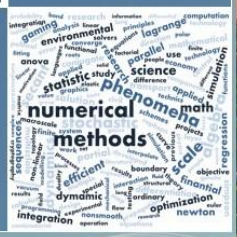
WCDANM | 2026

X Workshop on Computational Data Analysis and Numerical Methods

June
11-13 / 2026

University of Minho

(Guimarães)



https://w3.math.uminho.pt/WCDANM2026/

Book of Abstracts

University of Minho
Guimarães, Portugal

BOOK OF ABSTRACTS
X WCDANM

University of Minho
Portugal
June 11–13, 2026

WELCOME TO THE X WCDANM | 2026

Dear participants, colleagues and friends,

On behalf of the Executive and Organizing Committees of the X WCDANM (Workshop on Computational Data Analysis and Numerical Methods), we are delighted to welcome you to this workshop, hosted by the University of Minho at the Azurém campus in Guimarães, Portugal. Guimarães is widely recognized as the "Birthplace of the Portuguese Nation," as it was the setting for the key political and military events that led to Portugal's independence in the 12th century. In recognition of its remarkable historical heritage and exceptionally well-preserved architecture, the city's historic center was designated a UNESCO World Heritage Site in 2001. This year, the event is also supported by research centers from several Portuguese universities, including the University of Minho, the University of Porto, the University of Aveiro, and the University of Évora. We are committed to delivering an exceptional experience and to exceeding the expectations of all participants, sponsors, and organizers. Distinguished international scholars, including Anuj Mubayi (Intercollegiate Biomathematics Alliance, IL & South Mountain Community College, AZ, USA), Carlos Braumann (University of Évora & CIMA, Portugal), Jörg Henseler (University of Twente, the Netherlands & NOVA University Lisbon, Portugal), and Milan Stehlík (University of Valparaiso, Chile & University of Applied Sciences Upper Austria, Austria), will deliver keynote lectures at the conference. The program will include around 60 presentations of scientific papers spanning diverse research areas, providing a vibrant platform for discussion and the exchange of ideas. The active participation and engagement of the scientific community are essential to the success of this event. This year, the X WCDANM also features two additional specialized courses. Anuj Mubayi and Dharmendra Tripathi (National Institute of Technology, Uttarakhand, India) will conduct a course entitled "Numerical Methods in Health", while Jörg Henseler will coordinate the course "Structural Equation Modeling with Latent Variables and Formative Constructs." We would like to express our sincere gratitude to the Professors/Researchers for kindly accepting our invitation to teach these courses and thus contributing to the success of this year's workshop. We would also like to express our sincere gratitude to the members of the Executive, Scientific, and Organizing Committees for their dedication and commitment to the success of the X WCDANM. In particular, we acknowledge the invaluable contributions of Anuj Mubayi, Luís Meira-Machado, and Milan Stehlík from the Executive Committee. We also extend our appreciation to the members of the local Organizing Committee for their tireless

efforts, namely A. Manuela Gonçalves (Local Chair, University of Minho), Adelaide Freitas (University of Aveiro), Helena Luzia Grilo (Open University), Inês Sousa, Luís Meira-Machado, Marta Ferreira, Soraia Pereira, and Susana Faria (University of Minho, Portugal).The scientific contributions presented at the meeting are expected to be published in special issues of the Journal of Applied Statistics and Research in Statistics (Taylor & Francis Group). We anticipate that the X WCDANM will provide an inspiring platform for intellectual exchange, scientific collaboration, and the dissemination of innovative research. We extend our best wishes to all participants for a productive, engaging, and enjoyable workshop.

University of Minho, Guimarães, June 11-13, 2026.

Chairman of the Executive Committee of X WCDANM,

Luís Miguel Grilo
University of Évora, Portugal
CIMA (Research Center for Mathematics and Applications), University of Évora, Évora, Portugal
NOVAMath (Center for Mathematics and Applications), FCT NOVA, NOVA University of Lisbon, Portugal
Ci2 (Smart Cities Research Center), Polytechnic Institute of Tomar, Portugal

Committees

Organizing Committee

A. Manuela Gonçalves (Local Chair), University of Minho, Portugal
Adelaide Freitas, University of Aveiro, Portugal
Helena Luzia Grilo, Polytechnic Institute of Tomar, Portugal
Inês Sousa, University of Minho, Portugal
Luís Meira-Machado, University of Minho, Portugal
Luís Miguel Grilo, University of Évora, Portugal
Marta Ferreira, University of Minho, Portugal
Soraia Pereira, University of Minho, Portugal
Susana Faria, University of Minho, Portugal

Executive Committee

Luís M. Grilo (Chairman), University of Évora, Portugal
Anuj Mubayi, Intercollegiate Biomathematics Alliance, IL, USA; South Mountain Community College, AZ, USA; & NumericaIQ, AZ, USA
Luís Meira-Machado, University of Minho, Portugal
Milan Stehlík, Univ. of Appl. Sciences Upper Austria & Universidad de Valparaíso, Chile

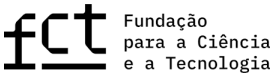
Scientific Committee

Ana Belén Nieto Librero, University of Salamanca, Spain
Ana Isabel Borges, Polytechnic Institute of Porto, Portugal
Ana Rodrigues, University of Évora, Portugal
Anabela Afonso, University of Évora, Portugal
Anuj Mubayi, Intercollegiate Biomathematics Alliance, IL, USA; South Mountain Community College, AZ, USA; & NumericaIQ, AZ, USA
Arminda Manuela Gonçalves, University of Minho, Portugal
Adelaide Figueiredo, University of Porto, Portugal
Adelaide Freitas, University of Aveiro, Portugal
Aldina Correia, Polytechnic Institute of Porto, Portugal
Carla Santos, Polytechnic Institute of Beja, Portugal
Carlos Agra Coelho, NOVA University of Lisbon, Portugal
Carlos Braumann, University of Évora, Portugal
Carlos Ramos, University of Évora, Portugal
Catarina Nunes, University of Aberta, Portugal
Clara Grácio, University of Évora, Portugal
Cristina Dias, Polytechnic Institute of Portalegre, Portugal
Cristina Lopes, Polytechnic Institute of Porto (ISCAP), Portugal
Dharmendra Tripathi, National Institute of Technology, Uttarakhand, India
Dora Gomes, NOVA University of Lisbon, Portugal
Dulce Gomes, University of Évora, Portugal
Dulce Pereira, University of Évora, Portugal
Eliana Costa e Silva, Polytechnic Institute of Porto, Portugal
Fernanda Figueiredo, University of Porto, Portugal
Fernando Carapau, University of Évora, Portugal
Flora Ferreira, University of Porto, Portugal

Frederico Caeiro, NOVA University of Lisbon, Portugal
Inês Sousa, University of Minho, Portugal
Luís Meira-Machado, University of Minho, Portugal
Luís M. Grilo, University of Évora, Portugal
Manuela Oliveira, University of Évora, Portugal
Marco Costa, University of Aveiro, Portugal
Maria do Rosário Ramos, University of Aberta, Portugal
Marília Pires, Universidade de Évora, Portugal
Marta Ferreira, University of Minho, Portugal
Milan Stehlík, Univ. of Appl. Sciences Upper Austria & Universidad de Valparaíso, Chile
Padmanabhan Seshaiyer, George Mason University, USA
Rita Gaio, University of Porto, Portugal
Rui Pereira, University of Minho, Portugal
Russell Alpizar-Jara, University of Évora, Portugal
Sarada Ghosh, Department of Statistics, Kolkata, India
Soraia Pereira, University of Minho, Portugal
Susana Faria, University of Minho, Portugal
Tiago Dias Domingues, University of Lisbon, Portugal

Sponsored by

Universidade do Minho
Escola de Ciências



Technical Specifications

Title
Book of abstracts of the X Workshop on Computational Data Analysis and Numerical Methods

Web-page
`https://w3.math.uminho.pt/WCDANM2026/`

Editor
Universidade do Minho. CMAT - Centro de Matemática
Campus de Gualtar
4710 - 057 Braga, Portugal

Editors
A. Manuela Gonçalves (University of Minho & CMAT, Portugal)
Luís Meira-Machado (University of Minho & CMAT, Portugal)
Luís Miguel Grilo (University of Évora & CIMA & NOVA Math & Ci2, Portugal)
Marta Ferreira (University of Minho & CMAT, Portugal)
Soraia Pereira (University of Minho & CMAT, Portugal)

Authors
Many authors.

Published in a PDF format by:
University of Minho, Portugal

ISBN: 978-989-95122-5-2

Contents

Contributed Talks

Posters

Invited Speakers

Mechanistic Intelligence in Healthcare: Scientific Computing, Evidence Generation, and Agentic AI Systems

Anuj Mubayi[1,2,3],

[1]Intercollegiate Biomathematics Alliance, IL, USA
[2]South Mountain Community College, AZ, USA
[3]NumericaIQ, AZ, USA

E-mail addresses: *anujmubayi@yahoo.com*

Despite major advances in artificial intelligence, many healthcare AI systems continue to struggle with biological realism, uncertainty, interpretability, and real-world decision integration. Healthcare is not merely a prediction problem; it is a complex mechanistic, computational, and evidence-driven system involving dynamical processes, sparse data, operational constraints, and economic decision-making.

In this talk, we will discuss the emerging concept of mechanistic intelligence, the integration of mechanistic modeling, scientific computing, uncertainty-aware inference, healthcare evidence generation, and Agentic AI systems. It will highlight how computational models, simulation, and evidence workflows remain central to complex healthcare decision-making. The talk further explores how Agentic AI systems may serve as orchestration layers for autonomous evidence synthesis, simulation, validation, and healthcare decision support.

The central argument is that the future of healthcare AI may depend not on replacing mechanistic science with black-box models, but on combining scientific computing, evidence generation, and intelligent orchestration into adaptive computational healthcare ecosystems.

Fishing in a random environment with parameter uncertainty

Carlos A. Braumann[1,2] and Nuno M. Brites[3]

[1]Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática, Portugal
[2]Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Portugal
[3]ISEG Research, ISEG Lisbon School of Economics & Management, Universidade de Lisboa, Portugal

E-mail addresses: *braumann@uevora.pt; nbrites@iseg.ulisboa.pt*

Fishing in a randomly varying environment is modeled by a stochastic differential equation. Variable effort optimal harvesting policies can be computed numerically through Monte Carlo simulations and the HJB equation. These policies have severe implementation problems, and so, using real fisheries data, we also study constant effort policies and show that they are almost as good in terms of profit. Since model parameters are estimated with error, we study how this affects efforts and profits.

Keywords

stochastic differential equations, population growth and fishing models, optimal harvesting, parameter uncertainty, effects on efforts and profits

For a harvested population in a randomly varying environment described by a general stochastic differential equation (SDE) model, we have previously studied profit optimisation using variable effort harvesting policies, which involve solving the Hamilton-Jacobi-Bellman (HJB) equation. We have applied the models to real fishery data, namely the Pacific halibut with logistic population growth [2] and the Bangladesh shrimp with Gompertz population growth [3], for both the optimal variable effort policy and the optimal constant effort policy, using a Crank-Nicolson scheme for the numerical solution of the HJB equation and Monte Carlo simulations for the SDE. We have shown that variable effort policies pose significant implementation and social problems, while, on the contrary, constant effort policies do not face such limitations and yield profits that are almost as good. Furthermore, constant effort policies yield, under appropriate conditions [4], a sustainable stochastic equilibrium.

We have also studied other general and specific models, namely contemplating populations with Allee effects, as well as intermediate harvesting policies, e.g. stepwise or penalised policies, which, due to time constraints, will not be addressed here.

For both variable and constant effort harvesting policies, estimation errors in biologic, economic and technologic parameters of the model and the profit function, either inherent to the estimation procedures or due to structural climatic and socioeconomic changes, affect harvesting efforts and future expected profits.

We examine these effects, illustrating them with the Bangladesh shrimp fishery ([1]) and the Pacific halibut fishery (forthcoming paper). With regard to future expected profits, we compare the *Ideal Profit* (obtained if the true parameter value were known) with the *Predicted Profit* (our prediction based on the incorrectly estimated parameter values) and the *Real Profit* (the profit actually obtained by applying harvesting efforts based on the

estimated parameter values to a population evolving under the true parameter values). Sensitivity indices measure the impact of improved parameter estimation on harvesting efforts, future predictions, and real profits, thereby providing guidance on priorities for improving parameter estimation.

