, University of Évora, Portugal, mrlucas@uevora.pt
[3] CETRAD (Centre for Transdisciplinary Development Studies), University of Trás-os-Montes e Alto Douro, Vila Real, Portugal, amarta@utad.pt
[4] Department of Economics, Sociology and Management, University of Trás-os-Montes e Alto Douro, Portugal
[5] Department of Economics, University of Évora, Portugal, pdamiao@uevora.pt
[6] Department of Mathematics, University of Évora, Portugal, luis.grilo@uevora.pt
[7] CIMA (Research Center in Mathematics and Applications), University of Évora, Portugal
[8] NOVA Math (Center for Mathematics and Applications), NOVA University of Lisbon, Portugal

---

Cooperatives are recognized as key drivers of organic production, sustainability, and inclusive business, but little is known about the values connecting producers to cooperatives. To assess these values in São Tomé e Príncipe's organic cocoa cooperatives, a Structural Equation Model was proposed with "functional value" as an exogenous construct, and "social," "emotional," and "monetary" values as endogenous. Using Partial Least Squares and a convenience sample of cocoa producers, the model reveals key relationships between these values.

Keywords: organic cocoa, PERVAL, São Tomé e Príncipe, structural equation model, survey

---

Cooperatives play a critical role in discussions on inclusion, rural development, food security, and agricultural sustainability, as well as their impact on income, poverty reduction, and producers' livelihood strategies [2]. However, limited research explores the value constructs that connect rural producers to cooperatives and their membership experiences. Traditionally, this value has been examined through monetary or social dimensions [4]. Yet, other dimensions from the PERVAL scale [3], encompassing functional, emotional, monetary, and social values, can be explored. This study examines how these dimensions' influence perceptions of cooperative membership among organic cocoa producers in São

Tomé e Príncipe. Data were gathered through face-to-face questionnaires conducted between June and December 2021 with a convenience sample of 400 organic cocoa producers, representing 12,2% of the total population of 3,274 producers. Participants were affiliated with the two main cooperatives—CECAB and CECAC11—or smaller residual organizations. Of the respondents, 32% were women, 61.5% were aged 40–60, and 78,5% had primary school education. These cooperatives act as vital intermediaries, linking farmers to national and international chocolate markets while providing resources, information, and support to cocoa producers. In this study, research hypotheses were formulated based on the scientific literature and also considering the empirical knowledge of cooperative leaders. A reflective Structural Equation Model (SEM) was then proposed, where "functional perceived value" served as a latent exogenous construct and the endogenous constructs were "emotional", "monetary" and "social" values. We used the consistent Partial Least Squares (PLSc) estimator, which has been used to handle complex models, with Likert-type scale (ordinal variables) and non-normal data [1]. The Bootstrap resampling process [1] was used to assess the model's goodness-of-fit and confirmed that its parameters are statistically significant at a significance level of 5%. In the model estimated with PLSc-SEM, the largest positive direct effect was found between "functional perceived value" and the "emotional" (0.742; $p < 0.001$), where the amount of variance explained (coefficient of determination) was $R^2 = 55.1\%$. The insights from this research contribute to a better understanding of cooperative value and its implications for governance strategies in organic cocoa production.

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/04011/2020 (`https://doi.org/10.54499/UIDB/04011/2020`), UIDB/05183/2020 (`https://doi.org/10.54499/UIDB/05183/2020`; `https://doi.org/10.54499/LA/P/0121/2020`) and UIDB/04674/2020, DOI 10.54499/UIDB/04674/2020 (`https://doi.org/10.54499/UIDB/04674/2020`). Funding was additionally provided by FCT under the research contract PRT/BD/152273/2021 to Ibrahim Prazeres.

References

[1] 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.* Cham: Springer, Switzerland, 2022.

[2] I. Prazeres, M. R. Lucas, A. Marta-Costa, and P. Henriques. Organic cocoa farmer's strategies and sustainability. *Bio-based and Applied Economics*, 12(1):37–52, 2023.

[3] J. C. Sweeney and G. N. Soutar. Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, 77:203–220, 2001.

[4] M. Wrede. Social benefits of cooperatives - an economic perspective. *Zeitschrift für das Gesamte Genossenschaftswesen*, 73(4):232–238, 2023.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Analysis of consumer perceived value in food products using PLSc-SEM

Eunice Venâncio[1], Maria Raquel Lucas[1,2], Ana Marta-Costa[3,4], Pedro Damião Henriques[1,5], Luís M. Grilo[6,7,8]
[1] MED (Mediterranean Institute for Agriculture, Environment and Development) & CHANGE (Global Change and Sustainability Institute), Portugal, eunycassia@gmail.com
[2] Department of Management, University of Évora, Portugal, mrlucas@uevora.pt
[3] CETRAD (Centre for Transdisciplinary Development Studies), University of Trás-os-Montes e Alto Douro, Vila Real, Portugal, amarta@utad.pt
[4] Department of Economics, Sociology and Management, University of Trás-os-Montes e Alto Douro, Portugal
[5] Department of Economics, University of Évora, Portugal, pdamiao@uevora.pt
[6] Department of Mathematics, University of Évora, Portugal, luis.grilo@uevora.pt
[7] CIMA (Research Center in Mathematics and Applications), University of Évora, Portugal
[8] NOVA Math (Center for Mathematics and Applications), NOVA University of Lisbon, Portugal

This study examines "consumer perceived value in food" choices using a Structural Equation Model considering the dimensions — "functional/quality", "emotional", "social", "price" and "environmental" — as latent constructs, and where "satisfaction" and "consumption" are the target constructs. Based on a sample of consumers, obtained through survey adapted from the PERVAL scale, a model was estimated with consistent Partial Least Squares estimator. Preliminary results show that "functionality/quality" and "price" have the strongest influence on "consumer satisfaction".

Keywords: consumer, food, PERVAL, PLS predict, survey

The overall perception of value in food products has gained increasing importance, driven by concerns about quality, emotional responses, social influences, price, and environmental practices. According to existing literature, each of these dimensions contributes uniquely but complementarily to perceived value, impacting both satisfaction and purchase or consumption decisions [1, 2]. To investigate these relationships, a Structural Equation Model (SEM) was proposed, postulating that the five dimensions—"functional/quality", "emotional", "social", "price", and "environmental"—act as exogenous latent constructs predicting "satisfaction", which in turn is directly linked to food product consumption. A convenience sample (444 valid responses) was collected via an online questionnaire based

on the PERVAL scale [4], conducted in Portugal in 2024. The dataset is relatively heterogeneous in terms of age, education, and food consumption habits: 71.6% were women, 50% were between 31 and 50 years old; 34.2% of respondents were employed and 42.3% were single. The 5 point Likert scale measured participant agreement with each indicator (observed variable). The consistent Partial Least Squares estimator (PLSc) [3], which does not assume multivariate normal distribution and is recommended for complex and also early-stage models, was used to estimate and validate the proposed model. Initial evaluation of the measurement model showed satisfactory convergent validity, with outer loadings exceeding 0.70, composite reliability (CR) between 0.82 and 0.90, and Average Variance Extracted (AVE) values ranging from 0.56 to 0.64. Preliminary results from the structural model indicated positive and statistically significant effects (standardized path coefficients) of all five perceived value dimensions on "satisfaction". "Functional/quality" (0.42, $p < 0.001$) and "price" (0.38, $p < 0.001$) had the greatest impact, followed by "emotional" (0.31, $p = 0.033$) and "environmental" (0.29, $p = 0.021$) dimensions. These relationships explained 64% of the variance in "satisfaction" ($R^2 = 64\%$), highlighting the significant role of perceived value in shaping "satisfaction". The robustness of the estimates, obtained with the PLSc-SEM, was confirmed through bootstrap resampling, while the predictive capacity of the model was assessed using the PLSpredict algorithm and Root Mean Square Error (RMSE). None of the "consumption" indicators exceeded the prediction errors of a simple linear model, indicating strong predictive validity of the estimated model.

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/04011/2020 (`https://doi.org/10.54499/UIDB/04011/2020`), UIDB/05183/2020 (`https://doi.org/10.54499/UIDB/05183/2020`; `https://doi.org/10.54499/LA/P/0121/2020`) and UIDB/04674/2020, DOI 10.54499/UIDB/04674/2020 (`https://doi.org/10.54499/UIDB/04674/2020`). Funding was additionally provided by FCT under the research contract PRT/BD/152339/2021 to Eunice Venâncio.

References

[1] L. L. Chang and B. King. The role of perceived value in understanding tourist experience and post experience at heritage destinations. *Jurnal Siasat Bisnis*, 26(1):36–49, 2022.

[2] J.-H. Cheah, H. Ting, T. Ramayah, M. A. Memon, T. H. Cham, and E. Ciavolino. A comparison of five reflective–formative estimation approaches: Reconsideration and recommendations for tourism research. *Quality and Quantity*, 54(4):1421–1455, 2020.

[3] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM) (3rd ed.)*. Thousand Oaks, CA: Sage Publications, Switzerland, 2022.

[4] J. C. Sweeney and G. N. Soutar. Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, 77(2):203–220, 2001.

4 April, 11:30 - 12:00, Hall of Grande Auditório

The role of social media and influencers in restaurant decision-making

Carla Henriques[1], **Suzanne Amaro**[2], **Ana Desiderati**[3]
[1] Polytechnic Institute of Viseu and CMUC, Portugal, carlahenriq@estgv.ipv.pt
[2] Polytechnic Institute of Viseu and CISeD, Portugal, samaro@estgv.ipv.pt
[3] Polytechnic Institute of Viseu, Portugal, anabdesiderati@hotmail.com

This study seeks to understand how Portuguese and Brazilians are influenced by social media and digital influencers when choosing a restaurant, using data from 169 Portuguese and 338 Brazilian respondents.

Keywords: social media, restaurants, influencer, data analysis

In today's digital age, the widespread reliance on information from social media in the purchasing decision process is undeniable, and the restaurant sector is no exception [1]. Social media platforms serve not only as a source of information for decision-making but also as a means of engaging with favourite restaurants. This study investigates the influence of social media and digital influencers on restaurant selection, using data collected through a questionnaire. The sample comprised 169 Portuguese and 338 Brazilian respondents, allowing for cross-national comparisons of consumer behaviour. Our findings revealed that over 60% of respondents follow their favourite restaurants on social media (65.7%, 95% CI: 61–70%) and similarly, over 60% often follow digital influencers (66.3%, 95% CI: 62–70%). These results underscore the importance of understanding how these platforms can be leveraged for customer acquisition and retention. Data were analysed using factor analysis for dimensionality reduction, regression models, and non-parametric tests for group comparisons.

The percentage of Brazilian respondents who report following their favourite restaurants on social media at least occasionally (69%) is significantly higher than that of Portuguese respondents (59%; $p = 0.017$). This propensity is also more prevalent among women than men (70% vs. 52%, $p < 0.001$) and appears to exhibit a non-linear relationship with age, increasing until approximately 30 years old and then decreasing. It was also found that individuals who tend to follow restaurants on social media are not necessarily those who spend more time daily on these platforms; that is, this behaviour does not appear to be driven by a potential addiction to social media engagement. However, it is noteworthy that following favourite restaurants on social media is associated with higher restaurant expenditure (median of 20–30 Euros vs 15–20 Euros, per-person per visit, $p = 0.005$). In other words, these individuals tend to spend significantly more money per restaurant visit. Following restaurants on social media is also associated with following digital influencers

(70% follow digital influencers, compared to only 57% in the group that does not report following their favourite restaurants on social media, p = 0.005).
A considerable proportion of respondents (70%, 95% CI: 66–74%) acknowledged that information obtained on social media influences their restaurant selection decisions. This influence is more pronounced among Brazilians than Portuguese (75% vs. 60%, $p < 0.001$) and among women than men (75% vs. 54%, $p < 0.001$), but no association was found with restaurant visit frequency or expenditure. Thus, while many individuals use social media to aid in their decision-making process, this behaviour was not significantly associated with those who frequent restaurants more often or spend more.
No significant differences in Brazilian nationality, gender, age, or education level were observed between respondents who reported being potentially influenced by digital influencers and those who did not. This suggests that the influence of digital influencers on restaurant decision may now be widespread across various demographic groups. However, a significant difference emerged in restaurant expenditure, with the influencer-influenced group reporting higher median spending (median of 20–30 Euros vs. 15–20 Euros, per person per visit, p = 0.004).
Using specific photo and video posts included in the questionnaire, respondents were asked about their intention to visit the featured restaurant. Respondents were randomly assigned to view either a restaurant-generated post or an influencer-generated post. Intention to visit was significantly higher for restaurant-generated photo posts compared to influencer-generated photo posts (p = 0.002); no significant difference was observed for video posts. Thus, in this context, influencers did not offer an advantage over restaurant-generated content. Regarding positive attitudes toward the posts, no significant differences were found between restaurant and influencer video posts; however, for photo posts, restaurants received significantly higher ratings ($p < 0.001$). Influencers demonstrated a relative advantage only in perceived post credibility for video content (p = 0.006). It is important to acknowledge that these findings may be influenced by the specific stimuli used in the questionnaire. While these results offer preliminary insights into the comparative impact of influencer and restaurant-generated content, further research employing diverse stimuli and methodologies is necessary to confirm these observations.

