PROGRAMME and BOOK of ABSTRACTS

JOCLAD 2026

9 - 11 APRIL

PORTALEGRE, PORTUGAL

Associação Portuguesa de Classificação e Análise de Dados

PORTALEGRE POLYTECHNIC UNIVERSITY | POLITECNICO DE PORTALEGRE

XXXIII MEETING OF THE PORTUGUESE ASSOCIATION FOR CLASSIFICATION AND DATA ANALYSIS

XXXIII JORNADAS DE **CL**ASSIFICAÇÃO E **A**NÁLISE DE **D**ADOS

Programme and Book of Abstracts

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

9–11 April 2026

Portalegre, Portugal

https://sites.google.com/view/joclad-2026-pt

Sponsors

Banco de Portugal
Câmara Municipal de Arronches
Câmara Municipal de Ponte de Sor
Câmara Municipal de Portalegre
DELTA cafés
Herdade da Amada
Instituto Nacional de Estatística/Statistics Portugal
Instituto Politécnico de Portalegre
PSE — Produtos e Serviços de Estatística

Organisers

Associação Portuguesa de Classificação e Análise de Dados (CLAD)
Instituto Politécnico de Portalegre

Programme and Book of Abstracts
XXXIII Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2026)

Editors: Rita Gaio, Adelaide Figueiredo, Irene Brito, Jorge Caiado, José G. Dias, Paula Brito, Raquel Sebastião

Publisher: CLAD - Associação Portuguesa de Classificação e Análise de Dados

Printed: Statistics Portugal

ISBN 978-989-35097-3-9

Depósito legal: 561624/26

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 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 33rd edition of JOCLAD takes place at Escola Superior de Tecnologia, Gestão e Design of the Instituto Politécnico de Portalegre. The institution offers diverse programs across multiple schools, including technology, management, and health, fostering interdisciplinary collaboration highly relevant to data-driven fields. Its manageable size and campus setting further encourage interaction among participants, making it an ideal environment for focused academic discussion and networking.

The scientific program of JOCLAD 2026 focuses on a broad spectrum of topics, including time series analysis, data mining, regression models, and machine learning, as well as practical applications in data analysis. It features 31 oral contributed talks and 20 poster presentations. A specialized short course by Francisco Herrera on Imbalanced Data and Explainable AI provides participants with valuable tools for classification, data analysis, and data science. In addition, three keynote plenary lectures by distinguished international experts address the latest advances in automated data science (Adalbert Wilhelm), self-organizing maps and multidimensional projections (Victor Lobo), and imbalanced data (Francisco Herrera), highlighting key developments in these evolving fields.

The thematic sessions at JOCLAD 2026 reflect the ongoing collaboration between CLAD and key institutions advancing data science and its applications. We are pleased to feature sessions from the Bank of Portugal, which explores Numbers with a story; Statistics Portugal, which addresses Official statistics, innovation, and academia; and the Portuguese Statistical Society session, which focuses on the many faces of statistics.

In addition to these institutional contributions, the thematic sessions also feature perspectives from the corporate sector. Kyndryl and PSE discuss turning data into decisions in innovative applications of machine learning and data analytics, highlighting both the challenges and opportunities.

CLAD's dedication to nurturing young talent continues with the award of two scholarships

for students presenting high-quality work at JOCLAD 2026. These scholarships, granted to a promising master's student and to a doctoral candidate, reflect our commitment to supporting the next generation of researchers in the data science community. The jury for the CLAD Scholarships 2026 comprised Rita Gaio (University of Porto, Chair), Anabela Afonso (University of Évora), and Filomena Teodoro (Escola Naval).

Each abstract published in this volume has been evaluated by the Scientific Committee, composed of Rita Gaio, Adelaide Figueiredo, Irene Brito, Jorge Caiado, José G. Dias, and Raquel Sebastião, whose work significantly contributed to the quality of the JOCLAD 2026 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 2026, with a special mention to our long-lasting partners, Statistics Portugal and the Bank of Portugal. Finally, we convey our thanks to the members of the Local Organising Committee, Cristina Silva Santos, Sérgio Correia, Isabel Borges, Orlanda Póvoa, Sofia Catarreira, and João Cordeiro, whose enthusiasm and tireless work have made JOCLAD 2026 come true.

In addition to the scientific program, JOCLAD 2026 offers a rich and engaging social programme. Participants will visit the Tapestry Museum in Portalegre, where the Alentejo wine of honour will be served. The social dinner is offered in a cosy space of Marvão, a dramatic hilltop village perched on a granite ridge in Portugal's Serra de São Mamede. These moments of cultural immersion and social interaction are essential.

As we come together in Portalegre, JOCLAD 2026 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.

Portalegre, April 2026

Chair of the Scientific Programme Committee
Rita Gaio

Conference Chair
Cristina Silva Dias

President of CLAD
Paula Brito

Organisation

President of CLAD

Paula Brito

Chair of JOCLAD 2026

Cristina Silva Dias (Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Portalegre)

Local Organising Committee

Cristina Silva Dias (ESTGD – IPPortalegre)
João Cordeiro (UBI & NOVA LINCS)
Sérgio Correia (ESTGD – IPPortalegre)
Isabel Borges (ESTGD – IPPortalegre)
Orlanda Póvoa (ESBE – IPPortalegre)
Sofia Catarreira (ESTGD – IPPortalegre & Escola Secundária de Campo Maior)

Chair of the Scientific Programme Committee

Rita Gaio (Faculdade de Ciências – Universidade do Porto & CMUP)

Scientific Programme Committee

Rita Gaio (Faculdade de Ciências – Universidade do Porto & CMUP)
Adelaide Figueiredo (Faculdade de Economia – Universidade do Porto & INESC TEC)
Irene Brito (Escola de Ciências – Universidade do Minho & CMAT)
José G. Dias (ISCTE-IUL & BRU-Iscte)
Jorge Caiado (ISEG – Universidade de Lisboa & CEMAPRE)
Raquel Sebastião (Instituto Politécnico de Viseu & IEETA – Universidade de Aveiro)

Contents

Programme Overview

Thursday, 9 April – Campus Politécnico

08:30	Registration	BioBIP Building
09:00	**Short Course**	Room Nexus
10:30	Coffee Break	Room E113
11:00	**Short Course** (cont.)	Room Nexus
12:30	Lunch Time	Refectory, Main Building
13:30	Registration	Hall of ESTGD, Main Building
14:00	**Opening Session**	Auditorium ESTGD
14:30	**Keynote Lecture I**	Auditorium ESTGD
15:30	Coffee Break	Hall of ESTGD
16:00	**Parallel Sessions I**	Auditorium ESTGD
		Lecture Theatre E1, Anexo Building
18:00	**Visit to the Tapestry Museum**	Portalegre
19:00	**Alentejo de Honra**	Tapestry Museum, Portalegre

Friday, 10 April – Campus Politécnico

08:30	Registration	Hall of ESTGD
09:00	**Parallel Sessions II**	Lecture Theatre E1, Anexo Building
		Auditorium Francisco Tomatas, Main Building
10:10	**Parallel Sessions III**	Lecture Theatre E1, Anexo Building
		Auditorium Francisco Tomatas, Main Building
11:20	Coffee Break	Hall of ESTGD
11:30	**Poster Session I**	Hall of ESTGD
12:00	**Keynote Lecture II**	Lecture Theatre E1, Anexo Building
13:00	Lunch Time	Refectory, Main Building
14:30	**Thematic Session I – Bank of Portugal**	
		Lecture Theatre E1, Anexo Building
15:30	**Thematic Session II – Statistics Portugal**	
		Lecture Theatre E1, Anexo Building
16:30	Coffee Break	Hall of ESTGD
16:50	**Thematic Session III – CLAD 2026 Scholarships**	
		Lecture Theatre E1, Anexo Building
17:40	**Parallel Sessions IV**	Lecture Theatre E1, Anexo Building
		Auditorium Francisco Tomatas, Main Building
18:30	General Assembly of CLAD	Lecture Theatre E1, Anexo Building
19:00	Bus to Marvão	
20:00	**Social Dinner**	Restaurant 'Varanda do Alentejo', Marvão

Saturday, 11 April – Campus Politécnico

09:00	**Parallel Sessions V**	Auditorium Francisco Tomatas, Main Building Lecture Theatre E1, Anexo Building
10:30	**Thematic Session IV – SPE**	Lecture Theatre E1, Anexo Building
11:30	Coffee Break	Hall of ESTGD
11:40	**Poster Session II**	Hall of ESTGD
12:15	**Portalegre@CKATHON Awards**	Lecture Theatre E1, Anexo Building
12:45	Lunch Time	Refectory of ESTGD
14:30	**Thematic Session V – CLAD Corporate**	Lecture Theatre E1, Anexo Building
15:30	**Keynote Lecture III**	Lecture Theatre E1, Anexo Building
16:30	**Closing Session**	Lecture Theatre E1, Anexo Building

Programme

Thursday, 9 April

08:30 Registration – BioBIP Building

09:00 **Short Course** – Room Nexus
Imbalanced learning meets explainable human–AI interaction
Francisco Herrera, p. 5

Chair: Rita Gaio

10:30 **Coffee Break**

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

12:30 **Lunch Time**

13:30 Registration – Hall of ESTGD, Main Building

14:00 **Opening Session** Auditorium ESTGD

14:30 **Keynote Session I** – Auditorium ESTGD
Advances in automated data science: integrating deep learning and explainable AI for robust decision-making
Adalbert Wilhelm, p. 9

Chair: Paula Brito

15:30 **Coffee Break**

16:00 **Parallel Sessions I**

	Auditorium ESTGD **Time series analysis** Chair: Jorge Caiado	Lecture Theatre E1, Anexo Building **Machine learning** Chair: Pedro Duarte Silva
16:00	**Generalized model-based approach for count time series clustering**, Luís Sousa, Magda Monteiro, Isabel Pereira, Dimitris Karlis, p. 51	**Principal components for distributional data**, Sónia Dias, Paula Brito, p. 59
16:20	**Statistical modelling and time series clustering of retail imports and distance sales in Europe**, Magda Monteiro, Marco Costa, p. 53	**Kernel K-means clustering of distributional data**, Amparo Baíllo, José Ramón Berrendero, Martín Sánchez-Signorini, p. 63
16:40	**Building environmental justice indicators based on time series of counts models**, Adriano Gomes, Ana Martins, Sónia Gouveia, p. 55	**Model-selection in topic modeling**, José G. Dias, p. 61
17:00	**Outliers in state-space models: a robust approach to parameter estimation and Kalman filter**, A. Catarina Freitas, A. Manuela Gonçalves, Marco Costa, p. 57	**Analysis of classification models under reduced experimental designs in gene expression data**, José Febra, Paula Faria, João Meneses, Carlos Grilo, p. 65

18:00 **Tapestry Museum** – Portalegre

19:00 **Alentejo de Honra** – Tapestry Museum, Portalegre

Friday, 10 April

08:30 Registration – Hall of ESTGD

09:00 **Parallel Sessions II**

	Lecture Theatre E1 **Regression models I** Chair: Sónia Dias	Auditorium Francisco Tomatas **Data analysis I** Chair: Cristina Lopes
9:00	**Reproduction characterization of an ant population: an interdisciplinary data modelling**, Laura Machado, Filipe Ribeiro, Dulce Gomes, p. 67	**Modeling the association between type 2 diabetes and liver stiness in MASLD patients**, Ana Matos, Carla Henriques, Paula Mesquita, Armando Carvalho, Adélia Simão, p. 71
9:20	**How efficient are Portuguese beef farms?**, Cândida Santos, José G. Dias, M. Rosário Oliveira, Pedro Reis, p. 69	**Evaluating a numerical discomfort scale for immobilized trauma victims**, Carla Henriques, Ana Matos, Mauro Mota, p. 73
9:40		**Students' attitudes towards Mathematics: the influence of individual and academic characteristics in the 1st cycle**, Ana Felizardo Henriques, Adelaide Freitas, Fernando Sebastião, João Marôco, p. 75

10:10 **Parallel Sessions III**

	Lecture Theatre E1 **Regression models II** Chair: A. Manuela Gonçalves	Auditorium Francisco Tomatas **Data analysis II** Chair: Susana Faria
10:10	**Recent advances in model checking for logistic partially linear models**, Rui Costa-Miranda, Wenceslao González-Manteiga, Rita Gaio , p. 77	**Using data analysis for understanding the role of social virtual reality (VR) in stroke rehabilitation**, Carlos Ferreira, Mariana Leite, Sérgio Oliveira, Bernardo Marques, Beatriz Sousa Santos, p. 83
10:30	**Clusterwise linear regression for symbolic density-valued data**, Rui Nunes, Paula Brito, Sónia Dias, p. 79	**Statistical analysis for predicting harvest dates based on the maturation control cycles in the Vinho Verde wine region**, Mafalda T. Costa, Óscar Pereira, Bruno Leitão, António Seabra, Alexander Cornejo, Catarina Beth Dantas, Maria J. Polidoro, p. 85
10:50	**Impact of standardization methods on quantile regression models: a comparative study**, Dulce G. Pereira, Anabela Afonso, p. 55	**Tools for cohesion policies: a composite indicator of socio-spatial vulnerability for the municipalities of Portugal**, Aitor Varea Oro, Guilherme Vara, Sílvia Jorge, Rita Gaio, Pietro Brites, p. 87

11:20 **Coffee Break**

11:30 **Poster Session I** – Hall of ESTGD

Chair: Isabel Borges

12:00 **Keynote Lecture II** – Lecture Theatre E1
Update on the use of self-organizing maps and multidimensional projections
Victor Lobo

Chair: Cristina Silva Dias

13:00 **Lunch Time**

14:30 **Thematic Session I – Bank of Portugal** – Lecture Theatre E1
Numbers with a story: insights into trade, wealth and engagement

Chair: Luís Teles

14:30 **When statistics go social: meeting people where they scroll**, Ana Castor, p. 15
14:50 **How are Portuguese services traded? Analysing Banco de Portugal's data on internacional trade in services by mode of supply**, Inês Gouveia, Tiago Castro, p. 17
15:10 **Soaring housing prices: spillovers to finantial sectos and household's wealth**, André Oliveira, Diogo Guerreiro, p. 19

15:30 **Thematic Session II – Statistics Portugal** – Lecture Theatre E1
Official statistics, innovation and academia

Chair: Pedro Campos

15:30 **Use of administrative tax data as a input for compiling the house price index**, Ângelo Teixeira, Vítor Mendonça, Helena Carvalho, p. 23
15:50 **From official statistics to scientific knowledge: the potential of secure access to microdata**, José A. Pinto Martins, p. 25
16:10 **Seasons in the algorithm: error-driven insights into Winter and Spring crops classification - an exploratory study by Statistics Portugal**, Cristina Gabriel, Isabel Gonçalves, p. 27

16:30 **Coffee Break**

16:50 **Thematic Session III – CLAD 2026 Scholarships** – Lecture Theatre E1

Chair: Anabela Afonso

16:50 **PerPCA applied on air pollution data**, Diana Vázquez Limón, Adelaide Freitas, p. 31

17:10 **A smooth approximation for the interval Fischer's discriminant analysis**, Diogo Vaz, M. Rosário Oliveira, p. 33

17:40 **Parallel Sessions IV**

	Lecture Theatre E1 **Latent variable models** Chair: José G. Dias	Auditorium Francisco Tomatas **Outlier detection** Chair: Adelaide Figueiredo
17:40	**Model selection with Group LASSO in finite mixture linear regression models**, Ana Moreira, Susana Faria, p. 89	**Detecting VAT Fraud: an approach based on temporal analysis of bank transactions**, Ana Helena Tavares, João Marques, p. 91
18:00	**Structural equation modeling to analyze the impact of the innovation incentive system on the internationalization of Portuguese companies**, Luis Grilo, Elsa Pereira, Jean Maidana, Milan Stehlíkh, p. 93	**The challenge of hidden outliers: a robust approach to panel data**, Anabela Rocha, Cristina Miranda, Manuela Souto de Miranda, p. 95

18:30 **General Assembly of CLAD**
Lecture Theatre E1

20:30 **Social Dinner** – Marvão
Restaurant 'Varanda do Alentejo'

Saturday, 11 April

9:00 **Parallel Sessions V**

	Auditorium Francisco Tomatas **Optimization** Chair: Irene Brito	Lecture Theatre E1 **Data Mining** Chair: Margarida Cardoso
9:00	**Were European funds well distributed through the Portuguese municipalities or was it unfair?**, Guilherme Vara, Aitor Varea Oro, Sílvia Jorge Jorge, Pietro Brites, Rui Barros, Rita Gaio, p. 97	**Discriminant analysis for a folded directional distribution**, Adelaide Figueiredo, Fernanda Figueiredo, p. 105
9:20	**A robo-advisor based on the Markowitz model and investor risk profiling**, Manuel Rodrigues, Conceição Amado, p. 99	**Simulating Rosenthal's fail-safe number**, Vanusa Rocha, Vera Afreixo, Miguel Felgueiras, p. 107
9:40	**Exact sparsity control for multiclass linear support vector machines**, Immanuel Bomze, Laura Palagi, Bo Peng, Pedro Duarte Silva, Federico D'Onofrio, Marta Monaci, p. 101	**Respondent-driven sampling as adaptive network sampling**, Manuela Maia, Pedro Campos, p. 109
10:00	**maxRgain: R package for optimizing genetic gains in the selection of groups of genotypes**, Sónia Surgy, Jorge Cadima, Elsa Gonçalves, p. 101	**A data-driven scoring framework for robotic movement complexity**, Daniel Rodrigues, Eliana Costa e Silva, Pedro Ribeiro, Inês Costa, Gianpaolo Gulletta, Luís Louro, Sérgio Monteiro, André Cardoso, Ana Colin, Estela Bicho, p. 109

10:30 **Thematic Session IV – SPE** – Lecture Theatre E1
Many faces of Statistics
Chair: Conceição Amado

10:30 **Extremes without independence: the other side of the history**, M. Cristina Miranda, p. 43

10:50 **A branching process approach on population survival: skeleton process and stochastic introgression**, Maria Conceição Serra, p. 45

11:10 **From multi-omics to prediction: the priority-elastic net framework**, Eunice Carrasquinha, p. 47

11:30 **Coffee Break**

11:40 **Poster Session II** – Hall of ESTGD

Chair: Orlanda Póvoa

12:15 **Portalegre@CKATHON Awards** – Lecture Theatre E1
Chair: Anabela Afonso

12:45 **Lunch Time**

14:30 **Thematic Session V – CLAD Corporate: Kyndryl, PSE**
Lecture Theatre E1
Chair: Sérgio Duarte Correia

14:30 **Turning data into decisions: real cases from PSE**, Nuno Gomes, Rui Almeida, p. 37

15:00 **Data analysis and processing in smart cities**, Carla Oliveira, Nuno Correia, Laura Guijarro, p. 39

15:30 **Keynote Lecture III** – Lecture Theatre E1
Francisco Herrera
Chair: Rita Gaio

16:30 **Closing Session** – Lecture Theatre E1

Abstracts

Short Course

9 April, 9:00 - 10:30, 11:00 - 12:30, Room Nexus – Edifício da BioBIP

Imbalanced learning meets explainable human–AI interaction

Francisco Herrera

University of Granada, Spain, herrera@decsai.ugr.es

The short course explores the topic of machine learning under class imbalance. After showing that standard learning procedures may mask systematic failures on critical cases, it covers mechanisms to address imbalance at the data level. It then highlights Explainable AI and human-in-the-loop systems as essential for interpreting models and ensuring responsible decisions in high-risk domains.

Keywords: imbalanced classification, minority class risk, data-centric AI, synthetic data generation, Explainable AI (XAI), human–AI interaction, human-in-the-loop

---

This course examines machine learning systems operating under minority class risk, where misclassifications of rare events can lead to disproportionate costs, ethical concerns, and real-world harm. It begins with a strong focus on the technical foundations of imbalanced classification, analysing how skewed data distributions undermine standard learning procedures, loss functions, and evaluation metrics, often masking systematic failures on critical cases. Attention is given to data-centric approaches, including resampling strategies and synthetic data generation, as primary mechanisms to address imbalance at the data level. Through the lens of cost-sensitive learning, participants explore how asymmetric error costs and minority risks can be explicitly incorporated into model design, training, and evaluation, moving beyond accuracy-driven optimization. The course also examines recent advances in tabular foundation models, discussing their potential and limitations in imbalanced settings, as well as the implications for robust evaluation and reliable deployment in high-risk domains.

Building on these technical foundations, the course then introduces Explainable Artificial Intelligence (XAI) as a necessary mechanism to understand, validate, and control model behaviour in imbalanced settings. XAI is presented as a key enabler of effective Human–AI interaction, allowing humans to interrogate model decisions, detect spurious patterns amplified by imbalance, and assess model reliability in high-risk cases. Finally, drawing on recent work on agentic information systems and ethical delegation from a stakeholder perspective, the course frames human-in-the-loop approaches as a fundamental requirement in high-risk domains—such as medical applications—where decisions may be partially delegated to AI systems. In this context, human expertise, informed by technical rigor and clear explanations, remains central to responsible, stakeholder-aware AI-driven decision-making.

References

[1] M. Carvalho, A. J. Pinho, and S. Brás. Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, 12 (71), 2025.

[2] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023.

[3] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera. *Learning from Imbalanced Data Sets*. Springer, 2018.

[4] X. Gao, D. Xie, Y. Zhang, Z. Wang, C. Chen, C. He, H. Yin, and W. Zhang. A comprehensive survey on imbalanced data learning. arXiv preprint arXiv:2502.08960, 2025.

[5] F. Herrera. Reflections and attentiveness on explainable artificial intelligence (xai). the journey ahead from criticisms to human–ai collaboration. *Information Fusion*, 121:103133, 2025.

[6] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[7] K. Saeed and V. R. Prybutok. When utility meets ethics: A stakeholder perspective on agentic information systems delegation. *International Journal of Information Management*, 86:102976, 2026.

Keynote Lectures

9 April, 14:30 - 15:30, Auditorium ESTGD

Advances in automated data science: integrating deep learning and explainable AI for robust decision-making

Adalbert F.X. Wilhelm[1], **Hameed Moqadam**[1], **Ehsan Mehdipour**[1]
[1] Constructor University, awilhelm@constructor.university,
hmoqadam@constructor.university, emehdipour@constructor.university

---

With the proliferation of large-scale, heterogeneous data sources there is a growing need for robust, automated, and interpretable analytical solutions allowing a smooth data fusion as well as a comprehensive usage of available data. This presentation synthesizes key findings and methodological advances from our latest research based on projects related to marine biology and glaciology, outlining both theoretical considerations and practical data-pipeline implementations.

Keywords: data pipeline, data fusion, image analysis, empirical orthogonal functions, internal reflection horizons

---

Automated data science workflows are increasingly central to extracting actionable information from large, heterogeneous environmental and geophysical datasets. Their deployment in high-stakes decision contexts remains constrained by concerns over robustness, interpretability, and operational scalability. In many applications, such as climate and hydrological modeling, hazard monitoring, and environmental risk assessment, decision-makers must rely on models that not only deliver high predictive performance but also communicate their internal logic and uncertainties in a transparent and diagnostically useful way. These workflows or scientific data pipelines are structured sequences of computational steps that transform raw data into analysis-ready products and interpretable results, with a strong emphasis on reproducibility, transparency, and methodological rigour in line with FAIR principles

As data sources are often heterogeneous with respect to format, resolution, uncertainty characteristics, metadata con- ventions, and geographical reference, a primary objective in data fusion and data pipeline implementation is the construction of coherent, analysis-ready datasets that are consistent in representation while preserving relevant infor- mation for interpretation.

In [1], satellite observations of phytoplankton chlorophyll-a (including several phytoplankton functional types) from Sentinel-3 OLCI are affected by gaps due to clouds and poor viewing conditions, which hampers robust trend analysis and other applications. Two gap-filling methods, DINEOF and the deep-learning-based DINCAE, are applied to three

years of Atlantic Ocean data to reconstruct complete fields, and these reconstructions are evaluated using withheld test data and in situ measurements from the RV Polarstern PS113 cruise. The results show that DINCAE provides better reconstructions than DINEOF, especially for transient features. Building on this, a data fusion pipeline integrates the gap-filled satellite Chla products with ACS spectrophotometer measurements to improve estimates of phytoplankton functional type composition. The approach combines gap-filling, EOF-based modeling, and optimal interpolation to generate spatially and temporally consistent chlorophyll-a fields, which are then validated against laboratory reference measurements to demonstrate enhanced accuracy and reliability of the fused products.
Once information-rich, interpretable data is produced, automatic analysis supports comprehensive analysis of the available input. [2] introduces a new method for automatically mapping the internal structure of ice sheets, known as internal reflection horizons (IRHs), using deep learning. IRHs are critical for understanding ice sheet behavior, which can help improve predictions of the ice sheet's past and present sea level rise contribution. Traditionally, tracing these layers required manual or semi-automatic methods, which were slow and labor-intensive. Our model, IRHMapNet, uses a machine learning approach known as a U-Net architecture to automatically trace these IRHs on radar data. By training the model on manually-labeled data and applying advanced image processing techniques, we were able to achieve high accuracy in detecting IRHs, even in complex ice sheet environments. The significance of this work is that it provides a faster, more accurate method for analyzing large amounts of radar data, which could greatly aid scientists in predicting future changes in ice sheets and global sea levels. Our method represents a step forward in automating processes that were previously extremely time-consuming, allowing for more efficient glaciological studies.

Acknowledgements EM's and HM's contribution was funded through the Helmholtz School for Marine Data Science (MarDATA), Grant HIDSS-0005. The authors are very grateful to Olaf Eisen and Astrid Bracher and her teams at AWI Bremerhaven.

References

[1] E. Mehdipour, H. Xi, A. Barth, A. Alvera-Azcárate, A. Wilhelm, and A. Bracher. Assessment of gap-filling techniques applied to satellite phytoplankton composition products for the Atlantic Ocean. *Geoscientific Model Development*, 19(4):1619–1643, 2026.

[2] H. Moqadam, D. Steinhage, A. Wilhelm, and O. Eisen. Going deeper with deep learning: Automatically tracing internal reflection horizons in ice sheets—methodology and benchmark data set. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2):e2024JH000493, 2025.

10 April, 12:00 - 13:00, Lecture Theatre E1 – Anexo Building

Update on the use of self-organizing maps and multidimensional projections

Victor Lobo

Escola Naval & Universidade Nova de Lisboa, vlobo@novaims.unl.pt

---

This talk reviews historical and theoretical aspects of Self-Organizing Maps (SOM), highlighting the work of Teuvo Kohonen. It outlines the basic algorithm and key applications, including clustering, anomaly detection, and visualization. Examples from diverse fields, particularly maritime contexts, are presented. Recent developments are discussed, including links to deep learning, explainable AI, and comparisons with techniques such as t-SNE and UMAP.

Keywords: keyword 1, keyword 2, keyword 3, keyword 4, keyword 5

Self-Organizing Maps (SOM), also known as Self-Organizing Feature Maps, Kohonen Neural Networks, or simply Kohonen Nets, are a neural network architecture, originally developed by Professor Tuevo Kohonen in 1982.

Almost all neural networks used today have multiple layers of neurons, and use supervised learning, based on the Error Backpropagation learning rule published in 1986 by Rumelhart et.al.. In fact, that algorithm had been independently developed by Paul Werbos in 1974, but his work had basically no impact on the scientific community because it was not published in a major scientific journal related to Computer Science. Today, multilayer neural networks are widely used in everyday engineering applications such as Optical Character Recognition (OCR), control, classification and regression.

However, SOM uses a much simpler architecture, with two fundamental differences from most neural networks. On one hand, it uses a single layer of neurons (or units), connected to each other only by neighbourhood relationships in a fixed grid. On the other hand, it uses unsupervised learning, with an update learning rule that is very similar to the well-known k-means algorithm.

The simplicity of the architecture and learning rule, allied with the lack of solid theoretical foundations (such as the minimization of a global energy function) and its apparent inability to perform supervised classification and regression tasks, led to a lack of generalized interest of the scientific community in the SOM, that has been usually used a “side show” to perform very specific tasks in a more complex architecture.

And yet, SOM has proven to be very useful and efficient in a very wide range of applications, and new ways of using it have continued to be developed in many different areas. A simple search in the Scopus database over the last 40 years, with SOM (or its other names) in the keywords, title, or abstract, reveals that 25.460 papers have been written. Surprisingly,

there was a "slow start" between 1986, when there was only 1 paper, and 1994, when 170 papers were published. This coincides with the excitement of the use Muti-layer Perceptrons (MLP). From there until 2006, when 1110 papers were published there was a sharp increase in interest. But ever since, for the last 20 years, there has been a "plateau" of around 1000 papers per year on this subject. In the face of the tremendous success of multilayer networks, why has research on SOM continued to be relevant and actively pursued by a large community ? One clear reason is that the original SOM performs very well a large set of tasks, and there is no need to complicate what is simple (keep it simple – KISS). Another is that it has continued to evolve into numerous variants, and to find new areas of application. The papers that have been published on SOM fall into four main categories:

1. New algorithms that have spawned numerous variants, closely or loosely related to the original SOM. These algorithms have explored different update rules, variable grids, new visualization techniques, and others.

2. Theoretical and experimental explanations of SOM behaviour, both in general and in the presence of specific statistical distributions, and issues such as the estimation of the magnification factor.

3. New and creative ways of using SOM for fundamentally different tasks, such as solving Traveling Salesman related Problems (TSP), performing feature extraction, sampling, exploratory clustering, ordering of multidimensional data, and data visualization.

4. Applying SOM, in sometimes very creative ways, to different use cases and areas, many time in conjunction with other techniques. Only 30

In this talk, we will start by reviewing historical, and theoretical aspects of SOM, with a special mention to the work of Professor Tuevo Kohonen that recently passed away. We will then give a brief explanation of the basic SOM algorithm. SOM has been explored by CLAD (the Portuguese scientific association of reference in the field of data science, and member of the International Federation of Classification Societies IFCS) in many JOCLAD conferences and in short courses, so this will be just an overview.
We will then critically review the tasks where SOM has traditionally performes well, such as "deterministic" clustering, "exploratory" clustering, anomaly detection, dimensionality reduction, sampling and anonymization, data visualization, non-linear multidimensional data projection, simple supervised learning, data ordering, and gating. At the same time, we will present a sample of very different application areas, despite a bias towards maritime applications due the authors personal experience.
Finally, we will examine recent "hot topics" of research, such as how SOM has been used in conjunction with Deep Learning methods, how it has contributed to Explainable AI, and how it compares to more recent multidimensional data visualization techniques such T-SNE and UMAP.

Thematic Session
Bank of Portugal

10 April, 14:30 - 14:50, Lecture Theatre E1

When statistics go social: meeting people where they scroll

Ana Castor

Banco de Portugal, acastor@bportugal.pt

Central banking communication has evolved from a tradition of secrecy to one of transparency and active public engagement. To support broader understanding of official statistics, Banco de Portugal launched *Numbers That Matter*, a social media initiative that transforms economic data into short, accessible visual narratives. This paper presents the strategy and impact of the campaign, showing how clear communication strengthens trust in modern central banking.

Keywords: central bank communication, social media, statistical literacy

---

Scientific work creates societal value only when its messages are understood. In many fields, however, complexity is still mistaken for quality, which can confine rigorous outputs to expert circles and limit their real-world impact. Central banking offers an instructive example of this broader science communication challenge, because credibility and trust are integral to policy effectiveness. Evidence from past economic and financial crises suggests a simple mechanism: it is difficult to trust what cannot be understood. And trust is essential for macroeconomic stability and for the transmission of policy through expectations [1]. At the same time, the information environment has shifted towards fast, scroll-based consumption on social platforms. Non-expert audiences rarely follow economic indicators systematically or seek out statistical releases. Communication cannot therefore assume sustained attention, technical vocabulary, or stable interest [1]. To be effective, communication must reduce the costs of acquiring and processing information, by making key messages clear, relevant, and connected to everyday concerns, while preserving methodological rigour. To address these constraints, Banco de Portugal implemented a user-centered communication approach on social media. The flagship initiative is *Numbers That Matter*, launched in 2025 to translate official statistics into short, accessible visual narratives. The initiative complements *BPstat*, the Bank's statistics portal, by adding a public-facing layer designed for discovery and comprehension. The strategy is user-centered and guided by five operational principles: clarity, empathy, timeliness, consistency and accessibility. Content is delivered as carousel posts on Instagram and LinkedIn, as shown in the Figure1.

Figure 1: Examples of social media posts for *Numbers That Matter*

Each post follows a repeatable question–answer template. It opens with a plain-language question, answers it using a small set of carefully selected figures, and ends with a concise takeaway and a link to the source, directing users seeking further detail to *BPstat*. The "one post, one message" rule reduces cognitive load and supports comprehension in scroll-based environments. Topics are selected based on public interest, maximizing relevance while remaining within the statistical mandate. Publishing across Instagram and LinkedIn also reveals evidence of audience heterogeneity. Although these are the Bank's most active social channels, they attract different segments and reward different framings. Posts with an everyday-life angle tend to perform better on Instagram, whereas themes with a clearer professional or economic framing tend to perform better on LinkedIn. Timing is similarly critical. A post on first-home government support measures for people under 35, widely debated in Portugal throughout 2025, became one of the most viewed items on Instagram, illustrating how impact increases when statistical messages coincide with moments of heightened public interest. Performance is monitored through reach and engagement, complemented by stronger indicators of usefulness such as saves and shares. In 2025, the campaign exceeded 300,000 visualizations and appeared among Instagram's featured topics, suggesting that social-native formats can broaden the visibility of official statistics among non-expert audiences and complement traditional dissemination channels. The broader implication extends beyond central banking. In an ecosystem marked by low statistical literacy and pervasive misinformation, simplification should be understood as an inclusion mechanism rather than a loss of rigour. By lowering the cognitive cost of interpretation while preserving traceability to sources and methods, institutions increase the likelihood that accurate evidence competes effectively with misleading narratives. Communication design therefore becomes part of scientific impact and public accountability, not an optional add-on.

References

[1] I. Gameiro, R. Duarte, and I. Abreu. Reaching out to the general public: a challenging journey for central banks. Technical report, Banco de Portugal, Economics and Research Department, 2024.

10 April, 14:50 - 15:10, Lecture Theatre E1

How are Portuguese services traded? Analysing Banco de Portugal's data on international trade in services by mode of supply

Inês Gouveia[1], **Tiago Castro**[1]
[1] Banco de Portugal, igouveia@bportugal.pt, tcastro@bportugal.pt

---

This paper focuses on Banco de Portugal's experience on compiling the new Modes of Supply (MoS) report, a new statistical output which describes how services are supplied internationally and complements core statistics focused solely on value traded. Banco de Portugal's compilation process integrates granular Balance of Payments (BoP) data with Foreign Affiliates Statistics (FATS) data. The paper presents the main MoS results for Portugal, offering fresh insight into the dynamics of trade in services and equipping regulators with a valuable tool for policymaking in increasingly internationalised markets.

Keywords: international trade in services, modes of supply, balance of payments, FATS

The Modes of Supply (MoS) report provides essential information on how services are delivered across borders. These data are becoming increasingly important, as services play a central role in modern economies, yet remain more challenging to measure than trade in goods. Understanding the channels through which services are supplied and consumed internationally allows policymakers to evaluate market-access conditions and design more targeted, effective trade policies.

This paper describes the four modes of supply of services, defined in the WTO General Agreement on Trade in Services (GATS) [1], based on the location of suppliers and consumers: **cross-border supply (mode 1)**, both the supplier and consumer remain in their respective countries (e.g., services supplied via the Internet); **consumption abroad (mode 2)**, the consumer travels to another country to acquire the service (e.g., travel); **commercial presence (mode 3)**, the supplier establishes a local branch or subsidiary in another country to deliver services; and **presence of natural persons (mode 4)**, an individual temporarily moves abroad to supply a service.

Banco de Portugal's statistical process for meeting this new data requirement follows internationally recommended methodological practices [2]. The main sources for compiling MoS data are the granular BoP database, FATS, and inputs from the Portuguese National Statistics Institute (INE).

The methodology for compiling modes 1, 2 and 4 – cross-border supply of services, consumption abroad and presence of natural persons – centred on the services account of the Portuguese BoP. In line with international guidelines, non-commercial services (e.g., those supplied by embassies) were excluded, whereas distribution services (distributive services provided by wholesale and retail industries, classified under goods in the BoP) were added. Furthermore, the value of goods included in service items such as travel and construction was deducted. The resulting services total was then allocated across the three modes of supply according to the Eurostat-WTO model.

The compilation of services supplied through commercial presence abroad (mode 3) relied mainly on FATS data, capturing services provided by branches of domestic firms operating overseas. Inward-FATS[1] constituted the primary source for service imports and outward-FATS[2] for service exports.

Our results showed that, when measured through the modes of supply framework, Portuguese services exports exceed imports. This is particularly pronounced in mode 2, due to the strong prominence of travel exports in the Portuguese BoP. Mode 3 is the only supply mode where imports exceed exports, reflecting Portugal's relative exposure to branches of foreign enterprises operating domestically.

Table 1: Portuguese international trade in services by mode of supply, 2023 (million EUR and share of total)

	Mode 1	**Mode 2**	**Mode 3**	**Mode 4**	**Total**
Exports (million EUR)	21 988	21 374	27 983	3 868	75 213
Exports (%)	29%	28%	37%	5%	100%
Imports (million EUR)	16 073	6 584	42 608	1 910	67 175
Imports (%)	24%	10%	63%	3%	100%

Disclaimer: The analyses, opinions and conclusions presented in this paper are solely those of the authors and do not necessarily reflect the views of Banco de Portugal or the Eurosystem. Any errors or omissions remain the full responsibility of the authors.