Acknowledgements: C.A. Braumann is a member of the Centro de Investigação em Matemática e Aplicações, supported by Fundação para a Ciência e a Tecnologia – FCT (Foundation for Science and Technology), Portugal, Project UID/04674/2025, `https://doi.org/10.54499/UID/04674/2025`. N.M. Brites was partially funded by national funds through FCT in the framework of the project UID/06522/2025.

References

[1] N. M. Brites and C. A. Braumann. Optimal harvesting in randomly varying environments: Sensitivity of profit and effort to population and economic parameters. In: M. Bezzeghoud, F. Carapau, F. Minhós, and A. Vaidya (Eds.), *Advances in Mathematical Modeling in Science, Engineering and Social Sciences*, Springer, 2026. doi: 10.1007/978-3-032-10281-2_27 (in press).

[2] N. M. Brites and C. A. Braumann. Fisheries management in random environments: Comparison of harvesting policies for the logistic model. *Fisheries Research* **195**, 238–246, 2017. doi: 10.1016/j.fishres.2017.07.016.

[3] N. M. Brites and C. A. Braumann. Fisheries management in randomly varying environments: Comparison of constant, variable and penalized efforts policies for the Gompertz model. *Fisheries Research* **216**, 196–203, 2019. doi: 10.1016/j.fishres.2019.03.016.

[4] C. A. Braumann. Harvesting in a random environment: Itô or Stratonovich calculus? *Journal of Theoretical Biology* **244**, 424–432, 2007. doi: 10.1016/j.jtbi.2006.08.029.

Design Science vs. Behavioral Research: Implications for Computational Data Analysis

Jörg Henseler[1]

[1]Faculty of Engineering Technology (ET), University of Twente, Enschede, The Netherlands

E-mail address: *j.henseler@utwente.nl*

Research methodologies in the social and cognitive sciences broadly divide into two paradigms: design science research, which focuses on the creation and evaluation of artifacts, systems, or interventions, and behavioral research, which seeks to observe, measure, and explain naturally occurring or experimentally elicited human behavior. While both paradigms generate rich empirical data, they impose different demands on computational data analysis — demands that are frequently underappreciated when methods are borrowed across paradigms.

This talk examines the differences between design science and behavioral datasets and their downstream consequences for data analysis. Design science research data is often small-scale, iterative, and qualitative or mixed-methods in nature, foregrounding interpretive validity and contextual richness over statistical power. Behavioral research, by contrast, tends to prioritize large samples, controlled conditions, and inferential rigor, making it amenable to classical hypothesis testing and machine learning approaches. These differences have implications for choices in preprocessing, feature engineering, model selection, and the interpretation of computational outputs.

I argue that computational data analysis researchers must make paradigm awareness an explicit methodological consideration, and I propose a set of guiding principles for selecting and adapting analytical tools to match the epistemological commitments of the underlying research tradition. The talk concludes by identifying opportunities for cross-paradigm synthesis, including hybrid methodologies that leverage the generative flexibility of design science research alongside the inferential strength of behavioral research methods, and the role that computational data analysis can play in bridging these traditions productively.

Revolutions in Neural Networks and Data Science: Introducing SPOCU and DExPSO.

Milan Stehlík[1]

[1] Institute of Statistics, Universidad de Valparaíso, Valparaíso, Chile

E-mail addresses: *milan.stehlik@uv.cl*

In this talk, I will introduce our adaptive transfer function SPOCU [1], which we developed with collaborators to address the insufficiency of standard transfer functions in properly processing real data flows. SPOCU is a revolutionary improvement in the speed and innovation of adaptation strategies, filling a gap in existing technology. Activation functions are crucial in deep learning for extracting complex data patterns, and traditional functions like ReLU, Selu, among others, have limitations in adapting to specialized tasks. Standard transfer functions have limitations in complex setups, thus necessitating the development of robust approaches like large-scale self-normalizing neural networks. See, for example, [1], [2], and [3]. To address this, we propose a novel trainable adaptive activation function based on SPOCU construction. Dynamical networks face challenges with big and irregular data. Optimal activation function selection and hyperparameter management are crucial. The SPOCU transfer function offers flexibility and superior performance in machine learning tasks (see [4], [5] or [6]). Experimental results show improvements in cancer diagnosis and pollutant adsorption dynamics. Developing adaptive algorithms for hyperparameter selection is essential, and our milestone result of DExPSO [7] is giving a chance to avoid recurrent premature failures of standard neural networks. In [8], we showed how Cobetia bacteria adaptation to large temperature ranges can be modeled by an SPOCU-based neural network, which has a principal application also to food systems and image recognition. During the talk, we will discuss how the SPOCU prototype adaptive function has been created and explore new ideas for optimizing hyperparameters in adaptive transfer functions like SPOCU for real-world data flows, improving methodologies in different application areas.

Keywords: SPOCU Prototype, DExPSO, optimization , hyperparameters, transfer functions, dataflows, complexity, networks, ecology.

Acknowledgements

This work was partially supported by Proyecto Exploracion ANID GRT-AC: Nr. 13220184.

References

[1] J. Kiselak, Y. Lu, P. Svihra, M. Szepe, and M. Stehlik. "SPOCU": scaled polynomial constant unit activation function. *Neural Comput and Applic* **33**, 3385-3401, 2020.

[2] B. Subramanian, R. Jeyaraj, R. Akhrorjon, and K. Ugli. APALU: A Trainable, Adaptive Activation Function for Deep Learning Networks. *arXiv preprint arXiv:2402.08244*, 2024.

[3] Z. Chen, W. Zhao, L. Deng, Y. Ding, Q. Wen, G. Li, and Y. Xie. Large-scale self-normalizing neural networks. *Journal of Automation and Intelligence* **3(2)**, 101-110, 2024.

[4] A. Bamimore, N. B. Sobowale, A. S. Osunleke, et al. Offset-free neural network-based nonlinear model predictive controller design using parameter adaptation. *Neural Comput and Applic* **33**, 10235–10257, 2021.

[5] V. Vives-Boix and D. Ruiz-Fernández. Fundamentals of artificial metaplasticity in radial basis function networks for breast cancer classification. *Neural Comput and Applic* **33**, 12869–12880, 2021.

[6] Y. Mesellem, A. A. E. Hadj, M. Laidi, et al. Computational intelligence techniques for modeling of dynamic adsorption of organic pollutants on activated carbon. *Neural Comput and Applic* **33**, 12493–12512, 2021.

[7] M. Stehlík, Ch. Ping-Yang, W. K. Wong, and J. Kiseľák. A Double Exponential Particle Swarm Optimization with non-uniform variates as stochastic tuning and guaranteed convergence to a global optimum with sample applications to finding optimal exact designs in biostatistics. *Applied Soft Computing*, `https://doi.org/10.1016/j.asoc.2024.111913` 2024.

[8] A. Dinamarca, M. Stehlik, et al. Comprehensive and deep learning classification for analyses of biological complexity of growth and biofilms of Cobetia marina under different temperature growths. *PLOS ONE* **20(12): e0336575**, `https://doi.org/10.1371/journal.pone.0336575` 2025.

Short Courses

Numerical Simulation and Inference for PDE-Based Biological Models

Anuj Mubayi[1,2,3],

[1]Intercollegiate Biomathematics Alliance, IL, USA
[2]South Mountain Community College, AZ, USA
[3]NumericaIQ, AZ, USA

E-mail addresses: *anujmubayi@yahoo.com*

Mechanistic partial differential equation (PDE) models are widely used to study biological systems involving transport, diffusion, interaction, and spatial heterogeneity. This course introduces computational frameworks for simulation and inference in PDE-based biological models, motivated in part by a recent SIAM News article on SARS-CoV-2 transport and establishment of infection in tissue.
Topics include transport–diffusion–reaction PDEs, numerical methods for biological simulations, PDE-constrained inverse problems, inference, uncertainty quantification. The course emphasizes how mechanistic PDE models can be integrated with probabilistic inference frameworks to estimate uncertain biological parameters from sparse or indirect data.

Structural Equation Modeling with Latent Variables and Formative Constructs

Jörg Henseler[1]

[1]Faculty of Engineering Technology (ET), University of Twente, Enschede, The Netherlands

E-mail address: *j.henseler@utwente.nl*

Structural equation modeling (SEM, c.f. Bollen 1989) is a powerful multivariate statistical framework that enables researchers to test theories by simultaneously examining complex relationships among observed and unobserved variables. This four-hour workshop provides a comprehensive, hands-on introduction to SEM with a particular focus on two advanced and often misunderstood components: latent variables and formative constructs (Henseler 2021).

The workshop begins by establishing the conceptual and statistical foundations of SEM, contrasting it with traditional regression-based approaches and introducing key model components including measurement models, structural models, and fit evaluation. All methods are implemented in R using the lavaan package (Rosseel 2012), an actively maintained package that offers flexible and transparent model specification through an intuitive model syntax. Participants will work directly with lavaan code throughout the session, gaining practical experience they can immediately apply to their own research.

A central theme of the workshop is the distinction between reflective measurement models of latent variables and composite models of formative constructs — a critical methodological choice that profoundly shapes model specification, validation, and interpretation. While reflective indicators are caused by the underlying latent construct, formative constructs are defined by or made up of their indicators Participants will learn how to specify both types of models in lavaan, interpret their output, and understand the constraints each imposes.

The workshop closes with an open session for questions and discussion, offering participants the opportunity to reflect on how the methods covered apply to their own research contexts, and to seek guidance on specific modeling challenges they may face.

References

[1] K. A. Bollen. *Structural equations with latent variables*. Wiley, 1989.

[2] J. Henseler. *Composite-based structural equation modeling: Analyzing latent and emergent variables*. Guilford Press, 2021.

[3] Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* **48(2): 1–36**, `https://doi.org/10.18637/jss.v048.i02` 2012.

PSE Session

Turning Data into Decisions: Real Cases from PSE

Nuno Gomes[1]

[1]Head of Data Science at PSE

E-mail address: *nuno.gomes@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.

Session in Memory of Jerzy Filus

On contributions of Jerzy Filus to dependence modeling with applications to medicine

Milan Stehlík[1,2]

[1]Institute of Statistics, Universidad de Valparaíso, Valparaíso, Chile
[2]University of Applied Sciences Upper Austria, Austria

E-mail addresses: *milan.stehlik@uv.cl*

Keywords

dependence models, pseudo-exponential model, joiner

Jerzy Filus and his wife Lidia have been researching dependency modeling for decades. The pseudo-exponential model (specifically in the context of bivariate distributions where one marginal is exponential and the conditional distributions of the second variable are also of exponential form) was introduced by [3]. They presented these models in a series of papers starting in the early 2000s. Many important results by Filus-Filus have been developed in the parameter dependency paradigm. See, for example, [11], [12], and [13]. Also, refer to [4] for further discussion and references. In [1], we received a representation of pseudo-dependency models, including the particular copula representations. Several principal applications have shown the feasibility of parameter dependence modeling, including those in ecology [5], finance [7], biomedical modeling [6], [1], [8], [9], [10]. In this presentation, we will discuss Filus and Filus's contributions to Dependence Modeling and the significance of 'Filus-Filus Dependency Models'.

References

[1] M. Stehl´ do, J. Filus, L. Filus, P. Hermann, P. Jordanova, S. Stehlikova, I. Mozos, B. C. Arnold, Y. Lu, and J. Lopez-Fidalgo. Statistical modelling of asymmetric dependencies in medicine. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2026.2629843, 2026.

[2] J. K. Filus, L. Z. Filus, and B. C. Arnold. Families of Multivariate Distributions Involving "Triangular" Transformations. *Communications in Statistics - Theory and Methods* **39(1)**, 107–116, 2009. `https://doi.org/10.1080/03610920802687793`

[3] J. K. Filus and L. Z. Filus. Pseudoexponential and Semi-Pseudoexponential Models in Medical Prognostics. *AIP Conf. Proc.* **1281**, 1877–1880, 2010.

[4] J. Filus, L. Filus, Y. Lu, P. Jordanova, B. C. Arnold, L. N. Soza, S. Stehlikova, and M. Stehlík. On parameter dependence and related topics: The impact of Jerzy Filus from genesis to recent developments (with discussion). In: I. Vonta, M. Ram, and B. Roca (Eds.), *Reliability Engineering: Theory and Applications*, 1st Edition, CRC Press, pp. 143-169, 2018.