Acknowledgements Partially supported by the Centre for Mathematics of the University of Coimbra (funded by the Portuguese Government through FCT/MCTES, DOI 10.54499/UIDB/00324/2020).

References

[1] C. J. F. Anjos, S. Marques, and A. Dias. The impact of instagram influencer marketing in the restaurant industry. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 13:1–20, 2022.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Organisational climate and employee well-being

Vera Valente[1], Cláudia Amanajás[1], Hugo Carvalho[1], Cristina Lopes[2], Isabel Vieira[2], Lurdes Babo[2], Cristina Torres[2]
[1] ISCAP, Instituto Politécnico do Porto, 2202143@iscap.ipp.pt, 2240334@iscap.ipp.pt, 2191144@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, cristinalopes@iscap.ipp.pt, mivieira@iscap.ipp.pt, lbabo@iscap.ipp.pt, ctorres@iscap.ipp.pt

This study analyses factors related to organisational climate and employee well-being. Results show no gender differences in job satisfaction. However, men and women have different perceptions of work-life balance, with men tending to feel they have a better balance. The study's findings underscore the significance of key predictor variables such as the general working environment, fair payment, work recognition, and an inclusive and diverse work environment. These variables collectively account for approximately 71% of the observed variance in job satisfaction.

Keywords: satisfaction, organisational climate, well-being, generation, hypothesis testing

The labour market is in constant transformation, driven by generational changes that redefine mentalities and priorities. The organisational climate and employees well-being have come to play a prominent role when it comes to hiring and retention and are, in fact, interconnected and essential elements for the performance and longevity of companies.
To analyse these elements, a questionnaire was developed with 17 closed questions (evaluated in a 5-level Likert scale) and one open question, and 129 responses were collected between October and November 2024. The survey included questions such as "How would you rate the general working environment in the company?"; "Does your company promote an inclusive and diverse working environment?"; "Do you think your pay is fair compared to your peers?"; "Do you feel that you are valued for your work in the company?".
The majority of the sample were female (69%), and had a higher education degree (74.4%) or secondary education (22.5%). The majority of respondents worked face-to-face (75.2%), 20.2% worked on a hybrid basis and only 4.6% worked remotely. The sample covered different generations currently in the labour market, with 54.3% of participants from Generation X, followed by 33.3% of Millenials, 9.3% of Generation Z and 3.1% Baby Boomers. Fig. 1 shows that the generations who have been in the labour market for a longer period of time are more satisfied than the younger generations.
The Kruskal Wallis test (p-value = 0.019) confirms the existence of significant differences in job satisfaction between the generations. This conclusion is in line with various generational studies [1] which indicate that baby boomers are less focused on balance and more willing

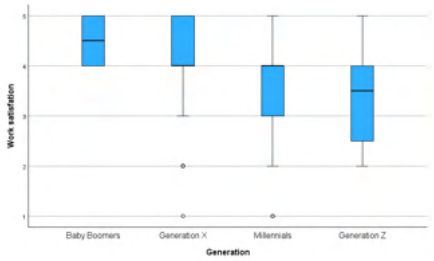

Figure 1: Job satisfaction by generation

to work longer hours, and that younger generations expect a healthy work-life balance, with a strong preference for flexibility and well-being.
Various factors were analysed to assess organisational climate and employee well-being. The analysis showed that there are no significant differences in job satisfaction between genders (Mann-Whitney test: p-value $= 0.389$). However, men and women have different perceptions of the balance between personal and professional life. Men tend to feel they have a better balance, although the difference is not very significant (Mann-Whitney test: p-value $= 0.086$). The linear regression model (Table 1) shows that 70.8% of overall job satisfaction varies according to the work environment, fair pay compared to peers, professional recognition and an inclusive work environment, while the remaining 29.2% is unexplained, i.e. due to other factors not included in the model.

Table 1: Overall job satisfaction linear regression model

Variable	Coefficient	p-value
Constant	0.217	0.332
General working environment	0.322	< 0.001
Fair payment	0.186	0.002
Work recognition	0.182	0.008
Inclusive atmosphere	0.302	< 0.001

The findings of this study indicate that organisations must implement inclusive and diverse working environments, fair remuneration, and the appreciation of employees in order to retain high-calibre personnel.

Acknowledgement: This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., as part of project UID/05422: Centro de Estudos Organizacionais e Sociais do Politécnico do Porto.

References

[1] L. J. Brown. *Millennials and Work-life Balance: Comparisons Across Generations*. PhD thesis, Liberty University, 2023.

4 April, 11:30 - 12:00, Hall of Grande Auditório

A support vector machine model for stock risk classification

Faustino Sachimuco[1], **Gaspar J. Machado**[2], **Irene Brito**[3]
[1] Universidade do Minho, faustino.sachimuco@uninbe.ao
[2] Centro de Matemática e Departamento de Matemática da Universidade do Minho e Centro de Física das Universidades do Minho e do Porto, gjm@math.uminho.pt
[3] Centro de Matemática, Departamento de Matemática, Universidade do Minho, ireneb@math.uminho.pt

---

In this work a support vector machine model is proposed that can be used for stock risk classification, which is relevant for building efficient portfolios in investment decision problems. The model is constructed on the basis of various risk-related attributes from stock returns over a given time period. The method is applied to data from the Portuguese stock market index.

Keywords: support vector machine, risk assessment, risk measures, stock classification

---

Recently, different machine learning approaches were proposed for assessing the risk of a portfolio. Most of these approaches combine a machine learning model for the selection of stocks with the classical mean-variance optimization model, see e.g. [2], or with modifications of this model, see e.g. [3], where the optimization was based on maximizing the Sharpe ratio. Risk is in these cases expressed by variance. On the other hand, methods were developed where risk is assessed by other measures, see for example [1], who used tail risk measures, such as value at risk and expected shortfall.
The aim of the present work is to develop a support vector machine (SVM) model to classify stocks, where the purpose is to assess risk based on the most relevant risk attributes for an efficient portfolio construction. Given a set of stocks with returns collected over a given in-sample period, the method consists first in assessing the stocks considering a combination of the mean with one of the following risk measures: variance, upper semivariance, loss probability, entropy, value at risk, conditional value at risk. Regarding the combination of risk measures, the experiments will be performed using the weighted sum model and the weighted product model. For each combination of risk measures, 50% of the best classified stocks are selected forming a set with lowest risk. Then, equally weighted portfolios are formed with the best classified stocks. The performance of these portfolios is evaluated in an out-of-sample period using cumulative returns, Sharpe ratio, Sortino ratio and Beta value. The performances are compared with the benchmark portfolio's performance. An aggregated ranking permits selecting the most efficient portfolio. The associated combination of risk measures is identified. These risk measures are the most relevant ones for an efficient portfolio construction and serve as attributes for the construction of the SVM

model. The proposed SVM classifier separates the two-dimensional risk attribute space in two regions, a high risk region and a low risk region. The methodology is applied to stocks of the Portuguese stock market PSI using data from January 2021 to December 2023 for the in-sample period and data from January 2024 to December 2024 for the out-of-sample period. Figure 1 presents the SVM classifier and the stocks of the PSI index belonging to the high risk region and to the low risk region considering the risk attributes mean and upper semivariance.

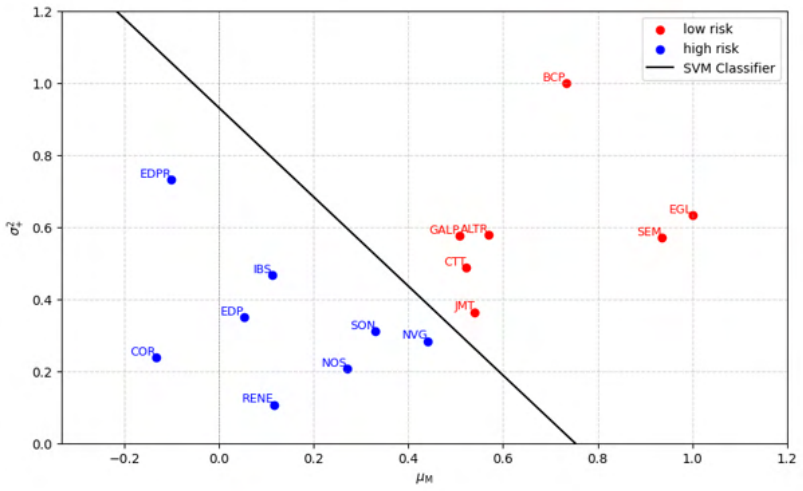

Figure 1: SVM classifier for risk attributes mean and upper semivariance

The SVM classifier will be used in subsequent works for updating the portfolio considering future investment periods, by selecting stocks belonging to the low risk region.

Acknowledgements
IB and GJM thank support from FCT through the projects UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020)
UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020) and GJM also through the FCT project UIDB/04650/2020.

References

[1] P. Cheridito, J. Ery, and M. Wüthrich. Assessing asset-liability risk with neural networks. *Risks*, 8(1), 2020.

[2] F. Paiva, R. Cardoso, G. Hanaoka, and W. Duarte. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, 115:635–655, 2019.

[3] N. Silva, L. de Andrade, W. da Silva, M. de Melo, and A. Tonelli. Portfolio optimization based on the pre-selection of stocks by the support vector machine model. *Finance Research Letters*, 61:105014, 2024.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Lab-grown diamond price estimates: an interpretation

Margarida G. M. S. Cardoso[1], **Luís Chambel**[2]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), margarida.cardoso@iscte-iul.pt
[2] Sínese, luischambel@sinese.pt

This study aims to understand how lab-grown diamond prices are formed. We rely on the results of Supervised Learning techniques and resort to specific measures of the relative importance of predictors and Shapley values.

Keywords: supervised learning, relative importance of predictors, Shapley values, lab-grown diamonds, diamonds' prices

Addressing the relative importance of predictors is relevant in many practical prediction tasks. While some supervised models can easily provide a ranking of predictors - e.g., Multiple Linear Regression or Classification and Regression Trees (CART), [2] -, using more complex models such as Ensembles - e.g. [3] - requires a different approach.
In this work, we resort to Shapley values to address the relationship between predictors and estimates obtained using an Ensemble approach. Shapley values explain the difference between the prediction and the global average prediction. They quantify the contribution of feature values to the estimates obtained, when in conjunction with different subsets of predictors -[1].
We specifically analyze the role of the main physical characteristics of diamonds - the 4 Cs (Carat, Color, Clarity, and Cut) - when estimating the unit price of lab-grown diamonds using Multiple Linear Regression, CART, and Random Forests. While the relationships between these predictors and prices are reasonably known for natural diamonds, the results obtained are expected to shed light on the same relationship when referring to synthetic diamonds. Shapley values associated with individual observations (Queries) can also bring understanding on the price formation for certain diamonds - e.g. rarer diamonds.
We analyze a data set with 44443 observations collected from the website of an online retailer of lab-grown diamonds (https://www.1215diamonds.com, accessed September 2022). We note that the TwelveFifteen site closed around September 2023 and now the company offers its products through its sister company, Diamond Nexus.
We resort to R packages "tree" and "shapr" to implement the proposed approach.

Acknowledgements This research was supported by Fundação para a Ciência e a Tecnologia, grant UIDB/00315/2020 (DOI: 10.54499/UIDB/00315/2020).

References

[1] K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.

[3] U. Sarmah, P. Borah, and D. K. Bhattacharyya. Ensemble learning methods: An empirical study. *SN Computer Science*, 5(924), 2024.

4 April, 11:30 - 12:00, Hall of Grande Auditório

Variable selection in low and high-dimensional mixtures of linear regression models

Ana Moreira[1], **Susana Faria**[1]
[1] Departamento de Matemática, Centro de Matemática (CMAT), Universidade do Minho, id10866@uminho.pt, sfaria@math.uminho.pt

The selection of variables is a critical step in constructing a regression model, as it dictates which covariates will be considered to explain or predict the response variable. In this study, we investigate the problem of variable selection in mixtures of linear regression models using penalized maximum likelihood estimation with the Expectation-Maximization (EM) and Classification Expectation-Maximization (CEM) algorithms. We conduct a simulation study to compare the performance of variable selection methods and apply these methods to real datasets.