References

[1] *GATS: General Agreement on Trade in Services, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1B.* 1994.

[2] Publications Office of the European Union. *European business statistics compilers guide for European statistics on international supply of services by mode of supply - 2023 edition.* European Union manuals and guidelines, 2023.

[1]Data on the domestic operations of subsidiaries, branches or affiliates controlled by foreign investors.

[2]Data on operations of a country's enterprises abroad through subsidiaries, branches or other foreign affiliates.

10 April, 15:10 - 15:30, Lecture Theatre E1

Soaring housing prices: spillovers to financial sector and households' wealth

André Oliveira[1], **Diogo Guerreiro**[1]
[1] Banco de Portugal, ansoliveira@bportugal.pt, dguerreiro@bportugal.pt

Using data from the Central Credit Register (CCR) and the Distribution Wealth Accounts (DWA), this paper examines the impacts of rising housing valuations on housing credit and wealth distribution. We find that recent market dynamics have pushed upwards the median and the upper tail of the new credit distribution. We further assess how these developments have influenced household wealth inequality and conclude that housing valuations are the main factor behind the poorest households' rising share of total wealth in Portugal.

Keywords: housing market, credit market, wealth distribution, Portuguese households

This study assesses how housing price valuations spill over into mortgage credit and the wealth distribution of Portuguese households. Since 2020, total credit granted for house purchases has followed a strong upward trajectory, with only a temporary interruption during the COVID-19 pandemic and the subsequent period of increasing interest rates (2022-2023). Since then, housing credit has expanded considerably, and by 2025Q3 the amount of new credit granted to households was double the observed at the end of 2019 (see Figure 1a).

The rise in housing credit may stem from two effects; a **volume effect**, reflecting an increase in the number of new borrowers; and, a **price effect**, associated with housing valuations. Although both the number of debtors and housings prices have increased, the data clearly show that price growth has far outpaced debtor growth. While both rising prices and higher new credit volumes appear correlated, caution is warranted: new credit is also influenced by factors such as borrowers' income capacity, macroprudential measures, and broader financial conditions.

The effects of housing valuations are clearly visible in the distribution of new housing credit. Since 2019, the median amount of new loans for house purchase has increased from EUR 100,000 EUR to EUR 160,000, a rise of 60%. The central 50% of contracts (interquartile range), has also trended upwards. However, the distance between the first (Q1) and third quartiles (Q3) has widened, and the median has moved closer to Q3, indicating a pronounced shift towards higher-value loans (Figure 1b). This upward skew is significant: in 2019, around 8% of contracts exceeded EUR 300,000, whereas by 2025 this share had risen to 20%.

Using DWA data, we explore how housing valuations affect household wealth. For the poorest 50% of households, housing constitutes nearly 80% of total assets, compared with

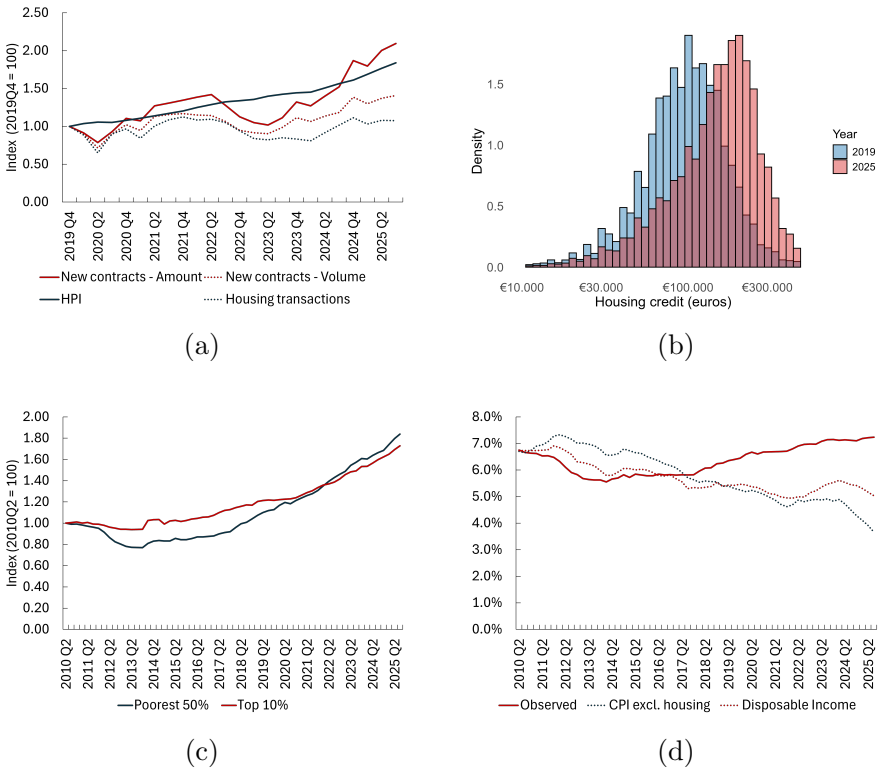

Figure 1: New credit refers to pure new loans for the purchase of main residence, excluding renegotiations and credit transfers. New transactions refer to newly transacted dwellings based on fiscal data. In panel (b) data is truncated below EUR 10,000. Source: Banco Portugal and INE

just 30% for the wealthiest 10%. It is therefore expected that changes in housing valuation disproportionately affect lower-wealth households.

Indeed, total net wealth has risen across all wealth groups, but in recent years the net-wealth growth rate of the poorest households has exceeded that of the richest. This pattern is largely driven by housing, which not only dominates the asset structure of poorer households but has also experienced significant valuation gains (Figure 1c).

To isolate the role of price valuations, we simulate alternative paths for housing prices. If housing prices had grown at the pace of the Consumer Price Index (excluding housing) or disposable income, the poorest households' wealth share would have followed a declining trend (Figure 1d). This suggests that housing valuations have been a crucial factor supporting the relative wealth position of poorer households.

Acknowledgements The authors are grateful for the comments of Sónia Mota, Pedro Pinto, Débora Martins and Cloé Magalhães.

Thematic Session
Statistics Portugal

10 April, 15:30 - 15:50, Lecturer Theatre 1

Use of administrative tax data as an input for compiling the house price index

Ângelo Teixeira[1], **Vitor Mendonça**[2], **Helena Carvalho**[3]
[1] Statistics Portugal, angelo.teixeira@ine.pt, vitor.mendonca@ine.pt, helena.carvalho@ine.pt

---

Tax administrative data, due to its availability, comprehensiveness, accessibility and cost-free nature, constitutes the primary data source to produce price indices based on hedonic methods capable of describing price developments in the national residential real estate market.

Keywords: housing, price index, hedonic regression, administrative data

---

Given the real estate market's economic relevance, as exposed by the recent financial and subprime crisis, there is a need for robust statistical indicators capable of accurately reflecting trends in property prices, particularly for housing. In fact, Europe, and Portugal in particular, has been facing a significant rise in property prices over the recent years. It is therefore particularly important to accurately assess the behavior of the property market. Understanding the dynamics of house price trends requires taking into account the complexity of the 'product' in question, which can be described by a multitude of characteristics, many of which are subjective and dependent on the individual preferences of each owner. Think about the dwelling you currently live in. Is it a flat? A detached house? Is it near the historic city centre? Or, on the contrary, is it in a rural area? Does it have a parking space? Perhaps a swimming pool? Does it show signs of ageing and need some repairs? As you can see, describing a property requires knowledge of a wide range of features.

In this sense, the price of each property sold at any given time will depend on factors such as the size, quality and type of construction, but also the location and other additional amenities, such as the availability of private parking, a lift or a swimming pool, for example. Taking into account the methodology referred to in international manuals on the calculation of this type of indicator [1], as well as that to which the various Member States are subject in the context of implementing the regulations laid down by the European Commission (Eurostat), measuring changes in house prices requires the construction of price indices that take into account the effects of differences in the quality of the dwellings sold. It is necessary to ensure that the price variation observed is the result of actual price changes and not of changes in the characteristics of the dwellings sold between the points in time under analysis. It follows from the above that compiling these indices requires databases containing as many details as possible about dwellings, as well as characteristics of the

transactions carried out. It is in this context that administrative tax databases emerge as a primary data source for the purpose in question, insofar as they cover the entire range of transactions, hold information on a large number of details characterizing the properties and the participants, whilst also being readily accessible.

The House Price Index (HPI) – published by INE since 2014 – is the official statistic that measures the evolution of prices for dwellings purchased by Households in Portugal. This indicator is compiled using data from the Municipal Property Transfer Tax (IMT) and the Municipal Property Tax (IMI), both provided by the Tax and Customs Authority (AT). The payment of IMT, which is a necessary condition for the conclusion of a contract for the transaction of a dwelling, makes it possible to identify, amongst other aspects, the dwelling that was transacted, the timing of the transaction, the buyer(s) and seller(s) and, most importantly, the value of the transaction. On the other hand, the IMI provides a wide range of property details, such as location, floor area, age of the building, type of construction, and whether a garage, swimming pool or lift is available, for example. By matching the two data sources, using an identifier for each property present in both sources, it is possible to build a database containing information on the transaction, but above all on the characteristics of the properties transacted at any given time.

The methodology used to compile the HPI [2] is based on estimating a functional relationship between the (natural) logarithm of dwelling transaction prices and their characteristics, falling within the scope of probabilistic 'hedonic price' models. This estimation is carried out quarterly using the (adjacent) time-dummy method, with data from two consecutive quarters, for the entire set of transactions; this makes it possible to account for qualitative differences in the dwellings traded and to estimate a rate of price change adjusted for changes in quality.

The construction of the HPI, with the technical characteristics outlined, clearly demonstrates the crucial role that available administrative data can play in the production and dissemination of official statistics. In this context, the systematic and consistent use of such data made it possible to develop a historical series, starting in 2009, capable of describing the evolution of dwelling sales prices, by category, as well as quantifying the scale of the residential market through the recording of both the number and value of transactions. This outcome reinforces the relevance and potential of administrative data to support robust statistical analysis and to enhance the quality, usefulness, and continuity of the information made available to the public.

References

[1] Eurostat. Residential Property Price Indices Handbook. Technical report, Eurostat, 2013. `https://ec.europa.eu/eurostat/`.

[2] Statistics Portugal. Índice de Preços da Habitação - Documento Metodológico. Technical report, Statistics Portugal, 2022.

10 April, 15:50 - 16:10, Lecturer Theatre 1

From official statistics to scientific knowledge: the potential of secure access to microdata

José A. Pinto Martins

Statistics Portugal, pinto.martins@ine.pt

This paper analyses how secure access to official microdata supports the transformation of data into scientific knowledge, presenting the access framework implemented by Statistics Portugal and evidencing its growing impact on research activity.

Keywords: official statistics, microdata, safe centre, confidentiality, scientific research

---

Official statistics play a central role as a public good, providing reliable, impartial and accessible information that supports decision-making, transparency and democratic participation. In a context marked by rapid digital transformation, increasing data availability and growing demand for evidence-based analysis, National Statistical Offices face the challenge of enhancing both the accessibility and usability of statistical information, while safeguarding confidentiality.

Within this framework, access to microdata has become a key enabler of scientific research. The increasing complexity of economic and social phenomena, combined with the need for granular and multidimensional analysis, requires access to detailed datasets that allow researchers to explore behavioural patterns, structural relationships and emerging trends beyond aggregated indicators.

This paper presents the model adopted by Statistics Portugal (SP) for providing access to microdata for scientific purposes. This model is grounded in a balanced approach that combines openness and accessibility with strict legal, ethical and technical safeguards. Different access modalities are considered, including public use files (PUFs), which provide fully anonymised datasets for general use, and more restricted forms of access for accredited researchers. The accreditation process, conducted through national systems aligned with European statistical regulations, ensures that access is granted exclusively to qualified users and for clearly defined research purposes.

A central component of this framework is the secure access infrastructure (safe centre), designed to enable the use of highly detailed microdata under controlled conditions. The new SP Safe Centre represents a significant investment in technological capacity and institutional capability. It provides a dedicated environment equipped with high-performance computing resources, specialised analytical software and strict operational protocols, including physical access control, secure data environments and rigorous output validation

procedures. This ensures that confidentiality is fully preserved while maximising the analytical potential of the data.

The paper also presents empirical evidence on the use of microdata for research purposes, highlighting a growing number of active projects, an expanding community of accredited researchers and an increasing volume of datasets being used. In particular, the analysis shows a tendency towards the combined use of multiple data sources within the same project, reflecting the need for integrated and multidimensional approaches to address complex research questions.

These developments illustrate the evolving role of official statistics in the data ecosystem, moving from the provision of aggregated outputs towards enabling data-driven knowledge production. Secure access to microdata emerges as a strategic instrument to strengthen collaboration between statistical institutions and the scientific community, foster innovation and enhance the societal value of official statistics.

10 April, 16:10 - 16:30, Lecturer Theatre 1

Seasons in the algorithm: error-driven insights into Winter and Spring crops classification - an exploratory study by Statistics Portugal

Cristina Gabriel[1], **Isabel Gonçalves**[1]
[1] Statistics Portugal, cristina.gabriel@ine.pt, isabel.goncalves@ine.pt

---

Statistics Portugal (INE) is enhancing its Spatial Data Infrastructure (SDI) by integrating Artificial Intelligence (AI) technologies to address emerging challenges in statistical production. To this end, INE has ben participating in the Eurostat grant "AIML4OS – Artificial Intelligence and Machine Learning for Official Statistics", specifically in Work Package 7 (WP7), which focuses on the use of Earth Observation (EO) data.

Keywords: artificial intelligence, official statistics, earth observation, crop types

---

The project aims to develop methodological and implementation guidelines for generalising research findings in official statistics. This initiative supports the transition from experimentation to operational production in AI and Machine Learning (ML) projects, ensuring that EO-based solutions provide valid and comparable results across different countries and time periods.

As part of the project, this study presents a comprehensive error analysis of a crop type model (CTM). The original implementation distinguishes multiple crop types—such as winter crops, spring crops, and rapeseed (the latter not applicable to the Portuguese context)—using Sentinel-1 radar time series combined with advanced object-based classification software. This study validates the model's performance in the Oeste e Vale do Tejo (PT11D) NUT2 region of Portugal, using the 2025 Land Parcel Identification System (LPIS) as the ground-truth reference. The methodology involved cross-referencing data from nearly 84,000 LPIS parcels with the CTM, focusing specifically on 79,600 hectares of arable crops, of which 56% correspond to winter and spring crops.

Validation results indicate that, although the algorithm identified 91,500 hectares, only 74% of this classified area overlaps with LPIS arable crop data, and 93% of that overlapping area is specifically classified as winter and spring crops. Within this intersection, the algorithm achieved a 96% match for spring crops and a 68% match for winter crops, resulting in an Overall Accuracy (OA) of 89% for distinguishing between seasonal cycles.

Comparative analysis shows that, while the algorithm is highly effective at identifying the spatial footprint of spring cycles, it tends to over identify seasonal cycles in areas not registered as temporary arable land in the LPIS. Moreover, a closer examination of the

CTM's performance for specific crops reveals that, although the algorithm identifies more than 90% of the area dedicated to maize, tomato and rice, these same crops also account for the largest share of the area missed by the classification (omission errors).

This research concludes that CTM approaches show strong potential for identifying seasonal crop cycles. The project is still under active development, and these findings reflect the model's current state. The implementation of robust spatial masks and refined segmentation is crucial. These steps are essential to ensure that EO-based solutions meet the rigorous quality standards required for the production of official statistics.

Thematic Session
CLAD 2026 Scholarships

10 April, 16:50 - 17:10, Lecture Theatre E1

A smooth approximation for interval Fisher's discriminant analysis

Diogo Vaz[1], **M. Rosário Oliveira**[1,2]
[1] Instituto Superior Técnico, Universidade de Lisboa, Portugal, diogovaz318@gmail.com
[2] CEMAT rosario.oliveira@tecnico.ulisboa.pt

In this work, we propose a smooth approximation of the Interval Fisher's Discriminant Analysis objective function based on a differentiable surrogate of the absolute value, leading to a smooth Fisher ratio with explicit gradients. The proposed formulation enables stable optimization while preserving the discriminant structure. Theoretical results establish approximation bounds and convergence of the solutions, and empirical results on real and simulated data confirm the effectiveness of the approach.

Keywords: symbolic data analysis, interval-valued data, multigroup classification, Mallows distance, absolute value function

---

Interval-valued data are a fundamental component of Symbolic Data Analysis (SDA), where observations are described by intervals rather than single values, allowing both location and variability to be captured [2]. A common representation expresses each interval through its centre and range, and distances between such objects are often defined using the Mallows distance, which incorporates both components [3].

Within this framework, Interval Fisher's Discriminant Analysis (IFDA) extends conventional approach to interval-valued data by maximizing a ratio of between-class to within-class variability [4]. A key difficulty in IFDA arises from Moore's projection of intervals, which introduces componentwise absolute values in the optimization problem. As a result, the objective function is not smooth and cannot be reduced to a standard generalized eigenvalue problem, which complicates both theoretical analysis and numerical optimization.

To address this limitation, we propose replacing the absolute value with a smooth approximation of the form $g_\mu(x) = x \tanh(x/\mu)$, applied componentwise [1]. This leads to a smooth version of the IFDA objective function, which retains the original structure while becoming differentiable everywhere. The resulting formulation enables the use of efficient gradient-based optimization methods and avoids the combinatorial complexity associated with sign configurations in the original problem.

We establish theoretical error bounds for the proposed approach. In particular, the smooth objective uniformly approximates the original IFDA objective, with an approximation error proportional to the smoothing parameter. Furthermore, under reasonable assumptions, the maximizers of the smooth problem converge to the true IFDA discriminant directions as the

smoothing parameter tends to zero. Convergence rates can also be derived under standard growth assumptions, ensuring that the smooth solutions remain close to the original ones. The empirical performance of the method is evaluated on both real and simulated datasets. The results of a real interval-valued temperature dataset show that, for sufficiently small smoothing parameters, the smooth IFDA achieves classification performance identical to the original formulation. A comprehensive simulation study also demonstrates robustness at different levels of class separation, class imbalance, and microdata distributions. In all scenarios, the discriminant directions obtained from the smooth formulation closely match those of the original IFDA.

In addition to preserving classification performance, the proposed approach offers significant computational advantages. The original IFDA requires evaluating all possible sign configurations, leading to exponential complexity in the number of variables. However, the smooth formulation replaces this combinatorial problem with a continuous optimization task, resulting in substantially reduced computation times and improved scalability in higher-dimensional settings.

Overall, the proposed smooth approximation provides an alternative to IFDA, maintaining its discriminant power while enabling different and more efficient optimization techniques.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia, Portugal, through the projects [UID/04621/2025

References

[1] Y. J. Bagul. A smooth transcendental approximation to $|x|$. *International Journal of Mathematical Sciences and Engineering Applications*, 11:213 – 217, 2017.

[2] P. Brito. Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4):281–295, 2014.

[3] M. R. Oliveira, D. Pinheiro, and L. Oliveira. Location and association measures for interval-valued data based on Mallows' distance. https://arxiv.org/abs/2407.05105, 2025.

[4] D. Pinheiro, M. R. Oliveira, I. Kravchenko, and L. Oliveira. Interval Fisher's discriminant analysis and visualisation. https://arxiv.org/abs/2512.11945, 2025.

10 April, 17:10 - 17:30, Lecture Theatre E1

PerPCA applied on air pollution data

Diana Angélica Vázquez-Limón[1], **Adelaide Freitas**[1,2]
[1] Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal.
dangelica.vazquezl@ua.pt
[2] Center for Research and Development in Mathematics and Applications (CIDMA),
adelaide@ua.pt

The study of environment is relevant to humanity as we can see a rapid change due to the trends in consumption. These changes come with more availability of data with possible high-dimensionality and heterogeneous data sources. The aim of this work is to present a recently proposed statistical technique (Personalized Principal Component Analysis) that tackles these challenges and that could be explored for analysis of air pollutants, their characterization and visualization.

Keywords: dimensionality reduction, data visualization, air pollution, principal component analysis

---

The increasingly rapid urban growth from the past decades has involved an increment in anthropogenic activities. Many of these activities lead to air pollution which refers to the accumulation of substances in the atmosphere; when the concentration exceed certain levels, it endangers the environment, ecosystems, and also presents health risks. There is growing interest in understanding air pollutants, studying their composition as well as its effects on health. The study of air pollutants for their characterization is in general hard and time consuming as it is needed to consider many factors relevant for interpretation. The objective of this work is to demonstrate the potential of a recent statistical technique, called PerPCA [3], to provide insights into the composition of air pollutants.

A statistical technique useful for visualizing and identifying patterns in data is Principal Component Analysis (PCA). We could obtain a characterization of the chemical composition of particulate matter (PM) by using a representation of them in a reduced space obtained with PCA. One limitation of PCA is that it assumes that data comes from homogeneous sources Therefore, when samples are collected from different areas to obtain a general characterization of PM, PCA may capture variance that is not relevant to the underlying phenomenon of interest. Personalized PCA (PerPCA) is a novel approach that deals with this limitation by explicitly accounting for data heterogeneity, enabling the analysis of datasets originating from multiple, heterogeneous sources.

In this work we apply PerPCA to a data set with annual historical anthropogenic chemically reactive gases emissions (from the Community Emissions Data System (CEDS) [2]) to construct Global PCs which capture variance common for all the sources (countries recorded)

and Local PCs for the variance remaining to be explained in each of the sources.We used the data from 19 countries, each treated as a distinct source. For each country, ten emission-related variables were collected, with annual records spanning 30 years. The iterative algorithm used to execute PerPCA was implemented in Python by its authors in [3] , including the necessary preprocessing steps to conform the data to the structure required for this study.

After applying PerPCA to the data, we obtained a Global component which captures the variability for all the countries and a Local Component to captures part of the remaining variability at each country. The proportion of variance captured for each of the countries ia high (total explained variance above 75%) with most countries' variability captured by the first Global PC. We propose a visualization for the Global and Local Principal Components. The idea is to have a common ground to compare the countries' emissions, without loosing from our sight the original variables of interest (species of emissions). We relate the loadings for the first Global PC with the loadings for the first Local PC obtaining columns of markers, representing the value of a specie of emission for each country with respect to the Global and Local PCs. Namely, the variable with the highest absolute value in the first Global PC correspond to the one with most influence in the construction of the that Global PC. Moreover, for each variable we can also identify in the other axis its loadings for the the first Local PC corresponding to each of the studied countries, allowing the identification of countries with variable as influential globally and locally.

Using PerPCA, we demonstrated a methodology capable of capturing and disentangling shared and source-specific characteristics within heterogeneous datasets. In a single run, it revealed the emission species with the highest variability and highlighted country-specific patterns through the Local PCs. These findings provide insights relevant to air pollution regulation, while future work may explore alternative data perspectives, such as different spatial or temporal scales.

Acknowledgements This work is supported by CIDMA (https://ror.org/05pm2mw36) under the Portuguese Foundation for Science and Technology (FCT, urlhttps://ror.org/00snfqn58), Grants UID/04106/2025 (https://doi.org/10.54499/UID/04106/2025) and UID/PRR/04106/2025 (https://doi.org/10.54499/UID/PRR/04106/2025).

References

[1] C. A. Alves. Characterisation of solvent extractable organic constituents in atmospheric particulate matter: an overview. *Anais da Academia Brasileira de Ciências*, 80:21–82, 2008.

[2] R. M. Hoesly, S. J. Smith, L. Feng, K. Zbigniew, G. Janssens-Maenhout, T. Pitkanen, J. J. Seibert, L. Vu, R. J Andres, R. M. Bolt, et al. Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds). *Geoscientific Model Development*, 11(1):369–408, 2018.

[3] N. Shi and R. Al Kontar. Personalized PCA: Decoupling shared and unique features. *Journal of machine learning research*, 25(41):1–82, 2024.

Thematic Session
CLAD Corporate

11 April, 14:30 - 15:00, Lecture Theatre E1

Turning data into decisions: real cases from PSE

Nuno Gomes[1], **Rui Almeida**[2]
[1] Head of Data Science at PSE, nuno.gomes@pse.pt
[2] Data Scientist at PSE, rui.almeida@pse.pt

PSE is a Portuguese company specialized in consulting, analytics projects, and predictive and artificial intelligence solutions. With experience in predictive analytics since 1994, PSE supports organizations across multiple industries in transforming data into valuable knowledge for better decision-making. With a tailored approach, the company develops customized solutions for each client. This presentation provides a brief introduction to PSE and highlights the importance of data science through two real success cases from very different business sectors.

Keywords: predictive analytics, artificial intelligence, data science, decision-making, business value

---

PSE is a national company specializing in consulting, developing analytical projects, and implementing predictive solutions for organizations. With experience in predictive analytics since 1994, PSE's consulting team is the ideal partner to deliver successful, high-impact projects. We have extensive expertise across multiple industries — including telecommunications, banking, insurance, retail, distribution, energy, consumer goods, government, and healthcare — and across diverse functional areas such as marketing, risk, operations, logistics, quality, and finance. Our team excels at integrating predictive intelligence into organizational processes, turning data into actionable insights.
In today's data-driven world, leading companies know how to leverage information as a competitive advantage. Data Science enables the analysis of large datasets, the identification of patterns, the anticipation of trends, and faster, more accurate decision-making. Organizations that harness these capabilities can optimize processes, enhance customer experience, reduce risks, and continuously innovate, staying ahead in an increasingly dynamic and competitive market. In our presentation, we will highlight two examples from very different industries, where analytical solutions delivered substantial financial impact:

1. Distribution Sector: A client distributing newspapers and magazines to around 10,000 points of sale faced highly variable demand influenced by location, timing, seasonality, publication type, and cover. The main challenge was maximizing sales while minimizing returns and stockouts.

2. Textile Sector: A client producing over 300 types of fabrics needed to identify the key factors causing defects across the production chain. The objective was to optimize production workflows and reduce defects for each fabric type, improving overall efficiency and quality.

Acknowledgements It is with great pride and pleasure that we announce our participation in JOCLAD 2026, further strengthening our long-standing partnership.

11 April, 15:00 - 15:30, Lecture Theatre E1

Data analysis and processing in smart cities

Carla Oliveira[1], **Nuno Correia**[2], **Laura Guijarro**[3]
[1] Kyndryl IoT Technological Center Leader, carla.oliveira@kyndryl.com
[2] Kyndryl Network Edge Leader, nuno.correia1@kyndryl.com
[3] Kyndryl Network Edge Offering Manager, Laura.Guijarro.Iguacel@kyndryl.com

---

Factors like global climate change, population growth or complex demand for public services are pressuring and challenging cities. This paper exploits how smart cities technology- built on real-time continuous monitoring through IoT devices and advanced analytic data platforms- is helping cities move to predictive and optimized governance. Kyndryl smart cities rollouts have delivered tangible benefits, including energy savings of up to 74% in public lighting, 50% reduction in energy consumption across building management systems or the reuse of more than 1.5B liters.

Keywords: smart cities, data analysis, IoT, urban intelligence, sustainability

---

Traditional urban management often relies on isolated systems (silos) with limited visibility of real time conditions which are not compatible with the increasingly complex and dynamic conditions of modern cities. The rollout of large scale sensing, correlated data and advanced analytics change this model by giving cities continuous visibility over environmental, mobility, or infrastructure data, for example. These data improve overall urban operations, support governance and daily operational decisions that allow to reduce and optimize costs, emissions, resources, or improve anticipation and resilience to different risks.

The transformation towards smart cities relies on several key technological components working in a layered, secure architecture. Smart city infrastructures rely on IoT devices that capture information on air quality, energy usage, traffic, noise, and water systems, among others. Connected networks such as LoRaWAN, private 5G, fiber, and municipal wireless links ensure that these data streams remain reliable and energy efficient. Urban intelligence platforms collect and normalize data from different domains, enabling cross areas correlation and improving transparency. Machine learning and advanced analytics extract patterns, detect anomalies, and generate predictions, allowing cities to be proactive, anticipating issues and adjust operations (rather than reacting after problems emerge):

- In smart traffic management, real-time sensor data allows traffic signals to dynamically adjust to congestion patterns, with benefit such as reducing idle time, decreasing fuel consumption, or lowering carbon emissions.

- Smart energy grids increasingly rely on predictive analytics to manage energy demand and integrate renewable generation sources, improving grid reliability while reducing waste.
- Predictive urban planning uses advanced urban simulation platforms and digital twins to evaluate the impact of development plans before implementation, simulating traffic patterns, pollution levels, and energy consumption under different scenarios so that urban data becomes a strategic resource enabling evidence-based decisions.
- AI in an urban incident?management platform to streamline how cities handle reports by automatically classifying citizen requests, identifying duplicates, and routing each issue to the right department. It can assess severity to set priorities, analyze images to confirm occurrences on?site, and forecast where incidents are likely to occur based on historical trends.

The benefits of these capabilities are already being demonstrated in real deployments, like the ones implemented by Kyndryl. Data driven lighting systems combining LED technology with centralized management have achieved reductions in energy use of more than 70%, improving lighting quality and lowering operational costs with a faster response to incidents. Building management analytics has reduced energy consumption by over 50%, and advanced monitoring has enabled the reuse of more than 1.5 billion liters of water. Cities deploying these solutions operate thousands of connected devices to monitor conditions, detect failures, and respond more rapidly to incidents.
For instance, a Kyndryl deployment at Fundão illustrates how integrated data processing improves urban management across domains. A citywide LoRaWAN dedicated network supports more than 42,000 connected sensors that track air quality, weather conditions, noise levels, and mobility indicators. Dashboards with georeferenced data, anomaly detection, and automated alerting strengthen cross-areas relations and support faster decision making. Mobility monitoring using video analytics allows authorities to understand traffic patterns, count vehicles and people, and reduce congestion, contributing to lower emissions or prevent incidents. These analytical tools transform raw sensor data into operational insights that support planning, public safety, and sustainability goals.
As cities continue to evolve, future developments will depend on enhanced analytical capabilities. Private 5G networks will expand massive IoT connectivity; digital twins will allow real time simulation of operational scenarios; and predictive models will help anticipate infrastructure failures, environmental risks, or service demands. Urban intelligence platforms will consolidate data from multiple domains to support climate monitoring, strategic planning, and new citizen services. Through continuous analysis of real time data, cities will strengthen their ability to manage resources efficiently, adapt to environmental challenges, and support sustainable urban development.

Thematic Session
SPE

11 April, 10:30 - 10:50, Lecture Theatre E1

Extremes without independence: the other side of the history

M. Cristina Miranda

ISCA, CIDMA, University of Aveiro, cristina.miranda@ua.pt

This talk focuses on estimating the Extremal Index (EI), an additional parameter in Extreme Value Theory (EVT) that characterizes the clustering of exceedances, when we have stationary dependent sequences. Two real-world data sets illustrate the performance of the P-estimator using a train/test split sample methodology.

Keywords: extremal index, extreme value theory, robustness

How long will it last, the heavier rainy days, the low prices on stock markets? For how many days will we have energy consumption peaks? How long will the fisherman be stacked at land due to high waves of the sea? These are examples of situations where extreme values appear in clusters. The magnitude of such clusters is related to an extreme value distribution parameter, the extremal index. Under suitable local dependence conditions, extreme value theory may be applied to stationary sequences, turning possible to obtain the maxima limit distribution [4]. This extreme value distribution maintains the shape parameter, when compared to the one of the associated independent and identically distributed sequence, but is affected in its local and scale values.

The extremal index is a parameter that measures how much extreme events cluster in a dependent time series. When data are independent or asymptotically independent, extremes are typically isolated and the extremal index is one. With stronger dependence, extremes tend to appear in clusters, and the extremal index decreases toward zero as the clusters size increases. A common interpretation of the extremal index is the inverse of the average cluster size [1].

This presentation focuses on estimating the extremal index. Many existing estimators mainly differ in how they define and detect clusters of exceedances over a high threshold (e.g., runs, blocks, intervals, K-gaps). The authors highlight a recent approach, inspired by the intervals estimator [3], called the Proportion estimator (P-estimator) [5]. It uses a simple idea: look at the times between exceedances and estimate the extremal index as the proportion of "non-immediate" gaps (*i.e.*, gaps that indicate exceedances are separated rather than consecutive). Because it is based on a sample proportion, it is easy to compute and interpret.

A key contribution of the recently proposed estimator is robustness. The authors argue the P-estimator has attractive robustness properties (bounded impact of individual observations and resistance to contamination in certain models) [2].

The method is illustrated and its performance is measured using a train/test split, and comparing extremal index estimators across thresholds with the Predicted Root Mean Square Error (PRMSE). Two sets of real data are used: daily close price data from BTC, since January 2015 to March 2025 and Portuguese photovoltaic energy data (hourly solar generation potential, 1986–2015). In both applications, the P-estimator tracks the test-based reference closely and performs similarly to the runs estimator; the K-gaps estimator also shows good performance.
The authors conclude that, given its simplicity and robustness, the P-estimator is well suited for practical applications where the cluster structure is unknown, and they stress the importance of reporting confidence intervals because different EI estimators can produce noticeably different values.

Acknowledgements This work is supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58) Multi-Annual Financing Program for *R&D* Units, grants UID/4106/2025 and UID/PRR/4106/2025.

References

[1] J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, 2004.

[2] E. Cantoni and E. Ronchetti. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics*, 25:198–213, 2006.

[3] C. A. T. Ferro and J. Segers. Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B*, 65(2):545–556, 2003.

[4] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1983.

[5] M. Cristina Miranda, Manuela Souto de Miranda, and M. Ivette Gomes. SPECIAL ISSUE LinStat: a new proposal for robust estimation of the extremal index. *Journal of Statistical Computation and Simulation*, 95(5):949–966, 2025.

11 April, 10:50 - 11:10, Lecture Theatre E1

A branching process approach on population survival: skeleton process and stochastic introgression

Maria Conceição Serra

Center of Mathematics, University of Minho, mcserra@math.uminho.pt

---

We study how populations seemingly doomed to extinction can survive via advantageous individuals that may appear through adaptive mutations or through stochastic genetic introgression. Using multitype branching processes: i) we characterize the skeleton process that describes typical survival paths of near-critical Galton-Watson branching process; ii) we derive the introgression hazard rate, i.e., the per-time unit probability of an introgression event given it has not yet occurred.

Keywords: branching processes, multitype, near-criticality, skeleton process, stochastic introgression

---

The purpose of this work is to study how populations that, in principle, are doomed to extinction manage to survive. This is usually achieved through the appearance of certain types of individuals that give some "advantage". For example:

i) viruses placed in new and hostile environments often develop mutations which are better adapted to the new environments;

ii) genetic introgression (the permanent incorporation of genes of one population into the genome of another) has a stochastic nature and is often achieved through a sequence of advantageous changes.

In this talk we will describe how multitype branching processes can be used to model the evolution of such populations and provide answers to some relevant questions arising in them. In particular, for the examples described above:

i) We characterize the *skeleton process*, conditioned on the appearance of mutations. The so-called *skeleton process* is also a branching process that describes typical survival scenarios for a near-critical Galton-Watson branching process and it was first described in [3] for the slightly supercritical case. Our main contribution, in [4], is the extension to the critical or slightly subcritical populations, where non-extinction can occur due to the appearance of mutants.

ii) We derive the *introgression hazard rate*, i.e., the probability, per time unit, that an introgression event takes place given that it has not happened before. This is a quantitative measure of introgression risk that takes the stochasticity of several elements into account and, in [1] and [2], we present a methodology to calculate such hazard rates in different situations.

Acknowledgements This research of Maria Conceição Serra was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013/2025 (https://doi.org/10.54499/UID/00013/2025).

References

[1] A. Ghosh, M. C. Serra, and P. Haccou. Quantifying stochastic introgression processes in random environments with hazard rates. *Theoretical Population Biology*, 100:1–5, 2015.

[2] P. Haccou and M. C. Serra. Establishment versus population growth in spatio-temporally varying environments. *Proceedings of the Royal Society B: Biological Sciences*, 288:20202009, 2021.

[3] N. O'Connel. Yule process approximation for the skeleton of a branching process. *Journal of Applied Probability*, 30:725–729, 1993.

[4] S. Sagitov and M. C. Serra. Skeletons of near-critical Bienaymé-Galton-Watson branching processes. *Advances in Applied Probability*, 47(2):530–544, 2015.

11 April, 11:10 - 11:30, Lecture Theatre E1

From multi-omics to prediction: the Priority-Elastic Net framework

Laila Musib[1], **Helena Mouriño**[1], **Eunice Carrasquinha**[1]
[1] Faculdade de Ciências da Universidade de Lisboa and CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal,
eitrigueirao@ciencias.ulisboa.pt

Priority-Elastic Net is a structured regularisation method for integrating high-dimensional multi-block data such as multi-omics profiles. It extends Priority-Lasso with an elastic-net penalty, enabling variable selection while accounting for within-block correlations and a predefined data-source hierarchy. Implemented in the R package `prioritypelasticnet`, the method is illustrated through a binary classification application to glioma data, demonstrating interpretability and predictive stability.

Keywords: multi-omics data, hierarchical blocks, regularization techniques, priority-elastic net

---

The increasing availability of multi-omics data in biomedical research poses significant statistical challenges, particularly due to the high dimensionality of the data, strong correlations among variables, and the presence of multiple heterogeneous data blocks. Classical regression and machine learning methods often struggle in this setting, leading to unstable models and limited interpretability.

Priority-Elastic Net [3] was proposed to address these challenges by combining structured modelling with penalised regression. The method builds upon the Priority-Lasso framework [1], which introduces the idea of fitting data blocks sequentially according to a predefined order of importance. This hierarchy reflects prior biological or clinical knowledge and is incorporated directly into the modelling process. Priority-Elastic Net extends this framework by using an elastic-net penalty within each block, thereby balancing sparsity and grouping effects when predictors are highly correlated.