[5] M. Stehlík, J. Filus, S. Stehlikova, and L. Filus. On statistical aspects of advanced stochastic modeling of lava flows of volcano Lonquimay. *ICDQM 2020*, ISBN 978-86-86355-42-3, 2020.

[6] J. Filus, L. Filus, and M. Stehlík. Pseudoexponential modelling of cancer diagnostic testing. *Biometrie Und Medizinische Informatik Greifswalder Seminarberichte* (ISBN 978-3-8322-7481-8, SHAKER Publ.), Heft 15, 41-54, 2009.

[7] Z. R. Kuenstler, B. C. Merley, M. Stehlik, J. Filus, L. Filus, C. Navarro, J. P. Maidana, and F. Fuders. Consequences of COVID-19 in EUR/US exchange rates and economy. *Springer*, 2023.

[8] M. Stehlík, T. Mrkvička, J. Filus, and L. Filus. Recent development on testing in cancer risk: a fractal and stochastic geometry. *Journal of Reliability and Statistical Studies* **5**, 83-95, 2012.

[9] J. Filus, L. Filus, and M. Stehlík. Pseudoexponential models in medical trials: design, estimation and testing. In: A. Karagrigoriou (Ed.), *Proceedings of the 4th Intern. Conf. on Risk Analysis (ICRA4)*, Cyprus, pp. 75-83, 2011.

[10] J. Filus, L. Filus, and M. Stehl´ i k. Pseudoexponential modelling of cancer diagnostic: testing, estimation and design. In: C. P. Kitsos and C. Caroni (Eds.), *Proceedings of ICCRA 3*, Porto Heli, Greece, pp. 1-16, 2009.

[11] J. K. Filus and L. Z. Filus. A class of generalized multivariate normal densities. *Pakistan Journal of Statistics* **16(1)**, 11–32, 2000.

[12] J. K. Filus and L. Z. Filus. On some bivariate pseudonormal densities. *Pakistan Journal of Statistics* **17(1)**, 1–9, 2001.

[13] J. K. Filus and L. Z. Filus. On some new classes of multivariate probability distributions. *Pakistan Journal of Statistics* **22**, 21–42, 2006.

A Hierarchical Structural Equation Model for Student Burnout

Luís M. Grilo[1,2,3,4]

[1]Department of Mathematics, University of Évora, Portugal
[2]CIMA (Research Center for Mathematics and Applications), University of Évora, Évora, Portugal
[3]NOVA Math (Center for Mathematics and Applications), FCT NOVA, NOVA University of Lisbon, Portugal
[4]Ci2 (Smart Cities Research Center), Polytechnic Institute of Tomar, Portugal

E-mail address: *luis.grilo@uevora.pt*

In Structural Equation Modeling (SEM), variables measured on an ordinal scale, such as Likert-type items, are frequently used. Therefore, when estimating SEM parameters involving latent constructs—such as "perceived stress" and "student burnout" operationalized through ordinal manifest variables (or indicators)—it is recommended to adopt estimation methods specifically suited to these data. In this study, "student burnout" was modeled as a second-order construct, with "emotional exhaustion," "disbelief," and "personal efficacy" specified as first-order constructs. The proposed theoretical model was estimated using survey data collected from college students and two estimators grounded in different statistical paradigms: PLSc and DWLS. As expected, the findings revealed a strong positive direct effect of "perceived stress" on "student burnout," with "emotional exhaustion" emerging as the central component.

Keywords

DWLS estimator, higher-order construct, mental health, PLSc estimator.

As commonly observed in the social, behavioral, and health sciences, an online survey was conducted in a Portuguese institution to assess "perceived stress" and "student burnout". Data were collected using the Perceived Stress Scale (PSS) and the Maslach Burnout Inventory–Student Survey (MBI-SS) [4, 5]. In this study, the latent exogenous construct "perceived stress" was considered a predictor of "student burnout" [2], which was modeled as a construct of higher-order, allowing for a more parsimonious representation of the model [6]. Because the higher-order construct and all lower-order constructs were reflectively specified, the proposed model was estimated using consistent Partial Least Squares (PLSc), which is recommended for reflective common factor models and reflective–reflective hierarchical component models estimated using the disjoint two-stage approach [1, 3, 6, 7]. The PLSc estimator has demonstrated robust performance in the presence of ordinal data (e.g., Likert scales) and non-normal data, particularly in complex models (involving numerous indicators, constructs, and interrelationships), as well as in studies with limited sample sizes. The Diagonally Weighted Least Squares (DWLS) method was also used. DWLS is particularly appropriate for ordinal indicators and in situations where the assumption of multivariate normality is violated. Rather than fitting Pearson covariance matrices directly, DWLS relies on polychoric correlations. These estimators represent the two main approaches to structural equation modeling—variance-based SEM (VB-SEM) and covariance-based SEM (CB-SEM). Although these approaches

differ in their modeling objectives, their application in this study is discussed as statistically complementary.

Acknowledgements: This work was partially supported 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] C. M. Ringle, S. Wende, and J.-M. Becker. SmartPLS 4. Bönningstedt: SmartPLS, 2024. Retrieved from `https://www.smartpls.com`

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

[3] J. F. Hair, M. Sarstedt, C. M. Ringle, and S. P. Gudergan. *Advanced Issues in Partial Least Squares Structural Equation Modeling (PLS-SEM).* 2nd ed., Sage, 2024.

[4] J. Marôco, M. Tecedeiro, P. Martins, and A. Meireles. O Burnout como factor hierárquico de 2ª ordem da Escala de Burnout de Maslach. *Análise Psicológica* **4 (XXVI)**, 639–649, 2008.

[5] J. Marôco and M. Tecedeiro. Inventário de Burnout de Maslach para Estudantes Portugueses. *Psicologia, Saúde & Doenças* **10 (2)**, 227–235, 2009.

[6] M. Sarstedt, J. F. Hair, J.-H. Cheah, J.-M. Becker, and C. M. Ringle. How to Specify, Estimate, and Validate Higher-Order Constructs in PLS-SEM. *Australasian Marketing Journal* **27(3)**, 197–211, 2019.

[7] T. K. Dijkstra and J. Henseler. Consistent Partial Least Squares Path Modeling. *MIS Quarterly* **39(2)**, 297–316, 2015.

On a modelling approach to Principle Component Analysis

Martin Schlather[1]

[1]Institut für Mathematik, University of Mannheim, Mannheim, Germany

E-mail addresse: *martin.schlather@uni-mannheim.de*

Principal Component Analysis simplifies high dimensional data by means of an orthogonal projection onto a lower dimensional hyperplane. An orthogonal projection can be defined through the (Euclidean) scalar product or, equivalently, as the minimizer of the sum of squared distances between the data and their approximations. In order to transfer the PCA to extreme values, we just need to find the natural scalar product and/or the natural distance measure for extreme values. The talk gives an insight into a general approach, how scalar products and distances might be defined, when fundamental requirements for a standard scalar product are violated. We illustrate this approach by many examples and new perspectives on some well-known concepts. Although the approach is heavily based on algebraic objects such as monoids and quasi-semi-rings, the talk will only deal with their properties, such as the validity of the associative law.

Tribute to Jerzy Filus and his work

Lidia Z. Filus[1]

[1]Mathematics Department, Northeastern Illinois University, Chicago, USA

E-mail addresse: *l-filus@neiu.edu*

The presentation pays homage to the unique talents and contributions of Jerzy Filus. It summarizes Jerzy's research, specially in the area of dependence modeling, its importance and originality.

Contributed Talks

Dimensionality reduction for compositional data vectors: an application to the codon space

Adelaide Freitas[1,2]

[1]Department of Mathematics, University of Aveiro, Portugal
[2]Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

E-mail address: *adelaide@ua.pt*

In several real studies, observations are compositional data or the relative structure of the parts is more informative than their absolute values. Recently, a strategy for applying Principal Component Analysis on data descreved by a composition of compositional data was proposed. To illustrate its interpretation, the codon space related to the genes of 147 bacterial species are analyzed and graphical methods are used to reveal patterns in the first principal component, itself a composition.

Keywords

principal component analysis, compositional data, codon, bacteria.

Projecting high-dimensional datasets into lower-dimensional spaces has proven effective for extracting relevant information across many fields. In several applications, observations are compositional data or the relative structure of the parts is more informative than their absolute values.

Compositional data consist of multivariate observations that describe the relative contribution of parts within a whole. Formally, a D-dimensional vector is considered compositional when all its components are positive and only the relative information among the D parts (such as proportions or percentages) is meaningful. In this setting, the relevant information is encoded in ratios between components rather than in their absolute differences.

This concept can be extended to observations defined by compositions of p-parts of D-parts compositional data (i.e., p vectors each one with D compositional components). This type of multivariate observation is referred to as a p-dimensional compositional data vector. Recently, [1] proposed a new approach to Principal Component Analysis (PCA) for modelling this type of data.

For living organisms, the codon space is defined as a 12-dimensional vector comprising the frequencies of the nucleotides A (adenine), C (cytosine), G (guanine), and T (thymine) at each of the three codon positions (first, second, and third) in the coding regions of their genome. The first four coordinates of the vector correspond to the frequencies of the four nucleotides at the first codon position, the next four coordinates to the second position, and the last four coordinates to the third position. The sum of the frequencies of the four nucleotides at any codon position equals the total number of codons in the genome; consequently, the sum of the 12 components of the vector equals three times the total number of codons. Thus, the codon space can also be viewed as a composition of 4-part compositional data across the three codon positions, i.e., a 3-dimensional compositional data vector with $D = 4$.

Investigations have shown that dimensionality reduction techniques such as PCA, when applied to datasets containing the codon space of living organisms, provide useful biological insights for complex biological systems [2, 3]. However, they have not been analyzed in terms of compositional vectors.

Coding regions (genes) of complete genome sequences from 147 bacterial species, of which 82 are Gram-positive (31 Actinobacteria, 9 Bacillales, 18 Lactobacillales, 14 Clostridia, and 10 Mollicutes) and 65 are Gram-negative, were downloaded from NCBI GenBank (https://www.ncbi.nlm.nih.gov/nuccore/). The nucleotide frequencies at the three codon positions were calculated using the package 'PERL 1.0' developed by [2]. Finally, a 147×12 data matrix representing the codon space of these bacteria was obtained.

Based on this dataset, the PCA approach proposed in [1] is applied, and bar graphs are used to explore patterns in the first principal component, itself a composition. Using this case study, strengths and limitations of this approach are discussed, and directions for future work are suggested.

Acknowledgements: This work is supported by CIDMA (https://ror.org/05pm2 mw36) under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snf qn58), 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] H. Wang, L. Shangguan, R. Guan, and L. Billard. Principal component analysis for compositional data vectors. *Computational Statistics* **30:4**, 1079-1096, 2015. doi: 10.1007/s00180-015-0570-1.

[2] Z. H. Qi, and R. Y. Wei. A combination dimensionality reduction approach to codon position patterns of eubacteria based on their complete genomes. *Journal of Theoretical Biology* **272:1**, 26-34, 2011.

[3] F. Takeuchi, Y. Futamura, H. Yoshikura, and K. Yamamoto. Statistics of trinucleotides in coding sequences and evolution. *Journal of Theoretical Biology* **222:2**, 139-149, 2003.

Multivariate Characterization of Statistical Literacy and Attitudinal Profiles in University Students

Ana B. Nieto-Librero[1,3,4], Nerea González-García[1,3,4], Antonio Blázquez-Zaballos[1,4] and Ana B. Sánchez-García[2,3,4]

[1]Departamento de Estadística, Universidad de Salamanca, España
[2]Departamento de Didáctica, Organización y Métodos de Investigación, Universidad de Salamanca, España
[3]Centro de Investigación en Derechos Humanos y Políticas Públicas (Diversitas), Universidad de Salamanca, España
[4] Grupo de Investigación en Perspectiva Multivariante en Investigación Educativa y Social (GIR PMIES), Universidad de Salamanca, España

E-mail addresses: *ananieto@usal.es; nerea_gonzalez_garcia@usal.es; abz@usal.es; asg@usal.es*

Multivariate analysis is applied to assess statistical attitudes and literacy in university students. The approach identifies latent structures and student profiles based on multidimensional data. Results show significant differences across disciplines, with higher anxiety in Social Sciences, lower perceived utility in Health Sciences, and distinct patterns in software acceptance.