Keywords: CEM algorithm, EM algorithm, penalized maximum likelihood estimation, simulation study, mixtures of linear regression models

Finite Mixture Regression (FMR) models provide a flexible tool for modeling data that arise from a heterogeneous population, where the relationship between the dependent variable and the explanatory variables varies among the various subpopulations. In the applications of these models, a large number of explanatory variables is often considered, for this reason, variable selection assumes great relevance for mixture models.
All subset selection methods, such as the Best Subset Selection, Backward and Forward Stepwise selection, and their modifications, have been widely investigated in the context of FMR models. However, all subset selection methods are computationally intensive. In order to overcome this problem, more efficient methodologies were developed such as, for example, methods based on penalty functions.
The methods of variable selection in the linear regression model can be extended to the models of mixtures of linear regressions. Variable selection through penalized maximum likelihood has attracted great attention in recent literature. For example, in [3] the authors propose a new method, based on the penalized likelihood function, penalizing not only the regression coefficients, but also the mixing proportions in mixtures of regression models.
This work addresses the problem of variable selection in mixtures of linear regression models with a large number of explanatory variables, using a penalized likelihood approach and employing the Expectation-Maximization (EM) and Classification Expectation-Maximization (CEM) algorithms for parameter estimation. Different selection methods based on penalization functions are studied, specifically: the Least Absolute Shrinkage and Selection

Operator (LASSO) method (see [2]), the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) method (see [4]) and the Relaxed Least Absolute Shrinkage and Selection Operator (RLASSO) method [1]) and their performance in the selection of explanatory variables compared. We apply the developed methodologies to real datasets, specifically to both low-dimensional and high-dimensional examples. The low-dimensional case is represented by a baseball dataset available on the website of the *Journal of Statistics Education* (https://jse.amstat.org/jse_data_archive.htm), while the high-dimensional case is illustrated using a genomic dataset: the NCI-60 cancer cell panel dataset. Our study reveals how different scenarios affect the performance of these algorithms and penalization functions in selecting explanatory variables. The results indicate that the ALASSO variable selection method demonstrates superior overall performance, leading us to strongly recommend its use. Regarding algorithm choice, while the CEM algorithm is considerably less computationally demanding than the EM algorithm, our findings show that the EM algorithm delivers better overall performance.

Acknowledgements The research at CMAT was partially financed by Portuguese funds through the FCT (Fundação para a Ciência e a Tecnologia), under the PhD scholarship with reference number 2022.12256.BD.

References

[1] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579 – 592, 2020.

[2] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[3] C. Yu and X. Wang. A new model selection procedure for finite mixture regression models. *Communications in Statistics-Theory and Methods*, 49(18):4347–4366, 2020.

[4] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Trends in mean sea level and extreme events: particular case of Leixões

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

---

Climate change has intensified its impact on coastal regions, including the rise in mean sea level and the occurrence of extreme tidal events. This study analyzes these dynamics in Leixões, using statistical methods and data classification techniques. The results indicate an increasing trend in mean sea level and the frequency of extreme events, correlated with adverse meteorological conditions. The findings highlight the importance of monitoring and modeling to mitigate impacts on coastal regions and enhance resilience to climate change.

Keywords: climate change, coastal events, extreme value theory, ocean monitoring, time series

---

The rise in mean sea level and the occurrence of extreme tidal events are critical issues in assessing the impacts of climate change on coastal regions. This study focuses on analyzing the evolution of mean sea level in Leixões, one of Portugal's main ports, with special attention to extreme tidal events, which can have significant economic and environmental impacts [1].

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 [3].

The data used in this work were obtained from sea level measurements in Leixões, covering a period of 65 years. Statistical analysis methods were applied to identify long-term trends, and data classification techniques were used to characterize extreme events. The methodology included: statistical modeling of time series using e.g., seasonal decomposition, ARIMA models; trend estimation through the Mann-Kendall test and linear regression; extreme event analysis based on Extreme Value Theory (EVT).

The results indicate an increasing trend in mean sea level in Leixões over the studied period, with an average rate of 2.28 mm/year; the upward trend is consistent with the findings of Mendes et al [2]. The analysis revealed a significant correlation between extreme tidal events and meteorological conditions such as storms and strong winds. The Gumbel model, with a variable scale parameter, provided the best fit to the observed surges in Leixões, leading to the study of return levels (Figure 1).

For this model, a rising trend in return levels over time was identified, with the highest predicted surge being 1.2 meters for an 80-year return period.

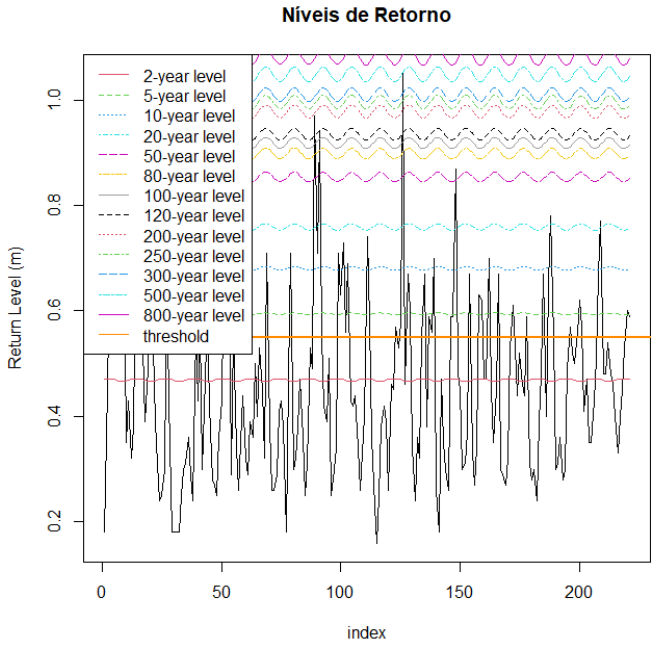

Figure 1: Return levels for the monthly maxima in Leixões

The findings underscore the importance of monitoring and modeling sea levels to predict extreme events and mitigate their impacts. The application of data classification techniques revealed useful patterns for predicting and categorizing future extreme events. This study provides a comprehensive overview of the trends in mean sea level and the analysis of extreme tidal events in Leixões, contributing to the adaptation and resilience of coastal regions to climate change. The methods presented can be replicated in other locations to broaden the understanding of global impacts.

References

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

[2] V. Mendes, S. Barbosa, and 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.

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

5 April, 11:30 - 12:00, Hall of Grande Auditório

Titanium prostheses in middle ear reconstruction: a statistical analysis of audiometric outcomes

Ana Matos[1], **José Marques Santos**[2], **Javier Gavilan**[3]
[1] Research Centre in Digital Services, Instituto Politécnico de Viseu, Portugal, amatos@estgv.ipv.pt
[2] Department of Otorhinolaryngology, Hospital CUF Viseu, Portugal, marquesdossantos14@gmail.com
[3] Department of Otolaryngology, IdiPAZ Research Institute, La Paz University Hospital, Madrid, Spain, jgavilan@coagavilan.es

Titanium prostheses are a cornerstone in the surgical reconstruction of the middle ear, renowned for their biocompatibility, durability, and effectiveness in improving hearing outcomes. This study retrospectively examines 156 middle ear surgeries performed between 2017 and 2020, where different types of titanium prostheses were employed. Inferential methods were used to quantitatively assess postoperative audiometric improvements, evaluate extrusion rates and causes, and analyze success rates across prosthesis types and surgical techniques.

Keywords: statistical inference, titanium prostheses, middle ear

The study included 156 patients (88 females, 68 males) aged 15 to 71 years, with an average age of 41. All participants were patients with an indication for ossicular reconstruction with chronic simple or cholesteatomatous otitis media or with a rupture of the ossicular chain without inflammation. All patients underwent tonal audiometry preoperatively and 12 months after surgery. A post-surgical air-bone gap (ABG) ≤ 20 decibel (dB), based on the frequencies 0.5, 1, 2 and 4 kHz, was considered as the criterion for success [1].
In ear surgeries for ossicular reconstruction, different types of titanium prostheses were applied: Partial Ossicular Replacement Prosthesis (PORP), Total Ossicular Replacement Prosthesis (TORP), and ANGULAR Clip. The success rate varied with prosthesis type: PORP achieved an 87.1% success rate, TORP 69.5%, and ANGULAR Clip 100%. Statistically significant reductions in ABG were observed across all prosthesis types (paired sign test, $p < 0.005$). The ANGULAR Clip prostheses demonstrated the highest auditory improvement (p=0.037, Kruskal-Wallis; p=0.047, multiple comparison test with Bonferroni correction).
When comparing 44 patients submitted to Canal Wall Down (CWD) versus Canal Wall Up (CWU) tympanomastoidectomy, CWU procedures were associated with significantly higher audiometric success (success rate of 79.2%; 95% CI, 62.5%-91.7%; p=0.007). Some of these cases involving reconstruction in one or two stages. Statistical analysis indicated

that single-stage surgeries were more effective than two-stage surgeries (p=0.041). Only one case of extrusion (0.64%) was recorded, underscoring titanium's excellent biocompatibility. In conclusion, titanium prostheses provide a robust solution for middle ear reconstruction, combining superior biocompatibility, durability, and auditory benefits. Success rates were significantly influenced by the type of prosthesis used and surgical timing. These results provide valuable guidance for optimizing surgical techniques and prosthesis selection in otologic procedures.

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. UIDB/05583/2020. Furthermore, we would like to thank the Research Centre in Digital Services (CISeD) and the Instituto Politécnico de Viseu for their support.

References

[1] J. Santos. *Ganancia Auditiva, Biocompatibilidad y Extrusión de Prótesis de Titanio en la Reconstrucción Quirúrgica del Oído Medio.* PhD thesis, Universidad Autónoma de Madrid, 2021.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Statistical approach to establishing geochemical baselines for metal concentrations in marine sediments: a case study of Madeira Island

Dora Carinhas[1], **Sandra Moreira**[2], **Anabela Oliveira**[2], **Carla Palma**[3], **Aurora Rodrigues**[2]

[1] Instituto Hidrográfico; IIFA/Universidade de Évora, dora.carinhas@hidrografico.pt

[2] Instituto Hidrográfico; Instituto Dom Luiz, sandra.moreira@hidrografico.pt, anabela.oliveira@hidrografico.pt, aurora.bizarro@hidrografico.pt

[3] Instituto Hidrográfico, carla.palma@hidrografico.pt

Establishing reference values for metal concentrations in marine sediments is crucial for identifying temporal trends and quantifying anthropogenic impacts on environmental changes. This study focuses on Madeira Island, where concentrations of certain metals in marine sediments exceed legislative thresholds for non-contaminated sediments. A regional geochemical baseline for the concentrations of chromium, nickel, copper, and zinc in the southern shelf sediments of the island will be established using two approaches: regression analysis and the relative cumulative frequency (RCF) method.

Keywords: anthropogenic influence, environmental contamination, regression analysis, relative cumulative frequency, trace metals

---

Madeira Island is a volcanic island, in which the marine sediments that accumulate up to approximately 120 meters deep result from coastal and marine erosion of basaltic rocks, naturally rich in high concentrations of certain metals.

This study uses a geochemical database of surface sediments from Madeira [1] to evaluate the most effetive statistical aproaches, such as regression analysis or relative frequency (RCF), for establishing regional baseline values for metals. The analysis is based on a dataset comprising the results of chemical analyses of 166 sediment samples collected in 2002 and 2007 during oceanographic surveys conducted by the Hydrographic Institute, in collaboration with the Madeira Regional Government (Figure 1).