The modelling procedure fits each block sequentially, including the linear predictor from previous blocks as an offset in subsequent models. This strategy ensures that information extracted from higher-priority blocks is preserved while still allowing additional blocks to contribute to the prediction. As a result, the final model is more interpretable and often more stable than approaches that treat all variables equally.

The methodology is fully implemented in the R package `priorityelasticnet`, available on CRAN [2]. The package offers a complete workflow, including model estimation, regularisation parameter selection via cross-validation, and graphical tools for inspecting selected variables and block-wise contributions. It supports several types of outcomes, including continuous, survival, and binary responses.

In this work, we focus on a binary outcome application to glioma data, demonstrating how Priority-Elastic Net can integrate multiple omics layers while enforcing biologically meaningful priorities. The example illustrates the practical advantages of the method in terms of prediction performance, stability, and interpretability, making it a valuable tool for applied statisticians and data scientists working with complex biomedical data.

Acknowledgements This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under CEAUL Research Unit, UID/00006/2025, DOI: https://doi.org/10.54499/UID/00006/2025.

References

[1] S. Klau, V. Jurinovic, R. Hornung, T. Herold, and A.-L. Boulesteix. Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19(322), 2018.

[2] L. Musib, E. Carrasquinha, and H. Mouriño. `priorityelasticnet`: Comprehensive analysis of multi-omics data using an offset-based method. R package version 1.x, CRAN. 2025.

[3] L. Musib, R. Coletti, M.B. Lopes, H. Mouriño, and E. Carrasquinha. Priority-elastic net for binary disease outcome prediction based on multi-omics data. *BioData Mining*, 17(45), 2024.

Contributed Sessions

9 April, 16:00 - 16:20, Auditorium ESTGD

Generalized model-based approach for count time series clustering

Luís Sousa[1,2], **Isabel Pereira**[1,2], **Magda Monteiro**[1,3], **Dimitris Karlis**[4]
[1] CIDMA - Center for Research and Development in Mathematics and Applications, luissousa2@ua.pt
[2] Department of Mathematics - University of Aveiro, isabel.pereira@ua.pt
[3] ESTGA - Águeda School of Technology and Management, University of Aveiro, msvm@ua.pt
[4] Department of Statistics, Athens University of Economics and Business, Athens, Greece, karlis@aueb.gr

---

The main goal of this study is to group similar time series, with emphasis on the clustering of discrete-valued time series with zero inflation and differing levels of dispersion. This research adopts a model-based clustering approach using INAR(p) finite mixture models. It includes an expectation-maximization algorithm for parameter estimation, as well as two methods for selecting a representative series for the final clusters. A simulation study was conducted across three increasingly challenging scenarios and a practical case is presented.

Keywords: time series, model-based clustering, finite mixture, zero-inflation, INAR models

---

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. We assume that each cluster of time series comes from an INAR(p) process, such that [1]

$$X_t = \sum_{j=0}^{p} \alpha_j * X_{t-j} + \epsilon_t \tag{1}$$

where each $\alpha_j \in [0,1]$, $\sum_{j=0}^{p} \alpha_j < 1$, the innovations component, ϵ_t, is a sequence of any chosen i.i.d. distribution (for example, Poisson, Negative Binomial or Zero-Inflated Poisson [2]), and $*$ is the thinning operator, which may be any chosen distribution (for example, binomial or negative binomial). The idea is to have a generalized model which allows for a flexible class of models. The conditional distribution is assumed to be a convolution between the p chosen distributions for the thinning operation and the innovations' chosen distribution [3].
The data can be seen as a finite-mixture model, such that [4]

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

where π_g refers to the mixing proportion of each process and θ_g represents the group-specific parameters.
To implement the EM algorithm, initial parameter values are required. A k-means clustering is first applied to obtain an initial partition of the time series, where k is the number of initial clusters. The initial mixing proportions π_g are computed as the proportion of series assigned to each cluster. The parameters α_g are estimated by solving the Yule–Walker equations for each series, while the remaining distribution-specific parameters are obtained via the method of moments using these estimates. The initial model parameters are then defined as the averages of the corresponding estimates across all time series.
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 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 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 obtatin updated parameter estimates. Finally, the last step of the algorithm is to select the combination of INAR processes which are most fit for the data. Then, when applying the EM algorithm to all of the orders and model combinations, one would choose the combination which had the best performance measures (BIC and ICL).
Finally, two methods are presented for choosing the representative series of the final clusters. One which minimizes the euclidean distance of each series to the mean profile of the cluster, and another which chooses the series with the maximum value of the final $\hat{z}_{ig}$ as the representative of cluster g.

Acknowledgements. This study was partially supported by *NEXUS: Pacto de Inovação – Transição Verde e Digital para Transportes* (L-00000059 — Project no. 53), financed by PRR - Plano de Recuperação e Resiliência, with NextGenerationEU funds, at the University of Aveiro, and by CIDMA (`https://ror.org/05pm2mw36`), under the Portuguese Foundation for Science and Technology (FCT, `https://ror.org/00snfqn58`), through grants UID/04106/2025 (`https://doi.org/10.54499/UID/04106/2025`) and UID/PRR/04106/2025 (`https://doi.org/10.54499/UID/PRR/04106/2025`).

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. D. Lai. First-order integer valued ar processes with zero-inflated poisson innovations. *Journal of Time Series Analysis*, 33:954–963, 2012.

[3] D. Karlis, A. Chutoo, N. Mamode Khan, and V. Jowaheer. The multilateral spatial integer-valued process of order 1. *Statistica Neerlandica*, 78(1):4–24, 2024.

[4] T. Roick, D. Karlis, and P.D. McNicholas. Clustering discrete-valued time series. *Advances in Data Analysis and Classification*, 15:209–229, 2021.

9 April, 16:20 - 16:40, Auditorium ESTGD

Statistical modelling and time series clustering of retail imports and distance sales in Europe

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

---

This study examines retail import dynamics and distance sales in the European Union using Eurostat data. Intra-EU imports (2014–2023) are analysed through time series clustering, identifying heterogeneous growth patterns across country groups. Quarterly distance sales (2013–2023) are modelled using regression and Holt–Winters methods, with exponential smoothing delivering stronger forecasts. The findings highlight the value of advanced analytics to support retail planning and supply chain resilience in Europe.

Keywords: clustering, Holt-Winters method, linear models, retail imports, time series analysis

---

The retail sector plays a central role in the European economy, both as a driver of domestic consumption and as a key node in international supply chains. Understanding the dynamics of retail imports and distance selling is therefore essential for evidence-based decision-making, forecasting, and policy design. This study develops a statistical framework for analysing retail import patterns in the European Union (EU) and distance sales dynamics in the European Economic Area (EEA), using harmonised data from Eurostat, [1]

The analysis is conducted in two complementary stages. First, annual intra-EU retail import values (excluding motor vehicles and motorcycles) for 24 EU countries over the period 2014–2023 are examined. Intra-EU imports are shown to dominate retail trade across most countries, with average shares frequently exceeding 75%. To account for heterogeneity across countries, a hierarchical time series clustering approach based on Euclidean distance and Ward's linkage is applied. The resulting clusters reveal distinct import trajectories, with Germany and France emerging as single-country clusters due to their large import volumes and markedly different growth rates. Beyond differences in growth rates, the identified clusters reflect structural disparities in scale, volatility and post-2014 trajectories, with large economies exhibiting distinct import intensities and persistence patterns compared to more homogeneous country groups. Building on this segmentation, group-level linear mixed models are estimated to capture common trends while allowing for country-specific variation and temporal correlation.

Since the magnitude of retail import values differs substantially across groups, the logarithm of import values is considered and a fixed-effects linear mixed model is applied in

order to satisfy the assumptions associated with these models, particularly the normality of the error terms. The model can be written as

$$W_{it} = \beta_{01} + \beta_{11}\, t + \sum_{g=2}^{5} \text{Group}_{ig}\, (\beta_{0g} + \beta_{1g}\, t) + \varepsilon_{it}, \tag{1}$$

where, for each country $i = 1, \ldots, 24$, W_{it} denotes the logarithm of the intra-EU retail import value of country i in year t, with $t = 1, \ldots, 10$. The variable Group_{ig} is a dummy indicator that takes the value 1 if country i belongs to group g and 0 otherwise, for $g = 2, \ldots, 5$. The coefficients β_{0g} and β_{1g}, $g = 2, \ldots, 5$, represent, respectively, the change in intercept and slope relative to group 1, which serves as the baseline model. The error term ε_{it} captures all other factors not explicitly included in the model for each country i and time period t. The results indicate heterogeneous growth dynamics across clusters, with average annual growth rates of approximately 12.5% for Germany, 3.7% for France, and around 7% for the remaining groups.

In the second stage, quarterly distance sales volumes (mail order and internet) for EEA countries from 2013–2023 are analysed as indices normalised to 2021. The series show clear trend and seasonality, with structural changes around 2020. Regression models and Holt–Winters smoothing are applied and evaluated using RMSE and residual diagnostics. The multiplicative Holt–Winters approach outperforms regression, delivering lower forecast errors and more satisfactory residual behaviour.

Finally, the estimated trend and seasonal components from the Holt–Winters models are used to cluster countries according to similarities in distance sales behaviour, providing further insights into structural differences across European retail markets. The multiplicative Holt–Winters decomposition reveals pronounced fourth-quarter seasonal peaks and post-2020 structural shifts in distance sales, and although results should be interpreted in light of data constraints and pandemic-related disruptions, they provide actionable insights for retailers and policymakers in designing resilient supply chain and digital retail strategies. The study concludes by discussing future research directions, highlighting the growing role of artificial intelligence, machine learning, and real-time monitoring systems. These advanced approaches, when combined with robust statistical benchmarks, have the potential to enhance retail forecasting accuracy, improve supply chain resilience, and support strategic decision-making in an increasingly volatile global trade environment.

Acknowledgements This work is supported by CIDMA (`https://ror.org/05pm2mw36`) under the Portuguese Foundation for Science and Technology (FCT, `https://ror.org/00snfqn58`), Grants UID/04106/2025 (`https://doi.org/10.54499/UID/04106/2025`) and UID/PRR/04106/2025 (`https://doi.org/10.54499/UID/PRR/04106/2025`).

References

[1] Eurostat. (2025). Database. `https://ec.europa.eu/eurostat/databrowser/`, acessed on 12 july 2025

9 April, 16:40 - 17:00, Auditorium ESTGD

Building environmental justice indicators based on time series of counts models

Adriano Gomes[1], **Ana Martins**[1,2], **Sónia Gouveia**[1,2]
[1]IEETA and DETI, University of Aveiro, Aveiro, Portugal,
[2]LASI - Intelligent Systems Associate Laboratory, Portugal
aog@ua.pt,a.r.martins@ua.pt,sonia.gouveia@ua.pt

Environmental factors affect health outcomes in ways that are neither temporally independent nor spatially uniform. The unequal distribution of these effects across population raises concerns about environmental justice (EJ). Statistical modeling of health data must account for temporal dependence before meaningful comparisons of environmental impacts across regions can be made. Thus, this work proposes a statistical framework based on time series modeling of environmental health effects to build population-based indicators for EJ assessment.

Keywords: INGARCH, time series of counts, clustering, spatial patterns

It is known that environmental factors, such as temperature and air pollutants, are associated with health outcomes. However, their impacts on health are not uniformly distributed across space, raising the question of whether specific populations are disproportionately affected by these factors, which motivates the study of environmental justice (EJ). Thus, assessing EJ requires a statistical framework that accounts for the temporal dependence on health outcomes, the spatial heterogeneity of environmental effects, and the socioeconomic characteristics of populations. Building on previous research [4], this work considers daily counts of respiratory hospital admissions in mainland Portugal modeled via INteger-valued Generalized AutoRegressive Conditional Heteroskedasticity models with exogenous covariates (INGARCH-X) [2]. Independent INGARCH-X models for each of the 18 districts are estimated with *tscount* R package [3] while accounting for environmental effects (temperature, dew-point temperature, $PM_{2.5}$, NO_2 and O_3) in addition to the temporal dependence of the data.

The district-level INGARCH-X estimates show spatial heterogeneity of the environmental effects. To explore whether this heterogeneity is associated with socioeconomic characteristics, districts were grouped using hierarchical agglomerative clustering based on census-derived socioeconomic variables. Demographic variables were processed to reflect the distribution of groups of population (e.g., percentage of population with ages 0-14, 15-64, and $\geq$65). Given the importance of including economic data, a proxy for PIB per capita was further considered. Several standardizations were tested, but some alternatives resulted in clusters driven almost exclusively by PIB. Thus, to mitigate this issue, the

min–max normalization was adopted. Also, multiple linkage functions (Ward.D2, Ward.D, complete, average, and McQuitty) were tested, resulting in similar clusters, which suggest that the socioeconomic profiles were consistent. The estimated environmental effects of INGARCH-X were then analyzed at the cluster level, providing an exploratory assessment of similarities and differences according to their socioeconomic characteristics. The results indicate limited differences in environmental effects between the clusters, suggesting that the observed spatial heterogeneity does not translate into clear socioeconomic patterns.
The limitations of district- and cluster-based comparisons for EJ assessment, lead to the consideration of a population-weighted aggregation of the district environmental effects. These weighted effects explicitly account for the spatial distribution of socioeconomic groups and shift the focus from administrative units to populations, providing an alternative and interpretable EJ indicator. Statistical inference for the proposed EJ metrics is supported by asymptotic normality results for the INGARCH models [1] and is complemented with a bootstrap approach to evaluate its robustness. Preliminary results suggest no statistically significant differences in weighted environmental effects across socioeconomic groups. Thus, at the population level, no vulnerable populations were identified. However, the existence of injustice cannot be excluded, since the generalization of the results maybe limited by data resolution.
In conclusion, a novel statistical framework integrating time series modeling, spatial patterns, and socioeconomic-based aggregation to study and quantify EJ is proposed. To the best of our knowledge, no other statistical framework is available to quantify EJ. Thus, this study provides an important step in environmental justice study given that this framework can be easily applied to other regions, pollutants, and health outcomes.

Acknowledgments The authors acknowledge Administração Central do Sistema de Saúde, IP (`https://www.acss.min-saude.pt/`) for providing the data.
This work was supported by the Foundation for Science and Technology (FCT, `https://www.fct.pt/`) through the contract `doi.org/10.54499/UID/00127/2025`. AG acknowledges the research grant in the scope of the FCT project "ALICE – Air polLution: a stressor for environmental justICE" (2022.04351.PTDC).

References

[1] V. Christou and K. Fokianos. Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, 35(1):55–78, 2014.

[2] R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

[3] T. Liboschik, K. Fokianos, and R. Fried. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *Journal of Statistical Software*, 82(5):1–51, 2017.

[4] A. Martins, M. Scotto, R. Deus, A. Monteiro, and S. Gouveia. Association between respiratory hospital admissions and air quality in Portugal: A count time series approach. *PLOS ONE*, 16(7):1–24, 07 2021.

9 April, 17:00 - 17:20, Auditorium ESTGD

Outliers in state-space models: a robust approach to parameter estimation and Kalman filter

A. Catarina Freitas[1], **A. Manuela Gonçalves**[2], **Marco Costa**[3]
[1] Department of Mathematics, Centre of Mathematics, University of Minho, Portugal, pg46704@alunos.uminho.pt,
[2] Department of Mathematics, Centre of Mathematics, University of Minho, Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, mneves@math.uminho.pt,
[3] Águeda School of Technology and Management, Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt

This study investigates the treatment of outliers in linear state-space time series models applied to environmental data, using robust statistical methods and robust Kalman filter algorithms. Outliers pose a major challenge in environmental analyses, as irregular measurements can undermine the accuracy of forecasting models. To address this issue, alternative robust methodologies are proposed. The standard Kalman filter is replaced with robust variants by employing different loss functions that dynamically adjust the weights assigned to residuals. The maximum likelihood estimator is robustified as well, including the Huber approach, the trimmed likelihood method, and an approach based on the Cauchy loss function. The performance of the proposed methodologies is assessed through simulation studies and applied to real data.

Keywords: time series, state-space models, Kalman filter, outliers, robust estimation

Time series analysis plays a fundamental role in a variety of scientific fields [4], [5]. However, these series are often affected by the presence of outliers, resulting from natural phenomena, measurement errors, or failures in data collection systems. Linear state-space models are used for a wide range of applications [3]. The standard approach for estimating the parameters of such models is computing the likelihood with the Kalman filter and maximizing it, assuming normality of the noise ([2]). The presence of extreme values, however, can compromise the effectiveness of classical estimation methods, such as the Kalman Filter, leading to biased estimates and unreliable forecasts [1]. The main objective of this work is to study, implement, and evaluate robust methodologies capable of mitigating the impact of atypical observations in dynamic time series models formulated in state-space form. To this end, robust variants of the Kalman Filter are proposed, employing different loss functions that dynamically adjust the weights assigned to residuals, thereby reducing the influence of outliers. In parallel, robust extensions of maximum likelihood estimation are considered, including the Huber approach, the trimmed likelihood method, and an

approach based on the Cauchy loss function. The proposed methodologies are evaluated through simulation studies under various scenarios, altering structural parameters and dimensions, and considering time series both with and without contamination. Finally, the methodologies are applied to real time series related to water quality in a watershed, where the occurrence of outliers is frequent.

Acknowledgements The research of A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013/2025 (https://doi.org/10.54499/UID/00013/2025). A. Catarina Freitas thanks CMAT for the research fellowship (BI) UMINHO/BIM/2024/131. Marco Costa is supported by CIDMA (https://ror.org/05pm2mw36) under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58), Grants UID/04106/2025 (https://doi.org/10.54499/UID/04106/2025) and UID/PRR/04106/2025 (https://doi.org/10.54499/UID/PRR/04106/2025).

References

[1] T. Cipra and R. Romera. Robust kalman filter and its application in time series analysis. *Kybernetika*, 27(6):481–494, 1991.

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

[3] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge, 2009.

[4] R. Shumway and D. Stoffer. *Time Series: A Data Analysis Approach Using R.* CRC Press, Taylor Francis Group, 2019.

[5] R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications: With R Examples.* Springer Cham, Springer Texts in Statistics, 2025.

9 April, 16:00 - 16:20, Lecture Theatre E1

Principal components for distributional data

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

This work proposes a principal component analysis method for histogram-valued data based on the linear combination definition of the Distribution and Symmetric Distribution model. Principal components are obtained as linear combinations of quantile functions representing correlated histogram-valued variables. Covariance and variance are defined using the Mallows distance. The first component is estimated by maximizing variance under non-negativity constraints. The method is illustrated in two applications.

Keywords: distributional data, principal component analysis, DSD regression model

The development of statistical methods for distributional data is growing [1], and most of the considered models are linear, such as regression, discriminant analysis, and principal component analysis. The Distribution and Symmetric Distribution (DSD) linear regression model proposed in [2] allows predicting the distribution of the target variable from other histogram-valued variables, and is defined using the representation of distributions by the corresponding quantile functions, under specific assumptions.
This work proposes a Principal Component Analysis (PCA) method for histogram-valued data based on the linear combination definition of the DSD model. Each principal component is obtained by a linear combination of the p correlated histogram-valued variables as follows:

$$\Psi_\epsilon(t) = \sum_{j=1}^{p} a_j \Psi_{X_j}(t) - b_j \Psi_{X_j}(1-t) \quad \text{with} \quad a_j, b_j \geq 0$$

where $\Psi_{X_j}(t)$ and $\Psi_{X_j}(1-t)$ represent, for each unit, the quantile function of the histogram X_j and the quantile function of the respective symmetric histogram, respectively.
For the first principal component (PC1), the vector $\gamma = [a_1\, b_1\, \ldots\, a_p\, b_p]$, of the non-negative parameters, is estimated maximizing the variance of the PC1, that is a quantile function, $\Psi_{\epsilon_1}(t)$.
The definitions of variance and covariance for histogram-valued variables used here were proposed by [3], and are based on the Mallows distance. The variance is defined as follows:

$$var(\Psi_{\epsilon_1}(t)) \;=\; \frac{1}{n}\sum_{i=1}^{n} D_M^2\left(\Psi_{\epsilon_1(i)}(t), \overline{\Psi_{\epsilon_1}}(t)\right)$$

where $\overline{\Psi_{\epsilon_1}}(t)$ is the barycenter of $\Psi_{\epsilon_1(i)}(t)$. Under the assumption of a uniform distribution within each sub-interval, we have

$$var(\Psi_{\epsilon_1}(t)) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{m} p_\ell \left[\left(c_{\epsilon_1(i)_\ell} - \overline{c}_{\epsilon_{1\ell}}\right)^2 + \frac{1}{3}\left(r_{\epsilon_1(i)_\ell} - \overline{r}_{\epsilon_{1\ell}}\right)^2\right]$$

where $\overline{c}_{\epsilon_{1\ell}}$, $\overline{r}_{\epsilon_{1\ell}}$ are the mean of the centers and of the half-ranges of the sub-intervals, respectively.
Similarly to the classical statistics but considering the definitions presented above, the parameters for the PC1 are obtained from

$$\begin{array}{ll} Maximize & var(\Psi_{\epsilon_1}(t)) \\ subject\,to & \gamma\gamma^T = 1 \text{ and } a_j, b_j \geq 0 \end{array}$$

Maximisation of the variance of the PC1 is obtained by solving this quadratic optimization problem. Note that in this case PC1 is a quantile function.
Since, as usual, the variables used in PCA are measured in different scales, the original histogram-valued variables should be standardized.
The proposed approach may be particularized to interval-valued variables, which constitute a special case of histogram-valued variables.
To analyse and interpret the behaviour of the results obtained for the PC1, two applications were studied. For a data set with 33 car models described by four strongly correlated interval-valued variables - price, engine capacity, top speed, and acceleration - the PC1 accounts for 91.35% of the total variance of the original variables. In the second study, scientific journals were aggregated in eight scientific areas described by five histogram-valued variables: number of published papers, impact factor, immediacy index, total citations, cited half-life. The conclusions were that the PC1 accounts for 54.2% of the total variance of the original variables. Moreover, the PC1 is strongly and positively correlated with impact factor, immediacy index, and total citations, while it is weakly (negatively) correlated with the number of published papers and cited half-life.
The results obtained in these applications show the importance and relevance of continuing the generalization of the method to p principal components.

Acknowledgements This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025).

References

[1] P. Brito and S. Dias. *Analysis of Distributional Data.* CRC Press, Taylor & Francis Group, 2022.

[2] S. Dias and P. Brito. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, 8(2):75–113, 2015.

9 April, 16:20 - 16:40, Lecture Theatre E1

Model selection in topic modeling

José G. Dias

BRU-IUL, Iscte – Instituto Universitário de Lisboa, jose.dias@iscte-iul.pt

---

The Latent Dirichlet Allocation (LDA) has become a very popular unsupervised machine learning model in topic modeling. Since the number of topics (K) needs to be specified in advance, various heuristics have been proposed to guide its selection. This research reviews the LDA model, discusses model selection, and sets up a Monte Carlo study to explore and compare the most popular measures.

Keywords: text mining, topic modeling, model selection, Bayesian analysis, Monte Carlo studies

---

The Latent Dirichlet Allocation (LDA) is a hierarchical Bayesian method for discovering a hidden thematic structure in a collection of documents by modeling each document as a mixture of topics, and each topic is characterized by a probability distribution over words. The Dirichlet distribution defines the prior for both the multinomial topic distributions per document and the word distributions per topic. Thus, the topic distribution for each document is drawn from a Dirichlet prior with hyperparameter α, and the word distribution for each topic is drawn from a Dirichlet prior with hyperparameter β. The generative process is: for every topic (k), a word distribution (ϕ_k) is sampled from a Dirichlet distribution parameterized by β. For each document (d), a topic distribution (θ_d) is sampled from a Dirichlet distribution parameterized by α. Then, for each word in the document, a topic is selected from (θ_d), and a word is sampled from the corresponding topic's word distribution (ϕ_k). This process defines a joint probability distribution over the observed documents and the hidden topic structure, expressed as $p(\theta, z, w|\alpha, \beta)$, where θ is the topic distribution, z is the topic assignment for each word, and w is the observed word.

Given the LDA specification, the number of topics must be set a priori. Different model selection approaches have been proposed to determine the optimal number of topics, usually using similarity between the profiles of the estimated pairs of topics. Four metrics are typically applied to select the number of topics:

- *Arunetal2010* [1] determines the optimal number of topics by looking at the distributions of the topic-terms and document-terms matrix outputs of LDA. The optimal number of topics is reached when the symmetric Kullback-Leibler divergence (KL divergence) of the distributions derived from these matrix factors is minimum. For a non-optimal number of topics, the divergence values are higher.

- *Caoetal2009* [2] assumes that the LDA model performs best when the average cosine distance between topics (T_i, T_j) reaches a minimum. The smaller the correlation (T_i, T_j) value, the more independent and disjoint the topics are.
- *Deveaueat2014* [3] is based on maximizing the differences between pairs of topics. Thus, the number of topics is estimated by maximizing the information deviation (D) between all pairs of LDA topics (k_i,k_j), i.e., the optimal number of topics for which LDA models the most dispersed topics.
- *GS2004* [4] uses model evidence to select the model. In Bayesian inference, for a model M_K with K topics, the model evidence is: $p(\text{data} \mid M_K)$. This is the probability of the entire corpus under the model with K topics, integrating over all unknown parameters (topic distributions, document-topic proportions, etc.).

This Monte Carlo study controls factors such as the number of documents, the number of words, and the level of noise. 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 R, Rstudio, and Stan. Important implications will be derived about relative performance of the selection criteria.

Acknowledgements This work was financially supported by Fundação para a Ciência e Tecnologia (UIDB/00315/2025).

References

[1] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'10, pages 391–402, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

[3] R. Deveaud, E. SanJuan, and P. Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014.

[4] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

9 April, 16:40 - 17:00, Lecture Theatre E1

Kernel K-means clustering of distributional data

Amparo Baíllo[1], **José R. Berrendero**[1], **Martín Sánchez-Signorini**[1]
[1] Universidad Autónoma de Madrid, amparo.baillo@uam.es, joser.berrendero@uam.es, martin.sanchez@uam.es

We approach the problem of clustering a sample of probability distributions from a random distribution on $\mathbb{R}^p$ by means of the maximum mean discrepancy [1]. Our proposed method maps the probability distributions to their kernel mean embeddings to then apply the K-means clustering algorithm in the corresponding reproducing kernel Hilbert space. We present simulation studies and illustrate the performance of our clustering method on Synthetic Aperture Radar (SAR) images [2].

Keywords: Functional data, maximum mean discrepancy, nonparametric, unsupervised classification

Our work focuses on the unsupervised classification of distributional data [3]. Distributional data refer to realizations of a random probability distribution which we denote as F, that is, a random element taking values in the space of probability distributions on $\mathbb{R}^p$. A distributional observation may arise, for example, as a summary of a very large sample (e.g. as produced by wearable devices), or of a more complex random object (e.g. an image) [4], as well as in the context of symbolic data analysis.

Popular tools for analyzing distributional data include to leverage the geometry of the Wasserstein space. The L^p distance for $1 \leq p \leq \infty$ between cumulative distribution functions or densities has also been studied [4].

The problem we will be considering is that of clustering a sample of iid probability distributions F_i, $i = 1, \ldots, n$ from the random distribution F. We assume that these distributions are supported on a common set $\mathcal{X} \subseteq \mathbb{R}^p$ and we denote by $\mathcal{M}(\mathcal{X})$ the space of probability distributions on $\mathcal{X}$. To proceed with the clustering task, we map the distributions F_i into a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ associated to a symmetric, positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then we apply the classical K-means algorithm in $\mathscr{H}$ to cluster the embedded distributions.

We propose to map the distributions F_i via their kernel mean embedding μ_{F_i} into the RKHS $\mathcal{H}$ associated to k [1]. Once the sample of distributions F_i has been mapped to their kernel mean embeddings μ_{F_i} in $\mathscr{H}$, the K-means algorithm is carried out in $\mathscr{H}$ by using the maximum mean discrepancy as the distance between the distributions F_i.

Our work proposes this K-means clustering algorithm as a straightforward, computationally feasible method for clustering distributional data in any dimension $p \geq 1$. Simulation studies have been conducted to provide insight into the choice of the kernel k and its tuning parameter. Particularly, we illustrate the performance of this clustering method on a real dataset consisting of a collection of Synthetic Aperture Radar (SAR) images [2] where each image is represented by a distribution of its pixel values.

Acknowledgements A. Baíllo and J.R. Berrendero are supported by the Spanish MCyT grant PID2023-148081NB-I00.

References

[1] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[2] C. Wang, A. Mouche, P. Tandeo, J. E. Stopa, N. Longépé, G. Erhard, R. C. Foster, D. Vandemark, and B. Chapron, "A labelled ocean sar imagery dataset of ten geophysical phenomena from sentinel-1 wave mode," *Geoscience Data Journal*, vol. 6, no. 2, pp. 105–115, 2019.

[3] P. Brito and S. Dias, *Analysis of Distributional Data*. CRC Press, 2022.

[4] L. Mora-López and J. Mora, "An adaptive algorithm for clustering cumulative probability distribution functions using the kolmogorov–smirnov two-sample test," *Expert Systems with Applications*, vol. 42, no. 8, pp. 4016–4021, 2015.

9 April, 17:00 - 17:20, Lecture Theatre E1

Analysis of classification models under reduced experimental designs in gene expression data

José Febra[1,2], **Paula Faria**[1,2], **João Meneses**[2], **Carlos Grilo**[1,3]
[1] School of Technology and Management, Polytechnic of Leiria, Leiria, Portugal, zefebra@gmail.com, paula.faria@ipleiria.pt, carlos.grilo@ipleiria.pt
[2] CDRSP – Centre for Rapid and Sustainable Product Development, Polytechnic of Leiria, Portugal, joao.p.meneses@ipleiria.pt
[3] CIIC – Computer Science and Communications Research Centre, Polytechnic of Leiria, Leiria, Portugal

The classification of gene expression data often relies on complex experimental designs involving multiple stimuli, conditions, or temporal measurements, increasing both experimental and computational cost. This work analyses, from a methodological perspective, how classification performance is affected when experimental designs are systematically reduced. Using gene expression datasets with different structures, the study evaluates whether reliable classification can be maintained under simplified experimental configurations.

Keywords: gene expression, classification, machine learning, data reduction

Gene expression datasets are typically high-dimensional, limited in sample size, and experimentally diverse, which poses challenges for both modelling and experimental design [2]. In this context, reducing the number of experimental conditions without compromising classification performance constitutes a relevant methodological problem.

In this study, several classification models, including machine learning and deep learning approaches, were initially evaluated on complete datasets to identify a suitable reference model for each classification task. The compared approaches included convolutional neural networks, multilayer perceptrons, long short-term memory networks, support vector machines, and gradient-boosted decision trees, following a comparative strategy commonly adopted in gene expression classification studies [4, 3]. Model selection was based on repeated training and evaluation. Classification performance was assessed using accuracy and F1-score on independent test sets, averaged across runs to obtain stable performance estimates and analyse variability [1].

After selecting a reference classifier, a systematic analysis of reduced experimental designs was conducted, examining structured reductions of stimuli, nutrient conditions, and temporal measurements to assess their impact on classification accuracy.

The experiments were performed on three public datasets representing distinct experimental designs. The first dataset consists of gene expression profiles measured in four yeast species belonging to the *Saccharomyces* family across five environmental stress stimuli,

where classification aims to distinguish between yeast species under different experimental conditions. This dataset enables the evaluation of performance as the number of available stress conditions is systematically reduced.
The second dataset is based on *Saccharomyces cerevisiae* gene expression measurements obtained in nutrient limitation experiments, comprising four nutrient conditions evaluated under two oxygen regimes (aerobic and anaerobic). Here, classification focuses on distinguishing between the two oxygen regimes, while nutrient conditions define the experimental environments over which reduced configurations are analysed.
The third dataset consists of gene expression profiles collected over the 24-hour embryogenesis period of *Drosophila melanogaster*, enabling the evaluation of reduced temporal sampling strategies. Together, these datasets cover static and time-dependent data structures, as well as binary and multiclass classification problems. Performance was evaluated through multiple repetitions to analyse variability as experimental information was progressively removed.
Table 1 summarises representative reduced experimental configurations for each dataset. The results show that, in many cases, performance remains close to that obtained with the full experimental design, even under substantial reductions in conditions or time points [4]. In particular, specific subsets of stimuli, nutrient environments, or temporally distributed measurements retain most of the discriminative information.

Table 1: Representative examples of reduced configurations and classification accuracy.

Dataset	Type	Reduced configuration	Acc. Full	Acc. Reduced
GSE3406	Stimuli	2 out of 5 stimuli	96.11%	90.96%
GSE1723	Nutrients	3 out of 4 nutrients	87.30%	86.03%
GSE6186	Temporal	Alternating time points	94.61%	92.01%

References

[1] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, D. Serdyuk, T. Arbel, C. Pal, G. Varoquaux, and P. Vincent. Accounting for variance in machine learning benchmarks. *arXiv preprint arXiv:2103.03098*, 2021.

[2] E. R. Dougherty. Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2(1):28–34, 2001.

[3] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13(S5), April 2020.

[4] N. I. Tripto, M. Kabir, M. S. Bayzid, and A. Rahman. Evaluation of classification and forecasting methods on time series gene expression data. *PLOS ONE*, 15(11):e0241686, November 2020.

10 April, 9:00 - 9:20, Lecture Theatre E1

Reproduction characterization of an ant population: an interdisciplinary data modelling

Laura Machado[1], **Dulce Gomes**[2], **Filipe Ribeiro**[3]
[1] MED-UÉ & CHANGE Global Change and Sustainability Institute, Instituto de Investigação e Formação Avançada (IIFA), Universidade de Évora (UÉ)
lauramagnani98@gmail.com
[2] Dep. de Matemática, ECT, CIMA, IIFA, UÉ, dmog@uevora.pt
[3] CIDEHUS-UÉ, Dep de Sociologia, ECS, UÉ, fribeiro@uevora.pt

---

The present work aimed to use biodemography analysis to characterize the reproduction behavior of Pharaoh's ants' colonies. We took advantage of Birch (1948) biodemography methods, Generalized Estimating Equations and classical demographic concepts on data from the Dryad project. The application of this interdisciplinary methodology highly contributed to obtain more accurate and directly driven results, being a novel approach extendible to other ants' colonies or even other species.

Keywords: biodemography, GEE, intrinsic growth rate, M. pharaonis ants

---

Monomorium pharaonis is one of the many invasive ant species which can be found associated with the endosymbiont bacteria of the genus *Wolbachia* [4]. This association is still being studied but is already known for giving the colony a reproductive advantage. Ants are social insects, which have colonies organized in different castes, reproductive (queens and males) and sterile (workers) [4]. The present work consisted of a longitudinal study and aimed to use biodemography analysis to check the effect of *Wolbachia* in the colonies of *M. pharaonis* ants, regarding its reproductive potential, as well as the application of bold methodologies, such as Generalized Estimating Equations and classical demography concepts to analyze the data. The data was collected from the Dryad database, an open data publishing platform, in December of 2023. The data used was from an experiment which compared egg-laying rates of queens in *Wolbachia*-infected and uninfected colonies. Thirty-one colonies of *M. pharaonis* ants were observed and population data and egg laying rates were collected each three days, consisting of 46 days of experiment. Twenty colonies were infected with *Wolbachia*, and eleven composed the uninfected group. The experiment started when the queens were four days old. The variables of number of queens, number of workers and number of eggs laid by the queens, infection status and queen age in days were used for the analysis.
The analytical biodemography framework followed the life table approach pioneered by Lotka [3] and Birch [1], adapted to insect populations. Age-specific egg ("birth") counts were transformed into fecundity rates (m_x) and "survival" probabilities (l_x), from which

force of fecundity (μ_x) was derived and the intrinsic growth rate (r) was calculated and compared between infected and uninfected colonies. To detect fecundity deceleration and potential plateaus at later times, we estimated the logarithmic derivative of the force of mortality, the demographic measure known as the log-aging rate (LAR). To quantify fecundity patterns and detect late-life fecundity deceleration in ants, we combined empirical life table construction with parametric mortality models. In the present work, we fit four parametric mean functions $\mu(x)$ representing per-queen fecundity at age x. These models provide alternative characterizations of age-specific reproductive schedules, and their parameters are estimated via maximum likelihood under a Poisson framework. To understand the influence of the presence of the bacteria and the time on the number of queens, eggs and workers, we used Generalized Estimating Equations (GEE), typical of marginal models, following the proposal by Liang & Zeger [2]. One marginal model for each life stage used for the biodemography analysis was constructed, showing different results regarding the influence of the *Wolbachia* for each caste of the ants' colony. The GEE models were constructed considering the family Poisson, with the logarithmic link function and the AR(1) constraint.

The results obtained show that the presence of *Wolbachia* increases the mean length and the colonies' capacity to grow in each generation as well as the intrinsic rate of increase. The demography analysis indicates that the infection, besides giving the colonies a later peak on the eggs-per-queen numbers, a higher intrinsic rate of increase, generation length and growth capacity, the queen's fecundity starts to decrease in earlier ages. This may indicate that the bacterial infection can act as a reproductive 'boost', leading to a rapid increase in the colony's reproductive capacity, followed by a decline in fecundity at a younger age compared to the absence of *Wolbachia*. The marginal models showed that the presence of *Wolbachia* influenced the numbers of eggs and queens, but not the worker numbers, which could indicate another evidence that *Wolbachia* is related to the reproductive capacity of the ants' colonies. The GEE model for the queens also showed a significant interaction between the bacteria infection and the days of the experiment. The results of the three areas of this study (biology, demography and statistics) proved to be complementary in providing a more holistic understanding of the effects of the endosymbiotic bacteria on both the distinct castes and the colony as an integrated unit.