Keywords

Multivariate analysis, Statistical Literacy in Higher Education, Attitudes toward Statistics, Technology Acceptance Model

The growing importance of data-driven decision-making has positioned statistical literacy as a fundamental competence in higher education, enabling individuals to interpret, critically evaluate, and communicate information based on data. As highlighted in the literature, statistical literacy encompasses both cognitive and communicative dimensions, integrating reasoning, interpretation, and critical thinking in real-world contexts [1]. However, persistent difficulties in learning statistics and negative attitudes toward the subject continue to be widely documented, particularly in non-technical disciplines, where anxiety and low perceived utility can hinder the development of statistical competence [2, 3].

In this context, this study analyses statistical attitudes, literacy, and technology acceptance in a sample of 375 university students from the University of Salamanca (2024–2025, 2025-2026), considering differences across areas of knowledge. Data were collected using validated instruments measuring attitudinal dimensions (interest, perceived utility, and anxiety), statistical literacy components (reasoning, interpretation, and critical understanding), and technology acceptance variables based on the Technology Acceptance Model (TAM) [4].

A multistep analytical framework was adopted. First, exploratory multivariate techniques were applied to identify global patterns and relationships among variables and disciplines. Correspondence analysis provided a low-dimensional representation of associations between academic areas and response profiles, revealing differentiated attitudinal and literacy patterns. Discriminant analysis further assessed the ability of these variables to classify students according to their disciplinary background.

Second, a confirmatory factor analysis (CFA) was conducted to validate the latent structure of the constructs under study, ensuring consistency across attitudinal, cognitive, and technological dimensions. Finally, a structural equation model (SEM) was specified to examine the relationships between these latent variables, allowing the estimation of direct and indirect effects among attitudes, statistical literacy, and technology acceptance.

The results reveal significant heterogeneity across disciplines. Students in Social Sciences show higher levels of statistical anxiety and lower confidence, whereas those in Health Sciences report lower perceived utility and weaker acceptance of statistical software. In contrast, students in Science and Engineering display more favourable attitudinal patterns and higher levels of perceived competence. The structural model highlights the central role of attitudes—particularly perceived utility and anxiety—in shaping both statistical literacy and the acceptance of analytical tools.

These findings illustrate the relevance of integrating exploratory and structural multivariate approaches to capture the complexity of educational data. The proposed framework provides a comprehensive perspective for analysing statistical competence and supports the development of discipline-sensitive strategies for improving statistical education in higher education.

Acknowledgements: This work was supported by the Teaching Innovation Project "Uso de metodologías activas para la alfabetización estadística: desarrollo de la herramienta LearnSTAT" (PID2024/212), funded by the University of Salamanca. The authors would like to thank all participating students for their collaboration and engagement, as well as the teaching staff involved in the development and implementation of the project.

References

[1] I. Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review* **70(1)**, 1–25, 2002.

[2] C. Batanero. Statistical sense in the information society. In *Proceedings of CISETC*, 28–37, 2019.

[3] J. Mondéjar-Jiménez, M. Vargas-Vargas, and J. Bayot. Medición de la actitud hacia la estadística en universitarios: Validación de una escala. *Revista de Educación* **347**, 407–431, 2008.

[4] U. Šebjan and P. Tominc. Impact of support of teacher and compatibility with needs on usefulness of SPSS by students. *Computers in Human Behavior* **53**, 354–365, 2015.

Comparison of group variable selection methods in mixture linear regression models via simulation study

Ana Moreira[1] and Susana Faria[1]

[1]Centre of Mathematics (CMAT), Department of Mathematics, University of Minho, Portugal

E-mail addresses: *id10866@uminho.pt; sfaria@math.uminho.pt*

Variable selection is a crucial step in model building, determining which explanatory variables are included to explain or predict the dependent variable. When categorical variables are involved, preserving the group structure is important. This study addresses variable selection in mixture linear regression models using penalized maximum likelihood estimation with Group LASSO, Adaptive Group LASSO, and Group SCAD penalties.

Keywords

group variable selection, mixtures of linear regression models, penalized maximum likelihood estimation, simulation study.

Finite Mixture Regression (FMR) models provide a flexible framework for analysing heterogeneous populations in which the relationship between the response and explanatory variables may differ across latent subpopulations. In practical applications, a large number of explanatory variables are often considered, making variable selection an important issue in FMR models.

When predictors are naturally grouped, such as dummy variables associated with categorical factors, group variable selection methods are especially useful because they allow the simultaneous selection or removal of related variables while preserving the group structure and interpretability of the model. In this context, [4] introduced several penalization methods for grouped variables, including the Group Least Absolute Shrinkage and Selection Operator (GLASSO), Group Least Angle Regression (GLARS), and the Group Garrote. To overcome the estimation bias and selection inconsistency associated with GLASSO, [3] proposed the Adaptive Group LASSO (AGLASSO), which incorporates adaptive weights to improve variable selection performance. In addition, [2] developed the Group Smoothly Clipped Absolute Deviation (GSCAD) penalty, which reduces shrinkage bias while performing group selection.

Furthermore, [1] studied variable selection in mixtures of linear regression models using penalized likelihood approaches, highlighting the relevance of penalization methods in the mixture modelling framework. In this study, we investigate group variable selection in finite mixtures of linear regression models using penalized maximum likelihood methods in both low- and high-dimensional settings. Specifically, we compare the performance of GLASSO, AGLASSO, and GSCAD in identifying the most relevant groups of explanatory variables.

To compare these methods, we conduct an extensive simulation study across a range of scenarios. In particular, we consider settings in which the number of groups increases with the sample size. We also examine the impact of varying the number of mixture components, the distribution of the explanatory variables, different levels of correlation

among the variables, the regression coefficients, and the dispersion level of the data. To evaluate the performance of the methods, we consider the following criteria: (i) the median number of groups correctly estimated with zero coefficients, corresponding to true zero coefficients correctly identified as zero; and (ii) the median number of groups incorrectly estimated with zero coefficients, corresponding to true nonzero coefficients incorrectly set to zero. Finally, we assess 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 version 4.5.3 of the R programming language. Several R packages were used in the analysis, including `flexmix`, `glmnet`, and `grpreg`, among others.

We conclude that AGLASSO generally yields better results in low-dimensional settings. However, in high-dimensional settings, GSCAD outperforms both GLASSO and AGLASSO in identifying the truly relevant groups.

Acknowledgements: This work has received funding from FCT - Fundação para a Ciência e a Tecnologia, I.P. under the project reference 2022.12256.BD and https://doi.org/10.54499/2022.12256.BDhttps://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/00013/2025https://doi.org/10.54499/UID/00013/2025`. This research was supported by FCT - Fundação para a Ciência e a Tecnologia, I.P. by project reference 2025.09550.CPCA.A0.

References

[1] A. Khalili, and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102(479)**, 1025–1038, 2007.

[2] L. Wang, G. Chen, and H. Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23(12)**, 1486–1494, 2007.

[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.

Decision Models Based on Extreme Value Theory for Ranking Extreme Risks

Beichen Liu[1,2], Marta Ferreira[1,2] and Irene Brito[1,2]

[1]Centro de Matemática, Universidade do Minho
[2]Departamento de Matemática, Universidade do Minho

E-mail addresses: *benjaminlm2000s@outlook.com; msferreira@math.uminho.pt; ireneb@math.uminho.pt*

An Extreme Value Theory-based framework is proposed for converting extreme observations into decision-oriented risk rankings. GPD model is combined with tail-sensitive measures and rank aggregation, and applied to fire insurance losses. The results show where sample-based summaries miss tail-driven changes in risk ordering.

Keywords

Extreme Value Theory, tail risk, GPD, risk ranking.

Extreme events are difficult to compare because their practical relevance is often determined by the tail rather than by the body of the distribution. In decision problems involving insurance, climate, finance, or infrastructure risk, the goal is not only to estimate the magnitude of rare events, but also to transform several tail-sensitive indicators into an interpretable ordering of alternatives. This work develops a computational decision framework that combines Extreme Value Theory (EVT) with rank-based aggregation for extreme risk assessment. Risk is treated as an orderable quantity, following the measurement perspective of perceived risk [1], while the tail modelling component follows the standard EVT framework for exceedances [2]. See also [3].

The proposed framework has three steps. First, extreme observations are extracted by a peaks-over-threshold procedure, leading to the Generalized Pareto Distribution (GPD) modeling. Second, fitted tail models are used to compute risk measures such as tail mean, Value-at-Risk and Expected Shortfall when applicable. Third, the resulting indicators are converted into ranks and aggregated into a single decision-oriented ordering. The aggregation step is deliberately simple and transparent: each risk measure provides an ordinal comparison, and the final score summarizes the multi-dimensional tail profile of each year or month.

One empirical application is considered, focusing on fire insurance claim costs using the `frecomfire` dataset in the `CASdatasets` R package [4]. Claim amounts are grouped by year and by month, and sample-based risk rankings are compared with GPD-based rankings. The results show broad agreement for moderate-risk periods, but also reveal targeted re-orderings when the far tail is material. In particular, the GPD framework provides additional interpretability through the estimated shape and scale parameters and improves the identification of periods where extreme claims dominate aggregate risk.

Overall, the results indicate that EVT-based decision models can change risk orderings precisely where sample summaries are least reliable: in the presence of rare, high-impact observations. By combining tail-sensitive modelling with rank aggregation, the proposed approach offers a transparent and portable pipeline for turning data on extremes into decisions for monitoring, capital allocation, and operational planning.

References

[1] H. W. Brachinger and M. Weber. Risk as a primitive: A survey of measures of perceived risk. *Operations Research Spectrum*, **19**(3):127–143, 1997.

[2] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

[3] I. Brito, and M. Ferreira. Extreme risk decision modeling: a new approach with an application to air pollution based on PM2.5. *Computational Statistics & Data Analysis*, 108413, 2026. `https://doi.org/10.1016/j.csda.2026.108413`.

[4] C. Dutang, A. Charpentier, E. Gallic, and J. Siharath. *CASdatasets: Insurance Datasets for Actuarial Science*. R package version 1.2.0, 2024. Dataset `frecomfire`.

Reliability and Construct Validity of a Discomfort Scale for Immobilized Trauma Victims

Carla Henriques[1,2], Ana Matos[1] and Mauro Mota[3,4]

[1] ESTGV, Instituto Politécnico de Viseu, Portugal
[2] Centre for Mathematics of the University of Coimbra (CMUC), Portugal
[3] ESSV, Instituto Politécnico de Viseu, Portugal
[4] UICISA: E/ESEnfC - Cluster at the Health School, Polyt. Inst. of Viseu, Portugal,

E-mail addresses: *carlahenriq@estgv.ipv.pt; amatos@estgv.ipv.pt; maurolopesmota@gmail.com*

In pre-hospital trauma care, immobilization is recommended to prevent harmful displacement. However, it may cause adverse effects such as discomfort, pain, and pressure injuries. Assessing discomfort is therefore as important as pain assessment. This study evaluates the reliability, temporal stability, and construct validity of the Discomfort Assessment Scale for Immobilized Trauma Victims (DASITV) using test–retest and multivariate statistical analyses.

Keywords

reliability, construct validity, discomfort scale.

The aim of this study was to evaluate the psychometric properties of the Discomfort Assessment Scale for Immobilized Trauma Victims (DASITV), developed to assess discomfort in trauma victims subjected to immobilization [3]. The study included 27 healthy volunteers assessed in two separate sessions (test and retest) conducted approximately two weeks apart. Discomfort was recorded at 30 anatomical locations over 12 time points, from 5 to 60 minutes of immobilization. To obtain a clinically meaningful global discomfort measure (DASITVg), the mean of the three highest discomfort scores reported by each participant at each time point was calculated, focusing on the anatomical locations where discomfort was most pronounced.