The regression analysis approach uses linear regression to establish a relationship between metal concentrations and a geochemical reference element, such as aluminium (Al) or iron (Fe) [1]. This method assumes that elemental concentrations follow a normal or lognormal distribution. Samples falling outside the 95% confidence interval are flagged as potentially influenced by anthropogenic activities. However, most variables in large datasets from regional geochemical and environmental surveys do not follow normal or lognormal distributions, even after applying transformation methods. These datasets often exhibit

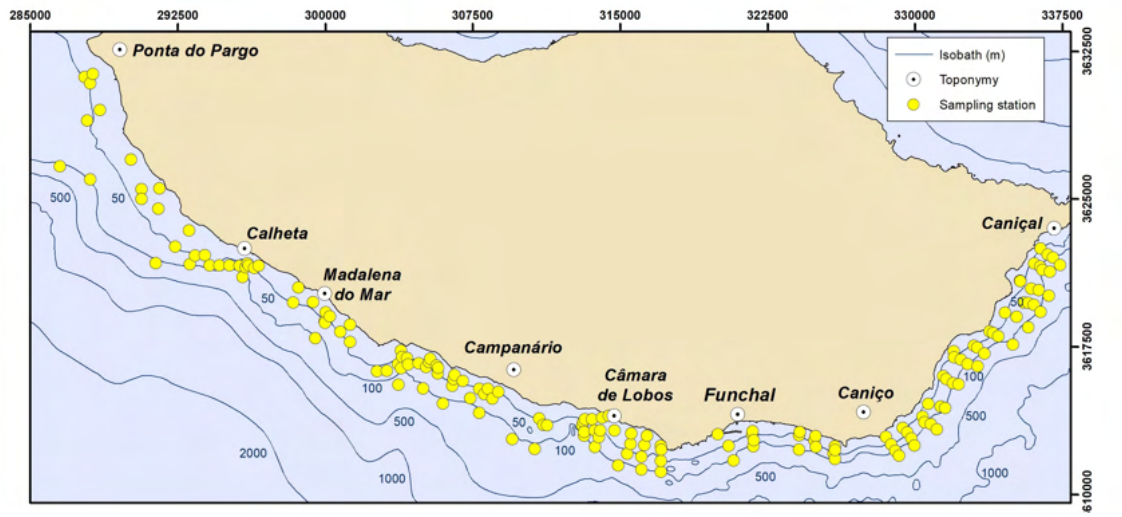

Figure 1: Location of surface sediment sampling stations on the southern shelf of Madeira Island

skewness, outliers, and the influence of multiple processes [3]. In contrast, the RCF method does not require specific distribution assumptions. It analyzes cumulative frequency curves for individual elements, with the first inflection point on the curve representing the upper limit of the geochemical baseline concentrations. This method is particularly useful for elements that lack a strong correlation with a reference element [2].
The application of these two distinct methods provides complementary approaches for establishing regional geochemical baselines, both of which are essential for monitoring environmental changes and supporting decision-making in coastal ecosystem management.

Acknowledgements Thanks to the Regional Government of Madeira for providing the data used in this study.

References

[1] S. Moreira, D. Carinhas, A. Rodrigues, A. Oliveira, and C. Palma. Valores de referência regionais de cr, ni, cu e zn para a plataforma sul da ilha da madeira. In *Atas das 3as Jornadas de Engenharia Hidrográfica*, pages 194–197, Cádiz, Outubro 2024.

[2] B. K. Newman, R. Ottenstein, and C. Reimann. Definition of baseline metal concentrations for assessing metal enrichment of sediment from the south-eastern cape coastline of south africa. *Water SA*, 33:675–691, 2007.

[3] C. Reimann and P. Filzmoser. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology*, 39:1001–1014, 2000.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Production of papayas in an aquaponics system

Fernando Sebastião[1,2,3], **Judite Vieira**[1,2,3], **Luís Cotrim**[1,2,3], **Ounísia Santos**[1,2,3], **Maria Rodrigues**[3], **Daniela Vaz**[1,2,4,5], **Vânia Ribeiro**[1,2,4], **Raul Bernardino**[1,2,6]
[1] Laboratory of Separation and Reaction Engineering-Laboratory of Catalysis and Materials (LSRE-LCM), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal
[2] ALiCE – Associate Laboratory in Chemical Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[3] School of Technology and Management, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, fsebast@ipleiria.pt, judite.vieira@ipleiria.pt, luis.cotrim@ipleiria.pt, ounisia.santos@ipleiria.pt, maria.l.rodrigues@ipleiria.pt
[4] School of Health Sciences, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal
[5] Coimbra Chemistry Centre, Institute of Molecular Sciences, Department of Chemistry, University of Coimbra, 3004-535 Coimbra, Portugal, daniela.vaz@ipleiria.pt, vania.ribeiro@ipleiria.pt
[6] MARE - Marine and Environmental Sciences Centre/ARNET – Aquatic Research Network, ESTM- School of Tourism and Marine Technology, Polytechnic Institute of Leiria, 2520-614 Peniche, Portugal, raul.bernardino@ipleiria.pt

---

Sustainability challenges increasingly require innovative solutions, like balancing aquaponics to maximise papaya production, boost economic value and utilize by-products. This work aims to analyse papaya production in aquaponics. Over 13 months, plant growth was assessed using Leca® and brick waste as substrates, comparing morphological characteristics and fruit yield for significant differences. Mathematical models were applied to stem growth, and fruit production was monitored until harvest.

Keywords: *Carica papaya*, *Clarias gariepinus*, plant growth, sustainability, statistical tests

---

In recent decades, aquaponics has become a sustainable alternative for food production. Rising temperatures and sea levels, soil salinity and changes in productivity resulting from climate change have a significant impact on food and fish production [2]. Although papayas (*Carica papaya*) are mainly grown in tropical and subtropical regions, due to their growing economic value and demand from a nutritional point of view in the human diet [3], some varieties can, under controlled conditions, be produced in greenhouses.
This project involves the production of papayas in two aquaponics systems containing African catfish (*Clarias gariepinus*), one of the most farmed fish around the world with great economic value. The main purpose is to study papayas' morphological growth and determine productivity levels, namely quantity and quality of the fruits.

The project began on May 2022, as did the respective morphological growth measurements of the young plants that were placed in the plant beds (in deep water culture) of the two aquaponics systems, in two different substrates: Leca® as conventional and broken waste brick as an alternative. Measurements continued monthly until May 2023 (13 months).
The morphological characteristics evaluated monthly were leaf number, biggest leaf length, foliage diameter, plant height, plant greenness, plant health and number and characteristics of the fruits. The normality (Kolmogorov-Smirnov test) and equality of variances (Levene's test), at a 5% significance level, of the number of leaves, the biggest leaf length and the diameter foliage, in each of the substrates were applied. It was concluded that in each month there were no significant differences in each characteristic depending on the substrate (T-test for independent samples). Plant height was compared for each of the 13 months of plant growth, for both the Leca® and brick substrates. Plants had an intermediate or high level of greenness until the 7th month of growth, and after that some had a low level of greenness. On the other hand, the health status of the plants, until November 2022 was generally strong and from that month onwards some plants showed slightly damaged leaves with fungus and some dry parts.
Two mathematical functions (Lundqvist-Korf and the Richards functions [1]) were used to model the biological growth of papayas, in particular stem growth, by measuring diameters in three different regions. The number of fruits was monitored until they ripened and were harvested. After ripening, basic statistics of the diameter, length and weight of the papayas were analysed. For each papaya, the skin was classified as smooth or rough, whether it had seeds or not and the colour of the seeds, and whether the flesh was tasty or not.
Some physical and chemical parameters were also measured daily in the water (temperature, pH, dissolved oxygen, oxidation reduction potential and total dissolved solids) as well as temperature and humidity of the air inside the greenhouse. It was also found that water remained of good quality and the fish were in good health. Although some of the papayas' leaves died due to the low temperatures, in general, the plants grew well.

Acknowledgements This work was supported by national funds through FCT/MCTES (PIDDAC): LSRE-LCM, UIDB/50020/2020 (DOI: 10.54499/UIDB/50020/2020), and UIDP/50020/2020 (DOI: 10.54499/UIDP/50020/2020); and ALiCE, LA/P/0045/2020 (DOI: 10.54499/LA/P/0045/2020).

References

[1] H. E. Burkhart and M. Tomé. *Modeling Forest Trees and Stands.* Springer Science and Business Media, 2012.

[2] I. J. Mirón, C. Linares, and J. Díaz. The influence of climate change on food production and food safety. *Environmental Research*, 216:114674, 2023.

[3] G. A. Paternina, F. V. Luna, and A. A. Bermudez. Nutraceutical, thermophysical and textural characteristics of papaya (Carica papaya L) and incidence for post-harvest management. *Heliyon*, 8, 4:e09231, 2022.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Domestic violence and economic and social vulnerability

Ana Ribeiro[1], Ana Martins[1], Daniela Fernandes[1], Juliana Barbosa[1], Lurdes Babo[2], Cristina Torres[2], Isabel Vieira[2], Cristina Lopes[2]
[1] ISCAP, Instituto Politécnico do Porto, Portugal, 2211158@iscap.ipp.pt, 2211958@iscap.ipp.pt, 2211562@iscap.ipp.pt, 2211449@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, Portugal, lbabo@iscap.ipp.pt, ctorres@iscap.ipp.pt, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt

This paper investigates the link between social and economic factors and domestic violence in Portugal. Data from 278 municipalities show that social factors, especially unemployment, significantly impact domestic violence, while economic variables like income and wage gap have less influence. Lower purchasing power index tended to be associated with higher levels of domestic violence. Although violence cases slightly declined from 2011 to 2021, the decrease was not significant. Higher population density and marriage/divorce rates also increased incidence. The study finds social conditions impact domestic violence more than economic factors.

Keywords: domestic violence, social factors, economic factors

Domestic violence is a complex and multifaceted phenomenon that affects thousands of people around the world. A wide range of factors, including cultural, social, economic and psychological elements, can contribute to the occurrence of violence. In order to analyse the relationship between domestic violence and economic and social vulnerability, this study considers 12 quantitative variables relating to 278 Portuguese municipalities [2].
The descriptive statistics revealed a slight decline in the average rate of domestic violence from 80.65 in 2011 to 76.70 in 2021. However, both variables exhibited high standard deviations (135.841 in 2021 and 161.690 in 2011), indicative of considerable variability in domestic violence rates across municipalities. Nevertheless, the T-test revealed that this observed difference was not statistically significant ($T = -1.123; p\text{-value} = 0.262$).
An exploratory factor analysis was conducted using the Principal Components extraction method with Varimax rotation ($KMO = 0.824$). Based on the Kaiser [3] criterion, two factors of correlated variables were obtained, which collectively accounted for approximately 89% of the total variance in the data. Factor 1 comprises social variables (population density, number of divorces, number of marriages, total females, total males, domestic violence 2021, male unemployed population, female unemployed population) and Factor 2 corresponds to economic variables (average monthly earnings for men, average monthly

earnings for women, purchasing power, wage gap). The 'Domestic Violence 2021' variable falls under Factor 1, suggesting that it correlates more with social than economic factors. The study found a positive correlation between the incidence of domestic violence cases and the rate of marriages and divorces within municipalities ($r = 0.890$ and $r = 0.957$, respectively). Furthermore, regions characterised by higher population density have been observed to experience a concomitant increase in domestic violence cases ($r = 0.705$). Furthermore, unemployment has been evidenced to play a significant role in the occurrence of domestic violence cases, as indicated by the high values of the Pearson correlation coefficient ($r = 0.980$ and $r = 0.964$, in the male and female population, respectively). Conversely, the economic variables in Factor 2 exhibited a lower correlation with domestic violence, specifically the variable wage gap ($r = 0.164$).

The municipalities were categorised according to three levels of the Purchasing Power Index (PPI): low purchasing power (values between 0 and 50), medium purchasing power (values between 51 and 100) and high purchasing power (values between 101 and 200). This categorisation facilitated the exploration of the relationship between local economic conditions and the prevalence of domestic violence. A One-Way Analysis of Variance (ANOVA) revealed statistically significant differences in the levels of domestic violence between the three PPI groups (p-value < 0.001). The magnitude of this effect was subsequently evaluated using metrics such as the Eta Square (0.284), which indicates that approximately 28.4% of the variability in levels of domestic violence at 2021 can be explained by the PPI categorisation. This is regarded as a moderate to large effect [1], thereby reinforcing the PPI's status as an explanatory factor for the observed disparities. The data indicates that municipalities with lower purchasing power exhibit higher levels of domestic violence on average.

The findings of this study suggest that, while economic factors must not be disregarded, enhancing social conditions in Portugal is imperative for the reduction of domestic violence.

Acknowledgement: This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., as part of project UID/05422: Centro de Estudos Organizacionais e Sociais do Politécnico do Porto.

References

[1] K. Backhaus, B. Erichson, S. Gensler, R. Weiber, and T. Weiber. *Multivariate Analysis: An Application-Oriented Introduction*. Springer, 2021.

[2] Fundação Francisco Manuel dos Santos. PORDATA estatísticas sobre Portugal e Europa, 2024. https://www.pordata.pt/.

[3] J. Hair Jr, W. Black, B. Babin, and R. Anderson. *Multivariate Data Analysis*. Prentice Hall, 2019.