References

[1] L. C. Birch. The intrinsic rate of natural increase of an insect population. *Journal of Animal Ecology*, 17:116–130, 1948.

[2] K. Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[3] A. J. Lotka. Relation between birth dates and death rates. science. *Science*, 26:21–22, 1907.

[4] R. Sing and T. A. Linksvayer. Wolbachia-infected ant colonies have increased reproductive investment and an accelerated life cycle. *Journal of Experimental Biology*, 9:223–229, 2020.

10 April, 9:20 - 9:40, Lecture Theatre E1

How efficient are Portuguese beef farms?

Cândida Santos[1,2], **José G. Dias**[3], **M. Rosário Oliveira**[1], **Pedro Reis**[2,4]
[1] CEMAT, Instituto Superior Técnico, University of Lisbon, Portugal
candida.santos@iniav.pt, rosario.oliveira@tecnico.ulisboa.pt
[2] Instituto Nacional de Investigação Agrária e Veterinária, I.P (INIAV), Portugal
[3] BRU-IUL, Iscte – Instituto Universitário de Lisboa, jose.dias@iscte-iul.pt
[4] Green-it, Bioresources4Sustainability, ITQB NOVA, Portugal, pedro.reis@iniav.pt

Using 2020 data from Portugal's Farm Accountancy Data Network (FADN) on specialised beef-cattle farms, we estimate a Cobb–Douglas stochastic production frontier. Output is most responsive to intermediate consumption, labour and capital, while the land effect is weaker and only marginally supported. Inefficiency dominates residual variation, indicating substantial scope to improve technical efficiency.

Keywords: farm efficiency, stochastic frontier model, beef cattle farms, FADN

Globalisation and market liberalisation have intensified competitive pressure in agriculture, pushing farms to restructure and improve efficiency while facing country-specific constraints such as agro-climatic conditions, resource endowments and inherited farm structures. Following Coppola et al. [2], this study treats efficiency as central to long-run farm viability and stresses the complementary roles of innovation and structural policies in sustaining farm persistence.

Stochastic frontier analysis (SFA) [3] is widely used to quantify production efficiency. Consider producer i with output y_i and k inputs $x_1, \ldots, x_k$. Then the production frontier is given by $y_i = f(x_{i1}, \ldots, x_{ik})exp(v_i - u_i)$, where f is the production frontier, $v_i \sim N(0, \sigma_v^2)$ captures random noise, and $u_i \sim \text{Half} - \text{Normal}(0, \sigma_{\text{u}}^2)$ is a non-negative random variable representing inefficiency economic factors. A common choice is the Cobb-Douglas function: $f(x_{i1}, \ldots, x_{ik}) = e^{\beta_0} \prod_{j=1}^{k} x_{ij}^{\beta_j}$.

The log-linear specification is: $\ln y_i = \beta_0 + \beta_1 \ln x_{i1} + \cdots + \beta_k \ln x_{ik} + v_i - u_i$. The model parameters are estimated by maximum likelihood, and the R package frontier [1] reports the parameter $\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$, the share of total variance due to inefficiency ($\gamma \in [0, 1]$). Thus, if $\gamma = 0$, inefficiency is absent and the whole variance comes from noise; if $\gamma = 1$, all variation is attributed to inefficiency, implying a deterministic frontier.

This study uses the 2020 FADN for Portugal, covering 367 specialised beef-cattle farms. The FADN includes market-oriented farms with standard output above €4,000 and is sampled to be representative by region, economic size, and farming type. The output is the total farm agricultural production value, including crops, livestock, and farm services.

Inputs are agricultural area (ha), labour (annual work units), capital (asset value), and intermediate consumption (feed, veterinary services, and other variable inputs). All variables are log-transformed to estimate a log-linear Cobb–Douglas production frontier.

Table 1: Estimates from the stochastic frontier model.

	Estimate	**Std. Error**	**z value**	**Pr($>|z|$)**
Intercept	5.5660	0.2720	20.4620	< 0.0001
ln(Land)	0.0410	0.0220	1.8730	0.0610
ln(Labour)	0.2300	0.0620	3.6830	0.0002
ln(Capital)	0.1900	0.0350	5.4260	< 0.0001
ln(Intermediate consumption)	0.3260	0.0270	12.1860	< 0.0001
σ^2	0.7460	0.0620	11.9580	< 0.0001
γ	0.9540	0.0110	91.5420	< 0.0001

Table 1 reports a positive significant intercept, consistent with a high baseline technology level in Portuguese specialised beef farms. All input elasticities are positive ($\hat{\beta}_i > 0$). Land has a marginally significant elasticity, suggesting limited output gains from expanding area alone. Labour, capital, and intermediate consumption have significant elasticities, reflecting the role of management, physical assets, and other inputs (feed, veterinary services, and related expenses) on the output.

Variance estimates imply that most residual variation is due to technical inefficiency rather than statistical noise: $\hat{\gamma}$ is close to one. The log-likelihood and mean technical efficiency (≈ 0.63) indicate substantial scope to raise output by closing the efficiency gap to the estimated frontier.

Acknowledgements This work was financially supported by Fundação para a Ciência e Tecnologia, through the projects UIDB/00315/2025 and UID/04621/2025 UID/4459/2025.

References

[1] T. Coelli and A. Henningsen. *frontier: Stochastic Frontier Analysis*, 2020. R package version 1.1-9.

[2] A. Coppola, M. Amato, D. Vistocco, and F. Verneau. Measuring the economic sustainability of Italian farms using FADN data. *Agricultural Economics (Zemědělská ekonomika)*, 68(9):327–337, 2022.

[3] J. Jondrow, C. A. K. Lovell, I. S. Materov, and P. Schmidt. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19(2):233–238, 1982.

10 April, 9:00 - 9:20, Auditorium Francisco Tomatas

Modeling the association between type 2 diabetes and liver stiffness in MASLD patients

Ana Matos[1], **Carla Henriques**[2], **Paula Mesquita**[3], **Armando Carvalho**[4], **Adélia Simão**[4]
[1] ESTGV, Instituto Politécnico de Viseu, Portugal and Research Centre in Digital Services (CISeD), Portugal, amatos@estgv.ipv.pt
[2]ESTGV, Instituto Politécnico de Viseu, Portugal and Centre for Mathematics of the University of Coimbra (CMUC), Portugal, carlahenriq@estgv.ipv.pt
[3]Internal Medicine Service, Local Health Unit of the Aveiro Region and Liver Disease Functional Unit – Internal Medicine Service, Local Health Unit of Coimbra, Portugal, alex.mesquita1@hotmail.com
[4]Liver Disease Functional Unit-lnternal Medicine Department, Local Health Unit of Coimbra e Faculty of Medicine, University of Coimbra, Portugal, aspcarvalho@gmail.com, adeliasimao@gmail.com

This study investigated the impact of type 2 diabetes mellitus (T2DM) on liver fibrosis in patients with metabolic dysfunction-associated steatotic liver disease (MASLD). Diabetic and non-diabetic groups were compared using non-parametric tests, and associations with liver stiffness and steatosis were evaluated through logistic and linear regression with robust standard errors. T2DM was linked to higher steatosis and more advanced fibrosis, supporting earlier diagnosis and targeted screening.

Keywords: logistic regression, linear regression

In Portugal, hepatic steatosis and metabolic dysfunction–associated steatotic liver disease (MASLD) represent a significant public health burden, with prevalence estimates of approximately 37.8% and 17%, respectively, according to data from 2020 [2]. In parallel, type 2 diabetes mellitus (T2DM) affects around 11% of the Portuguese adult population aged 20 to 79 years, as reported by the International Diabetes Federation [1]. Given the high and overlapping prevalence of these metabolic conditions, investigating the impact of T2DM on liver fibrosis in patients with MASLD is of particular clinical relevance. To address this gap, we conducted a retrospective study involving 252 patients diagnosed with MASLD who underwent transient elastography between July 2022 and February 2024 at Coimbra University Hospitals. Liver stiffness measurement (LSM, kPa) was the primary outcome, with fibrosis staged as: F0–F1 ($<$7.0 kPa), F2 (7.0–8.6 kPa), F3 (8.7–10.2 kPa), and F4 ($\geq$10.3 kPa, presumed cirrhosis).
The secondary outcome was the controlled attenuation parameter (CAP, dB/m), to assess hepatic steatosis. Patients were divided into two groups: Group 1, with type 2 diabetes (n

= 96; 38.1%), and Group 2, without diabetes (n = 156; 61.9%). Differences between groups were assessed using the Mann–Whitney U test for continuous variables and the chi-square test for categorical variables. Correlations between LSM and CAP were evaluated using Spearman's rho.

Linear regression models were used as the primary framework to assess the association between type 2 diabetes and continuous liver outcomes (ln-LSM and ln-CAP), preserving information across the full data distribution. Logistic regression was used as a complementary analysis to express results according to clinically relevant fibrosis and steatosis thresholds. Robust standard errors were applied to account for heteroscedasticity and ensure reliable inference. In linear regression models, with ln(LSM) as the outcome, diabetes showed coefficients of 0.47 (95% CI: 0.33–0.61) unadjusted model and 0.43 (95% CI: 0.28–0.57) adjusted -Table 1. After adjusting for age and BMI, the regression coeficient for diabetes corresponds to a 53.7% higher predicted LSM among patients with diabetes compared with those without diabetes.

Table 1: Estimates and p-values of the effect of diabetes on ln(LSM) and ln(CAP)

	Unadjusted Estimate		Adjusted Estimate	
Continuous Outcome	coefficient (95% CI)	p-value	coefficient (95% CI)	p-value
ln(LSM) (kPa)	0.47 (0.33–0.61)	< 0.001	0.43 (0.28–0.57)	< 0.001
ln(CAP score) (dB/m)	0.04 (0.01–0.07)	0.008	0.04 (0.01–0.07)	0.012

Logistic regression showed that patients with diabetes had significantly higher odds of advanced fibrosis (LSM $\geq$8.7 kPa; adjusted OR = 5.55; 95% CI: 2.80–11.00) and presumed cirrhosis (LSM $\geq$10.3 kPa; adjusted OR = 5.53; 95% CI: 2.45–12.51).

Diabetes was also independently associated with hepatic steatosis severity. In the adjusted linear model, diabetes was associated with a 4.1% higher CAP value ($\beta = 0.04$; 95% CI: 0.01–0.07) - Table 1. Moreover, patients with diabetes had nearly twice the odds of having severe steatosis (CAP $\geq$280 dB/m; adjusted OR = 1.96; 95% CI: 1.02–3.77), independent of BMI.

In conclusion, T2DM is independently associated with increased hepatic steatosis and more advanced liver fibrosis in patients with MASLD, including a higher risk of presumed cirrhosis. These results emphasize the importance of early MASLD detection in individuals with type 2 diabetes and support the implementation of routine screening strategies for liver-related complications.

References

[1] International Diabetes Federation. Portugal diabetes report 2000–2045. `https://diabetesatlas.org/data/en/country/159/pt.html`, 2026. Accessed jan 15, 2026.

[2] J. Leitão, S. Carvalhana, J. Cochicho, and A. P. Silva. Prevalence and risk factors of fatty liver in portuguese adults. *European Journal of Clinical Investigation*, 50:e13235, 2020.

10 April, 9:20 - 9:40, Auditorium Francisco Tomatas

Evaluating a numerical discomfort scale for immobilized trauma victims

Carla Henriques[1], **Ana Matos**[2], **Mauro Mota**[3]
[1] ESTGV, Instituto Politécnico de Viseu, Portugal and Centre for Mathematics of the University of Coimbra (CMUC), Portugal, carlahenriq@estgv.ipv.pt
[2] ESTGV, Instituto Politécnico de Viseu, Portugal and Research Centre in Digital Services (CISeD), Portugal, amatos@estgv.ipv.pt
[3] ESSV, Instituto Politécnico de Viseu, Portugal and UICISA: E/ESEnfC - Cluster at the Health School, Polyt. Inst. of Viseu, Portugal, maurolopesmota@gmail.com

Immobilization is applied to trauma victims to maintain proper anatomical alignment and prevent potentially harmful displacement. Despite being a frequent pre-hospital procedure, it often causes pressure-related discomfort and pain. The main objective of this study is to evaluate the psychometric properties of the Numerical Discomfort Scale (NDS), a patient-reported scale used to assess discomfort in immobilized patients.

Keywords: ICC, consistency, reliability, numerical scale

Data were collected from 27 volunteers immobilized in a vacuum mattress for 60 minutes [3]. Volunteers rated their discomfort with an integer from 0 to 10, where zero corresponds to non-discomfort and 10 indicates maximum discomfort. This procedure was conducted twice for each participant to facilitate test-retest reliability analysis, with sessions spaced approximately two weeks apart. Discomfort was monitored at five-minute intervals across 30 body locations. For each time point and participant, the mean of the three highest scores across the 30 locations was calculated as a summary measure reflecting the most clinically relevant discomfort.
Scores were compared across time points using Friedman's ANOVA. Post-hoc analyses revealed significant differences between any time points separated by more than 20 minutes. Given the lack of significance in nearly all shorter intervals, coupled with the negligible discomfort recorded in the early stages (mostly zero values), analysis was restricted to data from the 20-minute point onward.
A comparative analysis of the test and retest trials revealed lower discomfort values in the second trial, suggesting an expected familiarization effect wherein subjects' prior exposure to the immobilization procedure attenuated their perception of discomfort. Nevertheless, a progressive increase in discomfort over the 60-minute period was evident in both trials. To evaluate the consistency of this progression and provide support for the scale's reliability, paired samples t-tests and a Two-Way Repeated Measures ANOVA (factors: time and trial) were conducted. Bonferroni-adjusted paired t-tests revealed non-significant differences

between the two trials regarding the increments from 20 to 40 minutes, 40 to 60 minutes and 20 to 60 minutes (all adjusted $p > 0.1$). Furthermore, the Two-Way Repeated Measures ANOVA showed no significant interaction between time and trial ($p > 0.6$), suggesting that discomfort scores evolve consistently across both sessions.
Reliability was further assessed by estimating the Intraclass Correlation Coefficient (ICC). Specifically, a two-way mixed-effects model for consistency, ICC(3,k), was employed [2]. For this calculation, the mean score for each subject across all time points from 20 minutes onward was computed for both test and retest trials. The ICC for these aggregated scores was 0.729 (95% CI: 0.41–0.88), indicating fair-to-good reliability [1]. Additionally, ICC values for individual time points were found to range from 0.58 to 0.81.
Overall, the findings of this study reveal that the Numerical Discomfort Scale (NDS) exhibits consistent progression patterns with a fair-to-good reliability score. Thus, the NDS establishes itself as a valuable and reliable instrument for monitoring patient-reported discomfort during prolonged immobilization in both clinical and pre-hospital settings.

Acknowledgements The authors acknowledge financial support by the Centre for Mathematics of the University of Coimbra (CMUC, https://doi.org/10.54499/UID/00324/2025) under the Portuguese Foundation for Science and Technology (FCT), Grants UID/00324/2025 and UID/PRR/00324/2025.

References

[1] S. A. Doi and G. M. Williams (Eds.). *Methods of clinical epidemiology.* Springer, Berlin, 2013.

[2] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Chiropr Med.*, 15(2)::155–63, 2016.

[3] M. Mota, F. Melo, M. Castelo-Branco, R. Campos, M. Cunha, and M. R. Santos. Construction of the discomfort assessment scale for immobilized trauma victims (dasitv). *Int Emerg Nurs.*, 76:101501, 2024.

10 April, 9:40 - 10:00, Auditorium Francisco Tomatas

Students' attitudes towards Mathematics: the influence of individual and academic characteristics in the 1st cycle

Ana Felizardo Henriques[1], **Adelaide Freitas**[1,2], **Fernando Sebastião**[3], **João Marôco**[4]

[1] Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal, anac@ua.pt adelaide@ua.pt

[2] Department of Mathematics, University of Aveiro, Portugal.

[3] School of Technology and Management, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal & Laboratory of Separation and Reaction Engineering-Laboratory of Catalysis and Materials (LSRE-LCM), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal & ALiCE – Associate Laboratory in Chemical Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, fsebast@ipleiria.pt

[4] Intrepid Lab, ECEO, Lusófona University & CETRAD, University of Trás-os-Montes e Alto Douro, jpmaroco@gmail.com

---

The Student Questionnaire – Grade 4 is an international questionnaire used in TIMSS to assess the attitudes of 4th grade students towards mathematics. This questionnaire was used on a sample of 3rd and 4th grade students in central region of Portugal, and scores were estimated for four latent factors (enjoyment, clarity, disorder, and confidence) related to students' attitudes towards mathematics. The aim of this study is to analyse the four factors empirical distributions, conditioned by student characteristics (year of schooling, gender, negative towards mathematics and failure in the 1st cycle), in order to assess the effect of these characteristics on those factors.

Keywords: confirmatory factor analysis, mathematical education, TIMSS

TIMSS (Trends in International Mathematics and Science Study) is an international study conducted every four years that assesses the performance of 4th and 8th grade students in Mathematics and Science, comparing education systems in various countries [2]. The 2019 edition of TIMSS included the Student Questionnaire – Grade 4 (SQ4) [1], with 31 items on a 4-point Likert scale that assess students' attitudes towards Mathematics, including their enjoyment and confidence in Mathematics, the clarity of Mathematics teaching, and the disorderly behaviour during Mathematics lessons.

To analyse attitudes towards mathematics among 3rd and 4th year students in Portugal, the SQ4 questionnaire was conducted in early 2025, to 775 students (377 from the 3rd year

and 398 from the 4th year) from fifteen Portuguese primary schools in the central region, selected by non-probabilistic convenience sampling, although this may limit the generalisation of the results. Along with the SQ4, the following characteristics of all students participating in the study were also collected: C1 – year of schooling ("3rd" or "4th"); C2 – gender ("m" or "f"); C3 – whether they had ever failed a Mathematics test ("yes" or "no"); and C4 – whether they had ever failed Mathematics in the 1st cycle ("yes" or "no").

The results of the Confirmatory Factor Analysis (CFA) about the responses obtained in SQ4 showed that the four-factor model proposed in TIMSS in 2019 was appropriate for data, namely: Enjoyment (MS2: Enjoyment of learning Mathematics – 9 items), Clarity (MS3: Clarity in the teaching of Mathematics lessons – 6 items), Disorder (MS4: Disorderly behaviour during Mathematics lessons – 6 items) and Confidence (MS5: Confidence in Mathematics – 9 items).

Bartlett's sphericity test ($\chi^2 = 8573.04$, $p < 0.001$) and the KMO index (0.93) indicate that it is appropriate to proceed with a factor analysis. Cronbach's Alpha coefficients showed excellent internal consistency for the MS2 dimension (0.91), poor consistency for MS3 (0.61), acceptable consistency for MS4 (0.78), and good consistency for MS5 (0.82). CFA was conducted using the Diagonally Weighted Least Squares (DWLS) estimator, appropriate for ordinal items. The model demonstrated a good overall fit to the data, according to the fit indices: CFI (0.991), TLI (0.990), RMSEA (0.042), and SRMR (0.057). Based on the estimated scores for the four factors (MS2, MS3, MS4, and MS5), using Thurstone's method and the average item score, trends in the attitudes of 3rd and 4th year primary school students will be analysed and discussed in order to investigate the effects of individual characteristics C1–C4 on the four latent factors mentioned above.

Acknowledgements : This work is supported by CIDMA (https://ror.org/05pm2mw36) under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58), Grants UID/04106/2025 (https://doi.org/10.54499/UID/04106/2025) and UID/PRR/04106/2025 (https://doi.org/10.54499/UID/PRR/04106/2025).

References

[1] IEA - International Association for the Evaluation of Educational Achievement. TIMSS 2019: Student questionnaire - grade 4, 2018.

[2] IEA - International Association for the Evaluation of Educational Achievement. TIMSS 2019 methods and procedures: Creating context questionnaire scales – grade 4. Technical report chapter, TIMSS and PIRLS International Study Center, 2020.

10 April, 10:10 - 10:30, Lecture Theatre E1

Recent advances in model checking for logistic partially linear models

Rui Costa-Miranda[1], **Wenceslao González-Manteiga**[2], **Rita Gaio**[1]
[1] Faculty of Sciences of the University of Porto and Centre of Mathematics of the University of Porto, up201804962@up.pt, argaio@fc.up.pt
[2] Faculty of Mathematics of the University of Santiago de Compostela, Spain wenceslao.gonzalez@usc.es

Building on recent advances in testing goodness-of-fit for generalized partially linear models, we compare well-established methods with new hypothesis tests built from an integrated residual-marked Gaussian processes approach. The new tests make use of Neyman otrthogonality for local estimation robustness, and random projections for dimension reduction. Their performance is compared on a logistic partially linear model, investigating the effect of shell length and bed location on the infection of cockles from the Ria de Aveiro estuary.

Keywords: conditional moment restrictions, Gaussian processes, generalized partially linear models, goodness-of-fit, logistic regression

For $i = 1, \ldots, n$, consider the crude errors of a generalized partially linear model setting, given by $\varepsilon(A_i, S_i, \beta, m) = Y_i - g^{-1}(A_i^T\beta + m(S_i))$, where $g(\cdot)$ is a known link function, A and S are covariates, β is a vector of coefficients, and $m(\cdot)$ is a smooth function to be estimated. The goodness-of-fit of the adjusted model can be assessed by testing

$$H_0 : E[\varepsilon(A, S, \beta, m) \mid A, S] = 0 \quad \text{almost surely for some } \beta, m. \tag{1}$$

The conditional mean specification in (1) defines the model predictor, and is proved to be equivalent to a continuum of projected unconditional moment restrictions, so that the test can be based on an integrated approach over residual-marked empirical Gaussian processes (GP). The test statistic is defined in a Cramér-von Mises sense [4], and simplified into a finite sum as the sample equivalent of

$$T_K(\beta, m) := E_W\left[(R(W, \beta, m))^2\right] = E\left[\varepsilon(A, S, \beta, m)K\varepsilon(A, S, \beta, m)\right]. \tag{2}$$

where $R(W, \beta, m) := E\left[w(A, S)\varepsilon(A, S, \beta, m)\right]$, w is any measurable function, and K is the estimated covariance matrix of the GP W. Recently, a novel class of tests refines this methodology for finite-dimensional parametric models [3]. The key innovation is projecting W onto the orthocomplement of the conditional moment score subspace in L_2, the Hilbert space of square integrable measurable functions of A and S. The test statistic considers the estimated covariance-kernel matrix of the projected GP. This ensures that the test statistic remains robust to the estimation, allowing for fast-bootstrap procedures where there is no

need to fit the model in each iteration, thus resulting in more efficient computations. Using a local polynomial based approach, we extend this procedure to accomodate the infinite-dimensional parameter of the nonparametric component of the model, $m(\cdot)$.
To illustrate, data regarding the edible cockle *Cerastoderma edule* are considered. This species is an indigenous bivalve from semi-sheltered marine systems along the north-eastern coast of the Atlantic Ocean. For a better understanding of their ecology, samples of cockles were collected in four bed regions from the Ria de Aveiro (*REF*- Ovar, *1*- Espinheira, *2*- Ílhavo, *3*- Mira) over periods of one month each [1]. Factors like the shell lenght (S, in milimitres) and bed location (A_j, binary, $j = 1, \ldots, 3$) were considered for modelling the presence/absence of macroparasites (Y). Using cross-sectional data from a fixed month, the partially linear logistic model $logit(\pi_i) = \beta_1 A_{1i} + \beta_2 A_{2i} + \beta_3 A_{3i} + m(S_i)$ was fitted, and the previous model checking tests were applied. In Table 1, the baseline of the comparative study is a Hosmer–Lemeshow test, T_{HL}. The GP based procedures are labelled by T_E, considering a Escanciano-type covariance kernel [4], and T_D, for a distance covariance kernel [3]. Test statistics $T^{\perp}$ represent the proposed Neyman-orthogonal version of the latter, and T^{rp} labels a random-projections approach for dimention-reduction, adapted from [2]. Results show that the proposed orthogonal approach reduces time complexity without compromising the conclusion of the test. All tests fail to reject the hypothesized model, thus validating the corresponding association of the shell length and the location of the cockle beds to the presence of macroparasites.

Table 1: Comparison of the p-values and execution times obtained for each of the tests.

Test statistic	T_{HL}	T_E	T_D	T_E^{rp}	T_E^{rp}	$T_E^{\perp}$	$T_D^{\perp}$	$T_E^{rp\perp}$	$T_D^{rp\perp}$
p-value	0.238	0.732	0.665	0.691	0.758	0.580	0.542	0.613	0.642
time (seconds)	0.05	19.14	17.89	21.42	18.28	2.02	0.51	4.47	1.27

Acknowledgements Rui Costa-Miranda was granted a doctoral research fellowship financed by FCT - Fundação para a Ciência e Tecnologia, I.P., with reference 2024.03100.BD. Rita Gaio and Rui Costa-Miranda were partially supported by CMUP, member of LASI, which is financed by national funds through FCT, for the project with reference UID/00144.

References

[1] S. Correia. Count mixed-effects regression models in parasite ecology. Master's thesis, Faculty of Sciences of the University of Porto, Porto, Portugal, 10 2023.

[2] J. A. Cuesta-Albertos, E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga. Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, 47(1):439–467, 2019.

[3] J. C. Escanciano. A gaussian process approach to model checks. *The Annals of Statistics*, 52(5):2456–2481, 2024.

10 April, 10:30 - 10:50, Lecture Theatre E1

Clusterwise linear regression for symbolic density--valued data

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

We propose a clusterwise linear regression model for density-valued variables using quantile functions representations. Extending the Distribution and Symmetric Distribution regression model, it estimates cluster-specific relationships by minimizing the Mallows distance and jointly performs clustering and regression to capture heterogeneous distributional dependencies. Model adequacy is assessed via distributional goodness-of-fit and clustering indices. The proposed model is illustrated using Spotify audio-feature distributions.

Keywords: symbolic data analysis, density-valued data, clusterwise linear regression

Symbolic Data Analysis (SDA) [1] generalizes classical data analysis to units described by complex data objects, such as intervals, histograms, or distributions. In this work, we consider density-valued variables, where the densities are estimated via a non-parametric technique, the Kernel Density Estimator (KDE). Distribution and Symmetric Distribution (DSD) regression model [2] was extended to density-valued variables; however, assuming a single global regression relationship may be restrictive when heterogeneous subpopulations exhibit distinct distributional dependencies. To address this limitation, we propose a clusterwise linear regression (CLR) framework [3] for density-valued variables, which simultaneously partitions the data and estimates cluster-specific regression models. CLR allows different regression relationships across clusters, capturing structural heterogeneity. For each unit i, let us denote $Y_i(t)$ the response density-valued variable, $X_{ij}(t)$ and $-X_{ij}(1-t)$ the predictor and correspondent symmetric distribution for $j \in \{1, \ldots, p\}$. For each cluster k, we define the regression model [2]

$$\hat{Y}_i^{(k)}(t) = v_k + \sum_{j=1}^{p} \left(a_j^{(k)} X_{ij}(t) - b_j^{(k)} X_{ij}(1-t) \right) + \epsilon_i(t), \tag{1}$$

where $v_k \in \mathbf{R}$ and $a_j^{(k)}, b_j^{(k)} \geq 0$. The similarity between observed and predicted distributions is measured using the squared Mallows distance, $D_M^2(Y_i, \hat{Y}_i) = \int_0^1 \left(Y_i(t) - \hat{Y}_i(t) \right)^2 dt$.

Given a partition $\{C_1, \ldots, C_K\}$, clustering and parameter estimation are obtained solving:

$$\begin{aligned} &\text{Minimize} \quad \sum_{k=1}^{K} \sum_{i \in C_k} D_M^2\left(Y_i, \hat{Y}_i^{(k)}\right) \\ &\text{subject to} \quad a_j^{(k)} \geq 0, b_j^{(k)} \geq 0, v_k \in \mathbf{R} \end{aligned} \tag{2}$$

We use an alternating optimization scheme: (i) assign each unit i to the cluster yielding the smallest $D_M^2\left(Y_i, \hat{Y}_i^{(k)}\right)$; (ii) update $(v_k, a^{(k)}, b^{(k)})$ by constrained least squares on a fixed grid of quantile levels; iterate until the objective stabilizes. Model adequacy is assessed using a distributional explained-variability index [4]

$$\Omega = \sum_{k=1}^{K} \frac{n_k}{n} \Omega_k, \quad \Omega_k = \frac{\sum_{i \in C_k} D_M^2\left(\hat{Y}_i^{(k)}, \bar{Y}^{(k)}\right)}{\sum_{i \in C_k} D_M^2\left(Y_i, \bar{Y}^{(k)}\right)} \tag{3}$$

where $\bar{Y}^{(k)}$ denotes the point-wise mean function of cluster k. Clustering quality is further evaluated using a silhouette-type index, S, based on Mallows distances.
The method was applied to a Spotify dataset[1] aggregated by playlist name (104 symbolic units from 4,831 tracks). Regressors were *Energy*, *Valence*, and *Duration*, with *Danceability* as the response. For $K \in \{3, \ldots, 8\}$ we used 50 random initializations. A four-cluster solution provided a good compromise between fit and clustering quality ($\Omega \approx 0.87$, $S \approx 0.54$), revealing distinct cluster-specific effects, notably for Duration and Valence. Ongoing work includes automated selection of the number of clusters, improvements in robustness, and simulation studies.

Acknowledgements This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025).

References

[1] P. Brito. Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):281–295, 2014.

[2] S. Dias and P. Brito. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, 8(2):75–113, 2015.

[3] L. Qiang, B. Adil, S. Taheri, S. Nargiz, and X. Wu. Methods and applications of clusterwise linear regression: A survey and comparison. *ACM Transactions on Knowledge Discovery from Data*, 17(3), 2023.

[4] N. K. Suresh, P. Brito, and S. Dias. Prediction of pollution levels from atmospheric variables: a study using clusterwise symbolic regression. In *Proc. RECPAD 2021*, 2021.

[1] `https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset` (accessed 2025-12-04)

10 April, 10:50 - 11:10, Lecture Theatre E1

Impact of standardization methods on quantile regression models: a comparative study

Dulce G. Pereira[1,2], **Anabela Afonso**[1,2]
[1] Universidade de Évora, Centro de Investigação em Matemática e Aplicações, dgsp@uevora.pt, aafonso@uevora.pt
[2] Universidade de Évora, Escola de Ciências e Tecnologia

Quantile regression (QR) models offer robustness against heteroscedasticity, outliers, and non-normal errors, yet the impact of standardization remains understudied. We investigate how Z-score, Range, Box-Cox, and Yeo-Johnson affect fixed and mixed-effects QR across violation scenarios. We identify when these methods affect performance, stability, and significance. Our findings provide practical R-package guidance for balancing computational efficiency with statistical validity.

Keywords: quantile regression, standardization, mixed effects, Monte Carlo simulation, R packages

---

Quantile regression (QR) models [4] provide a robust distributional alternative to mean regression, enabling modeling of different parts of the conditional distribution. The extension to mixed-effects QR [3] accommodates hierarchical data structures. While standardization of predictors is common [2, 1], its impact on QR models remains largely unexplored. Recent work on standardization in linear models shows that transformation choice significantly affects inference, particularly coefficient significance and model stability. This motivates investigating whether similar patterns emerge in QR contexts.

This study aims to compare four standardization methods—Z-score, Range [0,1], Box-Cox, and Yeo-Johnson—within fixed and mixed-effects QR frameworks. We seek to identify specific scenarios where these methods improve, maintain, or degrade model performance, ultimately providing evidence-based recommendations for applied research.

A Monte Carlo simulation study is carried out to compare seven scenarios: (1) nonlinearity with weak heteroscedasticity and outliers; (2) nonlinearity with strong heteroscedasticity and outliers; (3) weak multicollinearity; (4) strong multicollinearity; (5) zero-inflated data; (6) heavy-tailed errors; and (7) small samples.

Models are evaluated at various lower and upper quantiles. We will use the most widely-used R packages to fit fixed and mixed-effects quantile regression (QR) models, considering different parameter estimation approaches. To compare the results, we will use several metrics such as RMSE, MSE, MAB (Mean Absolute Bias), power, and CI coverage.

The key questions that guided our analysis were: *i*) Does standardization change coefficient significance in QR at rates comparable to linear models? *ii*) Are there combinations of

violations (e.g., outliers + multicollinearity) where certain transformations consistently outperform others? *iii*) How do different packages handle numerical instability? *iv*) Do the benefits of the Yeo-Johnson transformation for zero-inflated data extend to QR?
Pilot simulations suggest that parametric transformations may interact unpredictably with QR's robustness—while addressing distributional issues in least squares, they might compromise outlier resistance. The impact seems to vary by quantile, with the effect of transformations in the tails ($\tau = 0.10, 0.90$) differing from that in central quantiles. Our package survey identified variation in implementation approaches, output formats, and efficiency.
This study fills the gap in standardization methods for QR models, focusing on defining best practices to maintain QR's robustness. By characterizing how transformations affect the tails versus central quantiles, we develop practical recommendations for choosing the appropriate R packages and estimation approaches. These guidelines aim to support researchers in balancing computational efficiency with statistical validity.

Acknowledgements This work was partially funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under project UID/04674/2025 (DOI: `https://doi.org/10.54499/UID/04674/2025`).

References

[1] J. Bring. How to standardize regression coefficients. *The American Statistician*, 48(3):209–213, 1994.

[2] A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, 2008.

[3] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

[4] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

10 April, 10:10 - 10:30, Auditorium Francisco Tomatas

Using data analysis for understanding the role of social virtual reality (VR) in stroke rehabilitation

Carlos Ferreira[1], **Mariana Leite**[1], **Sérgio Oliveira**[2], **Bernardo Marques**[2], **Beatriz Sousa Santos**[2]
[1] IEETA, DEGEIT, University of Aveiro, carlosf@ua.pt, marianaleite26@ua.pt
[2] IEETA, DETI, University of Aveiro, sergiodoliveira@ua.pt, bernardo.marques@ua.pt, bss@ua.pt

Social Virtual Reality (VR) can enhance stroke survivor's motivation and adherence to rehabilitation. To validate the role and effectiveness of this immersive technology, data analysis plays a central role. This work presents a user study in the stroke rehabilitation area.

Keywords: stroke rehabilitation, social virtual reality, non-parametric tests

Stroke is a leading cause of death and disability and can have profound consequences that require effective rehabilitation to support survivors' recovery and improve their quality of life [1]. Despite their value, traditional rehabilitation methods tend to be repetitive and lack variety, which challenges survivors' motivation [2]. Additionally, these rehabilitation sessions are usually solitary, leaving survivors to practice exercises alone, which can lead to physical setbacks and social isolation [3]. To help address these challenges, a social Virtual Reality (VR) framework has been proposed. Through its collaborative nature, it can involve multiple users in the same virtual space, from stroke survivors to healthcare professionals, giving them a common goal that they must join forces to accomplish. Four serious games were designed, focused on specific gestures related to the rehabilitation of the upper limbs, thus improving physical recovery and mental well-being. The design and development were guided by a Human-Centered Design (HCD) methodology that included survivors and professionals, resulting in a user study with a total of 53 participants, 18 from a rehabilitation center. The obtained results from the user study were analyzed using exploratory, descriptive, and inferential (non-parametric tests, due to the ordinal data of the assessed dimensions). The results indicate that this social VR tool effectively boosts motivation, social interaction, and engagement while maintaining an accessible and manageable level of physical and mental demand for stroke recovery, underscoring its suitability for stroke recovery. High median scores for satisfaction and motivation (both 5, maximum), together with very low frustration (median = 1, minimum) and minimal mental effort (median = 2), suggest a positive and manageable user experience. The Mann–Whitney test assessing equality of medians showed no significant differences between samples (controlled lab setting and rehabilitation center) across most

dimensions, with small effect sizes according to Cohen's criteria, supporting the framework's broad accessibility. Notably, motivation differed between samples, with a medium effect size, indicating stronger engagement in the rehabilitation-related group. To verify the hypothesis that all four serious games were equally usable and acceptable, a comparison of participants' ratings was conducted based on the Friedman test. The test rejected the hypothesis - equality of medians (p-value = 0.003), indicating participants' ratings differed across the games, although with a very small effect size, using Cohen's criteria, as measured by Kendall's W = 0.09. The use of post hoc tests (pairwise comparison) adjusted with Bonferroni correction highlighted significant differences, with an advantage for the 'Supermarket Shopping' game (median=5). Notwithstanding, the structure of all games appears especially effective in promoting social interaction, likely due to the way tasks were framed to require mutual engagement and coordinated effort, a reflection of the human-centered approach applied with domain experts in rehabilitation.

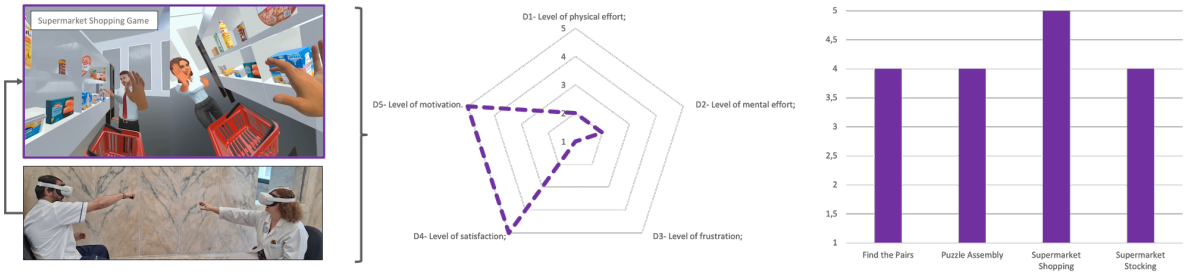

Figure 1: Social VR-based scenario (left) using shared serious games, fostering interaction, motivation, and teamwork during rehabilitation; Radar Chart representing the overall experience (center), using a Likert-type scale: 1- Low; 5- High; Bar chart representing participants' preferences among the VR-serious games (right).