Several complementary statistical approaches were applied. Friedman's ANOVA, with Bonferroni correction, was initially used to compare discomfort levels across time. Since no significant differences were observed during the first 20 minutes, only time points from 20 minutes onward were included in subsequent analyses. Test–retest reliability was assessed using paired-samples t-tests, a two-way repeated-measures ANOVA with time and trial as within-subject factors, and the Intraclass Correlation Coefficient ICC(3,k), calculated through a two-way mixed-effects model for consistency [2]. Additionally, multilevel models with three levels were estimated using restricted maximum likelihood (REML), considering time points nested within trials, and trials nested within participants.

Results showed a progressive increase in discomfort over time in both trials, although retest values were systematically lower, suggesting a familiarization effect. Nevertheless, the temporal pattern of discomfort progression remained highly consistent between test and retest. In fact, paired-samples t-tests revealed no significant differences between trials for discomfort changes during the intervals 20–40, 40–60, and 20–60 minutes. Similarly, the repeated-measures ANOVA showed no significant interaction between time and trial ($p=0.639$), supporting the temporal stability of the scale.

As for the reliability results, the ICC(3,k) was calculated with the mean DASITVg score for each participant across all time points from T4 (20 min) onward. The result, 0.739 (95% CI: 0.428–0.881), is suggestive of a fair-to-good reliability [1]. Multilevel modelling also produced high ICC values (0.78), demonstrating that most variability in discomfort scores was attributable to differences between participants and trials rather than to within-trial variability or residual error.

To identify latent patterns of discomfort across anatomical locations, an exploratory factor analysis (EFA) was performed using the original discomfort scores recorded at each anatomical location from T4 (20 minutes) to T12 (60 minutes) in both test and retest trials. Internal consistency of the resulting factors was assessed using Cronbach's alpha. The analysis revealed stable latent dimensions supporting the aggregation of anatomical locations into 11 clinically meaningful clusters, two of which consisted of a single anatomical location. Clusters 1–9 showed excellent internal consistency (Cronbach's $\alpha > 0.90$), supporting the reliability of the aggregated locations.

Regarding the anatomical clusters, the highest discomfort levels were observed in the occipital, calcaneus, and sacrum regions, which are clinically recognized as high-pressure areas during prolonged immobilization. These clusters also presented ICC values ranging from 0.68 to 0.80, further supporting the reliability of the DASITV for longitudinal discomfort monitoring.

Overall, the results support the DASITV as a reliable and clinically meaningful instrument for monitoring discomfort progression during prolonged immobilization in pre-hospital care 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.* Berlin: Springer, 2013.

[2] T. K. Koo, and M. Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine* **15(2): 155-163**,

[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). *International Emergency Nursing* **76:101501**, 2024.

Assessing Copula Models for Dependently Doubly Truncated Data with Interval Sampling

Carla Moreira[1], Jacobo de Uña-Álvarez[2]
[1] Faculty of Medicine, University of Porto, Porto, Portugal
[2] Department of Statistics and OR, University of Vigo, Spain

E-mail addresses: *carlamgmm@gmail.com; jacobo@uvigo.gal*

Statistical inference under dependent double truncation with interval sampling poses important methodological challenges, since the truncation mechanism induces selection on the observable sample and complicates the modeling of the underlying dependence structure. Recent advances by Moreira et al. [1] introduced a nonparametric framework for estimating marginal distributions and copula dependence models in this setting, providing a flexible basis for semiparametric dependence modeling. Motivated by the copula-based goodness-of-fit methodology of Emura and Pan [2] for dependently left-truncated data, we investigate procedures for assessing the adequacy of candidate copula models under dependent double truncation. The proposed methodology compares the empirical observable joint distribution with the fitted observable distribution implied by the estimated copula model through Kolmogorov–Smirnov and Cramér–von Mises type discrepancy measures. Inference is performed using bootstrap resampling directly from the fitted observable truncated distribution, yielding an efficient model-based assessment procedure that naturally accounts for the truncation mechanism. A simulation study is conducted to examine finite-sample performance under different truncation levels, dependence strengths, and sample sizes, and the practical usefulness of the proposed approach is illustrated through an application to real doubly truncated data.

References

[1] C. Moreira, R. Braekers, and J. de Uña-Alvarez. Nonparametric estimation of a distribution function from doubly truncated data under dependence. *Computational Statistics*, 2021.

[2] T. Emura and C.-H. Pan. Parametric likelihood inference and goodness-of-fit for dependently left-truncated data, a copula-based approach. *Statistical Papers* **61**, 479–501, 2020.

From Admissions to Bed-Days after Hip Fracture

Cecilia Castro[1]

[1]Centre of Mathematics, Universidade do Minho, Braga, Portugal

E-mail addresses: *cecilia@math.uminho.pt*

Bed-day burden depends on admission inflow, length of stay and survival status. Using 843,861 hip-fracture admissions in Brazil's public hospital system, we combine fixed-reference standardization and exact decompositions to separate admission, LOS and survival-composition effects.

Keywords

bed-day burden, decomposition analysis, hip fracture, in-hospital mortality, length of stay.

Hip fracture is a useful sentinel condition for studying hospital burden with routine administrative data. It is frequent in older and frail patients, is usually urgent, carries in-hospital mortality risk and follows an acute-care pathway in which discharge timing is operationally important. Shorter length of stay (LOS) may reflect improved flow, but under system pressure it may also indicate care compression. This distinction became especially relevant during the COVID-19 period, when hip-fracture pathways were maintained under hospital stress and mortality concerns were widely reported [3].

For service planning, the key question is not only how many admissions occur, but how many bed-days they generate. We propose a framework for acute hospital conditions recorded in administrative data, linking clinical safety and operational demand through admission timing, discharge status, LOS and basic case-mix adjusters. As a as a nationwide application, we analysed hip-fracture hospitalizations recorded in the Hospital Information System of the Brazilian Unified Health System (SIH/SUS) from January 2008 to December 2023; January–September 2024 was retained as an incomplete-year sensitivity extension [1]. Outcomes were in-hospital death and LOS, and admission-attributed mean daily bed-day burden was defined as admissions per calendar day multiplied by mean LOS.

Mortality was modelled by logistic regression with natural splines for time. Mean LOS was modelled separately for survivors and in-hospital deaths using Gaussian working generalized linear models with a March 2020 break. Models adjusted for age, sex, diagnostic subgroup, admission type, state of residence, seasonality and secular time, following interrupted time-series principles. Monthly predictions were standardized to a fixed pre-crisis reference population. Bed-day changes were decomposed into admission-inflow and LOS effects; LOS changes were decomposed into survivor, in-hospital death and survival-composition components.

The primary analysis included 843,861 admissions; the 2024 sensitivity extension added 60,081 records. The phase profile and bed-day decomposition are summarized in Table 1.

The results show why admission inflow and LOS must be interpreted together. In 2020–2021, shorter LOS partly offset the sharp rise in admissions, but bed-day burden still increased. Because this reduction in LOS occurred with standardized mortality above reference, the pattern is more consistent with crisis-period compression than with simple efficiency gain. In 2022–2023, admission inflow and LOS both increased burden, while

Phase	Adm./day	Obs. LOS	Obs. mort.	Mort. index	LOS index	Bed-days/day	Adm. effect	LOS effect
2008–2019	128.4	8.68	4.04%	1.000	1.000	1,114	—	—
2020–2021	177.7	7.34	4.39%	1.041	0.839	1,305	+395.2	-204.8
2022–2023	207.0	7.82	3.86%	0.938	0.887	1,619	+221.8	+92.2

Table 1. Phase profiles and bed-day decomposition. Indices relative to 2008–2019; effects in bed-days/day versus the previous phase.

standardized mortality and LOS were both below reference, indicating a different post-crisis profile. The LOS effect was driven mainly by survivor stays rather than by changes in the survivor/death mix.

The methodological contribution is a transferable administrative-data workflow that combines fixed-reference standardization, joint mortality–LOS interpretation, survival-status LOS decomposition and admission-based bed-demand decomposition. Although illustrated with Brazilian SIH/SUS records, the framework applies to other acute hospital conditions with routinely recorded admissions, LOS, discharge status and basic risk adjusters. The present study focuses on the admission-attributed burden layer, where changes in burden can be decomposed into admission and LOS components; complementary accepted work extends the framework to daily occupancy reconstruction from admission and discharge dates [5].

C. Castro acknowledges support from the FCT project UID/00013/2025 (`https://doi.org/10.54499/UID/00013/2025`).

References

[1] Brasil. Ministério da Saúde. DATASUS. Morbidade Hospitalar do SUS (SIH/SUS). `https://datasus.saude.gov.br/acesso-a-informacao/morbidade-hospitalar-do-sus-sih-sus/` n.d. (accessed in 02.May.2026).

[2] P. Nordström, Y. Gustafson, K. Michaëlsson, and A. Nordström. Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in Sweden. *BMJ* **350: h696**, 2015. doi: 10.1136/bmj.h696.

[3] S. Hwang, C. Ahn, and M. Won. Comparing the 30-Day Mortality for Hip Fractures in Patients with and without COVID-19: An Updated Meta-Analysis. *Journal of Personalized Medicine* **13: 669**, 2023. doi: 10.3390/jpm13040669.

[4] J. Lopez Bernal, S. Cummins, and A. Gasparrini. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology* **46**, 348–355, 2017. doi: 10.1093/ije/dyw098.

[5] C. Castro, A. J. Soares, A. P. Amorim, and A. Magalhães. A Daily-Grid Hazards-to-Dynamics Framework for Inpatient Flow. In: *Proceedings of the 26th International Conference on Computational Science and Its Applications (ICCSA 2026)*, Springer, Lecture Notes in Computer Science, 2026 (Accepted for publication).

Exact randomisation-based rank test for grouping factor levels with a multivariate extension

Célia Nunes[1,2], Kwaku Opoku-Ameyaw[2,3] and Manuel L. Esquível[4,5]

[1]Department of Mathematics, Universidade da Beira Interior, Covilhã, Portugal
[2]Center of Mathematics and Applications, Universidade da Beira Interior, Covilhã, Portugal
[3]Cocoa Research Institute of Ghana (CRIG), New Tafo-Akim, Ghana
[4]Department of Mathematics, NOVA School of Science and Technology, Universidade Nova de Lisboa, Monte da Caparica, Portugal
[5]NOVAMath - Center of Mathematics and Application, NOVA School of Science and Technology, Universidade Nova de Lisboa, Monte da Caparica, Portugal

E-mail addresses: *celian@ubi.pt; kwaku.opokuameyaw@crig.org.gh; mle@fct.unl.pt*

In this work, we propose an exact nonparametric test for assessing the homogeneity of predefined groupings of factor levels in designed experiments. The test is based on rank differences within groups, and its exact finite-sample distribution is derived by a randomisation method inspired by Fisher's randomisation principle. This approach avoids Monte Carlo simulation and yields exact quantiles for the test statistic under the null hypothesis of non homogeneity. We further extend the method to a multivariate framework by introducing a vectorised rank statistic, which enables the simultaneous testing of grouping homogeneity across several dependent variables. The methodology is illustrated using data from a cocoa breeding experiment in Ghana, where plant varieties were evaluated on multiple acidic soil types.

Keywords

Nonparametric test, grouping of factor levels, randomisation method, univariate and multivariate case, cocoa breeding experiment.

In many experimental contexts, the factor levels of a design can be partitioned into subgroups according to a known structural characteristic – such as genetic ancestry, geographical origin, or treatment similarity. A key statistical question is whether such a priori groupings are homogeneous with respect to one or more response variables. Testing this hypothesis provides insight into whether the assumed structure reflects real differences in the observed outcomes. Existing nonparametric approaches for factorial designs, including rank-based and pseudo-rank methods (see, e.g., [1, 2]), provide powerful tools for assessing treatment effects without distributional assumptions. However, these approaches usually rely on asymptotic results or simulation-based approximations. When sample sizes are small or the design structure is complex, exact inference may be preferable. In this work, following the study of Opoku-Ameyaw et al. (2023) [3], we introduce a rank-based exact randomisation test for assessing the homogeneity of factor-level groupings. The test statistic is constructed from within-group differences of ranks of aggregated observations, and its distribution under the null hypothesis is derived explicitly through combinatorial randomisation. This procedure, inspired by Fisher's randomisation method (see, e.g., [4]), provides exact quantiles and allows the test to be performed without resorting to simulation or asymptotic approximations, representing a clear advantage when compared

with the method proposed in [3]. Furthermore, we extend the methodology to the multivariate setting by proposing a multivariate extension of the test that allows simultaneous evaluation of grouping homogeneity across several dependent variables. To demonstrate the applicability of the proposed methodology, we analyse data from a cocoa breeding experiment in Ghana involving twelve plant varieties cultivated in four acidic soil types. The experiment provides a practical setting for evaluating the proposed tests, both in univariate and multivariate forms. The analysis confirms the capacity of the method to detect significant grouping effects with small sample sizes, and highlights its robustness across different soil conditions.