5 April, 11:30 - 12:00, Hall of Grande Auditório

The influence of socio-economic conditions on public health

Patrícia Pinto[1], Ivanise Gomes[1], Jéssica Martins[1], Lurdes Babo[2], Cristina Torres[2], Isabel Vieira[2], Cristina Lopes[2]
[1] ISCAP, Instituto Politécnico do Porto, Portugal, 2240343@iscap.ipp.pt, 2230232@iscap.ipp.pt, 2240348@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, Portugal, lbabo@iscap.ipp.pt, ctorres@iscap.ipp.pt, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt

This article aims to investigate the impact of socio-economic conditions on health conditions. The results revealed that material and social deprivation is related to higher rates of pre-obesity, indicating that a lack of adequate resources can negatively affect eating habits and health. Similarly, poverty is also associated with an increased prevalence of obesity, suggesting that, in addition to a lack of resources, the environment in which people live plays a crucial role. Women face more challenges in accessing medical examinations, highlighting the need for more equitable health policies.

Keywords: socio-economic conditions, health policies, obesity, equity

Despite progress in European health systems, inequalities persist that threaten the goals of equity and efficiency. This study analyses how socio-economic conditions impact health conditions in different European countries, focusing on the relationship between poverty and health, in line with Sustainable Development Goals (SDG) 1 (no poverty) and 3 (good health and well-being).

A quantitative methodology was used, with data collected from Eurostat and processed using SPSS software. Eight quantitative variables were analysed for 27 European countries. There was a moderate positive linear correlation ($r = 0.687$) between the rate of severe material and social deprivation and the rate of pre-obesity. This result is in line with the *Food and Agriculture Organization of the United Nations* (FAO) report [2], which states that socio-economic inequalities make it difficult to reduce food insecurity and malnutrition, i.e. the population has a limited diet, there is a lack of access to nutritious food, which can lead to overweight and obesity.

According to the *World Health Organization* [4], environments that promote the accessibility of unhealthy foods, along with limited opportunities for physical activity, can contribute to the problem and FAO [2] adds that the food environment, i.e. food insecurity and inadequate dietary patterns, contributes to an increase in pre-obesity. Our study found that there is a significant moderate positive linear correlation ($r = 0.530$) between the at-risk-of-poverty rate and the pre-obesity rate.

The non-parametric Wilcoxon test concluded that there are statistically significant differences between the unmet medical needs of men and women ($p-value < 0.001$). According to the *European Institute for Gender Equality* [1], gender is a determining factor in access to medical care. Society tends to discourage men from seeking diagnosis and treatment, and they are less likely than women to consult a doctor. On the other hand, because women spend more time caring for children and relatives, they make greater use of health services and may find it easier to obtain medical assistance.
In this study, a weak positive linear correlation ($r = 0.383$) was observed between the rate of risk of poverty or social exclusion in the city and suburbs and the rate of pre-obesity. There was also a moderate positive linear correlation between the rate of risk of poverty or social exclusion in rural areas and the rate of pre-obesity ($r = 0.643$). Rural areas therefore have a greater impact on the prevalence of pre-obesity. The fact that people living in rural areas often face difficulties in accessing health services, especially nutritional prevention and monitoring programmes, may result in less early detection of the problem [3]. Although agriculture is an important source of income for many people, productivity in the sector lags behind other sectors in Europe and Central Asia. This difference leads to higher rates of poverty and food insecurity in rural areas [2].
These findings highlight the need to address SDG 1 and 3, to promote the reduction of economic inequalities and to improve access to health services for the most vulnerable populations.
Acknowledgement: This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., as part of project UID/05422: Centro de Estudos Organizacionais e Sociais do Politécnico do Porto.

References

[1] European Institute for Gender Equality. *Gender and Intersecting Inequalities in Access to Health Services*. European Institute for Gender Equality, 2021.

[2] Food and Agriculture Organization of the United Nations. *Regional Overview of Food Security and Nutrition in Europe and Central Asia 2019: Structural Transformations of Agriculture for Improved Food Security, Nutrition and Environment*. Food and Agriculture Organization of the United Nations, 2019.

[3] Programa Nacional para a Promoção da Alimentação Saudável. *Obesidade: Otimização da Abordagem Terapêutica no Serviço Nacional de Saúde*. Direção-Geral da Saúde, 2017.

[4] World Health Organization. *WHO European Regional Obesity Report 2022*. World Health Organization Regional Office for Europe, 2022.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Weekend and holiday work by young people

Jéssica Nadais[1], **Catarina Ferreira**[2], **Isabel Gomes**[2], **Cristina Lopes**[1], **Lurdes Babo**[1], **Cristina Torres**[1], **Isabel Vieira**[1]
[1] CEOS.PP, ISCAP, Instituto Politécnico do Porto, 2240224@iscap.ipp.pt, cristinalopes@iscap.ipp.pt, lbabo@iscap.ipp.pt, ctorres@iscap.ipp.pt, mivieira@iscap.ipp.pt
[2] ISCAP, Instituto Politécnico do Porto, 2230227@iscap.ipp.pt, 2240222@iscap.ipp.pt

This paper presents an analysis of young people's tendency to work weekends and public holidays, based on data collected through an online questionnaire. It was concluded that young people's preference for this type of work is influenced by a number of factors, but there is no evidence that gender or job type has a direct impact on the number of hours worked or satisfaction. These data suggest that policies to promote work-life balance could be beneficial in improving the experience of young people working at weekends and on public holidays.

Keywords: young people, work, weekends, statistical analysis, hypothesis testing

Many young people choose to work during the weekend to supplement their income, gain work experience or to combine their studies with work. However, this practice can have a significant impact on quality of life, work-life balance and perceived job satisfaction. A healthy and effective reconciliation of roles requires individual, family and organisational efforts, together with appropriate management of time and professional responsibilities [1].
The aim of this study was to understand the phenomenon of young people working on weekends and public holidays. An online questionnaire was carried out and a sample of 132 people was obtained. The age range of the participants varies between 21 and 62 years, with a higher prevalence of women (74%). The majority of respondents live in the districts of Porto (81 responses) and Aveiro (33 responses). As regards their educational level, the majority have a university degree (53 responses), followed by secondary education (41 responses) and a master's degree (27 responses). Regarding their professional situation, the vast majority (100) are in full-time employment.
The descriptive analysis shows great variability in the number of hours worked on weekends and holidays. The mean number of hours worked per month by the subjects in the studied population was found to be between 21.21 and 32.29 hours, with 95% confidence, indicating a moderate commitment to weekend work.
Although a significant proportion of respondents are satisfied with their working hours, there is also a significant group who are dissatisfied with the impact of work on their personal lives. The analysis in Fig. 1 suggests that those who perceive the impact of work negatively may work longer hours than those who perceive it positively or neutrally.

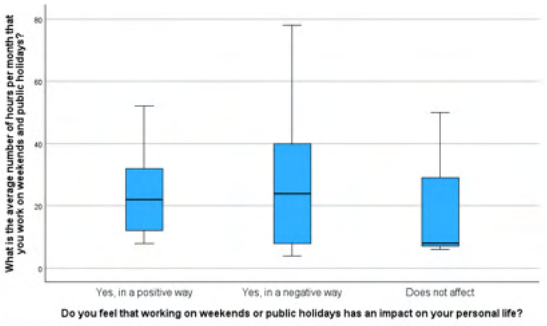

Figure 1: Working on weekends and public holidays and the impact on personal life

However, the ANOVA test obtained a test statistic of $F = 0.225$ and p-value $= 0.799$, so there was no statistically significant evidence that the groups differed in the average number of hours worked per month.
The chi-squared test shows that there are no statistically significant associations between employment status (full-time/part-time) and the perceived impact of working on weekends and public holidays (p-value $= 0.947$). In addition, Mann-Whitney tests show that the average number of hours worked on weekends and public holidays does not differ significantly by gender (p-value $= 0.462$) or by type of employment status (p-value $= 0.079$).
The results show that there is a strong positive correlation between salary level and the ability to reconcile work and family life ($r = 0.750$). Working hours are strongly related to the perception of work-life balance ($r = 0.821$).
In terms of satisfaction with working hours, there was a greater concentration of responses at the 'moderately satisfied' and 'very satisfied' levels. The study recommends that the promotion of flexible and balanced labour market policies is essential to reduce the negative impact and improve the quality of life of young workers.

Acknowledgement: This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., as part of project UID/05422: Centro de Estudos Organizacionais e Sociais do Politécnico do Porto.

References

[1] M. Matias and A. M. Fontaine. Managing multiple roles: Development of the work-family conciliation strategies scale. *The Spanish Journal of Psychology*, 17:E56, 2014.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Influence of physical readiness on fatigue variation during infantry officer training course - a case study

João Fonseca[1], **Rui Lucena**[1,2], **André Fonseca**[1], **Nuno Almeida**[1,2], **Paula Simões**[1,3]

[1] Military Academy Research Center (CINAMIL), fonseca.jca@academiamilitar.pt, fonseca.afp@exercito.pt

[2] Interdisciplinary Centre for the study of human Performance (CIPER), rui.lucena@academiamilitar.pt, almeida.nrc@academiamilitar.pt

[3] Center for Mathematics and Applications (NOVA MATH) - NOVA University of Lisbon, paula.simoes@academiamilitar.pt

The Infantry officer training course is the culmination of the practical instruction of Infantry Student Aspirants in the Military Academy. The goal of this research is to understand how physical preparation before the Internship influences the variation of fatigue during it. Measurements regarding heart rate variability before and during the internship as well as training load measurements were collected. Grades for basic physical training and physical training of military application in the 4th and 5th year of the Military Academy were also considered. It is conclude that the pre-internship physical training has positive effects, avoiding periods of overexertion.

Keywords: fatigue, military, heart rate variability, training load, hypothesis tests

Physical readiness has always been crucial for military performance. The Infantry is considered the most physically demanding branch of the military. The Infantry officer training course (IOTC) involves a heavy workload of practical and theoretical-practical instruction, which results in high physical exertion and subsequent fatigue, and physical difficulties. Monitoring fatigue and training status is crucial for in prescribing physical training and adequate rest [1]. The effective monitoring of heart rate variability (HRV) to address the course of adaptations of athletes and military populations is crucial for understanding whether functional (adapting positively) or non-functional (adapting negatively) overreaching is occurring [4]. The interpretation of changes in HRV must consider the parasympathetic saturation effect and the natural logarithm of the square root of the mean sum of the squared differences between R–R intervals (Ln rMSSD) to R-R interval (time in milliseconds of the intervals between successive heart beats) relationship [1].

Data regarding to physical training scores, average weekly training load, the weekly average of resting heart rate, HRV, the Ln rMSSD/RR ratio over time, were collected from the military participants during the 16 weeks of the IOTC. The data was analysed using various descriptive statistical methods using IBM SPSS Statistics 28 software. Statistical Inference

techniques and probability models were considered when assessing physical fitness in the military scope, evaluated according to the basic physical training and physical training of military application, in two different moments. Point estimation, interval estimation and implementation of various hypothesis tests (Parametric and non-parametric) for the parameters of interest (mean and median value) are considered with a significance level of 5% [2].

Results have shown that physical preparation before the IOTC avoid periods of overexertion. The IOTC generates high levels of fatigue due to the high training load and lack of sleep. Better physical preparation may contribute to the prevention of the accumulation of fatigue and ultimately to states of overreaching or overtraining. Furthermore, this study lends support to a possible relation between Ln rMSSD/RR ratio and individual fatigue. The Ln rMSSD/RR ratio over time reveals a tendency for athletes with higher levels of fatigue, to show a lower ratio. On the other hand, athletes who are not fatigued, seem to maintain or increase this ratio over time. Thus, analysing the behaviour of the ratio over time can be useful in understanding an individual's state of fatigue [3]. As expected, fatigue varies with training load and sleep hours. Increasing training load increases fatigue, while decreasing sleep hours also increases fatigue. Thus, the proper exercise prescription and control of fatigue status of the candidates is crucial for safety and health care but also for proper increases in performance.

References

[1] M. Buchheit and et al. Monitoring endurance running performance using cardiac parasympathetic function. *European Journal of Applied Physiology*, 108:1153–1167, 2010.

[2] G. Casella and R. Berger. *Statistical Inference*. CRC press, 2024.

[3] W. Hopkins, S. Marshall, A. Batterham, and J. Hanin. Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise*, 41(1):3, 2009.