Acknowledgments We thank everyone involved for their time and expertise. This work was supported by FCT through contract doi.org/10.54499/UID/00127/2025.

References

[1] B. Marques, S. Oliveira, H. Ferreira, P. Amorim, P. Dias, and B. Sousa Santos. The role of collaborative virtual reality engagement in stroke survivors' rehabilitation. *International Conference on Applied Human Factors and Ergonomics (AHFE)*, 121, 2024.

[2] S. Oliveira, B. Marques, P. Amorim, M. Leite, C. Ferreira, and B. S. Santos. Recovering through play: Studying the effects of collaborative vr serious games for stroke rehabilitation through a human-centered design methodology. *Computers & Graphics*, pages 1–15, 2025.

[3] S. Oliveira, B. Marques, P. Amorim, and B. Sousa Santos. Collaborative virtual reality serious games as a therapeutic medium for stroke survivors at a rehabilitation center. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 1051–1058, 2025.

10 April, 10:30 - 10:50, Auditorium Francisco Tomatas

Statistical analysis for predicting harvest dates based on the maturation control cycles in the Vinho Verde wine region

Mafalda T. Costa[1], **Oscar Pereira**[1], **Bruno Leitão**[1], **António Seabra**[1], **Alexander Cornejo**[2], **Catarina Beth Dantas**[3], **Maria J. Polidoro**[4]
[1] Comissão de Coordenação e Desenvolvimento Regional do Norte, anamafalda.costa@ccdr-n.pt, oscar.pereira@ccdr-n.pt, bruno.leitao@ccdr-n.pt, aseabra@ccdr-n.pt
[2] Comissão de Viticultura da Região dos Vinhos Verdes, alexander.cornejo@vinhoverde.pt
[3] Master Student Universidade Aberta, catarinacostadantas@gmail.com
[4] ESTG - Instituto Politécnico do Porto and CEAUL, mjp@estg.ipp.pt

This study evaluates the impact of extreme meteorological conditions on grape maturation in the Vinho Verde wine region over the 2015–2025 decade. The research focuses on the summer of 2025, which recorded temperatures 2.9 °C above the decade average. Using 2018 as the climatic reference year, the study aims to identify meteorological patterns and their influence on the physicochemical features of key grape varieties. The results demonstrate that the proposed model yields satisfactory predictive accuracy.

Keywords: vineyard maturation, meteorological conditions, Winkler index, regression

Following ongoing studies in the Vinho Verde wine region and the annual work conducted within the vineyard maturation control cycles, in partnership with the Comissão de Coordenação e Desenvolvimento Regional do Norte (CCDR-Norte Agricultura) and with the Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), an assessment of the last decade (2015–2025) was carried out in the Entre Douro e Minho region, considering the extreme meteorological conditions (high temperatures) recorded during the summer of 2025. The objective of this study was to identify years with similar meteorological patterns and to evaluate their influence on physicochemical features of the grapes in the Vinho Verde wine region, thereby providing additional information useful to producers for the 2025 harvest.
A preliminary analysis of meteorological data for August over the 2015–2025 decade confirmed that 2025 recorded exceptionally high mean temperatures, with a 2.9 °C increase above the decade average (21.8 °C), while August 2018 showed a 1.6 °C increase (Table 1). Using 2018 as the reference year, this study includes an assessment of the vineyard maturation control cycle during the 2018 harvest, intending to extrapolate potentially useful information for the 2025 harvest.

Table 1: Evolution of the mean, minimum, maximum, and range of the average temperatures during August from 2015 to 2025 (decade average 21.8 °C)

Year	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**
Maximum (°C)	27.3	31.6	30.3	**32.5**	28.1	28.4	28.6	29.8	30.1	30.0	**34.0**
Minimum (°C)	14.0	14.8	14.3	15.7	13.9	14.1	14.2	15.7	14.9	15.3	17.2
Mean (°C)	20.1	22.4	21.6	**23.4**	20.3	20.7	20.9	22.1	22.0	21.9	**24.7**
Range (°C)	13.3	16.7	16.0	16.8	14.2	14.3	14.5	14.1	15.2	14.7	16.8

In this study, Winkler's bioclimatic index, commonly used in viticulture, was used. The Winkler index or Growing Degree Days (GDD), calculated according to OIV guidelines [1], measures heat accumulation for vine growth between March 1 and September 30. It sums daily average temperatures above a base threshold of 10^{o}C.
In addition, from the vineyard maturation control cycle, it is possible to obtain information related to physicochemical features of the grapes: probable alcoholic content (in degrees), total acidity (in g/l), and malic acid (in g/l). These features are measured approximately every week from the moment that grapes start to change color until the harvest.
Regarding data organization, a cluster analysis was first performed to identify homogeneous groups of sub-regions based on physicochemical features. In parallel, councils and/or clusters of councils within the demarcated sub-regions were evaluated according to similarities or dissimilarities of their GDD growth curves, resulting in the identification of five sub-regions of interest: Minho and Lima, Cávado, Ave, Vouga, and Tâmega. In each sub-region and council (location), one or more grape varieties were studied (Alvarinho, Arinto, Trajadura, Loureiro, and Vinhão).
The statistical methodology adopted comprised two stages of analysis. The first stage involved the evaluation of GDD growth curves between March 1 and August 25 for the years 2018 and 2025, to compare the GDD values reached on August 25 of each year.
In the second stage, based on data obtained from the 2018 maturation control cycle, a linear regression for each physicochemical feature against GDD was fitted, and the quality of the fit was quantified using the coefficient of determination.
The methodology described yielded satisfactory results, allowing the conclusion that the adopted approach can be used to predict the evolution of vineyard maturation and harvest dates, provided that the maturation control plan is carried out at the beginning of August and that the GDD are calculated daily between March 1 and September 30.

Acknowledgements This work is funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under CEAUL Research Unit, UID/00006/2025, DOI: https://doi.org/10.54499/UID/00006/2025.

References

[1] OIV guidelines for studying climate variability on vitiviniculture in the context of climate change and its evolution. https://www.oiv.int/.

10 April, 10:50 - 11:10, Auditorium Francisco Tomatas

Tools for cohesion policies: a composite indicator of socio-spatial vulnerability for the municipalities of Portugal

Aitor Varea Oro[1], **Guilherme Vara**[2], **Sílvia Jorge**[3], **Pietro Meirelles Brites**[3], **Rui Barros**[4], **Rita Gaio**[5]
[1] Faculdade de Arquitetura da Universidade do Porto - Centro de Estudos Nuno Portas, aitorvarea@arq.up.pt
[2] CEGIST, Instituto Superior Técnico, Universidade de Lisboa, guilherme.vara@tecnico.ulisboa.pt
[3] CiTUA- IST-ID, silviajorge@tecnico.ulisboa.pt, pietrombrites@tecnico.ulisboa.pt
[4] Universidade Lusófona - Faculdade de Comunicação, Arquitetura, Artes e Tecnologias de Informação, p7024@ulusofona.pt
[5] Centro de Matemática da Universidade do Porto & Departamento de Matemática da Faculdade de Ciências da Universidade do Porto, argaio@fc.up.pt

---

This study proposes a composite indicator to evaluate socio-spatial vulnerability across 308 Portuguese municipalities. By synthesizing 49 variables into a hierarchical model, we compare expert-based weighting against Principal Component Analysis (PCA). The framework is integrated into an interactive platform to support evidence-based policy formulation and resource allocation.

Keywords: composite indicators, principal component analysis (PCA), territorial vulnerability, spatial analysis, decision support

---

Assessing territorial vulnerability is essential for effective regional planning and resource allocation. This paper proposes a composite indicator of vulnerability designed to evaluate and classify the 308 municipalities of Portugal. The study establishes a hierarchical aggregation model that synthesizes 49 variables into a single vulnerability score, enabling the ordinal classification of municipalities.

The framework is structured around eight sub-dimensions, ranging from socio-demographics, economics, and territorial cohesion to housing conditions and natural risks. These are further aggregated into two primary dimensions: (1) Territory, Population, and Housing, assessing the interplay between demographic dynamics, socio-economic status, and the adequacy of the housing stock; and (2) Environment, Energy, and Risks, evaluating the quality of the surrounding habitat, resource sustainability, and resilience against natural hazards.

To ensure robust results, the study compares two weighting strategies. The first relies on a panel of experts applying equal weights within each dimension. The second uses

Principal Component Analysis (PCA) to generate data-driven weights based on statistical patterns. This allows the research to contrast subjective expert opinions with objective mathematical profiles. A contribution of this research is the development of an interactive decision-support platform. This tool allows users to visualize the impact of these different weighting profiles or to dynamically adjust their own preferences, empowering political decision-makers to simulate scenarios based on either statistical evidence or subjective policy priorities. By visualizing how specific preferences impact the spatial distribution of vulnerability, the platform facilitates the extraction of actionable, tailored insights for public policy design.

The results indicate that PCA-based selections enhance separability between vulnerability profiles, reducing within-group dispersion and improving the interpretability of territorial clusters. This data-driven segmentation provides a statistically grounded baseline that can subsequently be adjusted through expert-driven weighting schemes. By identifying municipalities with coherent vulnerability structures, the framework enables the design of tailored policy interventions while preserving methodological transparency. Beyond its empirical application to Portugal, the proposed aggregation structure constitutes a replicable tool for cohesion policy design in multi-level governance contexts.

References

[1] B. Beccari (2016). A comparative analysis of disaster risk, vulnerability and resilience composite indicators. *PLoS Currents*, 8.

[2] V. Sebestyén, A. J. Trájer, E. Domokos, A. Torma and J. Abonyi (2024). Objective well-being level (OWL) composite indicator for sustainable and resilient cities. *Ecological Indicators*, 158, 111460.

[3] M. Rodrigues and M. Franco (2020). Measuring the urban sustainable development in cities through a composite index: the case of Portugal. *Sustainable Development*, 28(4), 507-520.

[4] Ş. Kılkış (2016). Sustainable development of energy, water and environment systems index for Southeast European cities. *Journal of Cleaner Production*, 130, 222-234.

10 April, 17:40 - 18:00, Lecture Theatre E1

Model selection with Group LASSO in finite mixture linear regression models

Ana Moreira[1], **Susana Faria**[1]
[1] Centre of Mathematics (CMAT), Department of Mathematics, University of Minho
id10866@uminho.pt, sfaria@math.uminho.pt

Variable selection is a crucial step in model building, as it determines which explanatory variables are included to explain or predict the dependent variable. When dealing with categorical variables, it is particularly important to preserve the inherent group structure of the data. In this study, we address the problem of variable selection in mixtures of linear regression models by employing penalized maximum likelihood estimation. Specifically, we consider group-based penalization methods, including the Group LASSO, and the Adaptive Group LASSO.

Keywords: group lasso, mixtures of linear regression models, penalized maximum likelihood estimation, simulation study

Finite Mixture Regression (FMR) models provide a flexible tool for modelling 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 and their contributions to explaining or predicting the dependent variable vary from component to component in the mixture model. For this reason, variable selection assumes great importance for mixture models, something particularly noticeable in recent years [1]. [2] proposed the Least Absolute Shrinkage and Selection Operator (LASSO) estimator, which penalizes the $\ell 1$-norm of the regression coefficients. Other commonly used penalty functions include the Smoothly Clipped Absolute Deviation (SCAD), the Adaptive LASSO, and the Elastic Net.

Among the methods previously discussed, group variable selection methods are particularly attractive when dealing with categorical data, as they preserve the inherent group structure of the predictors, that is, ensuring that entire groups of related variables are either selected or excluded together. [4] introduced several group penalization approaches, including the Group LASSO, the Group Least Angle Regression (Group LARS) and the Group Garrote. In addition, [3] introduced the Adaptive Group LASSO. This study focuses on the problem of variable selection in mixtures of linear regression models. We compare the performance of three penalization methods in identifying the most relevant groups of explanatory variables: the LASSO, the Group LASSO and the Adaptive Group LASSO. To compare these variable selection methods, we conduct an extensive simulation study

across a range of scenarios. Particularly, we consider scenarios where the number of groups increases with the sample size. We also examine the impact of varying the number of mixture components, the distribution used to generate the explanatory variable, incorporating different levels of correlation among the variables, the regression coefficients, and the level of dispersion in the data. To evaluate the performance of these methods, we consider the following criteria: (i) the median number of groups that were correctly estimated with zero coefficients, corresponding to the true zero coefficients set to zero; and (ii) the median number of groups incorrectly estimated with zero coefficients, referring to true nonzero coefficients that were incorrectly set to zero. Finally, we assess the predictive performance by computing the average twice negative log-likelihood (predictive log-likelihood loss) on an independent test set. The simulation study was conducted using the 4.5.2 version of R programming language. Several R packages were used in the analysis including `flexmix`, `glmnet`, and `grpreg` packages, among others.
We apply the developed methodologies to real datasets, specifically to both low-dimensional and high-dimensional cases.
We conclude that the Adaptive Group LASSO outperforms both the Group LASSO and the LASSO in variable selection across both settings, and it also achieves superior predictive performance in low-dimensional settings.

Acknowledgements This research at CMAT was supported by FCT - Fundação para a Ciência e a Tecnologia, I.P. by project reference 2022.12256.BD and https://doi.org/10.54499/2022.12256.BD identifier. The research of the authors was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013/2025 https://doi.org/10.54499/UID/00013/2025.

References

[1] Z. Lu and W. Lou. Bayesian approaches to variable selection in mixture models with application to disease clustering. *Journal of Applied Statistics*, 50(2):387–407, 2023.

[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] H. Wang and C. Leng. A note on adaptive group lasso. *Computational statistics & Data Analysis*, 52(12):5277–5286, 2008.

[4] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

10 April, 18:00 - 18:20, Lecture Theatre E1

Structural equation modeling to analyze the impact of the innovation incentive system on the internationalization of Portuguese companies

Luís M. Grilo[1,2,3], **Elsa J. Pereira**[4], **Jean P. Maidana**[5,6], **Milan Stehlíkh**[6,7]
[1]Department of Mathematics, University of Évora, Portugal, luis.grilo@uevora.pt
[2]CIMA (Center for Research in Mathematics and Applications), University of Évora, Évora, Portugal
[3]NOVA Math (Center for Mathematics and Applications), NOVA University of Lisbon, Portugal
[4]Polytechnic Institute of Tomar, Tomar, Portugal, Portugal; elsapereirafazendeiro@gmail.com
[5]Facultad de Ingeniería, Universidad Andres Bello, Valparaíso, Chile, jean.maidana@unab.cl
[6]Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile, mlnstehlik@gmail.com
[7]University of Applied Sciences Upper Austria, Austria

Small and medium enterprises (SMEs) in Portugal received innovation incentives under the QREN 2014–2020, aimed at supporting internationalization. In the proposed structural equation model, the latent constructs 'Organizational Innovation', 'Product Innovation', 'Marketing Innovation', and 'Working Conditions' are identified as potential predictors of 'companies' internationalization'. In the model estimated, using the consistent Partial Least Squares (PLSc) estimator, 'Working Conditions' exhibit the strongest impact, indicating that measures such as wage increases and the reduction of temporary employment have a highly significant direct effect on the internationalization of Portuguese SMEs.

Keywords: non-normal data, ordinal variables, semicontinuous covariance, tetra confirmatory analysis

---

In 2014, Portugal entered into a Partnership Agreement with the European Commission—known as Portugal 2020—for the 2014–2020 period. This agreement established the programming framework for implementing the Europe 2020 Strategy and defined the economic, social, environmental, and territorial policies required to promote a more competitive and prosperous country [1]. Previously, under the Innovation Strategy of the National Strategic Reference Framework (QREN) 2007–2013 [3], the promotion of innovation within the business models of Portuguese small and medium-sized enterprises (SMEs) had

already been identified as one of the key instruments contributing to business competitiveness and internationalization, particularly among SMEs. This study aims to analyze and understand the impact of the Innovation Incentive System on the internationalization of Portuguese SMEs. The proposed/theoretical Structural Equation Model (SEM) included reflexive and formative latent constructs (the tetra confirmatory analysis algorithm was used) and the sample data (120 companies) were collected through a questionnaire (using the 5-point Likert scale), considering the geographical heterogeneity and the different sectors of activity of the companies, as well as the fact that they all benefited from the same incentive system. The consistent Partial Least Squares (PLSc) estimator was applied, since it is a flexible approach (a non-parametric technique that makes no distribution assumptions and works well with small sample sizes). As main results, it was found that 'Organizational Innovation' (exogenous latent construct) and 'Product Innovation' demonstrated indirect effects on the 'Internationalization of Companies' (target construct of this study). With statistically significant direct positive effects, the constructs 'Marketing Innovation' and 'Working Conditions' emerged, with the latter showing the greatest impact on the 'Internationalization of Companies' [2]. The stochastic regularity conditions for the PLSc-SEM estimator were also discussed in this study from the perspective of semicontinuous covariance, where jumps in covariances can occur. These sudden changes/shifts are very common in real-world data (non-smooth data), especially involving human systems (such as SMEs). The estimator used adequately handles these discontinuities because it does not assume perfect continuity—it can accommodate these jumps or irregularities in the data structure [2, 4].

Acknowledgements This work is funded by national funds through the FCT - Foundation for Science and Technology, I.P., under the scope of the project UID/4674/2025, DOI https://doi.org/10.54499/UID/04674/2025.

References

[1] Governo de Portugal. *Acordo de Parceria 2014-2020 Portugal 2020*. Documentação oficial do governo de Portugal, Lisboa, Portugal, 2014.

[2] L. M. Grilo, E. J. Pereira, J. P. Maidana, and M. Stehlík. On stochastic aspects of impact modeling of the innovation incentive system and business internationalization: evidence from portuguese SMEs. *Stochastic Analysis and Applications*, 42(1):20–44, 2024.

[3] QREN. *Plano global de avaliação do QREN e dos programas operacionais 2007-2013*. QREN Portugal, Lisboa, Portugal, 2011.

[4] M. Stehlík, Ch. Helpersdorfer, P. Hermann, J. Šupina, L. M. Grilo, J. P. Maidana, F. Fuders, and S. Stehlíková. Financial and risk modelling with semicontinuous covariances. *Information Sciences 394-395*, 2017.

10 April, 17:40 - 18:00, Auditorium Francisco Tomatas

Detecting VAT fraud: an approach based on temporal analysis of bank transactions

Ana Helena Tavares[1,2], **João Marques**[3]
[1] ESTGA - Águeda School of Technology and Management, Portugal, ahtavares@ua.pt
[2] CIDMA - Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal
[3] Tax and Customs Authority, Portugal, joaoaraujomarques@gmail.com

---

This work proposes a statistical algorithm for anomaly detection in forensic contexts, focusing on bank movements that evidence VAT fraud and money laundering. The methodology combines time and value criteria to identify rapid capital circulation chains. Tested with simulated data from real-world patterns, the algorithm effectively filtered suspicious transactions and flagged complex structures, proving how data analysis supports the investigation of hard-to-detect financial crimes.

Keywords: VAT fraud, anomaly detection, network analysis, robust statistic, transactional data

---

VAT fraud, specifically through Missing Trader Intra-Community (MTIC) schemes or derivatives like *carousel fraud*, represents a critical challenge for tax authorities [2]. These schemes are extremely complex, involving multiple shell entities and billions of euros in tax losses. They are organized into sophisticated criminal networks that operate in a coordinated manner through thousands of simulated operations. These include the rapid circulation of capital through multiple bank accounts and other payment methods, creating highly complex transactional structures designed to hide the origin of funds and hinder conventional forensic analysis [3]. Recent literature emphasizes that detecting such patterns requires advanced network-based approaches to map the connectivity and flow between fraudulent agents[1].
This paper presents a methodology for the automated detection of suspicious networks by integrating exploratory data analysis and relational mapping. The procedure is structured into four distinct phases:

1. Robust data filtering: application of thresholding techniques based on the Interquartile Range to isolate monetary outliers and prune non-relevant transactions;

2. Relational structure detection: Identification of directed relationships between entities through recursive self-joins of transactional records;

3. Temporal latency analysis: integration of connectivity constraints to identify transaction pairs with immediate temporal proximity;

4. Path extraction and deduplication: isolation of multi-level circulation chains, followed by a sequence-hashing algorithm to eliminate structural redundancies and ensure model parsimony.

The proposed methodology was validated using a synthetic dataset modeled from empirical MTIC fraud typologies, with approximately 76,000 transactions and 99 bank accounts. The dataset incorporates heterogeneous behavioral profiles corresponding to specific operational roles (missing traders, conduits, buffers, and brokers) and ultimate beneficial owners (UBOs), who are usually involved in money laundering strategies. To evaluate the procedure's sensitivity, illicit patterns were embedded within a high volume of legitimate activities, effectively masking anomalous signals within the global distribution. This structure served to test the algorithm's capability to isolate latent risk structures from background noise in a complex transactional environment.
The procedure successfully identified 200 short-length chains and 8 high-complexity structures consisting of 10 sequential transactions. Statistical analysis of these networks revealed sophisticated layering patterns, with time intervals indicating deliberate coordination and transaction values significantly exceeding the overall average.
The approach proved effective in mitigating the dimensionality and scale problems inherent in financial big data, allowing investigators to focus on high-risk structures. Future developments include the integration of unsupervised clustering to identify *fraud families* and the use of multivariate outlier detection to enhance sensitivity to emerging patterns of economic crime.

Acknowledgements This work is supported in part by CIDMA (https://ror.org/05pm2mw36) under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58), Grants UID/04106/2025 (https://doi.org/10.54499/UID/04106/2025) and UID/PRR/04106/2025 (https://doi.org/10.54499/UID/PRR/04106/2025).

References

[1] A. Alexopoulos, P. Dellaportas, S. Gyoshev, C. Kotsogiannis, S. C. Olhede, and T. Pavkov. A network approach to detect value added tax fraud. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnaf205, 12 2025.

[2] P. Baert. Filling the gap: The EU's fight against VAT fraud. Briefing, European Parliament, January 2025. PE 757.632.

[3] European Public Prosecutor's Office. 2024 annual report: Investigation "moby dick". Technical report, EPPO, Luxembourg, 2024.

10 April, 18:00 - 18:20, Auditorium Francisco Tomatas

The challenge of hidden outliers: a robust approach to panel data

Anabela Rocha[1], **Cristina Miranda**[1], **Manuela Souto de Miranda**[2]
[1] ISCA, CIDMA, University of Aveiro, Portugal, anabela.rocha@ua.pt, cristina.miranda@ua.pt
[2] CIDMA, University of Aveiro, Portugal, manuela.souto@ua.pt

Real-world data often contain outliers, requiring robust methods for detection and accurate estimation. This study focuses on panel data, accounting for individual and temporal effects. The authors propose a robust estimator for random effects models by integrating robust procedures into the Feasible Generalized Least Squares (FGLS) approach. Simulations and a case study demonstrate this proposal's advantages over traditional methodologies.

Keywords: casewise outlier, cellwise outlier, FGLS, panel data, robust methods

Panel data (PD) are suitable for econometric studies as well as for research in other fields. PD is a set of observations of multiple variables for different units (e.g., countries, firms, regions, individuals), collected over several time periods (e.g., day, week, month). The statistical analysis of PD allows to identify and measure effects that would not be detected in a cross-sectional nor in a time-series analysis. Estimating the parameters of a panel data model is often performed using the Feasible Generalized Least Squares (FGLS) estimator. An outlier is an observation that deviates from the pattern followed by most of the data. A data matrix can contain two types of outliers [4]: Casewise outliers which are observed when a case is atypical, or Cellwise outliers, when most of the data cells in a row are similar, but some are atypical [1]. Detecting outliers and employing robust methods in panel data is crucial for reliable analysis and solid conclusions [2, 3]. We propose robust methods for outlier detection and for fitting panel data models.
The proposed robust method can be applied according to the RFGLS algorithm, given by:

- Estimate the pooled model parameters using the robust regression method - Least Trimmed Squares (LTS), and compute the residuals.
- Estimate the error covariance matrix using the robust covariance estimator - cellMCD, from the LTS residuals.
- Identify outliers in the original data matrix with the Detect deviating cells (DDC), and derive the imputed data matrix.
- Estimate the model parameters using FGLS, based on the imputed data matrix and the robust covariance matrix estimate obtained in the previous steps.

The performance of the proposed method was assessed based on the RMSE (root mean square error) and RMSEP (root mean square error of prediction) values, respectively, for a simulation study (contaminated samples with outliers by changing only y and both y and x, and with percentages of contamination 0%, 5% and 10%), and for a real data (Grunfeld data). The results obtained were as follows.

Table 1: Simulation study - RMSE values

	C0	C5-y	C5-y and x	C10-y	C10-y and x
FGLS	0.15	1.26	2.72	1.73	2.82
RFGLS	0.21	0.22	0.21	0.25	0.19

Table 2: Grunfeld data - RMSEP values

	FGLS	RFGLS
RMSEP	106.49	16.44

The simulation study and Grunfeld data showed that the RFGLS estimator improves upon FGLS, as expected in the presence of various types of outliers. Robust methods clearly outperform classical methods in both, outlier detection and parameter estimation in the presence of outliers.

Acknowledgements This work is supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58) Multi-Annual Financing Program for *R&D* Units, grants UID/4106/2025 and UID/PRR/4106/2025.

References

[1] A. Fatemah, Van A. Stefan, Victor J. Y., and Ruben H. Z. Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331, 2009.

[2] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, jun 2006.

[3] J. Raymaekers and P. J. Rousseeuw. Challenges of cellwise outliers. *Econometrics and Statistics*, 2024.

[4] P. J. Rousseeuw and W. Van Den Bossche. Detecting Deviating Data Cells. *Technometrics*, 60(2):135–145, apr 2018.

11 April, 9:00 - 9:20, Auditorium Francisco Tomatas

Were European funds well distributed through the Portuguese municipalities or was it unfair?

Guilherme Vara[1], Aitor Varea Oro[2], Sílvia Jorge[3], Pietro Meirelles Brites[3], Rui Barros[4], Rita Gaio[5]

[1] CEGIST, Instituto Superior Técnico, Universidade de Lisboa, guilherme.vara@tecnico.ulisboa.pt

[2] Faculdade de Arquitetura da Universidade do Porto - Centro de Estudos Nuno Portas, avoro@arq.up.pt

[3] CiTUA- IST-ID, silviajorge@tecnico.ulisboa.pt, pietrombrites@tecnico.ulisboa.pt

[4] Universidade Lusófona - Faculdade de Comunicação, Arquitetura, Artes e Tecnologias de Informação, p7024@ulusofona.pt

[5] Centro de Matemática da Universidade do Porto & Departamento de Matemática da Faculdade de Ciências da Universidade do Porto, argaio@fc.up.pt

---

This study analyzes the allocation of Portugal's Recovery and Resilience Plan (PRR) housing funds. Using a linear optimization model, we simulate distribution scenarios across 308 municipalities to balance stock expansion and rehabilitation. The models are integrated into an interactive platform, allowing decision-makers to visualize how different priorities impact budget efficiency and regional equity.

Keywords: optimization models, resource allocation, housing policy, regional disparities, decision support

---

The allocation of funds from the Recovery and Resilience Plan (PRR) represents a critical opportunity to mitigate the housing crisis in Portugal. However, distributing limited financial resources across 308 municipalities with distinct socio-economic realities requires navigating complex trade-offs. This paper presents an Integer Linear Programming (ILP) model developed to simulate and justify optimal strategies for allocating PRR housing funds under strict budgetary constraints.

The core of the mathematical model is formulated as:

$$\text{Maximize } Z = \sum_{i=1}^{308} \sum_{j=1}^{5} w_j x_{ij}$$

where x_{ij} represents the integer number of dwellings allocated to municipality i using housing operation j, and w_j denotes the policy-driven weight assigned to each operation type.

The objective function is subject to the following constraints:

Global budget constraint

$$\sum_{i=1}^{308}\sum_{j=1}^{5} PCost_{ij}\, x_{ij} \leq Budget$$

Local housing needs constraint

$$0 \leq x_{ij} \leq Needs_{ij}$$

Integrality condition

$$x_{ij} \in Z$$

The study examines a spectrum of distribution scenarios governed by adaptable objective weights, ranging from purely quantitative goals, such as maximizing the total number of new dwellings, to multi-criteria strategies. These strategies incorporate additional socio-economic rules, including: (i) restricting new construction in municipalities with high housing vacancy rates; (ii) enforcing wealth-based proportionality to ensure equitable investment between affluent and economically vulnerable municipalities; and (iii) imposing minimum allocation thresholds for high-pressure markets, namely the Metropolitan Areas, District Capitals, and the Algarve.

Through these mechanisms, the model balances housing stock expansion and rehabilitation targets. Each scenario is analyzed to reveal the trade-offs, efficiency gains, and potential distributive inequities embedded in alternative allocation logics.

To facilitate evidence-based policy formulation, the model is integrated into an interactive online platform. This tool enables decision-makers to visualize how different policy priorities alter capital distribution patterns, demonstrating that although no single optimal political solution exists, mathematical optimization can significantly enhance budget efficiency and territorial equity.

References

[1] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA, 1997.

[2] U. Hasanah, S. Putrawangsa, and D. T. Kumoro. Applying linear programming in business decision making: A case of profit maximization of a commercial housing development. *European Journal of Business and Management*, 11:19, 2019.

[3] R. Mansini, W. Ogryczak, and M. G. Speranza. *Linear and Mixed Integer Programming for Portfolio Optimization*. Springer, Cham, 2015.

11 April, 9:20 - 9:40, Auditorium Francisco Tomatas

A robo-advisor based on the Markowitz model and investor risk profiling

Manuel Rodrigues[1], **Conceição Amado**[2]
[1] IST, ULisboa, manuelmariavilasrodrigues@gmail.com
[2] CEMAT, IST, ULisboa, conceicao.amado@tecnico.ulisboa.pt

The growing accessibility of financial markets has allowed more individuals to invest in diverse instruments. However, this democratization has not been matched by increased financial literacy, often leading to poor risk management. In this context, robo-advisors have emerged as automated tools providing portfolio recommendations based on quantitative models. This work develops a robo-advisor using the Markowitz model, aligned with the investor's risk profile.

Keywords: robo-advisor, portfolio optimization, Markowitz model, risk profiling, asset allocation

The proposed robo-advisor combines principles from Modern Portfolio Theory with a structured assessment of investor risk preferences to deliver personalized portfolio recommendations. The core of the system relies on the mean–variance portfolio optimization model introduced by [1], in which portfolio selection results from balancing expected return and risk, measured through the variance of returns. To incorporate individual preferences, a risk-profiling questionnaire was developed, allowing investors to be classified along a spectrum ranging from conservative to aggressive. In this work, the investor's risk score obtained from the questionnaire is normalized to the interval $[0, 1]$ and then mapped linearly to a risk-aversion coefficient $\gamma \in [0.1, 5.0]$. This parameter enters a quadratic mean–variance utility function of the form

$$U(w) = w^\top \mu - \frac{\gamma}{2} w^\top \Sigma w,$$

which is maximized under standard portfolio constraints (non-negative weights summing to one). Here μ denotes the vector of expected asset returns, Σ the corresponding covariance matrix, and w the portfolio weight vector.

A diversified universe of financial assets was considered to reflect the main investment opportunities available in global markets. This universe includes individual equities, exchange-traded funds representing different asset classes such as equities, bonds, commodities and real estate, as well as a limited set of cryptocurrencies to capture higher-risk assets. Historical price data for all assets were collected from publicly available sources, namely [3], and processed to obtain consistent time series of returns. The data treatment phase includes

the estimation of expected returns and the covariance structure between assets, which together form the fundamental inputs of the optimization model.
Portfolio optimization is carried out by maximizing a utility function that balances expected return and portfolio risk, subject to standard constraints such as full capital allocation and non-negativity of asset weights. In addition, the model allows for user-defined constraints, including the exclusion of specific assets or the imposition of maximum allocation limits by asset class, increasing the flexibility and realism of the recommendations. The resulting portfolios are evaluated using commonly adopted performance indicators, namely expected return, volatility and the Sharpe ratio [2], which provides a measure of risk-adjusted performance.
To assess the behaviour of the robo-advisor, several investor profiles were simulated. The results show that the system consistently adapts its recommendations to the investor's level of risk aversion. Conservative profiles lead to portfolios characterized by lower volatility and more stable returns, while aggressive profiles are associated with higher expected returns at the cost of increased risk. Moderate profiles tend to achieve the most favourable balance between risk and return, in line with the predictions of Modern Portfolio Theory.
The robo-advisor was implemented as an interactive web application using Python, allowing users to input their preferences and visualize portfolio recommendations in a clear and intuitive manner. Despite inherent limitations related to the reliance on historical data and simplified modelling assumptions, the proposed framework demonstrates how classical portfolio optimization models can be effectively combined with modern computational tools to support informed investment decision-making.

References

[1] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952.

[2] W. F. Sharpe. Mutual fund performance. *The Journal of Business*, 39:119–138, 1966.

[3] Yahoo Finance. AAPL historical data, 2025. Accessed: 2025-05-21.

11 April, 9:40 - 10:00, Auditorium Francisco Tomatas

Exact sparsity control for multiclass linear support vector Mmchines

Immanuel Bomze[1], Laura Palagi[2], Bo Peng[3], Pedro Duarte Silva[4], Federico D'Onofrio[2], Marta Monaci[2]

[1] Faculty of Mathematics and Research Network Data Science, University of Vienna, immanuel.bomze@univie.ac.at

[2] Department of Computer, Control and Management Engineering, Sapienza University of Rome, laura.palagi@uniroma1.it

[3] VGSCO, University of Vienna, bo.peng@univie.ac.at

[4] Universidade Católica Portuguesa, Católica Porto Business School, Research Centre in Management and Economics, psilva@ucp.pt

---

We study the embedded feature selection problem in linear multiclass classification Support Vector Machines. We propose two explainable mixed-integer models with cardinality constraints. Novel semidefinite relaxations are introduced, and decomposed into equivalent scalable relaxations in a much smaller cone. An exact procedure is proposed as well as heuristics that use the information of its optimal solution. Numerical results on benchmarking datasets show the effectiveness of our approach.

Keywords: machine learning, multiclass classification, interpretable AI, fairness, zero-norm

---

Recent advances in Machine Learning have lead to the pervasiveness of AI applications that rely on highly accurate classification algorithms. In the particular case of multiclass classification problems, while state of art deep learning neural networks and kernel based Support Vector Machines often excel in minimizing expected error rates, they also work as uninterpretable black boxes, that cannot be use in many problems with strict transparency constraints. In this work, we will propose a sparse linear multiclass Support Vector Machines, given by decisions rules of form

$$\hat{y} = max_{c\in\mathcal{Y}}\, g_c(\mathbf{x}) \quad with \quad g_c(\mathbf{x}) := \mathbf{w_c}^\top \mathbf{x} + b_c \tag{1}$$

where $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is a training set of n examples, $\mathbf{x}_i$ is an attribute vector belonging to some domain, $\mathcal{X} \subseteq \Re^d$, and the label y_i is an integer belonging to the set $\mathcal{Y} = \{1, \ldots, k\}$, and all pairs $(\mathbf{x}_i, y_i)$ were independently generated from some unknown, but common, probability distribution, $P(\mathbf{X}, Y)$.

The learned parameters, $\mathbf{w_c}^\top$ and b_c are, respectively, the cth row and cth element of a matrix of coefficients $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]^\top \in \Re^{k\times d}$, and a vector of bias terms $\mathbf{b} = [b_1, \ldots, b_k]^T \in \Re^k$ given by solution of the following the optimization problem

$$min_{\mathbf{W}\in\Re^{k\times d},\mathbf{b}\in\Re^k} \quad \tfrac{1}{2}\,\lambda\,||\mathbf{W}||^2 + \textstyle\sum_{i=1}^{n} L(\mathbf{g}(\mathbf{x}_i), y_i) \tag{2}$$

$$\textit{subject to} \quad \textstyle\sum_{c\in\mathcal{Y}} \mathbf{w}_c = \mathbf{0_d} \tag{3}$$

$$\textstyle\sum_{c\in\mathcal{Y}} b_c = 0 \tag{4}$$

$$||\mathbf{W}||_{1,0} \leq B, \tag{5}$$

where λ is a regularization hyper-parameter, $||.||^2$ denotes the squared Frobenius norm, and $L(.,.) : \Re^k \times \mathcal{Y} \mapsto \Re$ is a multi-class large margin loss function (see *e.g.* [3]).
To promote column (i.e., feature) sparsity in matrix $\mathbf{W}$ we define the column sparsity pseudo-norm as: $||\mathbf{W}||_{1,0} := |\{j : \|\mathbf{w}_j\|_1 \neq 0\}|$ which counts the number of nonzero columns in $\mathbf{W}$.
Most existing approaches to reduce the number of features involved in the classifier, either add surrogate regularizers to the chosen SVM criteria, or employ heuristic algorithms such as multiclass generalizations (e.g. [4]) of the popular Recursive Feature Elimination (RFE-SVM) algorithm [2]. However, this approach sometimes does not work as intended.
In this presentation, following [1], we will discuss a class of sparse multiclass linear SVMs, that make a rigorous explicit control on the cardinality of feature set employed.
In particular, to handle the NP-hard problem (2)-(5), we first introduce two mixed-integer formulations for which novel semidefinite relaxations are proposed. Exploiting the sparsity pattern of the relaxations, we decompose the problems and obtain equivalent relaxations in a much smaller cone, making the conic approaches scalable. To make the best usage of the decomposed relaxations, we propose heuristics using the information of its optimal solution. Moreover, an exact procedure is proposed by solving a sequence of mixed-integer decomposed semidefinite optimization problems. Numerical results on classical benchmarking datasets are reported, showing the efficiency and effectiveness of our approach.