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/00212/2025 (https://doi.org/10.54499/UID/00212/2025) (Center of Mathematics and Applications of University of Beira Interior) and UID/00297/2025 (https://doi.org/10.54499/UID/-00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications – NOVA Math).

References

[1] W. J. Conover. *Practical Nonparametric Statistics.* 3rd Ed., New York: John Wiley & Sons, Inc., 1999.

[2] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods.* 3rd Ed., Hoboken, NJ: John Wiley & Sons, Inc., 2014.

[3] K. Opoku-Ameyaw, C. Nunes and M. L. Esquivel. CMMSE: a nonparametric test for grouping factor levels: an application to cocoa breeding experiments in acidic soils. *Journal of Mathematical Chemistry* **61**(3), j652-672, 2023.

[4] D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods.* Springer Texts in Statistics, New York: Springer, 2013.

Combining Neural Networks and Additive Models for Improved Credit Risk Prediction

Cristina Duarte[1], M. Rosário Ramos[1,2,3]

[1] Universidade Aberta, Portugal
[2] LE@D, Universidade Aberta and CEG
[3] CEAUL, Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal

E-mail addresses: *2102145@estudante.uab.pt ; mariar.ramos@uab.pt*

This work explores a Neural Additive Credit Risk Model (NACRM), an approach that combines the flexibility of neural networks with the interpretability of additive models such as GAM. The model is designed for applications where both predictive accuracy and explainability are essential, particularly in the financial sector, where decisions must be transparent to regulators and clients.The NACRM is applied to credit scoring, a domain that requires accurate and interpretable assessments of borrower risk. A key objective is to reduce false negatives, where high-risk borrowers are incorrectly classified as creditworthy, potentially leading to significant financial losses. The case study shows that the NACRM can create distinct clusters that effectively separate default and non-default borrowers, improving decision boundaries and lending decisions. In addition, the model estimates individual default probabilities, providing valuable insights into applicant risk profiles. The empirical analysis is based on the Freddie Mac loan-level dataset [4] as one of the most relevant publicly available datasets for credit risk research. All modelling and experiments were conducted using the R software.

Keywords

Credit scoring, false negatives rate, interpretable deep learning, probability of default

Acknowledgements: This work is partially funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, within the scope of the projects UID/4372/2025 and UID/PRR/04372/2025.

References

[1] Chen, Y., Calabrese, R. and Martin-Barragan, B. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research* **312(1)**, 357–372, 2023.https://doi.org/10.1016/j.ejor.2023.06.036

[2] Dastile, X., Celik, T., Potsane, M. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, **91**, 341-251, 2020. https://doi.org/10.1016/j.asoc.2020.106263

[3] Markov, A., Seleznyova, Z., Lapshin, V. Credit scoring methods: Latest trends and points to consider. *Journal of Finance and Data Science*, **8**), 180–201, 2022. https://doi.org/10.1016/j.jfds.2022.07.002

[4] Sousa, M. R., Gama, J., Brandão, E. A new dynamic modeling framework for credit risk assessment. 45, 341-351. *Expert Systems with Applications*, **45**, 341-251, 2016. https://doi.org/10.1016/j.eswa.2015.09.055

Global Disaster Trends and Impacts: A Statistical Analysis of EM-DAT Records (1920–2025)

Rita Martins[1], Cristina Lopes[2], Isabel Vieira[2], Lurdes Babo[2], Cristina Torres[2]

[1] ISCAP, Polytechnic of Porto, Portugal
[2] CEOS.PP, ISCAP, Polytechnic of Porto, Portugal

E-mail addresses: *2190904@iscap.ipp.pt; cristinalopes@iscap.ipp.pt;mivieira@iscap.ipp.pt; lbabo@iscap.ipp.pt; ctorres@iscap.ipp.pt*

This study analyses 26,456 disaster records (1920–2025) from the Emergency Events Database (EM-DAT) to examine patterns of occurrence and their human and economic impacts. Results indicate decreasing mortality despite rising disaster frequency, likely reflecting advances in preparedness, training, medical care, and humanitarian logistics. Economic losses correlate strongly with injuries and affected populations, but weakly with mortality, highlighting the complex relationship between fatalities and economic damage.

Keywords

disasters, humanitarian supply chain, correlations, hypotheses tests.

A disaster may be defined as a serious disruption to the functioning of a community or society caused by hazardous events, resulting in human, material, economic, or environmental losses that surpass the affected community's ability to cope using its own resources [1]. Over recent decades, the increasing frequency and severity of natural and technological disasters have posed significant challenges to societies worldwide, affecting human life, infrastructure, economies, and the environment [2]. In many cases, the scale of these events exceeds local response capacities, requiring national or international assistance [3].

This study aims to identify and analyse patterns in disaster occurrence and to examine the relationships between their human, material, and economic impacts. Secondary data were collected from the Emergency Events Database (EM-DAT) [4], comprising 26,456 disaster records worldwide from 1920 to February 2025. Statistical and correlation analyses were conducted to evaluate trends in mortality, injuries, affected populations, economic losses, declarations of state of emergency, and requests for international aid [5].

The results reveal that, although the absolute number of injuries remains substantially higher than the number of deaths, disaster-related mortality has shown a decreasing trend over the decades. This pattern may reflect improvements in disaster preparedness and mitigation strategies, including advances in early warning systems, population education, professional training, medical care, and humanitarian supply chain management. These developments appear to have contributed to reducing fatalities even as disasters continue to occur more frequently and with greater intensity.

Correlation analyses using Spearman's coefficient indicate strong associations between the number of injured individuals, affected populations, and economic losses. In contrast, total economic damage demonstrates only a weak relationship with mortality rates, suggesting that reductions in fatalities do not necessarily correspond to lower economic impacts. Furthermore, requests for international aid and declarations of states of emergency

are more strongly associated with injuries, affected populations, and financial losses than with the number of deaths.

Overall, the findings highlight the multidimensional nature of disasters and reinforce the importance of integrated disaster risk reduction policies. While progress has been achieved in reducing mortality, economic vulnerability and social disruption remain substantial challenges, emphasizing the need for continued investment in preparedness, resilience, and coordinated emergency response systems.

References

[1] UNDRR. The Sendai Framework Terminology on Disaster Risk Reduction. Available at: `https://www.undrr.org/drr-glossary/terminology7disaster`, accessed January 12, 2026.

[2] M. Ladds, A. Keating, J. Handmer, and L. Magee. How much do disasters cost? A comparison of disaster cost estimates in Australia. *International Journal of Disaster Risk Reduction* **21**, 419–429, 2017.

[3] M. T. Ortuño, P. Cristóbal, J. M. Ferrer, F. J. Martín-Campo, S. Muñoz, G. Tirado, and B. Vitoriano. Decision aid models and systems for humanitarian logistics: A survey. *Decision Aid Models for Disaster Management and Emergencies* **7**, 17–44, 2013.

[4] Centre for Research on the Epidemiology of Disasters. *EM-DAT: The International Disaster Database*. Université Catholique de Louvain, 2025.

[5] 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.

Bootstrap-based fisher scoring in contaminated state-space models

F. Catarina Pereira[1], A. Manuela Gonçalves[2], Marco Costa[3]

[1]University of Minho and Centre of Mathematics, Campus de Azurém, 4800-058 Guimarães, Portugal
[2]University of Minho, Department of Mathematics and Centre of Mathematics, Campus de Azurém, 4800-058 Guimarães, Portugal
[3]University of Aveiro, ESTGA - Águeda School of Technology and Management and CIDMA – Centre for Research and Development in Mathematics and Applications, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

E-mail addresses: *catarina.cardoso@math.uminho.pt; mneves@math.uminho.pt; marco@ua.pt*

State-space models are widely used for dynamic time series analysis, although outliers may affect parameter estimation and forecasting performance. This work proposes a robust parameter estimation approach based on a modified Fisher Scoring algorithm combined with bootstrap techniques. The method improves numerical stability and estimation reliability under contaminated scenarios, showing promising results in simulation studies and temperature forecasting applications.

Keywords

state-space models, outliers, bootstrap, fisher scoring, parameter estimation.

State-space models (SSMs) are widely used to describe dynamic systems in fields such as environmental sciences, economics, and engineering [1, 2]. Their flexibility allows the modelling of latent processes and the recursive updating of predictions through the Kalman filter. However, parameter estimation in SSMs may become unstable in the presence of outliers or small sample sizes, affecting convergence and inference reliability [3].

This work presents the Boost Fisher Scoring (BF) algorithm, proposed in [4], which combines the Fisher scoring method with bootstrap techniques to approximate the Fisher information matrix. The methodology aims to improve numerical stability and the estimation of standard errors while preserving the original Gaussian state-space framework. A robust extension, denoted by BFout, was specifically developed for contaminated time series by performing bootstrap resampling after removing extreme innovations.

The performance of the proposed methods was assessed through simulation studies under different scenarios of sample size, variance, and autocorrelation. The results indicate that BF and BFout improve convergence behaviour and provide competitive parameter estimates with lower computational cost than full bootstrap procedures.

The practical usefulness of the methodology was further illustrated through an application to short-term maximum temperature forecasting in Northern Portugal, where the proposed approaches showed improved robustness and prediction accuracy in the presence of contaminated observations.

Acknowledgements: The research of F. Catarina Pereira and 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). Marco Costa was supported by CIDMA (https://ror.org/05pm2mw36) under the 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] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: with R Examples*. 5th Ed., Springer, 2025. https://doi.org/10.1007/978-3-031-70584-7

[2] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[3] M. Auger-Méthé, C. Field, C. M. Albertsen, A. E. Derocher, M. A. Lewis, I. D. Jonsen, and J. M. Flemming. State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. *Scientific Reports* **6**, 26677, 2016. https://doi.org/10.1038/srep26677

[4] F. C. Pereira, M. Costa, and A. M. Gonçalves. A bootstrap-enhanced Fisher scoring algorithm for parameter estimation in state-space models. *Computational Statistics* **41**, 65, 2026. https://doi.org/10.1007/s00180-026-01746-2

Exploring topological features of gait dynamics in Fabry disease

Jhonathan Barrios[1], Wolfram Erlhagen[1], Miguel F. Gago[2,3], Estela Bicho[4] and Flora Ferreira[5,6]

[1]Centre of Mathematics (CMAT), University of Minho, Portugal
[2]Neurology Department, Hospital da Senhora da Oliveira, Portugal
[3] Life and Health Sciences Research Institute (ICVS), University of Minho, Portugal
[4]Algoritmi Centre, School of Engineering, University of Minho, Portugal
[5] School of Economics and Management, University of Porto, Portugal
[6] CMUP, Faculty of Sciences, University of Porto, Portugal

E-mail addresses: *id10605@uminho.pt; wolfram.erlhagen@math.uminho.pt; miguelgago@hospitaldeguimaraes.min-saude.pt; estela.bicho@dei.uminho.pt; flora.ferreira@fep.up.pt*

Fabry disease (FD) is a rare multisystemic disorder in which gait impairments are increasingly reported. This work investigates whether topological data analysis of gait time series identifies patterns associated with central nervous system lesions (CNSL). Gait data from 41 subjects (controls, FD with CNSL, FD without CNSL) were analyzed using persistent homology. FD patients with CNSL showed increased H0 lifetime variability and maximum lifetime in posture and speed variables, suggesting less compact and more heterogeneous gait dynamics.

Keywords

Fabry disease, gait time series, nonlinear dynamics, Topological Data Analysis, persistent homology

Fabry disease (FD) is a rare, genetic, progressive, and multisystemic disease associated with renal, cardiac, and neurological involvement. Previous studies have shown that gait variables contain discriminative information. Fernandes et al. demonstrated that multiple regression normalization improved the classification of FD based on gait characteristics, highlighting differences in foot flat and pushing [1]. Braga et al. further investigated Fabry patients with and without white matter lesions, identifying gait variables as relevant discriminative features of gait [2]. Recently, Topological Data Analysis (TDA) has been applied to gait time series in parkinsonism, demonstrating that descriptors derived from persistent homology can capture nonlinear and structural information that complements traditional gait analysis [3].