[4] D. J. Plews, P. Laursen, J. Stanley, A. Kilding, and M. Buchheit. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. *Sports Medicine*, 43:773–781, 2013.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Comparative robustness of machine learning methods for genomic prediction

Vanda M. Lourenço[1], **Joseph O. Ogutu**[2], **Hans-Peter Piepho**[2]
[1] Department of Mathematics and NOVA Math, NOVA FCT, NOVA University of Lisbon, Portugal, vmml@fct.unl.pt
[2] Biostatistics Unit, University of Hohenheim, Germany, jogutu2007@gmail.com, hans-peter.piepho@uni-hohenheim.de

Accurate prediction of genomic breeding values is key to genomic selection in plant and animal breeding. Genomic prediction uses thousands of molecular markers, requiring methods able to handle high-dimensional data. We compare the predictive performance and robustness of supervised machine learning methods, including regularized, ensemble, and instance-based approaches, using simulated data. This study evaluates their accuracy and errors under clean and contaminated phenotypic data scenarios.

Keywords: genomic prediction, SNPs, machine learning, robustness, breeding studies

The accurate prediction of genomic breeding values is pivotal in genomic selection for plant and animal breeding studies. Genomic prediction (GP) relies on thousands of molecular markers distributed across the genome, requiring computational methods capable of efficiently handling high-dimensional data. In this context, machine learning (ML) methods have gained significant attention due to their flexibility and potential to address the complexity of GP.

While many studies have compared the predictive performance of individual ML methods, few have offered comprehensive evaluations of different groups of methods, and even fewer have examined how data contamination can affect their predictive performance. Understanding how different groups of ML methods perform under both clean and contaminated data scenarios is critical, as it can help identify robust methods and reveal their relative advantages and limitations compared to established classical approaches.

This study addresses these gaps by evaluating the predictive accuracy and robustness of several groups of supervised ML methods, including regularized, ensemble, and instance-based approaches, discussed in [1]. Using a simulated dataset derived from an animal breeding population (Table 1; [2]), we assess their performance in terms of predictive accuracy and prediction errors under varying data contamination conditions. These findings deepen our understanding of the merits and demerits of different ML groups, and offer valuable guidance for genomic prediction in real-world breeding contexts.

Table 1: Summary statistics for the animal quantitative trait dataset (trait T_1; $n = 3000$)

Trait	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Sd**
T_1	-584.993650	-116.244762	-1.711490	-0.000004	112.248515	587.189720	176.518911

Acknowledgements This work was partially 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 (NOVA Math - Center for Mathematics and Applications) and project REACTION - 2023.14934.PEX (https://doi.org/10.54499/2023.14934.PEX).

References

[1] V. M. Lourenço, J. O. Ogutu, R. A. P. Rodrigues, and H-P. Piepho. Genomic prediction using machine learning: A comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC Genomics*, 25(1):152, 2024.

[2] J. O. Ogutu and H-P. Piepho. Regularized group regression methods for genomic prediction: Bridge, mcp, scad, group bridge, group lasso, sparse group lasso, group mcp and group scad. *BMC Proceedings*, 8(5):1–9, 2014.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Exploring dynamic neural field self-organizing maps for dimensionality reduction, visualization and classification

Paulo Barbosa[1], **Flora Ferreira**[1], **Estela Bicho**[2], **Wolfram Erlhagen**[1]
[1] Centre of Mathematics, School of Sciences, University do Minho,
id10608@uminho.pt, {fjferreira, wolfram.erlhagen}@math.uminho.pt,
[2] Algoritmi Centre, School of Engineering, University of Minho,
estela.bicho@dei.uminho.pt

Dynamic Neural Fields are Recurrent Neural Networks modeling neural activity in a continuous feature space using localized activity regions called bumps. Despite their robustness to noise and low data requirements, DNFs struggle with high-dimensional inputs. The Dynamic Neural Field Self-Organizing Map overcomes this by integrating self-organization to create low-dimensional representations, using an adaptive learning rule for complex data and real-time applications. We illustrate its potential with a practical example for dimensionality reduction and classification.

Keywords:
dynamic neural field, self-organizing map, unsupervised learning, dimensionality reduction, classification

---

Dynamic Neural Fields (DNFs) [1] are Recurrent Neural Networks that describe the coarse-grained activity of populations of interacting neurons organized in a continuous feature space. Information in DNFs is represented by supra-threshold, localized regions of neural activity, commonly referred to as bumps. The recurrent interactions within the neural population give rise to dynamic behaviors, such as standing bumps (stable localized activity) and traveling waves (propagating activity), which have been observed in various brain regions, namely in those involved in sensory processing, motor control, navigation, and memory formation. Furthermore, coupling multiple DNFs enables the implementation of higher cognitive phenomena, such as perception, working memory, decision-making, and forgetting.

These networks excel under regimes that demand continuous feedback, making them particularly well-suited for robotics, where agents operate in highly dynamic and changing environments [3]. Moreover, they are robust to noise and perturbations in external stimuli, which are ubiquitous in most real-world applications. Another significant advantage is that they typically require far less training data than most traditional Machine Learning or Deep Learning models, primarily due to their intrinsic structure and biologically inspired mechanisms.

However, arguably the most significant limitation of DNF-based architectures lies in their reliance on smooth, continuous variables, which renders them inefficient at processing the high-dimensional and complex inputs typically found in real-world datasets. Hence, it is essential to employ mechanisms that can transform high-dimensional inputs into low-dimensional, smooth representations compatible with DNFs, thereby extending their applicability to a broader range of problems.
The Dynamic Neural Field Self-Organizing Map (DNF-SOM) [2] integrates the biologically inspired principle of self-organization into DNFs to address the aforementioned limitation. Self-organization essentially creates a low-dimensional representation of high-dimensional data while preserving the original topology in a fully unsupervised manner. Unlike the original Self-Organizing Map (SOM), proposed by Teuvo Kohonen (1982) [4], the DNF-SOM, introduced by Giorgios Detorakis (2013), employs a modified learning rule capable of processing complex topologies, skewed data distributions, and performing online learning. This makes the DNF-SOM adaptive, allowing it to dynamically adapt to changes in the input distribution, making it ideal for real-time learning tasks.
We argue that the DNF-SOM holds significant potential for the machine learning community. To illustrate the potential of this framework in practical machine learning applications, we apply the DNF-SOM to a classification task. A key advantage of the DNF-SOM is its minimal need for precise hyperparameter tuning. Unlike methods like K-Nearest Neighbors, which require an explicit specification of the number of clusters, the self-organizing nature of the DNF-SOM enables it to automatically uncover meaningful structures in the data without prior knowledge of cluster counts.

Acknowledgements
The research was financed by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) through the doctoral grant with reference UI/BD/153737/2022 and 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. Work partially financed by FCT with national funds through the project I-CATER - Ref. PTDC/EEI-ROB/3488/2021.

References

[1] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

[2] G. I. Detorakis and N. P. Rougier. A neural field model of the somatosensory cortex: formation, maintenance and reorganization of ordered topographic maps. *PlOS ONE*, 7(7):e40257, 2012.

[3] W. Erlhagen and E. Bicho. The dynamic neural field approach to cognitive robotics. *Journal of Neural Engineering*, 3(3):R36, 2006.

[4] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Assessing risk: an extremal inference methodology

Marta Ferreira[1], **Elisa Moreira**[2]
[1] Centro de Matemática, Universidade do Minho, msferreira@math.uminho.pt
[2] Departamento de Matemática, Universidade do Minho, a103613@alunos.uminho.pt

Extreme Value Theory focuses on inference at the tails of distributions and can go beyond observations. It thus provides the natural framework for inferring risk measures with applications in diverse contexts such as finance, insurance, environment, engineering, among others. This paper presents a tail inference methodology with the aim of assessing risk in financial data.

Keywords: extreme value theory, tail index, POT method, coefficient of variation, threshold choice

The main result in extreme value theory establishes the possible limiting laws of the linearly normalized maximum, which can be described by the generalized extreme value (GEV) distribution function (df).
An alternative approach denoted "Peaks over Threshold" (POT) is to consider the largest observations above a high threshold, whose df can be well approximated by a Generalized Pareto (GP) distribution. Both models share a shape parameter, γ, called the tail index, which governs the type of tail: $\gamma < 0$ indicates a short tail with a finite right endpoint, $\gamma = 0$ corresponds to an exponential-type tail, and $\gamma > 0$ determines a heavy tail of polynomial-type, with an infinite right endpoint.

In modeling the excesses above a high threshold t with a GP, the main difficulty lies precisely in choosing t: if t is too high, we lose valuable information about the tail, where data tends to be scarce, complicating inference; if t is too low, we may include information that no longer corresponds to the tail, thus biasing the results.
A typical empirical approach is to consider the plot of the empirical mean excesses based on the mean excess function of a GP model, which corresponds to a linear function of the threshold t and is only defined for $\gamma < 1$. Thus, the method involves selecting the threshold t from which the plot exhibits linearity.

An alternative approach was proposed in [1, 2], based on the coefficient of variation (CV) of a GP model, which is a constant function depending only on γ:

$$CV = (1 - 2\gamma)^{-1/2},\ \gamma < 1/2.$$

The analysis of the empirical CV plot becomes simpler to apply than the empirical mean excesses plot. An automatic threshold selection methodology can be seen in [1, 2]. Although the CV is only defined for $\gamma < 1/2$, the aforementioned authors propose a data transformation that allows for any value of γ, thus eliminating any application restrictions.

In this work, we present the results of a simulation study on the application of this methodology. These results allow us to obtain guidelines for its application to real data. The approach is applied to the inference of the risk measures, Value-at-Risk and Expected-Shortfall, for the Portuguese stock index PSI20, during the period 2020-2024.

Acknowledgements The research of the first author 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, with references DOI 10.54499/UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020) and DOI 10.54499/UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020).

References

[1] J. D. Castillo and M. Padilla. Modeling extreme values by the residual coefficient of variation. *SORT Statist. Oper. Res. Trans.*, 40(2):303–320, 2016.

[2] J. D. Castillo, I. Serra, M. Padilla, and D. Moriña. Fitting tails by the empirical residual coefficient of variation: The ercv package. *The R Journal*, 11(2):56–68, 2019.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Analyzing a method for estimating the tail index

Marta Ferreira[1], **Liliana Monteiro**[2]
[1] Centro de Matemática, Universidade do Minho, msferreira@math.uminho.pt
[2] Departamento de Matemática, Universidade do Minho, pg49147@alunos.uminho.pt

The tail index is the primary parameter in inferring extreme values in a univariate random sample. In semi-parametric estimation of the tail index we have to choose the number k of top ordinal statistics to be considered. Such a choice is based on a balance between variance (small k) and bias (large k). We will address the maximum likelihood estimator in light of the application of a method for choosing k. The performance will be evaluated based on a simulation study. An illustration with real data concludes this work.

Keywords: extreme value theory, semi-parametric estimation, maximum likelihood

Extreme Value Theory (EVT) aims to characterize the tails of a distribution, where extreme values are found. The Extremal Types Theorem establishes the limiting models for the linearly normalized maximum of a random sample, these being Weibull (of maxima), Gumbel and Fréchet. The Generalized Extreme Values (GEV) model unifies the writing of the three, defining itself as:

$$G_\gamma(x) = \begin{cases} \exp\left\{-\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\gamma}\right\} & , \gamma \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & , \gamma = 0 \end{cases}.$$

The real constants μ, $\sigma > 0$ and γ are, respectively, the location, scale and shape parameters. In the context of EVT, γ is called the tail index and plays the main role, as it dictates the type of tail of a distribution: if $\gamma > 0$ we have a heavy tail of polynomial type and therefore we are in the Fréchet maximum domain of attraction, if $\gamma = 0$ we have an exponential tail corresponding to the Gumbel domain and if $\gamma < 0$ we have a short tail respecting the Weibull domain (of maxima).
In parametric inference, we fit extreme models to the data and estimate their parameters, including the tail index. For example, in a random sample of maxima, a typical approach is to fit the GEV model, using for example the maximum likelihood method. In semi-parametric inference, it is simply assumed that the distribution underlying the data is in one of the three maximum domains of attraction, with the estimators of γ defined from the top k ordinal statistics of the sample. The choice of k involves a trade-off between bias and variance, where larger values of k correspond to larger bias and smaller variance. There are several methodologies in the literature for choosing k. See, for example, Beirlant et al. [1]. In this work we consider the Reiss and Thomas method in [3] applied to the

semi-parametric maximum likelihood estimator (Smith [4]) which we denote ML. This method seeks the value of k that minimizes an average distance with a penalty factor for the observations furthest from the tail. A study involving several estimators of γ is found in Neves and Fraga Alves [2]. However, to the authors' knowledge, there is still no analysis of the method involving the semi-parametric ML estimator. This is the objective of this work, namely, to find penalty values that contribute to the best performance of the method. An application to real data is also presented.