Acknowledgements Financial support from the Fundação para a Ciência e Tecnologia (through project UID/00731/2025: Research Centre in Management and Economics and DOI https://doi.org/10.54499/UID/00731/2025) is gratefully acknowledged.

References

[1] I. Bomze, F. D'Onofrio, L. Palagi, and B. Peng. Feature selection in linear support vector machines via a hard cardinality constraint: A scalable conic decomposition approach. *European Journal of Operational Research*, 2025.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

[3] D. Ürün, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(1):1550–1581, 2016.

[4] X. Zhou and D. Tuck. Msvm-rfe: extensions of svm-rfe for multiclass gene selection on dna microarray data. *Bioinformatics*, 23(9):1106–1114, 2007.

11 April, 10:00 - 10:20, Auditorium Francisco Tomatas

maxRgain: R package for optimizing genetic gains in the selection of groups of genotypes

Sónia Surgy[1], **Jorge Cadima**[2], **Elsa Gonçalves**[1]
[1] LEAF Research Center, Instituto Superior de Agronomia, Universidade de Lisboa, soniasurgy@isa.ulisboa.pt, elsagoncalves@isa.ulisboa.pt
[2] Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal, jcadima@isa.ulisboa.pt

Polyclonal selection aims to maximize overall genetic gain across multiple traits through the selection of genotype groups. This work presents the R package maxRgain, which implements an integer programming method for multi-trait polyclonal selection based on predictors of genotypic effects. The package includes real data and reproducible examples, facilitating the application of the method in breeding programs.

Keywords: integer programming, polyclonal selection, genetic selection, genetic gains

Polyclonal selection is a methodology developed in Portugal within the framework of genetic improvement of traditional grapevine varieties, consisting of the selection of a group of clones that maximizes overall genetic gain across multiple traits. By considering the collective performance of the group, this approach preserves genetic diversity and enhances stability in the face of environmental variability. In contrast to traditional selection based on indices applied to individual genotypes, polyclonal selection requires methods that simultaneously integrate multiple criteria and constraints. In this context, the problem is formulated as a mathematical optimization based on empirical best linear unbiased predictors (EBLUPs) of genotypic effects obtained from the fitting of linear mixed models, aiming to maximize the predicted genetic gain of the group, subject to user-defined constraints. The objective function is [2] $z = \sum_{i=1}^{n} \sum_{k=1}^{p_o} y_{k_i}\, x_i$ subject to $\sum_{i=1}^{n} x_i = s$ and

$$\sum_{i=1}^{n} y_{k_i} x_i \geq l_k \text{ (or } \leq l_k \text{)} \qquad \text{for } k = 1, \ldots, p_c \text{ and } l_k = R_k \times s/100$$

where n is the total number of genotypes, y_{k_i} is the normalized EBLUPs of genotypic effects of clone i in trait k, x_i is a binary variable (1 if genotype i is selected; 0 otherwise), s number of genotypes to be selected, R_k is the minimum desired genetic gain for trait k. The maxRgain package (https://CRAN.R-project.org/package=maxRgain) implements this method efficiently and accessibly, using the lpSolve package [1] for optimization. A real dataset of the Gouveio grapevine variety is included, with EBLUPs of genotypic effects for 150 genotypes, ensuring reproducibility of the examples. The package provides three main functions (Table 1), allowing the selection of groups of different sizes and flexible definition

of constraints and gain criteria. A brief example of *polyclonal()* is shown below, selecting two groups (7 and 8 clones) with minimum gains of 0 for yield (yd) and 10 berry weight (bw); the output displays genetic gains ("$gain") and selected clones ("$selected").

Table 1: Main functions of package maxRgain

Function	Description
polyclonal()	Maximizes the predicted genetic gains in the selection of genotype groups based on predictors of genotypic effects.
rmaxp()	Returns the maximum possible gain for each trait in each group. Only the constraint regarding the number of genotypes to be selected is applied.
rmaxa()	Returns the maximum possible gain for each trait in each group, without causing any loss in the others. Constraints are applied to the various traits with $l_k \geq 0$.

```
polyclonal(traits = c("yd", "bw"), ref = "Clone", clmin = 7, clmax = 8,
      dmg = data.frame( lhs = c("yd", "bw"), rel = c(">=", ">="), rhs = c(0, 10) ),
      meanvec = c(yd = 3.517, bw = 1.653), criteria = c(yd = 1, bw = -1),
      data = Gouveio )

Predicted genetic gains as a % of the overall mean
$gain
Group.Size       yd       bw
         8 25.10892 10.13044
         7 27.11125 10.13657

Selected genotypes (per group size)
$selected
    8     7
GV038 GV038
GV080 GV094
```

Further examples are included in the package documentation illustrating the practical application of the method, including full output descriptions and interpretation of the results, supporting transparency and reproducibility.

Acknowledgements To the Fundação para a Ciência e Tecnologia (FCT), for BD https://doi.org/10.54499/2020.07338.BD and projects UID/04129/2025 and 'BioGrape-Sustain' (C644866286-011, PRR – Agendas Mobilizadoras, B6.1).

References

[1] M. Berkelaar and G. Csárdi. lpsolve: Interface to lp solve v. 5.5 to solve linear/integer programs. R package version 5.6.20. 2024.

[2] S. Surgy, J. Cadima, and E. Gonçalves. Integer programming as a powerful tool for polyclonal selection in ancient grapevine varieties. *Theoretical and Applied Genetics*, 138:122, 2025.

11 April, 9:00 - 9:20, Lecture Theatre E1

Discriminant analysis 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

---

We study directional data restricted to the positive orthant of the unit hypersphere, for which a folded von Mises-Fisher distribution is appropriate. We consider the maximum likelihood estimation for this distribution using the EM algorithm and evaluate the properties of the resulting estimates through simulation. Additionally, we compare the performance of the Bayes rule for this distribution for several datasets using both EM-based estimates and usual maximum likelihood estimates of von Mises-Fisher distribution parameters.

Keywords: Bayes rule, directional data, EM algorithm, folded directional distribution

---

Directional data analysis deals with unit vectors lying on the surface of a hypersphere and has applications in many areas such as machine learning, text analysis, bioinformatics, genetics and neurology. When such data are restricted to the positive orthant of the unit hypersphere, a folded directional distribution is more appropriate than a standard directional model. Since the von Mises-Fisher distribution is widely used for modeling directional data (see, for example, [1] and [3]), a folded von Mises-Fisher distribution can be considered in this context. This model, introduced in [2], extends the classical von Mises-Fisher distribution to accommodate data restricted to the positive orthant of the unit hypersphere.
The von Mises-Fisher distribution, denoted by $M_p(\boldsymbol{\mu}, \kappa)$ has probability density function given by

$$f(\mathbf{x}|\boldsymbol{\mu}, \kappa) = c_p(\kappa)\, \exp(\kappa \boldsymbol{\mu}'\mathbf{x}) \quad \mathbf{x} \in S^{p-1}, \quad \boldsymbol{\mu} \in S^{p-1},\ \kappa > 0\ ,$$

where the normalizing constant is

$$c_p(\kappa) = \kappa^{\frac{p}{2}-1} / [(2\pi)^{p/2}\, I_{p/2-1}(\kappa)]$$

and $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind and order ν. The parameter $\boldsymbol{\mu}$ represents the mean direction, while κ controls the concentration around $\boldsymbol{\mu}$. This distribution is rotationally symmetric about $\boldsymbol{\mu}$. The maximum likelihood estimates (MLEs) of the parameters based on a random sample $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are obtained as

follows. Let $\overline{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ denote the sample mean vector and $\overline{R} = \|\overline{\mathbf{x}}\|$ the mean resultant length. The maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ is the sample mean direction, $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}/\|\overline{\mathbf{x}}\|$. The MLE of the concentration parameter κ is the solution of $A_p(\kappa) = \|\overline{\mathbf{x}}\|$, where $A_p(\kappa) = c_p'(\kappa)/c_p(\kappa) = I_{p/2}(\kappa)/I_{p/2-1}(\kappa)$. For more details about these estimates, see [4], p. 198.

In this study, we first consider the maximum likelihood estimation for the folded von Mises-Fisher distribution. Due to the complexity of its density function, we use an Expectation-Maximisation (EM) algorithm to obtain the maximum likelihood estimates.

Next, we analyse the estimates obtained using the EM algorithm with data generated from folded von Mises-Fisher distributions in various scenarios, comparing them with the usual MLE of the von Mises-Fisher distribution parameters. Our results show that estimating the mean vector is not substantially affected by using the MLE for the von Mises-Fisher distribution rather than the MLE obtained using the EM algorithm. However, the maximum likelihood estimator of the concentration parameter obtained using the EM algorithm exhibits lower bias than the respective MLE for the von Mises-Fisher distribution.

Finally, we consider the Bayes classification rule for the folded von Mises-Fisher distribution, and evaluate its performance on several datasets, using both EM-based estimates and usual MLE of the von Mises-Fisher distribution parameters.

Acknowledgement This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025) and UID/00006/2025.

References

[1] A. Figueiredo. Discriminant analysis for the von Mises-Fisher distribution. *Communications in Statistics - Simulation and Computation*, 38:1991–2003, 2009.

[2] A. Figueiredo and F. Figueiredo. Classification for a folded von Mises-Fisher distribution. *Research in Statistics*, 3(1), 2025.

[3] P. L. López-Cruz, C. Bielza, and P. Larranaga. Directional naive Bayes classifiers. *Pattern Analysis and Applications*, pages 225–246, 2015.

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

11 April, 9:20 - 9:40, Lecture Theatre E1

Simulating Rosenthal's fail-safe number

Vanusa Rocha[1], **Vera Afreixo**[1], **Miguel Felgueiras**[1,2]
[1] Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal, vanusa@ua.pt, vera@ua.pt, mfelg@ipleiria.pt
[2] ESTG, Polytechnic Institute of Leiria and CEAUL – Centro de Estatística e Aplicações, Universidade de Lisboa, Lisbon, Portugal

---

This study examines bootstrap confidence intervals for Rosenthal's fail-safe number, truncating the estimator at zero and accounting for the correction term ε. Simulation results show that these modifications substantially improve coverage accuracy, particularly for small sample sizes and skewed distributions.

Keywords: fail-safe number, simulations, bootstrap, confidence interval, coverage probability

---

Publication bias threatens the validity of meta-analytic findings by selectively suppressing studies with no significant or even negative results, leading to inflated effect estimates. One of the most widely used tools to assess the robustness of meta-analytic significance against unpublished null studies is Rosenthal's fail-safe number (FSN), which is defined as follows [2]

$$\hat{N}_R = \left(\sum_{i=1}^{k} Z_i / Z_\alpha \right)^2 - k, \tag{1}$$

k is the number of studies and Z_α is the critical value associated with the significance level α for a one-tailed Z-test. Despite its popularity, the statistical properties of Rosenthal's FSN are still poorly understood. Although a method has been proposed to construct confidence intervals [1], the associated simulation study has important limitations, particularly the inclusion of negative FSN values, which lack substantive interpretation. The coverage probability was evaluated by comparing the confidence intervals with the true value of Rosenthal's FSN, for fixed k, defined as

$$E(\hat{N}_R) = (k^2\mu^2 + k\sigma^2)/Z_\alpha^2 - k + \varepsilon, \tag{2}$$

where $\varepsilon = (\phi(\lambda^*)/\Phi(\lambda^*)).(k\sigma(\sqrt{k}\mu + Z_\alpha)/Z_\alpha^2)$, $\lambda^* = (\sqrt{k}\mu - Z_\alpha)/\sigma$, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function of a standard normal distribution, respectively. The original study [1] ignores the correction term ε, assuming that $\Phi(\lambda^*) \approx 1$ for a significantly large k. However, this approximation does not hold for non-Gaussian distributions or for Gaussian distributions with small values of k. In this study, we re-evaluate the bootstrap coverage of confidence intervals for Rosenthal's FSN, given the limitations identified in the original simulation study. For each configuration, the simulation procedure was repeated 10,000 times with a fixed significance level of $\alpha = 0.05$

Table 1: Coverage probabilities of bootstrap confidence intervals and proportion of positive FSN values by number of studies (k)

Distribution of Z_i		$k = 5$	$k = 10$	$k = 30$	$k = 50$
Standard normal	Original	0.926	0.985	1.000	1.000
	Proposed (T)	0.934	0.977	1.000	1.000
	Proposed (T+ ϵ)	0.997	1.000	1.000	1.000
	p	0.103	0.101	0.100	0.100
Half normal	Original	0.784	0.862	0.915	0.926
	Proposed (T)	0.974	0.894	0.915	0.926
	Proposed (T+ϵ)	0.943	0.872	0.915	0.926
	p	0.564	0.937	1.000	1.000
Skew normal with negative skewness	Original	0.840	0.867	0.875	0.896
	Proposed (T)	0.919	0.980	0.993	0.959
	Proposed (T+ϵ)	0.981	0.991	0.979	0.936
	p	0.213	0.338	0.717	0.899

T - truncation at zero; ε - correction term; p - proportion of a positive FSN

and bootstrap confidence intervals were constructed using 1,000 resamples. In particular, we consider truncating the estimator at zero and incorporating the correction term ε into the definition of the true FSN.

Our simulation results show that truncation at zero consistently improves coverage relative to the original estimator, particularly for asymmetric distributions. The correction term ε produces additional gains for small k, but its effect diminishes as the number of studies increases. In both the original results and the proposed approach, the interval widths exhibit a similar behavior, increasing with the number of studies.

These findings highlight the impact of these methodological modifications on the performance of the confidence intervals.

Acknowledgements This work is supported by CIDMA (`https://ror.org/05pm2mw36`) under the Portuguese Foundation for Science and Technology (FCT, `https://ror.org/00snfqn58`), Grants `https://doi.org/10.54499/UID/04106/2025` and `https://doi.org/10.54499/UID/PRR/04106/2025`, within project PALOP/BD/155077/2024.

References

[1] K. Fragkos, M. Tsagris, and C. Frangos. Publication Bias in Meta-Analysis: Confidence Intervals for Rosenthal's Fail-Safe Number. *International Scholarly Research Notices*, 2014:1–17, 2014.

[2] R. Rosenthal. The " File Drawer Problem " and Tolerance for Null Results. *Psychological Bulletin*, 86(3):638–641, 1979.

11 April, 9:40 - 10:00, Lecture Theatre E1

Respondent-driven sampling as adaptive network sampling

Manuela Maia[1], **Pedro Campos**[2]
[1] Instituto Superior Politécnico Gaya, manuela.maia2014@outlook.com
[2] School of Economics and Management, University of Porto, LIAAD INESC TEC, and Statistics Portugal, pcampos@fep.up.pt

Respondent-Driven Sampling (RDS) is widely used to study hidden populations by exploiting peer-to-peer recruitment along social networks. This paper frames RDS as a specific form of adaptive network sampling, in which inclusion probabilities evolve as network information is progressively revealed. Using a migrant population as an illustrative example, we show how degree-based weighting naturally arises from this adaptive design, clarifying the theoretical foundations and relevance of RDS for populations lacking conventional sampling frames.

Keywords: adaptive sampling, respondent-driven sampling (RDS), migration

Migrant populations, particularly recent arrivals or individuals with irregular legal status, are often weakly covered by administrative registers and household sampling frames. As a result, conventional probability sampling designs are difficult to implement. Respondent-Driven Sampling (RDS) [1] has emerged as a practical alternative, relying on social networks within the target population to drive recruitment through peer referral.

While RDS is frequently motivated using metaphors of chain referral or random walks on networks, it can also be situated more formally within the framework of *adaptive network sampling*. In adaptive designs [2], the selection of new units depends on information observed during the sampling process, particularly on links between units. This paper develops this connection and illustrates it in the context of migrant population studies, following other similar approaches (eg. [3]). Let the target migrant population be represented by an undirected social network $G = (V, E)$, where $V = \{1, \ldots, N\}$ denotes individuals and E represents social ties such as kinship, friendship, or co-national connections. The population size N is unknown.

Each individual $i \in V$ has an associated characteristic of interest y_i (e.g., employment status or access to health services). The network degree of individual i is defined as

$$d_i = \sum_{j \in V} I\{(i, j) \in E\},$$

where I is the indicator function that takes the value 1 if a link between i and j exists, and 0 otherwise. In RDS, d_i is typically measured via self-report and plays a central role

in estimation. RDS begins with an initial set of seeds $S_0 \subset V$, usually selected through community organizations or service providers. At each wave t, sampled individuals are given a limited number of coupons to recruit peers from their personal networks. Let S_t denote the set of sampled individuals after wave t. The sample expands according to

$$S_{t+1} = S_t \cup \{j \in V \setminus S_t : \exists i \in S_t such that (i,j) \in E\},$$

subject to recruitment limits and participation decisions. This mechanism is adaptive in the sense that the probability of selecting unit j depends on whether its links to previously sampled units are revealed during data collection. Inclusion probabilities are therefore not fixed in advance but are endogenously determined by the evolving observed network. The RDS process can be approximated by a random walk on G. In this setting, the first-order inclusion probability of node i satisfies $\pi_i \propto d_i$ implying that individuals with larger personal networks are more likely to be sampled. In this study, we consider a study of hypothetically recently arrived migrants in a large urban area aimed at estimating the proportion with formal employment. Based on synthetic data, we explain how migrants with many social connections are more likely to be recruited, leading to unequal inclusion probabilities. Degree-based weighting corrects, in expectation, for this imbalance and enables inference on population-level characteristics.

References

[1] D. D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199, 1997.

[2] S. K. Thompson. *Adaptive Sampling*. Wiley, New York, 2006.

[3] D. Tyldum. Surveying migrant populations with respondent-driven sampling. experiences from surveys of east-west migration in europe. *International Journal of Social Research Methodology*, 24 (3):341–353, 2021.

11 April, 10:00 - 10:20, Lecture Theatre E1

A data-driven scoring framework for robotic movement complexity

Daniel Rodrigues[1], **Eliana Costa e Silva**[1,2], **Pedro Ribeiro**[1], **Inês Costa**[1], **Gianpaolo Gulletta**[1], **Luís Louro**[1], **Sérgio Monteiro**[1], **André Cardoso**[3], **Ana Colim**[4], **Estela Bicho**[1]

[1] Center Algoritmi/Department of Industrial Eletronics, University of Minho, Portugal, pg47125@alunos.uminho.pt, b13160@alunos.uminho.pt, pg50435@alunos.uminho.pt, d6468@alunos.uminho.pt, luislouro@algoritmi.uminho.pt, sergio@dei.uminho.pt, estela.bicho@dei.uminho.pt

[2] CIICESI, ESTG, Instituto Politécnico do Porto, Portugal, eos@estg.ipp.pt

[3] Center Algoritmi/Department of Production and Systems, University of Minho, Portugal, andre.cardoso@alunos.uminho.pt

[4] DTx - Digital Transformation Colab, Azurém Campus of University of Minho, Portugal, ana.colim@dtx-colab.pt

---

The present work focuses on the data analysis techniques used in the complexity scoring framework for robotic movements recently proposed. It involves hierarchical clustering with Ward's method and a Spearman-based distance matrix to assign weights to individual kinematic metrics. The study correlates this score with human-likeness using non-parametric statistics and evaluates motion similarity via Dynamic Time Warping (DTW) on velocity profiles from industrial tasks. Results show that complexity increases with hand displacement and reorientation.

Keywords: hierarchical clustering, spearman correlation, movement complexity, human-likeness, dynamic time warping

---

In the context of Industry 5.0, enhancing Human-Robot Collaboration (HRC) requires robotic movements that are intuitive and human-like. While previous research focuses on evaluating human-likeness, the impact of task and movement complexity on trajectory generation remains under-explored.

A novel aggregated metric designed to quantify movement complexity prior to execution, using spatial and kinematic properties was recently proposed in [1]. Specifically, the movement's complexity score is defined as the weighted sum of individual metrics. As a case study the following metrics: **(a)** final cartesian coordinates (x_f, y_f, z_f). **(b)** reorientation angles $(\Delta\phi, \Delta\theta, \Delta\psi)$ **(c)** direct distance (Δd) between the start and end positions, of the robot's end-effector are used.

The importance of each metric, was found using hierarchical clustering with Ward's method, to optimize cluster compactness, and defining the distance matrix as $1 - |C|$, where C is

the Spearman correlation matrix. The aim was to group the metrics into uncorrelated clusters that will have the same weight on the aggregated complexity score.
For testing the framework, a dataset was built using two distinct environments, and including a diverse range of motion data from an antropomorphic (human-like) robotic arm collected from two tasks: an assembly window task and a restocking task. Since the individual complexity metrics were not normally distributed, Spearman correlation (C) was used to capture both linear and nonlinear monotonic relationships. Using this methodology, the seven metrics were grouped into four uncorrelated clusters, namely: **(a)** direct distance, Δd: received the highest weight (0.174), as it is the primary differentiator in movement difficulty; **(b)** height, z_f, and side-to-side reorientation of the end-effector, $\Delta\theta$: received the second highest weight (0.172 each); **(c)** frontal, x_f, and lateral, y_f, positions: assigned a weight of 0.122 each. **(d)** forward, $\Delta\phi$, and vertical, $\Delta\psi$, reorientations: received the lowest weight (0.119 each).
Additionally, the generated movements were evaluated against established human-likeness metrics, including the Normalized Jerk Score (NJS), Number of Movement Units (NMU), and the One-sixth Power Law. For the practical application of comparing human and robot motions, the Dynamic Time Warping (DTW) algorithm, which measures the similarity by finding the optimal temporal alignment between two velocity sequences, even if they differ in speed or synchronization, was used. This allowed to quantify how closely the robot's bell-shaped velocity profiles mirrored human repetitions.
The results revealed that the complexity score has a positive and significant correlation with joint displacement ($r \simeq 0.28$) and the jerkiness of the hand trajectory ($r \simeq 0.18$). Movements involving greater displacement and higher final positions (farther from the robotic base) were found to be significantly more complex and, consequently, less human-like. While the robot's movements achieved a unimodal, bell-shaped velocity profile consistent with human motor control, human subjects performed tasks up to five times faster. Thus, pre-evaluating movement complexity can help optimize robotic performance and prioritize simpler paths in shared workspaces.

Acknowledgements This work has supported by Intelligent Robotic Coworker Assistant for Industrial Tasks with an Ergonomics Rationale(ref.PTDC/EEI-ROB/3488/2021) and Smart Retail project(ref.PRR 02/C05-i01.01/2022.PC645440011-00000062). E. Costa e Silva has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through project UIDB/04728/2025(https://doi.org/10.54499/UID/04728/2025). We sincerely appreciate the support and collaboration of the research teams at MARLab (Univ. of Minho) and DTx, whose expertise and contributions were invaluable to this work.

References

[1] D. Rodrigues, E. Costa e Silva, P. Ribeiro, I. Costa, G. Gulletta, Monteiro S. Louro, L., A. Cardoso, A. Colim, and E. Bicho. A complexity scoring framework and its effect on the human-likeness of robotic arm movements. *The International Journal of Advanced Manufacturing Technology*, 2025.

Poster Sessions

10 April, 11:30 - 12:00, Hall of ESTGD

The impact of female board representation on ESG pillars

Carla Henriques[1], **Pedro Pinto**[2], **Joana Silva**[2]
[1] ESTGV, Instituto Politécnico de Viseu, Portugal and Centre for Mathematics of the University of Coimbra (CMUC) , Portugal, carlahenriq@estgv.ipv.pt
[2] ESTGV, Instituto Politécnico de Viseu, Portugal, spinto@estgv.ipv.pt, joanafcsilva00@gmail.com

Sustainability has gained increasing prominence in the corporate landscape, as sustainable management practices are becoming progressively more valued. Prior research has extensively examined the relationship between gender diversity on boards of directors and Environmental, Social, and Governance (ESG) performance and disclosure. Using a sample of 95 companies listed on the PSI, IBEX 35, and Euro STOXX 50, this study investigates the relationship between board gender diversity and ESG scores.

Keywords: linear regression, ESG, board gender diversity

The relationship between Board Gender Diversity (BGD) and Environmental, Social, and Governance (ESG) performance remains a debated topic in the literature, producing heterogeneous results. While some studies report a significant positive association, others find negative or non-significant relationships [1]. These divergent findings highlight the need for further empirical investigation.

In this study, a sample of 95 companies listed on three major indices—the PSI, IBEX 35, and Euro STOXX 50 — was explored with linear regression models to further investigate this issue. Data refering to 2022 were obtained from the LSEG database (LSEG, 2023) and complemented with financial information extracted directly from companies' reports and financial statements.

To assess the impact of the proportion of female directors on ESG scores, linear regression models were fitted with ESG as the dependent variable and the proportion of female directors as the main explanatory variable. The models included controls for firm age, profitability, and size, operationalized as the natural logarithm of age, Return on Assets (ROA) and the natural logarithm of total assets, respectively. In addition to ESG, the study also considered linear models having as dependent variables its individual pillars: Environmental (E), Social (S), and Governance (G). Furthermore, the sample was divided into two subsamples to distinguish between the Iberian market (PSI and IBEX 35) and the broader European context (Euro STOXX 50). Given the relatively small sample size, assessing the normality assumption is essential. This was addressed using the Kolmogorov-Smirnov test with Lilliefors correction and the Shapiro-Wilk test. Furthermore, in cases

where the assumption of homoscedasticity was violated, robust standard errors were employed to ensure the validity of the statistical inferences.
Regarding the overall ESG score, the results suggest that after controlling for firm size, age, and profitability (ROA), the percentage of women on Boards of Directors (BGD) shows a potential positive association with ESG ratings; however, the study sample does not provide sufficient statistical evidence to support this hypothesis (Table 1). In contrast, the proportion of female directors exerts a significant positive impact on the Governance pillar (G), as evidenced in both the global sample ($p = 0.002$) and the Iberian market subsample ($p = 0.002$). Interestingly, while the proportion of female directors is also statistically significant for the Social pillar (S) in the Iberian subsample ($p = 0.043$), the coefficient is negative. This negative relationship may suggest a 'tokenism' effect, where female inclusion on boards remains largely symbolic, failing to translate into substantive decision-making power that influences social performance.

Table 1: Estimated impact of female board representation on ESG, E, S and G scores

	Iberian market (PSI and IBEX 35)		**Broader European** (Euro STOXX 50)		**Global sample**	
Dependent Variable	**Coef.**	**p-value**	**Coef.**	**p-value**	**Coef.**	**p-value**
ESG (Overall)	0,038	0,827	0,054	0,479	0,037	0,587
E (Environmental)	-0,270	0,147	-0,041	0,769	-0,162	0,142
S (Social)	-0,370	0,043	-0,026	0,809	-0,168	0,091
G (Governance)	1,048	0,002	0,257	0,120	0,560	0,002

This research contributes to the ongoing debate by demonstrating that the "heterogeneous results" found in existing literature may stem from the different impacts BGD has on specific ESG dimensions. These results suggest that gender diversity should be viewed not just as a matter of social equity, but as a strategic driver of corporate governance excellence. Future research should explore larger samples to further clarify the potential for a "critical mass" effect in the European context.

Acknowledgements The authors acknowledge financial support by the Centre for Mathematics of the University of Coimbra (CMUC, https://doi.org/10.54499/UID/00324/2025) under the Portuguese Foundation for Science and Technology (FCT), Grants UID/00324/2025 and UID/PRR/00324/2025.

References

[1] M. G. Abdelkader, Y. Gao, and A. A. Elamer. Board gender diversity and esg performance: The mediating role of temporal orientation in south africa context. *Journal of Cleaner Production*, 440:140728, 2024.

10 April, 11:30 - 12:00, Hall of ESTGD

Real-time industrial data quality pipeline for enhanced analytics and decision-making

Teresa Peixoto[1], **Óscar Oliveira**[1], **Eliana Costa e Silva**[1,2], **Bruno Oliveira**[1], **Fillipe Ribeiro**[3]
[1] CIICESI, ESTG, Instituto Politécnico do Porto, Portugal, tmo@estg.ipp.pt; eos@estg.ipp.pt; bmo@estg.ipp.pt; oao@estg.ipp.pt
[2] Center Algoritmi, University of Minho, Portugal
[3] JPM Industry, Vale de Cambra, Portugal, fillipe.ribeiro@jpm.pt

In modern industrial environments, data-driven decision-making depends on high-quality data from integration of Internet of Things (IoT) sensors. The present work addresses a modular data quality pipeline for ingestion, profiling, validation, and data analysis. Data quality is accessed using quality that combine accuracy, completeness, consistency, and timeliness metrics of quality. A real manufacturing case study shows that continuous monitoring enables early detection of data issues, improving reliability and timeliness of decisions.

Keywords: data quality, data ingestion, real-time data analysis, industrial monitoring

The integration of IoT devices, smart sensors, and advanced manufacturing systems has led to the generation of vast amounts of real-time data [1]. However, the value of this data depends on its quality. In fact, decisions based on inaccurate, incomplete, inconsistent, or outdated data can have significant negative consequences.

While prior studies have proposed taxonomies for data quality techniques, they also highlight a prevalence of low industrial adoption [1]. In the context of smart factories, Liu et al. [2] identify accuracy, completeness, consistency, and timeliness as core data quality dimensions. Addressing these limitations, the framework proposed in [3] introduces dynamic, real-time metric computation and adaptive profiling, representing a significant advancement over static rule-based systems.

The proposed pipeline comprises modular components for data ingestion, profiling, validation, and continuous monitoring. It relies on key data quality dimensions, namely: accuracy - normalized metric using dynamic Hampel filter bounds; completeness - content and temporal metrics; consistency - rule-based correlation checks; and timeliness - age vs. volatility. Metrics are computed over one-minute sliding windows via streaming processing with Apache Kafka.

Data quality is continually monitored using three quality indices, namely: Weighted Quality Score (WQS), Longitudinal Weighted Quality Score (LWQS) with exponential decay ($\beta = 0.5$), and Quality Score Delta (QSD = WQS - LWQS), enabling trend detection. Further, sensor-specific weights prioritize critical dimensions (e.g., $w_{acc} = 0.7$, $w_{comp} = 0.3$).

Regarding infrastructure, data is stored in InfluxDB, and Grafana dashboards are used to visualize raw data, metrics, and alerts.
Compared to existing data quality assessment approaches in industrial and big data contexts, the proposed pipeline provides a unified and lightweight architecture that integrates ingestion, profiling, validation, and visualization in real-time. Although many solutions [1] rely on rule-based frameworks or static configurations, our architecture introduces dynamic metric computation and adaptive profiling mechanisms, allowing the system to react autonomously to changes in data patterns.
This work advances the state of the art by introducing a lightweight real-time processing pipeline with integrated profiling metrics, whose effectiveness has been demonstrated in an industrial environment. However, the current implementation exhibits limitations related to windowing delays and scalability under increasing computational loads. Future research will focus on extending the framework to multi-sensor scenarios, incorporating machine learning techniques for anomaly detection, and conducting comprehensive comparative benchmarking.

Acknowledgements This work was supported by the European Union under the Next Generation EU, through a grant from the Portuguese Republic's Recovery and Resilience Plan (PRR) Partnership Agreement, within the scope of the project PRODUTECH R3 - "Agenda Mobilizadora da Fileira das Tecnologias de Produção para a Reindustrialização". Total project investment: EUR 166.988.013,71; total grant: EUR 97.111.730,27. This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through project UIDB/04728/2025 (https://doi.org/10.54499/UID/04728/2025).

References

[1] 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*, 55:1–38, 2023.

[2] C. Liu, G. Peng, Y. Kong, S. Li, and S. Chen. Data quality affecting big data analytics in smart factories: Research themes, issues and methods. *Symmetry*, 13(8):1440, 2021.

[3] T. Peixoto, Ó. Oliveira, E. Costa e Silva, B. Oliveira, and F. Ribeiro. A data quality pipeline for industrial environments: Architecture and implementation. *Computers*, 14(7):241, 2025.

10 April, 11:30 - 12:00, Hall of ESTGD

Quality of work life in a textile company in northern Portugal

Francisco Cardoso[1], **Cristina Torres**[2], **Adalmiro Pereira**[2], **Cristina Lopes**[2], **Lurdes Babo**[2], **Isabel Vieira**[2]
[1] ISCAP, Instituto Politécnico do Porto, Portugal, 2180532@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, Portugal, ctorres@iscap.ipp.pt, adalmiro@iscap.ipp.pt, cristinalopes@iscap.ipp.pt, lbabo@iscap.ipp.pt, mivieira@iscap.ipp.pt

This study analyses Quality of Work Life (QWL) perceptions among employees of a textile company in Northern Portugal. Exploratory factor analysis identified seven QWL dimensions, and cluster analysis revealed two distinct employee profiles. Results confirm QWL's multidimensional nature and highlight the importance of organisational conditions, equity, and work–life balance for employee well-being.

Keywords: quality of work life, exploratory factor analysis, cluster analysis, employee well-being

Quality of Work Life (QWL) has become a central concern in organisational research, particularly because it reflects how work conditions affect employees' well being and job satisfaction [2]. In industrial contexts characterised by demanding working conditions and strong productivity pressures, QWL assumes greater importance. In sectors such as the textile industry, where employees are frequently exposed to physical strain, rigid schedules, and limited autonomy, QWL is a critical determinant of well being, motivation, and organisational sustainability — improvements in QWL have been linked to higher employee commitment, reduced turnover, and enhanced organisational effectiveness [1]. Building on this, the present study investigates employees' perceptions of QWL within a textile company located in Northern Portugal, aiming to identify homogeneous profiles based on these perceptions and to explore the key organisational factors that shape well being, motivation, and job satisfaction.

Data were collected through a structured questionnaire composed of 50 items on a five-point Likert scale, yielding 104 valid responses. The sample was slightly female-dominated (53%), predominantly married or in a civil union (60%), and included participants across a range of educational levels (22% basic, 39% secondary, 39% higher education), hierarchical positions (from operational staff to managerial roles), and tenure (< 5 years: 33%, 6–10 years: 28%, > 10 years: 39%).

The instrument showed excellent internal consistency (Cronbach's $\alpha = 0.935$), and sampling adequacy was confirmed (KMO = 0.90; Bartlett's test of sphericity, $p < 0.001$),

supporting the application of Exploratory Factor Analysis (EFA). Using principal component extraction with varimax rotation, seven factors with eigenvalues greater than 1 were retained. Together, these factors explained 65.4% of the total variance, indicating a robust and interpretable factorial solution. The extracted dimensions were labelled as: Working Conditions and Well-Being, Equity and Organisational Justice, Organisational Identification and Professional Development, Organisational Support and Interdepartmental Relations, Active Participation and Social Support, Safety and Organisational Stability, and Work–Life Balance. All retained items exhibited satisfactory factor loadings (≥ 0.4) and communalities (≥ 0.5), confirming both the statistical adequacy and the conceptual coherence of the measurement model. Based on the factor scores, a cluster analysis using Ward's method and squared Euclidean distance was performed to identify groups of employees with similar QWL perception patterns. Two distinct clusters were identified: Cluster 1 (n = 36) and Cluster 2 (n = 68). Independent-samples t-tests were performed to compare the clusters on each QWL dimension, as the dependent variables were continuous and the sample sizes were sufficiently large to justify the use of parametric testing despite minor deviations from normality. Cluster 2 scored significantly higher than Cluster 1 in Working Conditions and Well-Being ($p < 0.001$), Organisational Identification and Professional Development ($p = 0.042$), Organisational Support and Interdepartmental Relations ($p = 0.042$), and Safety and Organisational Stability ($p < 0.001$). No significant differences were observed in Equity and Organisational Justice ($p = 0.061$), Active Participation and Social Support ($p = 0.591$), and Work–Life Balance ($p = 0.130$).

Chi-squared tests revealed that cluster membership was significantly associated with educational level ($p = 0.021$), suggesting that employees with higher educational attainment tended to adopt more critical evaluations of QWL. No significant associations were observed for gender, marital status, hierarchical position, tenure, or salary, indicating that perceptions of QWL were primarily shaped by organisational factors rather than demographic characteristics.

Overall, these findings confirm the multidimensional structure of Quality of Work Life and emphasise the importance of fair organisational practices, supportive management, and work–life balance policies. Despite limitations related to response rate and the single-case design, the study provides robust, statistically sound evidence that may inform both academic research and organisational decision-making in industrial contexts.

References

[1] A. H. Nuzulizzwan, N. R. A. Rahim, and M. F. Ramli. A concept of quality of work life in the textile manufacturing industry. *International Journal of Industrial Management*, 2025.

[2] D. Pereira, J. C. Leitão, and A. Gonçalves. Quality of work life and organizational performance: Workers' feelings of contributing, or not, to the organization's productivity. *International Journal of Environmental Research and Public Health*, 16(20):3803, 2019.

10 April, 11:30 - 12:00, Hall of ESTGD

Five social Europes? Empirical evidence and implications for the welfare state

Irene Oliveira[1], **Patrícia Martins**[2]
[1] Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro and CEMAT - Centro de Matemática Computacional e Estocástica , ioliveir@utad.pt
[2] Departamento de Economia, Sociologia e Gestão -UTAD and CETRAD, smartins@utad.pt

---

This study re-evaluates the validity of traditional welfare regime classifications by conducting a multidimensional analysis of Eurostat data from 2023 for the 27 EU Member States. Hierarchical clustering identified five distinct configurations of 'Social Europes', differentiated by poverty, inequality, labour-market fragility, and fiscal capacity. The results indicate that traditional typologies no longer capture Europe's heterogeneity, underscoring the need for structural welfare reconfiguration.