The aim of this work is to explore whether the topological features of step-by-step gait time series can reveal dynamic patterns associated with Fabry disease, with particular attention to central nervous system lesions (CNSL). Gait data were collected using Physilog sensors (GaitUp®) during a self-selected 60-meter walking protocol. The final dataset included 41 subjects: 15 healthy controls (CO), 15 FD patients with CNSL, and 11 FD patients without CNSL. Seventeen gait variables were analyzed (see details in [1, 2]). For each subject and gait variable, topological descriptors were extracted, and each step time series was reconstructed using Takens embeddings with delay $\tau = 1$ and embedding dimension $m = 3$, preserving the dynamics of consecutive steps and maintaining a sufficient number of embedded points. Persistent homology was computed from Vietoris-Rips

complexes in dimensions H_0 and H_1. Topological descriptors included total lifetime, mean lifetime, lifetime standard deviation, maximum lifetime, total persistence, and persistence entropy. Group comparisons were performed at the individual level using Mann-Whitney tests, permutation tests for median differences, bootstrap confidence intervals, Cliff's delta test, and Benjamini-Hochberg false discovery rate correction.

Preliminary results show that when comparing FD patients with CNSL to CO, specifically for H_0 features, an increase in the standard deviation and maximum lifetime of H_0 was observed for posture Takens embeddings, while an increase in the maximum lifetime of H_0 was also observed for speed. These results suggest that FD patients with CNSL may exhibit more heterogeneous gait dynamics, particularly in variables associated with stance phase and speed. Exploratory patterns, not statistically significant, in patients with FD without CNSL included lower minimum toe clearance, higher pushing, lower foot flat, and increased gait cycle time and stride length, consistent with previous studies of aggregated features in Fabry disease [1, 2].

Acknowledgements: The first and second authors acknowledge the support provided by national funds through FCT - Fundação para a Ciência e Tecnologia through the doctoral scholarship `2023.02242.BDANA` (`https://doi.org/10.54499/2023.02242.BDANA`) and the CMAT's project `UID/00013/2025` (`https://doi.org/10.54499/UID/00013/2025`). The last author was partially supported by CMUP, member of LASI, which is financed by national funds through FCT under the project with reference `UID/00144/2025` and associated DOI given by (`https://doi.org/10.54499/UID/00144/2025`).

References

[1] C. Fernandes, F. Ferreira, M. Gago, O. Azevedo, N. Sousa, W. Erlhagen, and E. Bicho. Gait classification of patients with Fabry's disease based on normalized gait features obtained using multiple regression models. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2288–2295, 2019.

[2] J. Braga, F. Ferreira, C. Fernandes, M. F. Gago, O. Azevedo, N. Sousa, W. Erlhagen, and E. Bicho. Gait characteristics and their discriminative ability in patients with Fabry Disease with and without white-matter lesions. *Workshop on Computational Data Analysis and Numerical Methods*, 2020.

[3] J. Barrios, W. Erlhagen, M. F. Gago, E. Bicho, and F. Ferreira. Topological descriptors of foot clearance gait dynamics improve differential diagnosis of Parkinsonism. *arXiv preprint arXiv:2603.06212*, 2026.

The Delta Approximation Method for Mixed SDE Models - a refined approach

Gonçalo Jacinto[1,2,3], Patrícia A. Filipe[3,5,6], Carlos A. Braumann[3,4] and Nelson T. Jamba[3,7]

[1] Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, Aveiro, Portugal
[2] Univrsidade do Algarve, Faculdade de Ciências e Tecnologia, Faro, Portugal
[3] Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Évora, Portugal
[4] Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática, Évora, Portugal
[5] Iscte Business School, Iscte-Instituto Universitário de Lisboa, Lisboa, Portugal
[6] Business Research Unit (BRU-IUL), Lisboa, Portugal
[7] Faculdade de Ciências Naturais, Universidade do Namibe, Namibe, Angola

E-mail addresses: *gjjacinto@ualg.pt; patricia.filipe@iscte-iul.pt; braumann@uevora.pt; nelson.jamba@uninbe.ao*

Accurately describing animal growth under environmental variability requires stochastic modeling approaches, since classical regression models fail to capture the underlying dynamic structure of growth and the temporal dependence introduced by environmental fluctuations. Stochastic differential equation (SDE) models address this need by incorporating random fluctuations directly into the growth dynamics and, through suitable transformations, can be expressed in a Ornstein–Uhlenbeck model.

In mixed-effects settings, key parameters are allowed to vary across individuals, capturing heterogeneity in growth trajectories. However, this approach leads to marginal likelihoods that lack closed-form expressions. To address this issue, we have previously considered a Delta approximation method. In this work, we introduce a refinement based on applying the second-order Taylor expansion to the argument of the exponential function rather than to the exponential itself.

The proposed approach is evaluated using both simulated data and real weight records from a large and heterogeneous sample of Mertolengo cattle. Simulation studies with one and two random parameters show that the refined method consistently outperforms the classical Delta method and achieves results comparable to exact likelihood approaches when available. The method remains computationally efficient and applicable to irregularly sampled data.

Keywords

Delta method, Maximum likelihood estimation, Mixed-effects models, Stochastic differential equations

Acknowledgements: This work was supported by FCT under projects 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), through CIDMA, and by CIMA – Centro de Investigação em Matemática e Aplicações, under Project UID/04674/2025 (https://doi.org/10.54499/UID/04674/2025).

References

[1] N. T. Jamba, G. Jacinto, P. A. Filipe, and C. A. Braumann. Likelihood function through the Delta approximation in mixed SDE models. *Mathematics* **10(3): 385**, 2022. DOI: https://doi.org/10.3390/math10030385

[2] N. T. Jamba, G. Jacinto, P. A. Filipe, and C. A. Braumann. Estimation for stochastic differential equation mixed models using approximation methods. *AIMS Mathematics* **9(4)**, 7866–7894, 2024. DOI: https://doi.org/10.3934/math.2024383

A unified imputation framework for interval-censored data: comparing AFT, RSF, and DeepSurv models

Gustavo Soutinho[1], and Luis Meira-Machado[2]

[1] Department of Science and Technology, Portucalense University, Portugal
[2] Centre of Mathematics, University of Minho, Portugal

E-mail addresses: *gustavo.soutinho@upt.pt; lmachado@math.uminho.pt*

Interval-censored data are common in longitudinal studies and pose challenges for time-to-event analysis. This work proposes a unified imputation-based framework for handling interval-censored data, where latent event times are iteratively generated within the observed censoring intervals and the censoring mechanism is handled externally through a scaled redistribution procedure. Within this framework, different predictive models—including AFT, Random Survival Forests, and DeepSurv—can be consistently compared through an iterative imputation scheme based on pseudo-event times within the observed intervals, followed by a common scaled redistribution procedure. Performance is assessed through simulations under varying censoring levels, interval widths, and hazard distributions, with extensions to nonlinear effects and high-dimensional covariates. Results are further validated using real-world clinical datasets.

Keywords

Interval-censored data, imputation-based framework, DeepSurv, random survival forests.

Interval-censored data represent a common challenge in biostatistics, reliability engineering, and longitudinal clinical trials. Unlike right-censoring, where the event is only known to occur after a given time, interval censoring arises when the event is only known to have occurred within a time window $[L, R]$. This setting is common in medical studies with periodic follow-up assessments, such as asymptomatic disease progression or dental health studies, where the exact transition time remains unobserved.

Let T denote a non-negative random variable representing the time to event. Under interval censoring, T is not directly observed; instead, one observes a pair (L_i, R_i) such that $T_i \in (L_i, R_i]$, where L_i and R_i denote the last time the event was known not to have occurred and the first time it was observed to have occurred, respectively. Among the available approaches for estimating the survival function, the Turnbull estimator [1] is a widely used nonparametric extension of the Kaplan–Meier method. However, it cannot incorporate covariates or estimate hazard ratios, limiting its use in regression settings. Imputation offers a practical alternative by replacing incomplete observations with plausible event times $\hat{t}_i \in (L_i, R_i]$, enabling the use of standard survival workflows.

Within the proposed framework, Accelerated Failure Time (AFT) models are used as predictive tools to guide event-time imputation. These models assume that covariates act multiplicatively on survival time and are commonly written as

$$\log(T) = X\beta + \epsilon,$$

where ϵ follows a specified distribution (e.g., Gumbel for Weibull models). Although interpretable, AFT models may be limited in capturing complex nonlinear relationships.

Recent developments in machine learning and deep learning provide greater flexibility for modeling intricate patterns while treating censoring intervals as constraints within an imputation framework. Random Survival Forests (RSF) construct ensembles of trees using survival-based splitting rules such as the log-rank criterion [3]. DeepSurv is a Cox-based deep neural network that replaces the linear predictor $\beta^T X$ with a nonlinear function $h_\theta(X)$ learned from the data [4].

The Scaled Redistribution method is a semi-parametric adjustment that reallocates predicted event times within the observed interval $[L, R]$, ensuring coherence with censoring bounds while preserving relative variability. It avoids strong distributional assumptions and remains computationally efficient for small-to-medium datasets. Despite the growing number of available methods, a unified benchmarking framework for comparing classical imputation approaches, AFT models, and modern predictive methods such as RSF and DeepSurv is still lacking. Building on previous research [5], this work addresses this gap through a comprehensive comparative evaluation of imputation strategies based on simulation studies and applications to real-world datasets.

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Programme Contracts UID/05105/2025, UID/00013/2025), and 2023.14897.PEX.

References

[1] B. W. Turnbull, The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society Series B (Methodological)*, **38**, 290–295, 1976.

[2] V. Kariuki, A. Wanjoya, O. Ngesa, M. M. Mansour, E. M. A. Elrazik, A. Z. Afify, The accelerated failure time regression model under the extended-exponential distribution with survival analysis, *AIMS Mathematics*, **9**, 15610–15638, 2024.

[3] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests, *The Annals of Applied Statistics*, **2**, 841–860, 2008.

[4] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment analysis for survival data using deep Cox proportional hazards networks, *BMC Medical Research Methodology*, **18**, 1–12, 2018.

[5] G. Soutinho, L. Meira-Machado, Imputation strategies for interval-censored data: from AFT models to machine learning and scaled redistribution, *AIMS Mathematics*, **11**, 5719–5737, 2026.

Optimization methods for trading off mean and loss probability in an additive risk model

Irene Brito

[1]Centre of Mathematics, Department of Mathematics, University of Minho, 4800-045 Guimarães, Portugal

E-mail addresses: *ireneb@math.uminho.pt*

Considering a two-component risk model that combines additively mean and loss probability through a trade-off parameter, optimization techniques adapted from criteria-weighting methods in multi-criteria decision-making are proposed to determine the optimal value of this parameter. The application of the risk model and the parameter optimization techniques are illustrated through a study on air pollution risk assessment.

Keywords

risk model, risk assessment, air pollution.

In risk–value models, the trade-off parameter controls the balance between risk and value, and it can be estimated based on decision-makers' preferences or attitudes toward risk in the economic context [1]. The mean-variance risk model formulated by Pollatsek and Tversky [2] is one of those models, where risk and expected return are linearly combined using a trade-off parameter. Adapting these models for risk assessment using historical data in different contexts, determining the trade-off parameter can be a complex task, since preference judgments are often difficult to obtain.

In this work, the Mean-loss probability risk model, that trades off mean and loss probability, is proposed for risk assessment together with different trade-off parameter optimization techniques. The optimization techniques are adapted from multi-criteria decision analysis, where they are used to assign criteria weights. Some of those techniques are based on maximizing the variation between the available choices [4], as e.g. the weighted principal component analysis method [3].

The Mean-loss probability risk model is applied to an air pollution risk assessment problem, where the different techniques are tested for determining the optimal parameter. The aim is to classify 27 EU capital cities in terms of risk for air pollution based on monthly averages of Nitrogen Dioxide concentrations measured from January 2018 to December 2025.

Acknowledgements: This research was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM).

References

[1] J. S. Dyer, J.Jia. Relative risk—value models. *European Journal of Operational Research* **103: 1**, 170-185, 1997.