Acknowledgements The research of the first author 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, with references DOI 10.54499/UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020) and DOI 10.54499/UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020).

References

[1] J. Beirlant, J. Goegebeur, Y.and Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, Chichester, 2004.

[2] C. Neves and M. I. Fraga Alves. Reiss and thomas' automatic selection of the number of extremes. *Computational Statistics & Data Analysis*, 47(4):689–704, 2004.

[3] R. D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Springer Science & Business Media, Berlin, 2007.

[4] R. L. Smith. Estimating tails of probability distributions. *Ann. Statist.*, 15:1174–1207, 1987.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Classification rules for a folded directional distribution

Adelaide Figueiredo[1], **Fernanda Figueiredo**[2]
[1] University of Porto, School of Economics and Management and LIAAD-INESC TEC, Portugal, adelaide@fep.up.pt
[2] University of Porto, School of Economics and Management and CEAUL, University of Lisbon, Portugal, otilia@fep.up.pt

A folded directional distribution is preferable for modeling data on the positive orthant of the unit hypersphere compared to a simple directional distribution. In this paper we consider Bayes classification rules based on a folded von Mises-Fisher distribution to assign data into predefined groups. We evaluate the performance of these classification rules through simulations across various dimensions of the hypersphere.

Keywords: Bayes rule, directional data, folded distribution

The statistics of directional data, which deals with unit vectors on the surface of the hypersphere S^{p-1}, has seen significant advancements in recent years. Numerous applications of directional data have emerged in the literature. Initially, most applications focused on the circle and the sphere, but more recently, there has been a growing interest in applications on the hypersphere.

Discriminant analysis for directional data has been explored in the literature, with notable contributions by Morris and Laycock [4] and El Khattabbi and Streit [1] for circular and spherical data. Additionally, several studies have investigated discriminant analysis based on directional distributions on the hypersphere. For instance, Figueiredo and Gomes [3] and Figueiredo [2] provided classification rules for the Watson and von Mises-Fisher distributions, respectively, and evaluated the performance of these classification rules in various scenarios.

The von Mises-Fisher distribution, denoted by $M_p(\boldsymbol{\mu}, \kappa)$, is one of the most commonly used distributions for modeling directional data. Its probability density function is defined by

$$f(\mathbf{x}|\boldsymbol{\mu}, \kappa) = c_p(\kappa)\, \exp(\kappa \boldsymbol{\mu}'\mathbf{x}) \quad \mathbf{x} \in S_{p-1}, \quad \boldsymbol{\mu} \in S^{p-1},\ \kappa > 0\ , \tag{1}$$

where the normalising constant is given by $c_p(\kappa) = \kappa^{\frac{p}{2}-1}/[(2\pi)^{p/2}\, I_{p/2-1}(\kappa)]$ and $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind and order ν, which is defined by $I_\nu(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \cos\nu\theta e^{\kappa\cos\theta}\mathrm{d}\theta$. The parameter $\boldsymbol{\mu}$ is the mean direction and κ is the concentration parameter around $\boldsymbol{\mu}$. This distribution is rotationally symmetric about $\boldsymbol{\mu}$.

However, when data fall on the positive orthant of the unit hypersphere, a folded von Mises-Fisher distribution should be used instead of a von Mises-Fisher distribution.
If $\mathbf{X} = (X_1, X_2, ..., X_p)'$ has a von Mises-Fisher distribution $M_p(\boldsymbol{\mu}, \kappa)$, then $\mathbf{Y} = |\mathbf{X}| = (|X_1|, |X_2|, ..., |X_p|)'$ has a folded von Mises-Fisher distribution on the positive orthant of the unit hypersphere S_+^{p-1} defined by $S_+^{p-1} = \left\{ (\upsilon_1, \upsilon_2, , ..., \upsilon_p)' : \upsilon_j \geq 0, \sum_{j=1}^{p} \upsilon_j^2 = 1 \right\}$.
In this study we present the classification rules for a folded von Mises-Fisher distribution to assign data into predefined groups and evaluate its performance in various scenarios.
The performance of the rules was analyzed using data generated from two von Mises-Fisher populations, considering both known and unknown parameters for various dimensions of the hypersphere and different parameter values. It was concluded that when the parameters are unknown, the performance of the folded von Mises-Fisher rule is not greatly affected by estimating the parameters using maximum likelihood. When the parameters are known, the results showed that for the considered dimensions of the hypersphere, for each concentration parameter value and each angle between the mean vectors, the folded von Mises-Fisher rule performs better than the von Mises-Fisher rule.

Acknowledgement This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020 and UIDB/00006/2020, DOI: 10.54499/UIDB/00006/2020.
https://doi.org/10.54499/LA/P/0063/2020
https://doi.org/10.54499/UIDB/00006/2020.

References

[1] S. El Khattabi and F. Streit. Identification analysis in directional statistics. *Computational Statistics and Data Analysis*, 23:45–63, 1996.

[2] A. Figueiredo. Discriminant analysis for the von Mises-Fisher distribution. *Communications in Statistics - Simulation and Computation*, 38:1991–2003, 2009.

[3] A. Figueiredo and P. Gomes. Discriminant analysis based on the Watson distribution defined on the hypersphere. *Statistics*, 40(5):435–445, 2006.

[4] J. E. Morris and P. J. Laycock. Discriminant analysis of directional data. *Biometrika*, 61(2):335–341, 1974.

5 April, 11:30 - 12:00, Hall of Grande Auditório

A comparative analysis of residential water consumption models and forecasting approaches

Eliana Costa e Silva[1,2], **Tatiana Cunha**[3], **Flora Ferreira**[4]
[1] CIICESI, ESTG, Politécnico do Porto, Portugal eos@estg.ipp.pt
[2] Centre Algoritmi, Universidade do Minho, Portugal
[3] ESTG, Politécnico do Porto, Portugal8180272@estg.ipp.pt
[4] Centre of Mathematics/Dept. of Mathematics, University of Minho, Portugal
fjferreira@math.uminho.pt

Efficient water resource management requires accurate modeling and forecasting of consumption patterns. In this work, different models and forecasting approaches are used in the analysis of hourly data from residential clients. The results show that TBATS demonstrated strong capabilities in modeling multiple seasonalities, while SARIMA excelled in simplicity and precision for specific users. The findings reveal distinct consumption patterns and emphasize the need for tailored approaches, guiding utilities toward sustainable water resource management.

Keywords: Times series forecasting, TBATS, SARIMA, sustainability

---

The increasing demand for water, coupled with its limited availability, underscores the importance of accurate forecasting models for consumption patterns. This research compares Seasonal Autoregressive Integrated Moving Average (SARIMA) with Trigonometric Box-Cox Transformation with ARMA errors, Trend, and Seasonal Components (TBATS), aiming to identify the more robust approach for individual consumption monitoring. SARIMA models is suited for linear seasonal patterns, while TBATS is capable of addressing complex multiple seasonality

A dataset with 8,760 hourly observations is used, and interpolation is applied for outlier detection. The data analysis was performed R software. The performance of each model and forecasting approach was assessed using Root Mean Squared Error (RMSE). SARIMA's ability to simplify short seasonal data contrasts with TBATS's advantage in handling multifaceted seasonal trends.

TBATS presented better RMSE values for all consumers expect one. Thus TBATS revealed higher accuracy in capturing intricate consumption behaviors. Nonetheless, SARIMA presented better Akaike Information Criterion (AIC) scores, which suggests greater model parsimony. Unique patterns of consumption came to light, with notable variations during peak hours and across consumers.

An additional evaluation of the forecasting models involved analyzing the percentage of the 80% prediction intervals that captured the real hourly consumption over 49 weeks

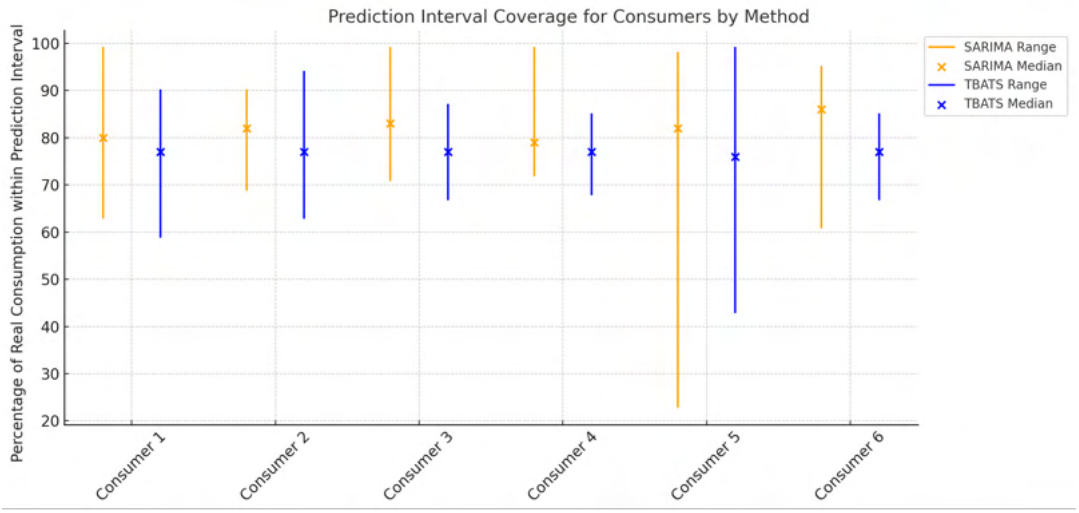

Figure 1: Percentage of prediction intervals containing the real hourly water consumption

(Fig. 1). Results demonstrated variability across consumers and methods. SARIMA consistently achieved higher median coverage percentages for consumers 1, 3, and 4, indicating slightly better reliability in capturing actual consumption. However, TBATS outperformed SARIMA for consumers 2 and 5, with intervals containing the real consumption in 95% and 99% of cases, respectively, compared to SARIMA's 90% and 98%. This suggests that while SARIMA provides stronger overall predictions for certain profiles, TBATS excels in scenarios with more complex seasonal patterns, offering greater robustness in coverage accuracy for specific consumers.
The comparative analysis provides actionable insights for integrating these models into monitoring systems, guiding water utilities in tailoring strategies to specific consumer profiles. Future work aims to expand the dataset and incorporate external factors like climatic variables for enhanced predictive accuracy.

Acknowledgements This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through projects UIDB/04728/2020, UIDP/00013/2020 and UIDB/00013/2020.

References

[1] T. Cunha, E. Costa e Silva, and F. Ferreira. Individual water consumption forecasting: Evaluating sarima and tbats methods. In *Proceedings of the 2024 7th International Conference on Mathematics and Statistics*, pages 57–62, 2024.

[2] P. I. Karamaziotis, A. Raptis, K. Nikolopoulos, K. Litsiou, and V. Assimakopoulos. An empirical investigation of water consumption forecasting methods. *International Journal of Forecasting*, 36:588–606, 2020.

5 April, 11:30 - 12:00, Hall of Grande Auditório

PAGEC: advancing attributed network analysis with joint embedding and clustering

Lazhar Labiod[1], **Mohamed Nadif**[1]
[1] Centre Borelli UMR 9010, Université Paris Cité
lazhar.labiod@u-paris.fr, mohamed.nadif@u-paris.fr

Representation learning in Attributed Networks is challenging, requiring both meaningful node embeddings and effective clustering. Traditionally, these tasks are addressed separately, missing potential benefits of joint optimization. We propose PAGEC (Power Attributed Graph Embedding and Clustering), a unified framework integrating embedding and clustering via a powered proximity matrix, capturing both structural and attribute-based affinities. Our theoretical analysis reveals strong links between this matrix and random walk theory.

Keywords: clustering, embedding, attributed graph

In recent years, Attributed Network Embedding (ANE) [1] has become a crucial task in applications like social networks, citation networks, and protein-protein interaction networks. ANE aims to create low-dimensional continuous representations of nodes while preserving both the topological structure and attribute proximities of the original network.
While Network Embedding (NE) has inspired numerous methods [11], research specifically focusing on ANE remains limited. Unlike NE, which relies solely on structural proximity, ANE incorporates both node connections and attribute affinities. These two sources of information differ significantly, rendering NE algorithms inadequate for ANE tasks. In fact, learned representations from ANE have proven beneficial for various tasks, including network clustering [10], node visualization [2], node classification [6], and link prediction [8]. However, ANE research faces critical challenges like high dimensionality, sparsity, and nonlinearity. Existing AN clustering methods often underperform due to two main limitations: (1) approximate continuous embeddings can deviate significantly from optimal discrete clustering solutions, and (2) the sequential nature of embedding and clustering leads to information loss. As a result, simultaneous embedding and clustering approaches have been proposed to bridge this gap [9, 3, 4, 5]. Here, we propose a novel simultaneous ANE and clustering framework [7]. Building on matrix decomposition, our approach jointly learns embeddings and discrete cluster labels by integrating node topology and attribute information. Unlike traditional methods, we enforce a discrete transformation on intermediate continuous embeddings, resulting in a tractable optimization problem with discrete solutions. Smooth transformations (e.g., rotations) map the relaxed continuous embedding to discrete clusters, addressing information loss in sequential methods. Experimental evaluations show that the PAGEC algorithm surpasses state-of-the-art algorithms in terms of both clustering accuracy and embedding quality across diverse attributed network datasets.