Keywords: welfare state, cluster analysis, Kruskal–Wallis, discriminant analysis, naive Bayes

---

The future of the European welfare state remains a key area of debate, particularly in light of demographic pressures, political fragmentation, international migration, austerity measures and the fiscal sustainability of public debt. The welfare state is conceived as a universal, publicly funded system that provides social protection, healthcare, education and a minimum income. However, the literature identifies three potential trajectories: dismantling, maintenance, or structural reconfiguration. In particular, classical regime typologies such as Esping-Andersen's (1990) [1] are mostly theoretical or institutional and have not been empirically validated using recent data or multivariate approaches.
This study provides a thorough empirical evaluation of the continued relevance of classical typologies in the European context, employing a multidimensional, data-driven statistical methodology. Based on Eurostat data from 2023 for the 27 EU Member States, a hierarchical cluster analysis using Ward's method was performed on eight indicators of living conditions: life expectancy; the proportion of people at risk of poverty or social exclusion; the in-work poverty rate; the pension replacement ratio; income inequality (S80/S20); inequality among older people; the effect of taxes and transfers on inequality; and housing cost overburden. Five clusters emerged: (1) Old European countries (including Ireland), (2) Central European countries (excluding Croatia), (3) Eastern European countries, (4) Small tourism-driven countries (Cyprus, Croatia and Malta), and (5) Mediterranean countries (including Luxembourg).

Social indicators show that cluster 1 performs best, with lower poverty and unemployment and higher life expectancy. Clusters 3 and 5 perform worst in terms of poverty, inequality and employment precarity. Cluster 2 stands out due to its considerably higher pension replacement ratio, while Cluster 4 shows the highest in-work poverty risk. Comparing inequality indicators before and after taxes and transfers suggests limited redistribution in group 5, indicating that fiscal tools have not been effective enough in reducing inequality. To statistically validate differences between clusters, 30 additional variables were analysed using the non-parametric Kruskal–Wallis test across six dimensions: demographics, macroeconomics, public finance, public expenditure and political context. Dunn post hoc tests (Bonferroni correction) confirmed significant differences in 16 variables. Public expenditure on health is significantly higher in cluster 1 than in cluster 3; public debt is significantly higher in cluster 5, exceeding an average of 100% of GDP, while cluster 3 remains near 35%. Real GDP growth varies from contraction in cluster (1) to marked expansion in cluster (4). Labour market vulnerability also varies, with long-term unemployment being significantly higher in cluster 5, whereas cluster 2 records the lowest unemployment rates. Expected demographic change (2023–2050) also separates the clusters, with cluster 1 projected to grow while cluster 3 is expected to shrink by more than 10%.

The robustness of the empirical classification was evaluated using stepwise discriminant analysis and a naïve Bayes classifier. The discriminant model correctly classified 74.1% of cases (63% cross–validated), whereas the naïve Bayes classifier achieved full accuracy when applied to variables with statistically significant differences. Public debt, tax burden and real GDP growth were the most influential predictors. This suggests that fiscal macro fundamentals are central to how European welfare states are differentiated today.

These findings demonstrate that the traditional three-model typology no longer adequately captures the current heterogeneity of Europe. Instead, five empirical configurations have emerged, shaped by a combination of living conditions, fiscal capacity, inequality, labour-market performance, and demographic dynamics. The empirical evidence supports the updating of welfare state typologies to reflect contemporary realities, with implications for long-term sustainability, policy targeting, and EU-level convergence debates.

Acknowledgements This work was supported by projects funded by the Portuguese Foundation for Science and Technology (FCT): UID/MULTI/04621/2020, doi: `10.54499/UIDB/04621/2020`, and UIDB/04011/2020, doi:url10.54499/UIDB/04011/2020.

References

[1] G. Esping-Andersen. *The Three Worlds of Welfare Capitalism.* Cambridge University Press, Cambridge, 1990.

10 April, 11:30 - 12:00, Hall of ESTGD

Predictive models for download counts

Beatriz Silva[1], **Susana Faria**[1,2]
[1] Department of Mathematics, University of Minho, pg59997@uminho.pt, sfaria@math.uminho.pt
[2] Centre of Mathematics (CMAT), Department of Mathematics, University of Minho

This study aims to predict the daily number of mobile app downloads using data from the Apple App Store in the United States over the period 2022–2024. Several machine learning methods were applied, including Decision Trees, Random Forest, and XGBoost, to evaluate the predictive performance of each model. The results show the importance of machine learning techniques for forecasting download volumes in marketing contexts.

Keywords: downloads, random forest, decision trees, *XGBoost*

Mobile applications are part of the daily lives of most people, ranging from the basic apps pre-installed on smartphones to highly innovative and creative products. Given their widespread use, the mobile app market continues to expand rapidly, increasing competitive pressure and prompting developers to use analytical and decision-support tools to differentiate their products. Digital marketing agencies play a crucial role by monitoring market dynamics and implementing evidence-based strategies to support app acquisition and growth.

Download counts on the Apple App Store and Google Play Store are indicators of an app's scale and can be used to infer whether it is in a phase of growth or decline. This metric is, therefore, a key indicator for app developers, marketing practitioners, and investors.

Model performance was assessed using Root Mean Square Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

To improve the models' predictive performance, variables capturing relevant events occurring in the United States during the study period were incorporated. The final models were then used to predict download volumes for selected mobile applications on new dates not included in the dataset.

The results showed that category-specific models achieved better predictive performance, with Random Forest models standing out when using rank, day of the week, week, and month as predictors.

The "Health & Fitness", "Lifestyle", and "Games" categories were the most difficult to predict in terms of download volume due to the high heterogeneity of the apps they comprise. Regarding variables capturing major U.S. events during the study period, events such as the Academy Awards, Super Bowl finals, and Black Friday were found to influence

app downloads. Nevertheless, incorporating these event-related variables did not yield substantial improvements in model performance.

Acknowledgements. The research of S. Faria was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013/2025 (https://doi.org/10.54499/UID/00013/2025).

References

[1] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley, Hoboken, 2 edition, 2004.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 26 of *Springer Series in Statistics*. Springer, New York, 2 edition, 2009.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning, with Applications in R*. Springer, New York, 2022.

10 April, 11:30 - 12:00, Hall of ESTGD

Does being born poor limit access to high income?

Beatriz Gouveia[1], Gabriel Neves[1], Rita Viana[1], Stephanie Jesus[1], Lurdes Babo[2], Cristina Torres[2], Isabel Vieira[2], Cristina Lopes[2]
[1] ISCAP, Instituto Politécnico do Porto, 2221107@iscap.ipp.pt, 2201101@iscap.ipp.pt, 2221124@iscap.ipp.pt, 2190955@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, lbabo@iscap.ipp.pt, ctorres@iscap.ipp.pt, mivieira@iscap.ipp.pt, cristinalopes@iscap.ipp.pt

This study examines the extent to which socioeconomic background conditions income attainment in adulthood and whether education functions as an effective mechanism of social mobility in a European context. The results indicate that, although access to higher education is not constrained by socioeconomic disadvantage at origin, adult income remains negatively correlated with it. The findings point to the existence of an imperfect meritocracy: while education mitigates the effects of origin-based inequality, it does not eliminate them, with persistent gender disparities.

Keywords: social mobility, socioeconomic background, gender inequality

In the context of debates on meritocracy, this research assesses whether education facilitates social mobility or if the 'scar' of poverty leaves a lasting impact on earnings and assets. Using 2023 Eurostat data [2] across 32 European countries, the study employs correlation analysis, mean comparison tests, and hierarchical cluster analysis. Key variables span multiple dimensions of inequality, including the Gini coefficient, intergenerational mobility, educational attainment, labor market participation, material deprivation, and gender.
Results indicate no significant association between national income inequality and social mobility rates, suggesting that wealth concentration alone does not account for opportunities for upward mobility. In contrast, socioeconomic origin—proxied by parental educational attainment—shows a consistently negative relationship with adult income, particularly among individuals with low educational attainment.
The analysis of men's socioeconomic trajectories demonstrates that higher education is a differentiating factor in terms of income and employability. The transition from low to high educational attainment is associated with a statistically significant economic increase ($t = -16.023, p < 0.001$, Cohen's $d = -2.789$), alongside a marked rise in employment rates ($t = -10.007, p < 0.001, d = -1.742$). Nonetheless, socioeconomic origin continues to shape income trajectories: the negative correlation between childhood poverty risk and adult earnings remains strong among low-educated men ($r_s = -0.609, p < 0.001$) and still present among the highly educated ($r_s = -0.49, p < 0.001$). These findings indicate that higher education reduces—but does not eliminate—the long-term effects of early-life deprivation.

Among women, educational attainment has an even stronger gradient in labor market integration. Women with low educational attainment have the lowest employment rate in the sample (40.37%), which rises to 76.66% among those with tertiary education - a highly significant difference ($t = -18.034, p < 0.001, d = -3.293$). Completing a tertiary degree is also associated with a large and statistically significant increase in annual income ($t = -15.509, p < 0.001, d = -2.700$). Nevertheless, highly educated women continue to earn less than equally educated men, suggesting persistent gender disparities beyond educational attainment.

Social mobility patterns reveal strong intergenerational persistence: 71.2% of women from low-education families remain in that group, while only 21.24% experience upward mobility. Socioeconomic origin also continues to depress income, with strong negative correlations among low-educated women ($r_s = -0.652$) and a still notable effect among the highly educated ($r_s = -0.494$).

A generational analysis of childhood financial deprivation indicates a progressive improvement in starting conditions. Women in the $45 - 59$ age group report a deprivation rate of 7.38%, while younger women ($25 - 29$ years) record 4.84%. Despite this progress, the translation of these initial improvements into effective income remains constrained by a labor market that is less favorable to women.

Hierarchical clustering using Ward's method with squared Euclidean distance identifies three European typologies. The 'Educational Mobility' cluster (e.g., Portugal, Spain) shows the highest upward mobility ($48 - 50\%$) but the lowest average income (€$17,000$), reflecting recent educational expansion with limited financial returns. Conversely, 'Consolidated Societies' (Central/Northern Europe) pair the highest incomes ($>$ €$27,000$) with the lowest inequality (Gini 27.8). Finally, the 'Social Deadlock' group (Eastern Europe/Balkans) exhibits high inherited poverty (34.68%) and inequality (Gini 33.92), resulting in a stagnant 'social elevator' ($4 - 5\%$ mobility). These results highlight that the education-economic success link is non-linear and context-dependent, echoing broader evidence that educational expansion does not automatically translate into poverty reduction or upward mobility [1].

Overall, the findings indicate that education substantially improves individual economic prospects, but does not fully offset the enduring influence of socioeconomic origin. The relationship between education and economic success therefore appears structurally heterogeneous and context-dependent across European settings.

References

[1] P. Brown and D. James. Educational expansion, poverty reduction and social mobility: Reframing the debate. *International Journal of Educational Research*, 100:101537, 2020.

[2] Eurostat. Eu statistics on income and living conditions (eu-silc). `https://ec.europa.eu/eurostat`, 2023. Accessed: 2025-12-10.

10 April, 11:30 - 12:00, Hall of ESTGD

Lab-grown diamonds population structure: an application

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

We analysed 18,775 lab-grown diamonds listed by an online retailer in 2025 by estimating a latent class model defined over carat, cut, colour and clarity. The segment comprising the heaviest stones (2%) displays excellent cut grades, a predominance of colour E, mid range clarity levels (VS2–VS1), and high price per carat. Future studies should consider segmentation methods purposefully tailored to model and predict price per carat as a function of multiple diamond grading attributes.

Keywords: lab-grown diamonds, diamond grading attributes, model-based clustering, finite mixture models, measurement models with mixed indicators

In this work, we analyse a database of lab-grown diamonds obtained in 2025 from a large online retailer that specialises in independently certified stones and offers tens of thousands of GIA and IGI graded diamonds through a web-based search platform. The retailer positions itself as a broker between manufacturers and final customers, emphasising transparency, international certification standards, and detailed gemological information for each stone. This commercial context makes it a suitable source for large-scale observational data on the characteristics of contemporary lab-grown diamonds. The database (including 18775 observations) was built by systematically extracting the full online inventory of lab-grown diamonds, yielding a tabular dataset with one record per stone and variables covering the standard "4Cs" and related grading attributes: carat, cut, colour, clarity, shape, polish, symmetry, fluorescence, depth, table, and length-to-width ratio. Each observation also includes transactional information, namely the list price in US dollars and the corresponding price per carat, enabling the joint study of physicochemical grading, geometry, and market valuation. Shapes represented in the file include, among others, Princess and Cushion cuts, with coverage across a wide range of carat weights and quality levels.
To understand the structure of lab-grown diamonds' population, we resort to the referred data and estimate a latent class mixture model in which the latent classes are identified using four lab-grown diamond characteristics: carat (continuous), cut (ordinal), colour (ordinal) and clarity (ordinal). Let k denote the unobserved class membership, taking values 1 to K. The model assumes that each individual belongs to one of the K latent classes, with class probabilities $\pi_1, \ldots, \pi_K$ estimated from the data. To determine K,

we use the BIC criterion [3]. Conditional on latent class membership, the four observed variables are assumed to be mutually independent except for a set of prespecified local dependencies, [2]. Specifically, the model includes four direct effects. These direct effects relax the usual local independence assumption by allowing the distribution of one variable to depend on the value of another, even after conditioning on the latent class. Carat is modelled as a normally distributed continuous variable with class-specific means and variances. Its mean is allowed to depend on the individual's levels of clarity, colour, and cut, and a direct effect is also specified between colour and cut. Cut, colour, and clarity are modelled using class-specific cumulative logit models (proportional odds regressions) with separate location parameters for each class.

The results indicate that the cluster with the highest mean weight (measured in carats, where 1ct = 0.2g) represents approximately 2% of the sample. Within this group, 84% stones exhibit an excellent cut, and the predominant colour and clarity categories fall within the mid-range levels. The price per carat, included as a model covariate, is on the higher end of its empirical distribution for this cluster.

In the future, further research is needed to understand the relationship between diamonds' characteristics and price per carat. Emerging developments in high-pressure high-temperature (HPHT) and chemical vapour deposition (CVD) synthesis techniques are reshaping the characteristics of lab-grown diamond populations, [1], creating new challenges for future research.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grant UID/315/2025 (DOI https://doi.org/10.54499/UID/00315/2025).

References

[1] S. Eaton-Magaña, M. F. Hardman, and S. Odake. Laboratory-grown diamonds: An update on identification and products evaluated at gia. *Gems & Gemology*, 60(2), 2024.

[2] J. H. M. Janssen, S. van Laar, M. J. de Rooij, J. Kuha, and Z. Bakk. The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling*, 26(2):280–290, 2019.

[3] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

10 April, 11:30 - 12:00, Hall of ESTGD

Adaptation and validation of the *Problematic TikTok Use Scale* in a sample of portuguese adolescents

Elisete Correia[1], **Ana Rita Monteiro**[2], **Susana Cardoso**[3],**Ana Paula Monteiro**[4]
[1] CEMAT, Dep. of Mathematics, UTAD, Vila Real, ecorreia@utad.pt
[2] Dep. of Education and Psychology, UTAD, Vila Real, anarita.monteiro2000@gmail.com
[3] CIDESD, Dep. of Education and Psychology, UTAD,Vila Real, susana.cardoso@utad.pt
[4] CIIE, Dep. of Education and Psychology, UTAD,Vila Real, apmonteiro.pt

With the growing popularity of *TikTok* among adolescents, concerns have arisen about its problematic use. The existence of an instrument to assess the problematic use of *TikTok* is essential for scientific research and the development of intervention strategies. This study aimed to adapt and validate the *Problematic TikTok Use Scale* (PTTUS) by Günlü et al.[2] in a sample of portuguese adolescents. For this purpose, 331 students completed the PTTUS and a sociodemographic questionnaire. The results showed good psychometric properties, allowing it to be used in both research and psychoeducational interventions.

Keywords: problematic use, psychometrics, scale, structural equation modeling, *TikTok*

At present, social networks, especially *TikTok*, play an increasingly relevant role in adolescents' daily lives [2]. This social network allows access to short-duration videos on a wide range of topics, thereby meeting users' needs for entertainment, expression, communication, and socialization, particularly among adolescents, who are the most active population on this application [1]. However, the growing adherence to this platform has raised concerns regarding the occurrence of excessive and problematic usage patterns. Evidence indicates that problematic *TikTok* use is associated with addiction-related symptoms, potentially compromising adolescents' well-being and quality of life.
Thus, the present research aimed to translate and adapt the PTTUS for the Portuguese population and to examine its psychometric properties. The *Problematic TikTok Use Scale* developed by Günlü et al. [2] consists of 16 items divided into three dimensions, namely Obsession, Escapism, and Loss of Control. The translation and back-translation procedures followed the guidelines recommended by Hambleton [3].

A total of 331 adolescents participated in the study, of whom 160 were female (48.3%) and 171 were male (51.7%), aged between 12 and 18 years, ($M = 15.20$, $DP = 1.48$). All participants completed the PTTUS and a sociodemographic questionnaire.

In this study, the reliability analysis of the PTTUS, assessed using Cronbach's alpha, revealed internal consistency values of .88 for the total scale, .77 for the Obsession subscale, .86 for the Escapism subscale, and .81 for the Loss of Control subscale.

The dimensionality and factor structure of the scale were analyzed using Structural Equation Modeling (SEM). The factorial model showed good fit indices ($\chi^2/df = 2.71,\ PGFI = .655,\ GFI = .909,\ CFI = .929,\ RMSEA = .072,\ p[RMSEA \leq .05] < .001$), according to the criteria recommended in the literature [4]. Most items showed high factor loadings and high individual reliabilities.

The findings suggest that this instrument makes a valuable contribution to research on problematic social media use. In particular, individuals' PTTUS scores may be analyzed in association with other psychological, behavioral, or sociodemographic variables, thereby enabling a more comprehensive understanding of protective and risk factors related to problematic *TikTok* use. In this way, the instrument offers not only an objective assessment of usage behavior but also creates opportunities for future research examining the mechanisms underlying problematic use of this platform.

References

[1] C. Bucknell Bossen and R. Kottasz. Uses and gratifications sought by pre-adolescent and adolescent *tiktok* consumers. *Young Consumers*, 21(4):463–470, 2020.

[2] A. Günlü, T. Oral, S. Yoo, and S. Chung. Reliability and validity of the problematic tiktok use scale among the general population. *Frontiers in Psychiatry*, 14, 2023.

[3] R. Hambleton. The next generation of the itc test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17(3):164–172, 2001.

[4] J. Marôco. *Análise de equações estruturais. Fundamentos teóricos, software e aplicações.* ReportNumbe, 2014.

10 April, 11:30 - 12:00, Hall of ESTGD

Human, material and economic impacts of disasters

Rita Martins[1], Cristina Lopes[2], Isabel Vieira[2], Lurdes Babo[2], Cristina Torres[2]

[1] ISCAP, Instituto Politécnico do Porto, 2190904@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

---

Over recent decades, the frequency and intensity of natural and technological disasters have increased, posing risks to life, health, and the environment, often exceeding local response capacities and demanding national or international assistance. This study analyses secondary data from the EM-DAT database to examine the human, material and economic impacts of disasters.

Keywords: disasters, humanitarian supply chain, correlations, hypotheses tests

---

A disaster is a serious, immediate or lasting, disruption of the functioning of a society due to hazardous events that can lead to human, material, economic and environmental losses and impacts. The effects of a disaster may pressure or exceed community's capacity to cope with the consequences by using its own resources [3].

This research aimed to identify and analyze patterns in disaster occurrence and examine the relationships between their human, material, and economic impacts on populations and countries. Records of 26,456 disasters worldwide (1920–Feb 2025) were collected from the Emergency Events Database (EM-DAT) [1]. To meet EM-DAT inclusion criteria, an event must involve at least least 10 deaths, or 100 injured or the country affected must either declare state of emergency or request international or national assistance [2].

Variables considered were: Assistance Requests and State of Emergency Declarations (annual frequency); Deaths (fatalities plus missing persons), Injured, Homeless, Affected, and Total Affected; Economic Losses (USD thousands) and inflation-adjusted Losses (CPI).

Figure 1 shows that although absolute injury numbers remain higher, deaths per decade follow a downward trend and consistently stay below injuries. This inversion in numbers is likely due to improvements in disaster preparedness, education of the population on how to behave in the event of a disaster, specific training for professionals involved in the preparation and response phases, evolution of medical care, and humanitarian supply chain efficiency.

Spearman correlation analysis reveals higher values between the number of injured, affected persons and economic losses. Total Damage is strongly correlated with all variables except Total Deaths ($\rho = 0.151$, $p > 0.05$), suggesting that economic loss and mortality are not tightly coupled and that reducing deaths does not necessarily reduce economic impact. The frequency of international aid requests is significantly correlated with the variables

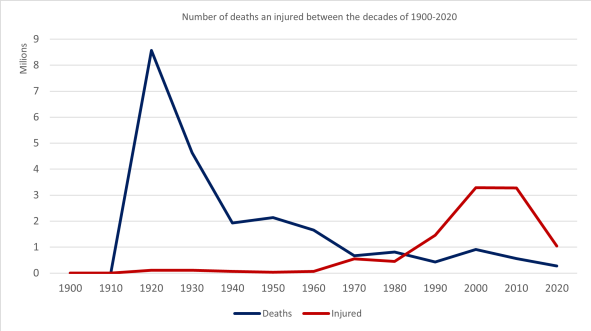

Figure 1: Comparison between the number of deaths and injured people after a disaster

injured, affected, and monetary damage, although more weakly with deaths ($\rho = 0.196$). A similar pattern is observed for declarations of state of emergency.

Binary indicators were created to represent whether a state of emergency was declared and whether international aid was requested in a given year. A chi-square test rejected the independence between these variables ($\chi^2 = 82.345$, $p < 0.001$), and Cohen's Kappa showed significant agreement (87.8%) between classifications.

To assess differences in disaster outcomes based on policy measures, Mann–Whitney tests compared the distributions of deaths, injuries, homeless persons, affected persons, total damage, and adjusted total damage between groups defined by the binary variables international aid and state of emergency. No statistically significant difference in the number of deaths was found for international aid ($p = 0.094$) or state of emergency ($p = 0.273$). However, both policy measures exhibited significantly higher ranks for injuries, homeless populations, and economic damage ($p < 0.001$).

These results suggest that while policy measures reflect event severity, they also facilitate countries' administrative and logistical capacity to recognize, collect data, and respond to the humanitarian and supply chain impact. However, neither factor significantly affects mortality, indicating that most of the fatalities may occur before response measures are enacted, underscoring the need to examine response times, resource capacity, medical infrastructure, and evacuation planning to better understand determinants of disaster mortality.

References

[1] Centre for Research on the Epidemiology of Disasters. EM-DAT: The international disaster database, 2025.

[2] D. Delforge, V. Wathelet, R. Below, C. L. Sofia, M. Tonnelier, J. A.F. van Loenhout, and N. Speybroeck. EM-DAT: the emergency events database. *International Journal of Disaster Risk Reduction*, 124:105509, 2025.

[3] UNDRR. The sendai framework terminology on disaster risk reduction. `https://www.undrr.org/drr-glossary/terminology7disaster`, 2017. Accessed: 2026-01-12.

10 April, 11:30 - 12:00, Hall of ESTGD

Municipality-level population projections in Portugal using regional forecasts

Francisco Branquinho[1], **Aitor Varea Oro**[2], **Rita Gaio**[1],
[1] Centro de Matemática da Universidade do Porto & Departamento de Matemática da Faculdade de Ciências da Universidade do Porto, up201804877@up.pt, argaio@fc.up.pt
[2] Faculdade de Arquitetura da Universidade do Porto - Centro de Estudos Nuno Portas, avoro@arq.up.pt

---

This work generates population projections for Portuguese municipalities using regional forecasts as a reference. Historical data were used to estimate each municipality's share of the regional population, which was then projected into the future using Exponential Smoothing models. The resulting municipal estimates provide higher-resolution insights into demographic trends, supporting the DataH project's mission of informed territorial planning.

Keywords: population projections, hierarchical time series, exponential smoothing, demographic modeling, territorial planning

---

The DataH – *Dynamic Analysis for Territorial Approaches to Housing* project [2] aims to improve public housing policies in Portugal using data-driven methods. By integrating housing, social, economic, and environmental data, DataH intends to provide insights into territorial diversity and inequalities, supporting a better understanding of local contexts. Part of this project consists of forecasting estimates for the number of inhabitants at the municipal level for the period of 2025–2100. Together with other municipal indices, this will allow for a precise and objective identification of the public housing needs in Portugal.
The projections of municipal populations were obtained from the national projections of INE [4] for 2025–2100 at the regional NUTS2 level, using the forecasting proportions method (Athanasopoulos et al., 2009 [1]). Historical population data collected from 1991 to 2024 were used to calculate each municipality's share of the regional population. These shares were modeled with Exponential Smoothing models and then multiplied by INE regional projections to estimate municipal populations. Unlike methods such as the Average Historical Proportions or Proportions of Historical Averages (Gross & Sohl, 1990 [3]), the previous approach uses forecasted rather than historical proportions, allowing gradual demographic changes to be accounted for. Modeling the proportions explicitly provides smooth structural evolution and remains consistent with the aggregate forecast.
The figure below presents the results obtained from the described methodology for the Madeira NUTS II region and for some municipalities of the North. In Madeira, the population of most municipalities tends to decrease with the exception of Porto Santo, with the sharpest decrease observed in the municipality of Funchal. In the North, Vila Nova de

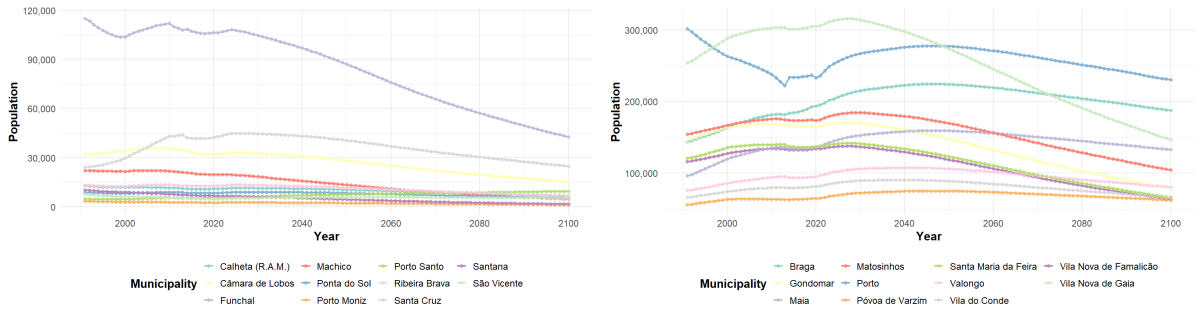

Figure 1: Population projections for Madeira (Left-hand side) and some municipalities of the North (Right-hand side).

Gaia seems to decrease too much. An adjustment to the methodology is therefore planned, through the inclusion of socio-economic variables.
The method's accuracy relies on the quality of the proportion forecasts, and since subpopulation counts are not forecast directly, local trends and cohort effects influence the projections only through proportions, which may result in a loss of detail at lower levels. Additionally, forecasting proportions independently does not account for interactions between municipalities. Critically, while these projections provide useful insights, there are important variables that could influence population dynamics but are not yet included in this method, such as the percentage of own revenue and other socio-economic or demographic indices.

References

[1] G. Athanasopoulos, R. A. Ahmed, and R. J. Hyndman. Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25:146–166, 2009.

[2] DataH Project Team. DataH – Dynamic Analysis for Territorial Approaches to Housing, 2025.

[3] C. W. Gross and J. E. Sohl. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9:233–254, 1990.

[4] Instituto Nacional de Estatística (INE). Projeções da população residente 2025–2100 por local de residência (nuts - 2024) [dataset]. Statistics Portugal, 2025.

11 April, 11:40 - 12:15, Hall of ESTGD

Analysis of financial indicators for small and medium-sized enterprises using ARIMA models

Débora Silva[1], Maria Almeida[1], Cristina Lopes[2], Cristina Torres[2], Isabel Vieira[2], Lurdes Babo[2]
[1] ISCAP, Instituto Politécnico do Porto, 2250336@iscap.ipp.pt, 2250344@iscap.ipp.pt
[2] CEOS.PP, ISCAP, Instituto Politécnico do Porto, cristinalopes@iscap.ipp.pt, ctorres@iscap.ipp.pt, mivieira@iscap.ipp.pt, lbabo@iscap.ipp.pt

This study forecasts key financial indicators for small and medium-sized enterprises based on monthly data from the Bank of Portugal. Seasonal ARIMA models were estimated and selected according to forecasting accuracy and residual diagnostics. Results indicate that the models capture temporal dynamics, with no significant residual autocorrelation, confirming the usefulness of ARIMA models for modelling seasonality and forecasting financial indicators.

Keywords: ARIMA models, financial indicators, seasonality, small and medium-sized enterprises, time series

The main objective of this study is to analyse the seasonality of financial indicators for small and medium-sized enterprises, using ARIMA models in IBM SPSS. Three indicators are analysed: Equity as a percentage of assets, Average Collection Period, and Financial Expense Coverage, using monthly data from the Bank of Portugal (2006-2025) [1].
Prior to model estimation, an exploratory analysis of the time series was conducted in order to identify trends, seasonal patterns and potential non-stationarity. Seasonality refers to a systematic intra-annual movement that is not necessarily regular and may be influenced by climatic changes, calendar effects, and decision-timing. Seasonality is a common feature of economic and financial time series and, if not properly modelled, may lead to biased inference and unreliable forecasts. Following the approach commonly adopted in the analysis of economic time series, seasonal behaviour was assessed and addressed through seasonal differencing and the inclusion of seasonal ARIMA components [2]. Based on this analysis, several seasonal ARIMA specifications were tested for each financial indicator, and the most appropriate models were selected according to their forecasting performance and residual diagnostics. Model selection was based on the Mean Absolute Percentage Error (MAPE), stationary R^2, the Bayesian Information Criterion (BIC), and the Ljung–Box statistic (to assess residual autocorrelation). Certain model specifications were excluded when the Ljung–Box test yielded values below 5%, indicating inadequate residual behaviour. Residual diagnostics were conducted using ACF and PACF plots. For all three financial indicators, the selected ARIMA models presented low error values, indicating good forecasting performance and supporting the adequacy of the chosen specifications.

Analysis of the model residuals showed no significant autocorrelation, confirming that the models provide reliable short-term forecasts for the analysed time series.
The results show that, for Equity Capital, ARIMA$(2,0,0)(1,1,0)_4$ outperformed other models, with $R^2 = 0.828$ and MAPE $= 1.319$. As shown in Figure 1 (left), the observed series and the corresponding forecasts reveal that the selected model is able to capture both the underlying trend and the seasonal component of the data. Forecasts indicate a structurally positive trajectory, consistent with sustained growth. Despite uncertainty, the overall outlook is favourable and suggests improved financial resilience.

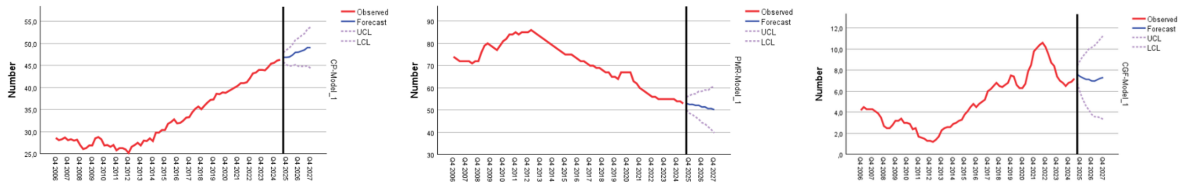

Figure 1: Equity Capital forecasts with ARIMA$(2,1,0)(1,1,0)_4$ (left), Average Collection Period forecasts with ARIMA$(1,0,0)(1,1,0)_4$ (center), and Financial Expense Coverage forecasts with ARIMA$(2,0,0)(2,1,0)_4$ (right)

For Average Collection Period, the chosen model was ARIMA$(1,0,0)(1,1,0)_4$ with MAPE $= 1.049$ and $R^2 = 0.815$. As shown in Figure 1 (center), the forecasts closely follow the historical evolution of the series, successfully reproducing its seasonal fluctuations, with relatively narrow confidence intervals. Projections point to a gradual decline, which may reflect continued improvements in firms' collection efficiency. Nevertheless, uncertainty remains, and future changes in commercial conditions may affect this trend.
For Financial Expense Coverage ARIMA$(2,0,0)(2,1,0)_4$ performed best, with good $R^2 = 0.893$ although with higher MAPE$= 8.409$, due to higher volatility and sensitiveness to financial conditions. As shown in Figure 1 (right), Financial Expense Coverage forecasts indicate relative stability with gradual adjustment, while widening confidence bands reflect increasing uncertainty over the forecast horizon. Given potential exposure to macroeconomic conditions, ongoing monitoring of this financial indicator remains advisable.
The forecasting results support SME risk assessment and financial health classification by providing forward-looking indicators of solvency, liquidity, and debt-servicing capacity, contributing to decision-making and financial stability monitoring and policy design. The identified quarterly seasonality suggests systematic financial cycles—such as recurring liquidity pressures and reporting effects—although future research could extend the univariate ARIMA framework to include macroeconomic and multivariate predictive models.

References

[1] Banco de Portugal. Séries estatísticas, 2019. https://bpstat.bportugal.pt/dados/explorer.

[2] A. Lopes. Sazonalidade em séries temporais económicas: uma introdução e duas contribuições. Technical Report 1001/2007, Instituto Superior de Economia e Gestão, 2007.

11 April, 11:40 - 12:15, Hall of ESTGD

Persistent homology analysis of unemployment dynamics

Flora Ferreira[1], **Jhonathan Barrios**[2]
[1] Centro de Matemática da Universidade do Porto & Faculdade de Economia da Universidade do Porto, flora.ferreira@fep.up.pt
[2] Centro de Matemática da Universidade do Minho, id10605@uminho.pt

---

We apply Topological Data Analysis (TDA) to monthly unemployment rates for the United States (USA) and the Euro-area. Persistent homology uncovers nonlinear dynamics and cross regional differences in labor market adjustment. Bottleneck and Wasserstein distances reveal slower and more persistent dynamics in the Euro-area and faster cyclical adjustment in the USA, showing that TDA can extract structural features from macroeconomic time series.

Keywords: persistence diagrams, persistent homology, topological data analysis, unemployment dynamics

---

Unemployment is an important macroeconomic indicator however characterizing the geometry of its dynamics remains a challenge. Traditional time-series models emphasize persistence and forecasting, but may obscure nonlinear structure or regime behavior. In contrast, Topological Data Analysis (TDA) provides a nonparametric and noise-robust framework that captures the intrinsic shape of the data without imposing functional form assumptions, allowing the detection of geometric patterns that remain stable under small perturbations. Recent applications of TDA in financial markets show that persistent homology can detect structural transitions and extreme events [1], suggesting that macroeconomic variables may also possess latent geometric features.

This work adopts TDA to examine unemployment dynamics time series in the United States (USA) and the Euro-area (EA20). We use monthly unemployment rates reported by the OECD for the USA and EA20, non-seasonally adjusted. The dataset covers $2000-2024$ and refers to individuals aged 25 years or over. We characterize the underlying dynamics by reconstructing the state space and examining the topological features of the embedded trajectories. If unemployment contains business cycle components or structural regimes, these should appear as geometric signatures in the reconstructed phase space. Because persistent homology is stable under perturbations and invariant to continuous deformations, it is particularly suitable for macroeconomic time series, which are often noisy and subject to measurement revisions. This robustness allows us to distinguish structural features from short-term fluctuations.

We reconstruct the state space using Takens' embedding $x(t) \mapsto [x(t), x(t+\tau)]$ with embedding dimension $d \in \{2, 3\}$ and time delays $\tau \in \{1, \ldots, 12\}$. The option $d = 2$ provides

planar phase portraits suitable for visualizing cyclic structure, while $d = 3$ incorporates inertial curvature that can reveal higher-order dynamics. All computations are implemented in Python using the giotto-tda library, alongside standard data-processing tools such as pandas and NumPy. Once the state space is reconstructed, persistence diagrams are built from Vietoris-Rips filtering. Here, persistent homology is calculated in connected components (H_0) and cycles (H_1), where persistent features H_1 indicate significant loops associated with cyclic dynamics, while fragmentation in H_0 may indicate hysteresis or structural breaks. Regions are then compared using bottleneck and Wasserstein distances. Bottleneck distances identify whether dominant cyclic features survive delays, while Wasserstein distances capture differences in the overall persistence distribution.

We observe a monotonic decline in both distances from $\tau = 1$ to approximately $\tau = 4$, indicating that the short horizon dynamics differs more sharply across regions than the medium horizon dynamics. For $d = 2$, bottleneck distances fall from roughly 2.99 at $\tau = 1$ to 0.99 at $\tau = 12$, while Wasserstein distances decline from 3.42 to 2.43. Raising the embedding dimension from $d = 2$ to $d = 3$ amplifies all distance measures, consistent with the inertia or hysteresis effects being more pronounced in EA20 relative to the USA. In addition, the results indicate that at short horizons the USA and EA20 operate under distinct transition dynamics, but at medium horizons they share a common business cycle geometry. The visual structure of the embeddings reinforces this interpretation: USA embeddings produce tighter and smoother loops, while EA20 embeddings appear more dispersed and distorted. Persistence diagrams further reveal stronger H_1 cycles for the USA and broader H_0 fragmentation for EA20, consistent with structural unemployment and crisis segmentation in the Euro-area.

This study demonstrates that TDA can extract significant structure from macroeconomic time series. The approach reveals differences in labor market adjustment and flexibility between the USA and EA20 and quantifies these differences across embedding scales. The proposed framework opens the door to topological analyses of macroeconomic dynamics, with extensions to multi country panels, multivariate attractors, and sliding window detection of structural change.