References

[1] H. Cai, V. W. Zheng, and K. C.-C. Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 30(9):1616–1637, 2018.

[2] Q. Dai, Q. Li, J. Tang, and D. Wang. Adversarial network embedding. In *AAAI*, pages 2167–2174, 2018.

[3] M. Febrissy, A. Salah, M. Ailem, and M. Nadif. Improving NMF clustering by leveraging contextual relationships among words. *Neurocomputing*, 495:105–117, 2022.

[4] C. Fettal, L. Labiod, and M. Nadif. Efficient graph convolution for joint node representation learning and clustering. In *WSDM*, pages 289–297, 2022.

[5] C. Fettal, L. Labiod, and M. Nadif. Simultaneous linear multi-view attributed graph representation learning and clustering. In *WSDM*, pages 303–311, 2023.

[6] X. Huang, J. Li, and X. Hu. Label informed attributed network embedding. In *WSDM*, pages 731–739, 2017.

[7] L. Labiod and M. Nadif. Power attributed graph embedding and clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1439–1444, 2022.

[8] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang. Adversarially regularized graph autoencoder for graph embedding. In *IJCAI*, pages 2609–2615, 2018.

[9] A. Salah and M. Nadif. Social regularized von mises–fisher mixture model for item recommendation. *Data Mining and Knowledge Discovery*, 31:1218–1241, 2017.

[10] C. Wang, S. Pan, G. Long, X. Zhu, and J. Jiang. Mgae: Marginalized graph autoencoder for graph clustering. In *CIKM*, pages 889–898, 2017.

[11] W. Yu, W. Cheng, C. Aggarwal, B. Zong, H. Chen, and W. Wang. Self-attentive attributed network embedding through adversarial learning. In *ICDM*, pages 758–767, 2019.

5 April, 11:30 - 12:00, Hall of Grande Auditório

Clustering, time series and risk analysis in surface water quality monitoring

A. Manuela Gonçalves[1], **Irene Brito**[1], **Ana Pedra**[2]
[1] Department of Mathematics and Centre of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt, ireneb@math.uminho.pt
[2] Centre of Mathematics, University of Minho, Portugal, pg46704@alunos.uminho.pt

This study combines clustering, time series forecasting and concepts from risk theory, to predict and analyze the surface water quality in a monitoring process. The aim is to analyse the risk for water pollution in the water monitoring stations clusters, proposing an average risk index predictor for water quality pollution, and to investigate if the overall ranking result (obtained in the in-sample period) can serve as a risk index for future water quality risk forecasts. Time series forecasting models will be applied to the data, to examine the ability of the risk measures and forecast the risk of water pollution. The methodologies are illustrated using a data set of surface water quality variables collected in the Douro River basin, in Portugal.

Keywords: clustering, time series, risk measures, risk ranking, surface water quality

The development of statistical methodologies for water quality variables analysis is a very important tool for the monitoring management of water pollution, such as in a river basin. The main objective of this work is to develop new methodologies by combining concepts from clustering analysis, risk theory and times series approaches in order to analyze, predict and forecast surface water quality. The methodologies are illustrated using a data set regarding the Douro River basin (in Portugal) in terms of environmental water quality variables, measured monthly by 18 monitoring stations and recorded in the period from January 2002 to December 2013 [2].
A cluster analysis was carried out to group homogeneous water monitoring stations in terms of surface water quality variables: dissolved oxygen (DO) and conductivity [3]. Since water quality depends on the flow variation, this approach differentiates and studies separately the time horizon from May to September (the dry period) and the time horizon from October to April (the wet period), after considering all the data. Several risk measures [1], such as value at risk, loss probability, variance, among others were determined for the considered clusters in order to assess the risk of water pollution for each one. Time series modeling approaches, such as SARIMA models and exponential smoothing methods, [5], were applied to the clusters to obtain forecasts for the last 12 months. The results were compared with the classifications obtained through risk measurements. This identifies clusters (of sampling stations) at the highest risk and concludes that pollution levels are

higher in the dry period [4]. Also, in order to analyze the risk of water pollution in the water station clusters, different risk measures, such as expected value and variance are calculated based on the DO and conductivity measurements in the in-sample period (from January 2002 to December 2012). The water station clusters are classified in terms of risk according to each risk measure, and a final ranking is established, considering the total, the dry and the wet periods. These results were compared with the ranking obtained from the average predictions and forecasts of the time series models, in the in-sample period and in the out-of-sample period (between January 2013 to December 2013), to analyze the performance of the risk index predictor using the time series forecasts.
The methodologies developed apply not only to the particular case of a hydrological basin, but are also a general procedure, which can be applied to any other situations where similar data are involved.

Acknowledgements A. Manuela Gonçalves and Irene Brito thank support from FCT through the projects UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020) and UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020. Ana Pedra thanks CMAT for the research fellowship (BI) UMINHO/BIM/2022/100.

References

[1] H. W. Brachinger and M. Weber. Risk as a primitive: A survey of measures of perceived risk. *Operations-Research-Spektrum*, 19:235–250, 1997.

[2] SNIRH. Sistema Nacional de Informação de Recursos Hídricos. `https://snirh.apambiente.pt/2023`.

[3] J. Ganoulis. *Risk Analysis of Water Pollution.* John Wiley & Sons, Weinheim, 2009.

[4] A. M. Gonçalves and M. Costa. Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, 25:151–163, 2011.

[5] R .H. Shumway and D. S. Stoffer. *Time Series Analysis and its Applications - with R Examples.* 4th edition, Springer, 2015.

Index

A. Catarina Gonçalves, 37
A. Manuela Gonçalves, 165
Adelaide Figueiredo, 159
Adelaide Freitas, 87
Alexandra Albuquerque, 95
Alexandra Oliveira, 71
Alexandre Cunha, 27
Ana Aida Sá, 87
Ana Desiderati, 125
Ana I. Melo, 37
Ana Marta-Costa, 121, 123
Ana Martins, 143
Ana Matos, 137
Ana Moreira, 133
Ana Paula Nascimento, 71
Ana Pedra, 165
Ana Ribeiro, 143
Anabela Oliveira, 139
André Costa, 21
André Fonseca, 149
Angela Montanari, 13, 77
António Branco, 11
Antonio Loría-García, 39
António Paulino, 111
Aurora Rodrigues, 139

Beatriz Ferreira, 35
Beatriz Silva, 115
Brígida Mónica Faria, 71
Bruno Oliveira, 99

Carla Afonso, 29
Carla Heriques, 125
Carla Palma, 139
Carlos Grilo, 117, 119
Carolina Costa, 19
Catarina Ferreira, 147
Christian Heumann, 79
Clara Viseu, 113
Cláudia Amanajás, 127
Cláudia Silvestre, 97
Conceição Rocha, 73
Cristina Gonçalves, 29
Cristina Lopes, 127, 143, 145, 147
Cristina Prudêncio, 71
Cristina Torres, 127, 143, 145, 147

Daniela Fernandes, 143
Daniela Vaz, 141
Diogo Santos, 45
Dora Carinhas, 135, 139

Eduarda Góis, 29
Eliana Costa e Silva, 99, 161
Elisa Moreira, 155
Elsa Soares, 107
Elysiario Santos, 119
Erick Alfredo Vásquez Murillo, 103
Estela Bicho, 41, 153
Eunice Venâncio, 123

Fábio Coutinho, 65
Fábio Gomes, 21
Faustino Sachimuco, 129
Fernanda Figueiredo, 159
Fernando Sebastião, 141
Fernando Silva, 33
Filipa Chambel, 25
Filipe Silva, 73
Flora Ferreira, 41, 61, 153, 161
Francisco Sousa Matos, 63
Francisco Tavares, 91

Gaspar J. Machado, 129
Genane Youness, 85
Gilbert Saporta, 85
Gonçalo Amado, 17

Hans-Peter Piepho, 151

Helena Bacelar-Nicolau, 71
Helena Figueiredo Pina, 97
Helena L. Grilo, 81
Helena Mouriño, 51
Hugo Carvalho, 127
Hugo Cipriano, 21

Ibrahim Prazeres, 121
Inês Braga, 95
Inês Rocha, 113
Inês Sousa, 107
Irene Brito, 105, 129, 165
Isabel Gomes, 147
Isabel Pedrosa, 111
Isabel Pereira, 69
Isabel Vieira, 127, 143, 145, 147
Ivanise Gomes, 145

Jacopo Bono, 47
Jaime Vale, 33
Javier Gavilan, 137
Jéssica Martins, 145
Jéssica Nadais, 147
Jhonathan Barrios, 41
Joana Leite, 65, 93, 111
João Fonseca, 149
João Pedro Araújo, 91
João Peixoto, 117
João Poças, 27
João S. Lopes, 25
José Brito, 73
José G. Dias, 83
José Marques Santos, 137
José Martins, 117, 119
Joseph O. Ogutu, 151
Judite Vieira, 141
Juliana Barbosa, 143
Juliana Castanheira, 87

Laura Anderlucci, 77
Lazhar Labiod, 163
Lídia Maria Galvão Rodrigues Praça, 101
Lígia Henriques-Rodrigues, 39
Liliana Monteiro, 157
Lucas de Souza, 83
Luís Chambel, 131
Luís Cotrim, 141
Luis Eduardo Amaya Briceño, 103
Luís M. Grilo, 81, 121, 123
Luís Sousa, 69
Luísa Novais, 91
Lurdes Babo, 127, 143, 145, 147

Magda Monteiro, 67, 69
Manuela Larguinho, 113
Marco Costa, 37, 67
Margarida G. M. S. Cardoso, 131
Maria Eduarda Silva, 33
Maria Gonçalves, 59
Maria Inês Vicente, 93
Maria Manuel Angélico, 55
Maria Manuel Pinho, 29
Maria Raquel Lucas, 121, 123
Maria Rodrigues, 141
Marta Ferreira, 155, 157
Marta Simões, 111
Mia Hubert, 5, 9
Miguel Gago, 41
Miguel Picoto, 135
Mohamed Nadif, 163
Mónica Vieira, 71
Mouhamadou Lamine Ndao, 85

Ndèye Niang, 85
Nuno Almeida, 149
Nuno Romão, 25

Óscar Oliveira, 99
Ounísia Santos, 141

Patrícia Pinto, 145
Paula Amaral, 75
Paula Brito, 89
Paula Carvalho, 95
Paula Simões, 149
Paulo Barbosa, 153
Paulo Barreira, 91
Paulo Infante, 135
Paulo Saraiva, 27
Pedro Campos, 39, 59
Pedro Cruz, 21
Pedro Damião Henriques, 121, 123
Pedro Duarte Silva, 63
Pedro Guimarães, 73

Pedro Rodrigues, 135
Peter J. Rousseeuw, 5, 9

Raquel Menezes, 35, 55
Raul Bernardino, 141
Renato Fernandes, 73
Rita Gaio, 79
Rui Costa-Miranda, 79
Rui Lucena, 149
Rui Nunes, 89

Sandra Moreira, 139
Sofia Rodrigues, 27
Sónia Dias, 89
Sónia Gouveia, 53
Soraia Pereira, 55
Susana Faria, 115, 133
Suzanne Amaro, 125

Tatiana Cunha, 161
Teresa Barros, 95
Tiago Dias, 75
Tiago Marques, 55
Tiago Pinho Pereira, 19

Vanda M. Lourenço, 151
Vanessa Freitas Silva, 33
Vânia Lopes, 21
Vânia Ribeiro, 141
Vera Valente, 127

Wenceslao González-Manteiga, 79
Wolfram Erlhagen, 41, 153

SPONSORS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


CEOS.PP
CENTRO DE ESTUDOS ORGANIZACIONAIS
E SOCIAIS DO POLITÉCNICO DO PORTO


P.PORTO
ISCAP


BANCO DE PORTUGAL
EUROSYSTEM


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


PSE
your data
specialists


NORTE


LEYA EDUCAÇÃO
www.leyaeducacao.com


matosinhos


Axis
HOTÉIS