Acknowledgements F. Ferreira was partially supported by CMUP - Centro de Matemática da Universidade do Porto, a member of LASI, financed through FCT-Fundação para a Ciência e a Tecnologia, I.P., project UIDB/00144/2025 (doi: 10.54499/UID/00144/2025). J. Barrios acknowledges the FCT support of the PhD grant 2023.02242.BDANA (doi: 10.54499/2023.02242.BDANA).

References

[1] A. Rai, B. Nath Sharma, S. Rabindrajit Luwang, M. Nurujjaman, and S. Majhi. Identifying extreme events in the stock market: A topological data analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(10):103106, 10 2024.

11 April, 11:40 - 12:15, Hall of ESTGD

Towards robust stochastic gradient boosting for genomic prediction in breeding studies

Beatriz H. Comparado[1], **João Lourenço**[2], **Vanda M. Lourenço**[1]

[1] NOVA Math & Department of Mathematics, NOVA FCT, Portugal,
b.comparado@campus.fct.unl.pt, vmml@fct.unl.pt

[2] NOVA LINCS & Department of Computer Science, NOVA FCT, Portugal,
joao.lourenco@fct.unl.pt

The presence of data contamination can compromise the performance of machine learning methods like Stochastic Gradient Boosting (SGB), in high-dimensional genomic prediction (GP). This work addresses response contamination and evaluates the robustness of SGB via simulations on a synthetic animal breeding dataset. Our findings show that contamination reduces accuracy and demonstrate how SGB adaptations improve robustness, offering practical guidance for GP in breeding studies with imperfect data.

Keywords: genomic prediction, SNPs, machine learning, robustness, breeding studies

---

Genomic prediction (GP) is an essential tool in plant and animal breeding, where accurate estimates of genomic breeding values guide selection decisions. Because GP relies on thousands of molecular markers, it requires computational methods capable of handling high-dimensional data effectively. Machine learning (ML) methods have become increasingly popular in this setting due to their flexibility and ability to capture complex patterns in the data.

However, many ML methods are sensitive to data contamination, even at moderate levels. Contamination arising from measurement errors, unusual environmental effects, or data recording issues, can distort prediction errors and affect the reliability of genomic breeding values. This motivates the evaluation of ML methods robustness and the development of strategies to improve their predictive performance.

In this study, we assess the predictive performance and robustness of the classical SGB method, along with robust counterparts based on response weighting, Huber-type loss and median-based initialisation. Using a simulated animal breeding dataset (Table 1; [1]), we compare their performance across different contamination scenarios. Our findings help clarify how contamination affects SGB and show which adaptations improve robustness, offering practical guidance for genomic prediction in breeding studies where imperfect data are unavoidable.

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	-584.99	-116.24	-1.71	-4.0×10^{-6}	112.25	587.19	176.52

Data contamination & Robust extensions of SGB
Data contamination is introduced via Huber's contamination model $(1-\varepsilon)\mathbf{F} + \varepsilon\mathbf{G}$ where $\mathbf{F} \sim N(\mu, \sigma^2)$ is the animal data distribution with μ and σ^2 estimated from the data, and the contaminant distribution $\mathbf{G}$ is drawn from the following Normal distritutions $N(\mu + k\sigma, \sigma^2)$ (*shift* contamination), $N(\mu, (s\sigma)^2)$ (*variance-inflated* contamination), $N(\mu, (\sigma/\gamma)^2)$ (*variance-deflated* contamination), with $\varepsilon = 2, 5, 10\%$, $k = 5, 7, 9$, $\gamma = 1000, 10000$ & $s = 5, 7$. Robust extensions of SGB consider: (i) model initialization with the median instead of the mean (SGB-`m`); (ii) l_1 loss (SGB-`l`); (iii) Huber loss (SGB-`h`); and (iv) observation weighting (SGB-`w`).

Performance assessment
The performance accuracy of SGB and its robust variants is evaluated using metrics

$$\text{Predictive Accuracy (PA)} = \text{cor}(y, \hat{y}), \quad \text{RMSPE} = \sqrt{\tfrac{1}{n}\textstyle\sum_i (y_i - \hat{y}_i)^2}, \quad \text{MAPE} = \tfrac{1}{n}\textstyle\sum_i |y_i - \hat{y}_i| \,.$$

Results
Preliminary results show that: (i) *variance-deflated contamination* is not harmful to SGB performance (best performing in this scenario); (ii) SGB performance degrades more sharply under *variance-inflated* contamination, followed by *shift* contamination; (iii) all robust variants improve upon SGB in the *shift* and *variance-inflated* scenarios; and (iv) SGB-`w` achieved the best prediction/selection compromise across all scenarios: high PA (good for selection; PA ≥ 0.7) and low PEs (good for prediction). These findings suggest that weighting constitutes a simple, transparent, and easily implementable strategy, making it particularly attractive for routine use in GP, while further methodological refinement of this approach appears to be a promising direction for future research.

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 UID/297/2025, UID/PRR/297/2025 (Center for Mathematics and Applications - NOVA Math), UID/04516/2025 - `https://doi.org/10.54499/UID/04516/2025` (NOVA LINCS), and project 2023.14934.PEX (REACTION) - `https://doi.org/10.54499/2023.14934.PEX`. This work is also supported by FCT I.P. under the project 2023.14934.PEX.F1 – at Deucalion supercomputer, jointly funded by EuroHPC JU and Portugal.

References

[1] 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.

11 April, 11:40 - 12:15, Hall of ESTGD

A comprehensive abstraction and classification tool to identify biomarkers for age-related macular degeneration onset and progression

Alina Humenyuk[1], **Luca Gherardini**[2], **Rita Coimbra**[1], **Cláudia Farinha**[3], **Patrícia Barreto**[1], **José Sousa**[4], **Rufino Silva**[3]
[1] AIBILI - Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal, ahumenyuk@aibili.pt, racoimbra@aibili.pt, pbarreto@aibili.pt
[2] Sano Centre for Personalised Computational Medicine, Krakow, Poland, l.gherardini@sanoscience.org
[3] Ophthalmology Department, Unidade Local de Saúde de Coimbra (ULS Coimbra), Coimbra, Portugal, claudia.farinha@hotmail.com, rufino.silva@oftalmologia.co.pt
[4] Multidisciplinary Institute of Ageing, MIA-Portugal, University of Coimbra, 3004-504, Coimbra, Portugal, j.sousa@sanoscience.org

Age-related macular degeneration (AMD) is a leading cause of blindness in older adults. Using CACTUS, an explainable artifical intellogence (AI) tool, we ranked biomarkers for AMD onset and progression in the Mira cohort of the Coimbra Eye Study. Key factors included age, BMI, diabetes, and genetic variants linked to lipid and complement pathways. CACTUS achieved good balanced accuracies, offering insights into AMD physiopathology and supporting personalised medicine.

Keywords: age-related macular degeneration, classification, explainable artificial intelligence

Age-related macular degeneration (AMD) is a neurodegenerative disease of the macula and the leading cause of blindness in individuals over 55 years in developed countries. Owing to its multifactorial etiology, advanced analytical approaches - such as explainable AI - are needed to identify and prioritize biomarkers associated with disease onset and progression. In this study, we applied the Comprehensive Abstraction and Classification Tool for Uncovering Structures (CACTUS) to rank AMD onset and progression biomarkers in the Mira cohort of the two-visit Coimbra Eye Study [1]. CACTUS abstracts continuous features using optimal cut-off values, preserves categorical variables, constructs class-specific knowledge graphs and ranks features based on distributional changes across classes [2]. Two models were analyzed: Model 1 included AMD cases at baseline ($n = 297$), classified as non-progressors, slow progressors (1 step), or fast progressors (≥ 2 steps). Model 2 included all participants at follow-up ($n = 1617$), categorized as no-AMD (stages 0a, 0b, 1b; $n = 1237$) or AMD (stages 1a, 2a–4; $n = 380$) according to the Rotterdam classification.

Data included demographics, lifestyle factors (Mediterranean diet adherence [mediSCORE 0-9], physical activity), ophthalmological assessments, and genetic profiles.
Similar balanced accuracies were obtained for both models using the PageRank approach (Model 1: 0.682; Model 2: 0.722). For AMD progression, fast progressors were primarily characterized by older age, 1 risk allele at rs73036519 (*APOE* gene), high body mass index (BMI), midsubfield drusens and diabetes. For AMD onset, controls were more likely to lack Subretinal Drusenoid Deposits (SDD), be younger, carry protective alleles at rs1410996 and rs10922109 (both at *CFH* gene) and have high BMI. Age, BMI, and diabetes ranked among the top 10 markers in both models. Key genetic variants for progressors were linked to the lipid pathway, whereas those related to AMD onset were predominantly associated with the complement pathway; extracellular matrix pathway variants contributed to both. CACTUS performed well in both Models, especially in Model 1 despite the small sample size. Future research should compare its performance to other state-of-the-art machine learning algorithms in order to validate these findings. The results are consistent with existing literature and provide valuable insights into AMD physiopathology by identifying critical pathways within our population. This work contributes to personalised medicine by demonstrating the algorithm's ability to uncover class-specific connectivity patterns across biomarkers, which warrants further investigation given the complex aetiology of AMD.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857524. This publication is partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement "Sano" No. 857533, and by the "Sano" project, carried out within the framework of the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13, co-financed by the European Regional Development Fund. The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

References

[1] C. V. L. Farinha, M. L. Cachulo, D. Alves, I. Pires, J. P. Marques, P. Barreto, S. Nunes, J. Costa, A. Martins, I. Sobral, I. Laíns, J. Figueira, L. Ribeiro, J. Cunha-Vaz, and R. Silva. Incidence of age-related macular degeneration in the central region of portugal: The coimbra eye study - report 5. *Ophthalmic Res.*, 61(4):226–235, 2019.

[2] L. Gherardini, V. R. Varma, K. Capała, R. Woods, and J. Sousa. Cactus: A comprehensive abstraction and classification tool for uncovering structures. *ACM Trans. Intell. Syst. Technol.*, 15:23, 2024.

11 April, 11:40 - 12:15, Hall of ESTGD

Clustering of residential characteristics influencing indoor air quality in Europe

Beatriz Saraiva[1], **Marta Gabriel**[2], **Rita Gaio**[3]
[1] Department of Mathematics, Faculty of Sciences, University of Porto & Institute of Science and Innovation in Mechanical and Industrial Engineering, Porto, Portugal, up202408844@edu.fc.up.pt
[2] LAETA - INEGI, Associated Laboratory of Energy, Transports and Aerospace - Institute of Science and Innovation in Mechanical and Industrial Engineering, Rua Dr. Roberto Frias 400, 4200-465, Porto, Portugal, mgabriel@inegi.up.pt
[3] Department of Mathematics, Faculty of Sciences, University of Porto & Centre of Mathematics of the University of Porto, Porto, Portugal, argaio@fc.up.pt

Air pollution is one of the greatest threats to global public health, accounting for 6.7 million deaths in 2019. In Europe, people spend about 90% of their time indoors, mainly in residential dwellings, whose characteristics vary across countries and influence pollutant concentrations. This study aimed to analyze housing characteristics in European countries using clustering methods.

Keywords: indoor air quality, housing characteristics, clustering

This study was based on questionnaire data from 205 households: 25 in the Czech Republic (CZ), Estonia (EE), Italy (IT), the Netherlands (NL), Portugal (PT), Sweden (SE) and the United Kingdom (UK) and 30 in Slovenia (SI). To group dwellings by characteristics, which included continuous and categorical variables, three clustering methods were applied: partitioning around medoids (PAM), hierarchical and finite mixture models (FMM). PAM identifies representative observations (medoids) to minimize the sum of dissimilarities between each observation and its nearest medoid [1]. This method allows any dissimilarity measure, so we used the DAFI-Gower distance, which balances continuous and categorical contributions [4]. For continuous variables, the dissimilarity was calculated using the Manhattan distance normalized by the interquartile range. Categorical variables were encoded as dummies and the dissimilarity was adjusted to align with the average contribution of continuous variables, preventing dominance. Importance was incorporated through weights derived from normalized mutual information. Hierarchical clustering successively splits or merges clusters using a dissimilarity measure, with the DAFI-Gower distance used in this study [2]. We applied an agglomerative approach with Ward's method, which minimizes total within-cluster variance, and complete linkage, which considers the maximum distance between clusters. FMM assumes that the data can be represented as a linear combination of a finite number of probability distributions, typically modeling continuous variables with normal distributions and categorical with independent multinomial distributions [3].

Model parameters were estimated by maximum likelihood and the number of components was chosen using a model selection criterion. For each method, the chi-squared, Fisher or Kruskal–Wallis test were applied to assess statistically significant differences in variable distributions across clusters. All analyses were performed in R version 4.4.2.

Table 1: Contingency table of hierarchical clusters by country for both linkage methods

	CZ	EE	IT	NL	PT	SE	SI	UK
Cluster 1	4	1	20	4	8	2	20	17
Cluster 2	20	24	2	15	17	23	9	8

Based on the average silhouette coefficient, hierarchical clustering showed better separation between clusters, with a value of 0.28 for both linkage criteria. Cluster 1 mainly consisted of households from IT, SI and the UK, while the remaining countries are grouped in cluster 2 (Table 1). A significant difference was observed in the use of electric cooling systems ($p < 0.0001$), with all individuals in cluster 1 using them and none in cluster 2. Regarding cooking systems, electricity was mostly used in cluster 2 ($p = 0.0004$), whereas natural gas was more frequently used in cluster 1 ($p = 0.0001$). For ventilation and climate control systems, usage was higher in cluster 1 compared to cluster 2 ($p < 0.0001$). The absence of mechanical ventilation systems was more frequent in cluster 1 ($p = 0.0102$). Water heating using bottled gas was more frequent in cluster 1 ($p = 0.0053$). The number of plants was significantly higher in cluster 2 ($p < 0.0001$) and indoor smoking occurred more in cluster 1 ($p = 0.0430$).
This study highlights differences in housing characteristics across European countries, which are important for assessing indoor air quality and guiding future research.

Acknowledgements The authors express their gratitude to all participating families and to their colleagues from the INQUIRE project Consortium (Grant No. 101057499), co-funded by the European Union. Rita Gaio was partially supported by CMUP, member of LASI, financed by FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project UID/00144/2025 and DOI: `https://doi.org/10.54499/UID/00144/2025`.

References

[1] S. Balzano, G. C. Porzio, R. Salvatore, D. Vistocco, and M. Vichi. *Statistical Learning and Modeling in Data Analysis*. Springer, 2021.

[2] A. Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017.

[3] F. Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004.

[4] P. Liu, H. Yuan, Y. Ning, B. Chakraborty, N. Liu, and M. A. Peres. A modified and weighted gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, 24(1):305, 2024.

11 April, 11:40 - 12:15, Hall of ESTGD

A variable influence analysis approach to risk assessment in age-related macular degeneration

Rita Coimbra[1], **Joana Martins**[2], **Cláudia Farinha**[3], **Patrícia Barreto**[1], **Rufino Silva**[3], **Eugénio Rocha**[2]

[1] AIBILI – Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal, racoimbra@aibili.pt, pbarreto@aibili.pt

[2] Department of Mathematics, University of Aveiro, Aveiro, Portugal, joanadirce@ua.pt, eugenio@ua.pt

[3] Ophthalmology Department, Unidade Local de Saúde de Coimbra (ULS Coimbra), Coimbra, Portugal, claudia.farinha@hotmail.pt, rufino.silva@oftalmologia.co.pt

Age-related macular degeneration (AMD) is a multifactorial disease influenced by genetic and environmental factors. We propose a Variable Influence Analysis (VIA) model, an extension of Bayesian Networks using Belief Propagation, to explore the relative influence of different risk factors on development and progression of AMD. The VIA model suggests that targeting key modifiable factors may help reduce AMD risk in high-risk individuals.

Keywords: age-related macular degeneration, variable influence analysis, bayesian networks

Age-related macular degeneration (AMD) is a complex multifactorial disease strongly influenced by a combination of genetic and environmental factors and it is the leading cause of severe vision loss in people over 55 years in developed countries [3].

Current approaches to explore the relative influence of the different risk factors in AMD development and progression have notable limitations: machine learning and deep learning models often lack interpretability, while Bayesian Networks (BNs) rely on independence assumptions and require large datasets, conditions that are not always satisfied in medical datasets.

To address these limitations, we propose a novel Variable Influence Analysis (VIA) model, an extension of BNs that applies the Belief Propagation algorithm on factor graphs. The VIA framework relaxes independence assumptions and incorporates diverse relationships - referred to as metrics - between patient characteristics and transitions across AMD stages. This approach enables the estimation of influence scores and the computation of individualized patient AMD risk scores, even in the presence of class imbalance.

Data to test the model were obtained from the Coimbra Eye Study (CES), a population-based epidemiological study designed to estimate AMD prevalence, 6.5-year incidence, and associated risk factors [1, 2]. Participants underwent multimodal retinal imaging for AMD phenotyping, and blood samples were collected for genetic analysis.

A total of 948 participants were included, of whom 243 were diagnosed with AMD. Adherence to the Mediterranean diet, physical activity, and body mass index emerged as causative factors influencing AMD progression, while baseline disease stage, most genetic variants, age and smoking were identified as predictive factors. As disease severity increased, smoking and diabetes gained greater relevance in predicting progression to advanced stages. Using the derived influence scores, a global AMD risk score was calculated for each individual, allowing the evaluation of how changes in modifiable risk factors could reduce the likelihood of disease progression.
In summary, the VIA model provides a flexible framework for identifying high-impact, modifiable risk factors that may help reduce AMD progression in individuals at elevated risk. It offers valuable insights for personalized risk assessment and preventive strategies in AMD. However, it relies on prior knowledge to guide metric selection, and its use of population-specific training data may restrict the generalization of the findings.

Acknowledgements This study was financially supported by Novartis.

References

[1] M. Cachulo, C. Lobo, J. Figueira, L. Ribeiro, I. Laíns, A. Vieira, S. Nunes, M. Costa, S. Simão, V. Rodrigues, N. Vilhena, J. Cunha-Vaz, and R. Silva. Prevalence of age-related macular degeneration in portugal: The coimbra eye study - report 1. *Ophthalmologica*, 233, 2015.

[2] C. V. L. Farinha, M. L. Cachulo, D. Alves, I. Pires, J. Marques, P. Barreto, S. Nunes, J. Costa, A. Martins, I. Sobral, I. Laíns, J. Figueira, L. Ribeiro, J. Cunha-Vaz, and R. Silva. Incidence of age-related macular degeneration in the central region of portugal: The coimbra eye study – report 5. *Ophthalmic Research*, 61:1–10, 2019.

[3] M. Fleckenstein, T. D. L. Keenan, R. H. Guymer, U. Chakravarthy, S. Schmitz-Valckenberg, C. C. Klaver, W. T. Wong, and E. Y. Chew. Age-related macular degeneration. *Nature Reviews Disease Primers*, 7:31, 2021.

11 April, 11:40 - 12:15, Hall of ESTGD

Chronic diseases and social determinants of health in a european context: a comparative spatial analysis

A. Caramez[1], **A. Nóbrega**[1], **J. Sousa**[1], **M. Leão**[1], **J. Mendonça**[1]

[1] Instituto Superior de Engenharia do Porto (ISEP), 1211327@isep.ipp.pt, 1211471@isep.ipp.pt, 1201196@isep.ipp.pt, 1210979@isep.ipp.pt, jpm@isep.ipp.pt

---

Keywords: chronic diseases, spatial analysis, social determinants of health, composite burden of disease index

---

Chronic noncommunicable diseases represent a primary public health challenge in Europe, contributing substantially to morbidity and persistent territorial inequalities. Understanding their spatial distribution and their relationship with sociodemographic determinants is essential for informing more equitable health policies [1, 2].

This study analyses the spatial distribution of four chronic diseases - diabetes, asthma, high blood pressure, and chronic depression - across European countries and examines their association with selected social and structural determinants. Inspired by the methodological framework of Benavidez et al. [3], the study adapts an integrated spatial analysis approach to the European context, characterised by marked socioeconomic heterogeneity and predominantly universal health systems.

National-level aggregated data on disease prevalence, income, education, unemployment, population, and health resources were obtained from official European sources [4].

The analytical strategy included the construction of choropleth maps to assess spatial patterns, boxplots to examine distributional differences, Pearson correlation matrices to evaluate statistical associations, and group comparison tests. To synthesise the overall burden of chronic disease, a composite index was constructed following a relative ranking approach. For each pathology, national prevalence values were classified into tertiles, assigning a score of 0 (low prevalence), 1 (medium prevalence), or 2 (high prevalence) according to the country's position within the distribution. The tertile scores for the four diseases were then summed, resulting in a composite index ranging from 0 to 8. Higher values indicate a greater cumulative burden of chronic disease relative to other countries. This procedure allows for a standardised comparison across pathologies with different prevalence scales and facilitates the identification of countries with systematically elevated disease burden.

Figure 1 presents a cartographic representation of the composite index, providing a visual summary of territorial disparities in chronic disease burden.

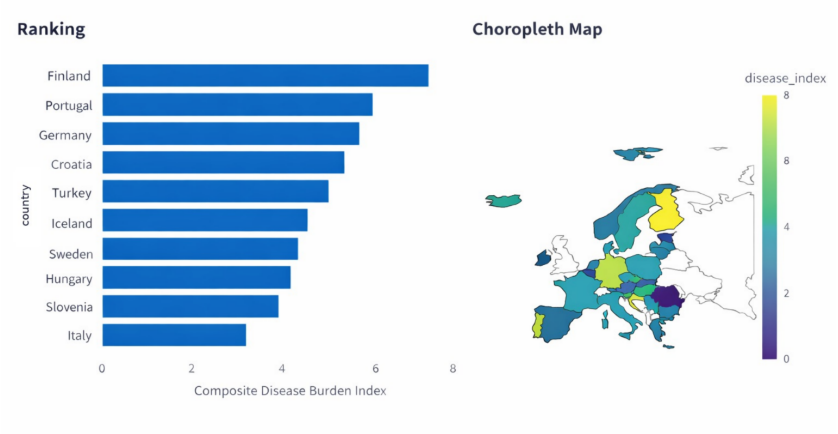

Figure 1: Composite disease burden index - ranking and choropleth map.

The results indicate that chronic disease prevalence is unevenly distributed across Europe, revealing distinct geographical patterns. Associations between diseases and determinants vary by pathology. Income and education emerge as relevant factors, although their effects differ depending on the disease analysed, while health resources show more specific and context-dependent associations.
Despite the limitations associated with aggregated national level data, the findings highlight the usefulness of spatial analysis and composite indices in identifying territorial inequalities in chronic disease burden. These approaches provide a structured framework for supporting geographically informed and socially responsive public health strategies.

References

[1] OECD and European Union. *Health at a Glance: Europe 2022: State of Health in the EU Cycle*. OECD Publishing, Paris, 2022.

[2] K. Marynia, J. Bhatt, Y. H. Park, N. A. Padrón, and A. Molefe. Quantification of neighborhood-level social determinants of health in the united states. *JAMA Network Open*, 3(1):e1919928, 2020.

[3] G. A. Benavidez, W. E. Zahnd, P. Hung, and J. M. Eberth. Chronic disease prevalence in the us: Sociodemographic and geographic variations by zip code tabulation area. *Preventing Chronic Disease*, 21:E14, 2024.

[4] Eurostat. Persons reporting a chronic disease, by disease, sex, age and educational attainment level (hlth_ehis_cd1e) [dataset]. Eurostat Data Browser, 2025.

11 April, 11:40 - 12:15, Hall of ESTGD

Spatio-temporal analysis of cancer mortality and incidence in Europe (1953–2023)

Ana Sofia Pinheiro[1], **Jéssica Pinto**[1], **Mariana Pinto**[1], **Sofia Nogueira**[1], **Jorge Mendonça**[1]
[1] Instituto Superior de Engenharia do Porto, 1210930@isep.ipp.pt, 1240085@isep.ipp.pt, 1210999@isep.ipp.pt, 1240085@isep.ipp.pt, 1210684@isep.ipp.pt, jpm@isep.ipp.pt

By the end of the 20th century, around 2.5 million cancer deaths had been registered in Europe. The cancers with the highest mortality rates are lung, colon, breast and pancreatic cancer. This study aims to conduct a statistical and graphical analysis of the mortality rates and the mortality-to-incidence ratio (MIR) for the four cancer types mentioned, comparing genders and identifying peak years. The data is explored by European regions . It also proposes measures to reduce mortality based on international examples.

Keywords: geographical analysis, cancer mortality, regional disparities, temporal trends

Cancer remains one of the leading causes of mortality in Europe. This is largely due to lifestyle risks, socioeconomic gaps, and environmental exposures, alongside a worrying shift toward earlier onset in several types of the disease [3]. The assessment of cancer trends relies on population-based registries, advanced statistical models, and spatial epidemiology to identify temporal patterns, geographic inequalities, and associated risk factors [2]. This study aims to analyze mortality rates and mortality-to-incidence ratios for the four most lethal cancers in Europe (lung, colon, breast, and pancreatic) using recent data, with a focus on gender disparities, identification of peak mortality years, regional variations, and the formulation of strategies to mitigate mortality.

Statistical analysis was performed using Age-Standardized Rates (ASR) to compare populations, complemented by the calculation of the Mortality-to-Incidence Ratio (MIR) as a proxy for survival [4]. Epidemiological data were obtained from the European Cancer Information System (ECIS) [1], providing harmonized information on cancer incidence and mortality across Europe. The dataset comprised 10,451 records covering 128 European regions over a 70-year period (1953–2023). Cartographic data were retrieved from Eurostat using the Nomenclature of Territorial Units for Statistics (NUTS), enabling spatial analysis at country and regional levels. Mortality and incidence data were analyzed using ASR per 100,000 inhabitants to ensure comparability across populations and time. Mortality trends were stratified by sex, the year of peak mortality was identified for each cancer type, and the MIR was calculated as an indicator of disease lethality and health system performance, supporting the assessment of long-term temporal trends.

Table 1 summarizes the obtained information aggregated by sex, highlighting the peak year for each cancer, as well as the associated age standardized rate. It also shows the variation in this rate and the MIR over the last 5 years. The country with the highest average mortality in recent years is also presented.

Table 1: Mortality trends and MIR for the selected cancers.

Cancer	Peak Year	ASR (Peak)	Δ Mortality (5y)	Δ MIR (5y)	Top Country (5y)
Lung	1995	70.66	[50.45 ; 34.55]	[0.63 ; 1.07]	Portugal
Colon	1978	29.38	[16.46 ; 20.92]	[0.30 ; 0.76]	Croatia
Breast	1989	27.26	[13.75 ; 17.92]	[0.15 ; 0.41]	Liechtenstein
Pancreatic	1973	19.44	[16.11 ; 19.07]	[0.92 ; 1.51]	Czech Republic

Lung cancer is the most concerning, with the highest mortality rate in Europe. In contrast, pancreatic cancer has the lowest value. Lung cancer has remained the most concerning pathology in recent years, and Portugal has the highest mortality rate in Europe. The lethality of pancreatic cancer is also particularly worrying, as it is approaching the peak recorded in 1973. Breast cancer has the lowest values for this ratio, and is the disease with the most encouraging prognosis.
A spatiotemporal analysis of cancer mortality and incidence in Europe (1953–2023) reveals a general downward trend in mortality rates for various cancers, with notable historical peaks occurring during the 20th century, particularly for lung cancer in men. However, rising mortality rates from colorectal and pancreatic cancers in both genders, alongside the growth of lung cancer mortality, represent critical public health challenges. Future research should include socio-economic and molecular factors to better understand long-term effects of personalised therapies on survival in Europe.

References

[1] ECIS - European Cancer Information System. European commission, joint research centre. `https://ecis.jrc.ec.europa.eu`, 2025.

[2] F. Giusti, C. Martos, R. N. Carvalho, V. Zadnik, O. Visser, M. Bettio, and L. Van Eycken. Facing further challenges in cancer data quality and harmonisation. *Frontiers in Oncology*, 14:1438805, 2024. `https://doi.org/10.3389/fonc.2024.1438805`.

[3] D. Majcherek, M. A. Weresa, and C. Ciecierski. A cluster analysis of risk factors for cancer across EU countries: Health policy recommendations for prevention. *International Journal of Environmental Research and Public Health*, 18(15):8142, 2021. `https://doi.org/10.3390/ijerph18158142`.

[4] H. Sung, C. Jiang, P. Bandi, A. Minihan, M. Fidler-Benaoudia, F. Islami, R. L. Siegel, and A. Jemal. Differences in cancer rates among adults born between 1920 and 1990 in the USA: an analysis of population-based cancer registry data. *The Lancet Public Health*, 9(8):583–593, 08 2024.

11 April, 11:40 - 12:15, Hall of ESTGD

Electrical current forecasting for monitoring industrial heat treatment processes

José Febra[1], **F. C. Batista**[1,2]
[1] CDRSP – Centre for Rapid and Sustainable Product Development, Polytechnic of Leiria, Portugal, zefebra@gmail.com
[2] School of Technology and Management, Polytechnic of Leiria, Portugal, fernando.batista@ipleiria.pt

---

This work presents a data-driven monitoring approach for industrial heat treatment based on short-horizon forecasting of three-phase electrical currents. A bidirectional LSTM is trained for one-step-ahead prediction, which defines a reference model of normal behaviour and supports error-based baselines. In parallel, the same model is used in an autoregressive roll-out to simulate full-cycle dynamics, enabling the analysis of error accumulation and the identification of critical process phases.

Keywords: predictive monitoring, time series forecasting, industrial processes, LSTM networks, heat treatment

---

Industrial heat treatment processes exhibit highly non-linear thermal–electrical dynamics, where electrical currents reflect both energy demand and process evolution [4]. Traditional monitoring approaches based on static thresholds or aggregated indicators are limited in capturing transient behaviour and early-stage deviations, motivating data-driven time series models for anomaly detection through forecasting errors [1, 3, 2].

In this work, the monitoring problem is formulated as a short-horizon forecasting task. A bidirectional Long Short-Term Memory (LSTM) network is trained to perform one-step-ahead prediction of three-phase electrical currents using sliding temporal windows of process variables. The key contribution of this work is the proposal of a unified monitoring framework that combines short-horizon prediction, autoregressive roll-out analysis, and percentile-based residual thresholding to establish adaptive, data-driven reference limits for industrial operation. This configuration enables the model to learn the local dynamics of normal operation while preserving interpretability of prediction errors. Prediction residuals obtained on the training set are summarised using percentile-based statistics, which define reference thresholds for subsequent analysis [2].

The experimental dataset comprises around two thousand industrial heat treatment cycles ranging from short exploratory runs to extended treatments exceeding twenty hours, spanning multiple cycle types with variability driven by material type, furnace load, and part geometry.

In autoregressive roll-out mode, the model reconstructs full-cycle evolution, revealing error amplification during abrupt transitions and dynamically sensitive phases. Latent bottleneck representations are further analysed to characterise process structure over time [3]. Full-cycle current forecasting also enables indirect energy consumption estimation, supporting early anomaly detection and proactive energy management in multi-furnace environments.

Error analysis confirms stable one-step-ahead performance under normal conditions, with residuals largely below the 95th percentile baseline. In contrast, errors obtained in autoregressive roll-out mode amplify during abrupt transitions, occasionally exceeding the 99th percentile threshold, highlighting sensitive phases and reinforcing the practical relevance of percentile-based limits for interpretable process monitoring.

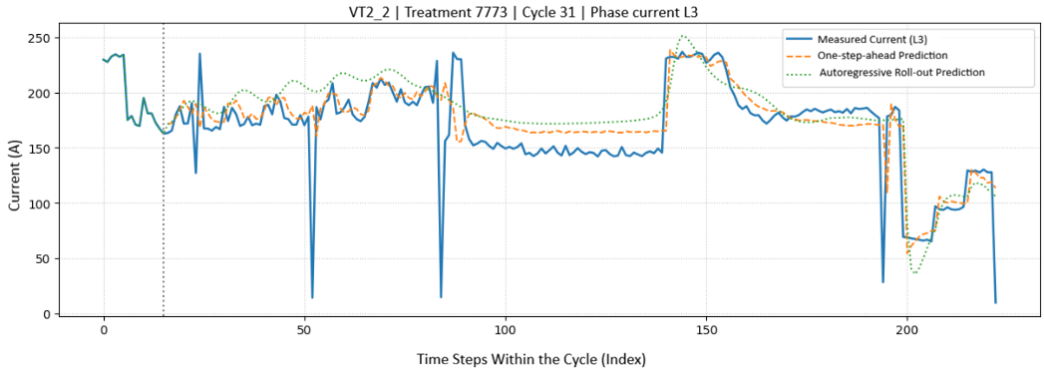

Figure 1: Measured and predicted phase current (L3) for a representative heat treatment cycle, comparing one-step-ahead prediction and autoregressive roll-out modes.

Acknowledgements This work was funded by Packaging of the Future ®, Green Agenda for Business Innovation, investment project nº 59, namely PPS16, funded by the PRR – NextGenerationEU.

References

[1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

[2] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 387–395, 2018.

[3] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

[4] G. E. Totten and M. A. H. Howes. *Steel Heat Treatment Handbook*. CRC Press, 2007.

11 April, 11:40 - 12:15, Hall of ESTGD

Clustering smartwatch heart rate features for subject and medication-state analysis

Jhonathan Barrios[1], **Miguel Gago**[2,3], **Wolfram Erlhagen**[1], **Estela Bicho**[4], **Flora Ferreira**[5]
[1] Centre of Mathematics, School of Sciences, University of Minho, id10605@uminho.pt, wolfram.erlhagen@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
[5] Centre of Mathematics of the University of Porto, flora.ferreira@fep.up.pt

We evaluate whether smartwatch-derived heart rate features allow clustering of Parkinson's disease subjects or medication state. Ten participants were monitored before and after treatment adjustment. Features from daytime and nighttime series were extracted, including variability, extremes, and Poincaré descriptors. Clusters mainly reflected subject identity, with no clear medication-state separation.

Keywords: digital phenotyping, heart rate time series, Parkinson's disease, clustering, wearable devices

The use of wearable devices has enabled the continuous recording of physiological signals in outpatient settings, offering new opportunities for longitudinal health analysis. In patients with Parkinson's disease (PD) and sleep disorders, such as REM sleep behavior disorder, the objective characterization of physiological patterns is particularly complex due to the coexistence of interindividual variability, medication effects, and daily fluctuations. In this context, heart rate (HR) is a signal of interest, as it indirectly reflects the interaction between autonomic control, arousal state, and sleep architecture [1]. However, the extent to which smartwatch-derived HR dynamics support unsupervised tasks, such as subject phenotyping or medication-state characterization, remains unclear. This work evaluates whether different HR feature subspaces carry subject-specific or medication-state structure in individuals with PD, considering both daytime and nighttime HR data.
HR time series was obtained with a commercial Amazon Fit smartwatch were analyzed in ten Parkinson's patients, who were evaluated before and after a therapeutic adjustment. Each subject contributed multiple daily series, each daily recording was treated as one

observational unit. Clustering was therefore performed across daily heart rate time series segments (daytime or nighttime), represented through feature vectors, rather than directly clustering raw time series or subjects as single entities. From each daily recording, two signals were defined: all-day series and nighttime series, extracted between 11:00 PM and 6:00 AM. Aggregate descriptors were extracted time dependency was handled implicitly through summary descriptors capturing temporal variability, extremes, and nonlinear dynamics. Thus, instead of aligning sequences pointwise, the approach maps each time series into a low-dimensional feature space that preserves information about its temporal structure from each series and five subsets of variables with explicit hypotheses were defined: baseline variability, suprathreshold activity, single extremes, full extremes profile, and nonlinear descriptors derived from the Poincaré plot (SD1 and SD2). k-means analysis was applied to each subset, selecting the number of clusters using the silhouette index. The alignment between clusters and external labels was assessed post-hoc using the Adjusted Rand Index (ARI) and cluster purity.

The results indicate that, although smartwatch derived HR time series contain internal structure, they do not provide sufficiently discriminating information to reliably separate medication states. Descriptors of baseline variability induce clusters aligned primarily with patient identity, especially in daytime recordings, reaching ARI values of approximately 0.25–0.30, while alignment with medication status is practically nil ($\text{ARI}_{status} \approx 0$), suggesting the capture of stable phenotypic signatures rather than therapeutic changes. Metrics based on suprathreshold and extreme events show high Silhouette Index values ($\approx$ 0.65–0.75), but generate highly unbalanced partitions dominated by rare episodes, without relevant clinical alignment ($\text{ARI} < 0.05$), which limits their usefulness to infer improvements associated with the intervention. The nonlinear descriptors SD1 and SD2, on the other hand, reveal a consistent dynamic structure (silhouette $\approx$ 0.5–0.6), but without clear separation by patient or medication status ($\text{ARI}_{patient} \approx 0.05$–$0.15$; $\text{ARI}_{status} \approx 0$), indicating that they characterize latent autonomic regimes rather than therapeutic effects. Taken together, these results suggest that, given the indirect, noisy, and limited nature of HR data from smartwaches, while useful for exploratory characterization, they have limited sensitivity to evaluate medication-related changes in PD associated sleep disturbances.

Acknowledgements We thank the Foundation for Science and Technology (FCT) for the doctoral scholarship 2023.02242.BDANA (doi: 10.54499/2023.02242.BDANA) and the support of Portuguese Funds through FCT within the Project UID/00013/2025 (doi: 10.54499/UID/00013/2025).

References

[1] P. Concheiro-Moscoso, B. Groba, D. Alvarez-Estevez, M. Miranda-Duro, T. Pousada, L. Nieto-Riveiro, F. Mejuto-Muiño, and J. Pereira. Quality of sleep data validation from the xiaomi mi band 5 against polysomnography: Comparison study. *Journal of Medical Internet Research*, 25:e42073, May 2023.

Index

Author Index