[2] A. Pollatsek, A. Tversky. A theory of risk. *Journal of Mathematical Psychology* **7: 3**, 540-553, 1970.

[3] S. B. Kim, P. Rattakorn. Unsupervised feature selection using weighted principal components. *Expert Systems with Applications* **38: 5**, 5704-5710, 2011.

[4] S. Chatterjee, S. Chakraborty. A study on the effects of objective weighting methods on TOPSIS-based parametric optimization of non-traditional machining processes. *Decision Analytics Journal* **11**, 100451, 2024.

Exploring Synthetic Data Generation for Count Time Series: Coverage Diagnostics and Wasserstein Feature Similarity

Isabel Silva[1], Maria Eduarda Silva[2] and Isabel Pereira[3]

[1]Faculdade de Engenharia da Universidade do Porto and CIDMA, Portugal
[2]Faculdade de Economia da Universidade do Porto and LIADD-INESC TEC, Portugal
[3]Departamento de Matemática, Universidade de Aveiro and CIDMA, Portugal

E-mail addresses: *ims@fe.up.pt; mesilva@fep.up.pt; isabel.pereira@ua.pt*

Research on count time series is limited by scarce real-world data, hindering model evaluation. This study addresses the issue by generating synthetic series using mixtures of INAR models and feature-based characterization. Observed and simulated data are compared in PCA space using KNN miscoverage and Wasserstein distance. The proposed framework yields diverse and realistic synthetic datasets with tunable characteristics, enabling more robust assessment of count time series models.

Keywords

Synthetic data; count time series; PoINAR(1) models; coverage diagnostics; Wasserstein distance

Synthetic generation of count-valued time series is increasingly useful for benchmarking and robustness evaluation of methods (e.g., forecasting and anomaly detection), for augmenting scarce machine-learning training data, and for privacy-preserving data sharing when raw series cannot be released. In public-health and reliability settings, in particular, realistic synthetic count trajectories can help evaluate methods under controlled but diverse scenarios.

We propose a feature-based generation framework build on mixtures of PoINAR(1) models. Feature-based approaches are attractive because they support both interpretable model control (via summary descriptors) and metric-based evaluation in a common representation space. The mixture structure, in turn, introduces heterogeneity across regimes and is designed to reproduce the diversity typically observed in real collections of count time series.

Observed and simulated count time series are mapped to an interpretable feature vector capturing key properties of count data (including dependence, dispersion, and sparsity) and then embedded into a common PCA space. We assess geometric overlap between observed and simulated feature clouds using nearest-neighbour miscoverage and reverse miscoverage, where the distance threshold is defined in a data-driven way as a percentile of the observed nearest-neighbour distance distribution. Complementarily, we quantify feature-level distributional agreement using Wasserstein distances computed on standardized features.

In a large simulation study with 20,000 randomly generated mixtures and numbers of components $k = 1, \ldots, 10$, the combined evidence from Wasserstein distances and the coverage-based diagnostics did not identify a single optimal value of k. Nevertheless, to balance model parsimony with overall performance across miscoverage, reverse miscoverage, and feature-wise Wasserstein discrepancies, we adopt $k = 3$ components as a pragmatic

trade-off. Finally, by comparing the synthetic output with commonly used benchmark datasets, we illustrate that the proposed generator can produce realistic yet diverse count time series, supporting more thorough evaluation and comparison of models for count-valued temporal data.

Acknowledgements: The first and third authors were partially supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT) Multi-Annual Financing Program for R&D Units, grants UID/4106/2025 and UID/PRR/4106/2025. The second author was partially supported by INESCTEC through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia) grant UID/50014/2025 (DOI:10.54499/UID/50014/2025) .

Estimating the nesting abundance of green sea turtles (*Chelonia mydas*) on Poilão Island

João Louro[1,2], Lisete Sousa[1,2], Ana Rita Patrício[1,3], and Castro Barbosa[4]

[1]Faculty of Sciences of the University of Lisbon, Portugal
[2]Research Center CEAUL, Portugal
[3]Research Center CE3C, Portugal
[4]Institute of Biodiversity and Protected Areas of Guinea-Bissau

E-mail addresses: *jllouro@ciencias.ulisboa.pt; lmsousa@ciencias.ulisboa.pt; arpatricio@ciencias.ulisboa.pt, castrobarbosa_2002@yahoo.com*

This study on Poilão Island, Guinea-Bissau, compared metrics to estimate the abundance of female green sea turtles. We found that counting emergences at high tide with a correction factor was the most effective method. GAM and GAMLSS models showed equal predictive performance. For 2020–2022, we estimated 29,529 unique nesting females and 46,006 clutches, aligning with previous literature. These findings refine monitoring and conservation for this globally significant nesting site.

Keywords

Green turtle, Poilão, Sea turtle, Generalized Additive Models, Generalized Additive Models for Location, Scale and Shape.

Poilão Island, in the Bijagós Archipelago of Guinea-Bissau, is one of the most important nesting sites for the green turtle (*Chelonia mydas*) in the world and a critical area for marine turtle conservation [2]. Despite its small size, the island supports a large nesting population, making reliable estimates of nesting abundance essential for monitoring and management [1]. In this study, we investigated which metrics best estimate the number of turtles emerging on Poilão to nest, and we assessed the performance of different approaches for predicting total emergence counts. We used data collected from four of the five beaches on Poilão between 2014 and 2022. The metrics evaluated included: emergence counts at high tide peak, high-tide peak counts corrected with a correction factor derived from 2019 data, track counts, and counts of turtles stranded on the rocks surrounding the island. In 2019 and 2020, we also implemented a more accurate counting method by marking turtles with non-toxic fluorescent, thereby avoiding double counts. High-tide peak counts consistently underestimated the total number of emergence counts, whereas applying the 2019 correction factor brought estimates closer to those obtained through the paint-marking method. This correction performed well across years with different nesting densities. We calculated absolute differences between each metric and the reference total derived from paint marking and ranked the metrics accordingly. Overall, the corrected high-tide peak count was the closest estimate in the years analysed. Generalized Additive Models (GAMs) and Generalized Additive Models for Location, Scale and Shape (GAMLSS) were fitted to predict total emergence counts using explanatory variables such as track counts and stranded turtle counts on the surrounding rocks. Models with track counts performed better than models with stranded counts. GAMLSS provided lower AIC values than GAM; however, both approaches showed similar predictive

performance. Across 2020–2022, the estimated number of individual emerging turtles was approximately 29,529, and nest abundance over the same period was estimated at 46,006, corresponding to an average of about 15,335 nests per year. Our findings suggest that a correction factor derived from 2019 can provide robust estimates under different nesting densities, and that track-based and model-based approaches are promising tools for long-term monitoring. Further data collection, especially during high-density nesting years, is needed to refine beach-specific correction factors and improve predictive accuracy.

Acknowledgements: This work was carried out as part of the “Tartarugas Marinhas" Conservation Project, Poilão, Guinea-Bissau, in collaboration with the Instituto da Biodiversidade e das Áreas Protegidas (IBAP).

References

[1] A. R. Patrício, M. R. Varela, C. Barbosa, A. C. Broderick, P. Catry, L. A. Hawkes, A. Regalla, and B. J. Godley. Climate change resilience of a globally important sea turtle nesting population. *Global Change Biology* **25**, 522–535, 2019.

[2] A. C. C. Barbosa, A. C. Broderick, and P. Catry. Marine turtles in the Orango National Park (Bijagós Archipelago, Guinea-Bissau). *Marine Turtle Newsletter* **81**, 6–7, 1998.

Affinity Coefficient vs. Euclidean Distance in Hierarchical Clustering of Patients with Alcohol Use Disorder

Leonor Bacelar-Nicolau[1], Áurea Sousa[2], Sónia Ferreira[3,4], Cristina Ribeiro[5], Ana Paula Nascimento[6] and Helena Bacelar-Nicolau[7]

[1]Center for Interdisciplinary Research in Health (CIIS), Católica Medical School, Universidade Católica Portuguesa, Lisboa, Portugal
[2]Faculty of Sciences and Technology, CEEAplA and OSEAN, Universidade dos Açores, Ponta Delgada, Portugal
[3]Unidade de Tratamento e Reabilitação de Alcoólicos, Unidade Local de Saúde de São José, Lisboa, Portugal
[4]Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[5]Instituto de Medicina Preventiva e Saúde Pública, Clínica Universitária de Medicina Geral e Familiar, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[6]RISE-Health, Center for Translational Health and Medical Biotechnology (TBIO), Escola Superior de Saúde (E2S), Polytechnic of Porto (P. Porto), Porto, Portugal
[7]Faculty of Psychology, Institute of Environmental Health (ISAMB/FM-UL), Universidade de Lisboa, Lisboa, Portugal

E-mail addresses: *lnicolau@ucp.pt; aurea.st.sousa@uac.pt; sonia.mferreira@sapo.pt; cpgomes@medicina.ulisboa.pt; ananascimento@ess.ipp.pt; hbacelar@psicologia.ulisboa.pt*

Hierarchical cluster analysis is applied to longitudinal neuropsychological and quality of life data from 67 patients with alcohol use disorder, comparing the affinity coefficient with the Euclidean distance on raw and difference score profiles. Partition quality and agreement between methods are assessed using the STAT coefficient, tanglegram and Adjusted Rand Index.

Keywords: Affinity coefficient, Hierarchical clustering, Complete linkage, Alcohol use disorder, Neuropsychological rehabilitation, Longitudinal profiles, Real-valued data.

Executive function deficits are well documented in alcohol use disorder (AUD) and represent a key target for neuropsychological rehabilitation (NR). In this work, we apply agglomerative hierarchical cluster analysis to a longitudinal dataset of 67 outpatients with AUD followed over 6 months at an Alcohology and New Addictions Service in Lisbon, Portugal. Patients were assigned to two treatment modalities: a weekly therapeutic group (WTG, $n = 44$) and a neuropsychological rehabilitation group (NRG, $n = 23$) [1]. At three assessment moments — baseline (M1), 3 months (M2) and 6 months (M3) — patients were evaluated on a battery of 19 neuropsychological and quality of life measures covering general executive function (FAB), cognitive flexibility (TMT, WCST), working memory (Letters & Numbers), information processing speed (Codes), planning (Zoo Map, Key Search), verbal fluency, inhibition (Stroop), and the four domains of the WHOQOL-Bref (physical, psychological, social relations, environment).

The study is structured in two parts, both using Complete Linkage as aggregation criterion. In Part 1, patients are described by the 57 raw score variables (19 measures $\times$ 3 time points), which are non-negative. Hierarchical clustering is performed using the standard affinity coefficient [2] and, for comparison, the Euclidean distance. In Part 2, patients

are described by 57 difference score variables, computed as pairwise differences between assessment moments (M1→M2, M1→M3, M2→M3) for each of the 19 measures. These difference scores are real-valued and may take negative values, motivating the use of the generalised affinity coefficient, extended to handle real-valued data [3]. Hierarchical clustering is again performed using both the generalised affinity coefficient and the Euclidean distance.

In both parts, the comparison between affinity-based and Euclidean distance-based solutions is conducted using the STAT coefficient to assess partition adequacy, the tanglegram and entanglement coefficient to compare dendrogram structures, and the Adjusted Rand Index (ARI) to quantify agreement between the best partitions from each method.

By relying on normalised row profiles and incorporating the sign of observations, the affinity coefficient emphasises similarities in temporal structure and directional evolution, reducing the influence of scale and mitigating redundancy induced by correlated variables. In contrast, Euclidean distance-based clustering is sensitive to magnitude and variance, potentially grouping patients with similar overall score levels but distinct recovery trajectories into the same cluster. This study illustrates the potential added value of the affinity coefficient — and its generalisation to real-valued data — in a health data context characterised by multidimensional, longitudinal and clinically relevant observations.

Acknowledgements: This work was partially supported by FCT – Fundação para a Ciência e Tecnologia, I.P., by project references UID/04279/2025 (DOI: `https://doi.org/10.54499/UID/04279/2025https://doi.org/10.54499/UID/04279/2025` – Centro de Investigação Interdisciplinar em Saúde), and UIDB/00685/2025 (Centre of Applied Economics Studies of the Atlantic – School of Business and Economics, University of the Azores).

References

[1] S. Ferreira, L. Bacelar-Nicolau, M. Oliveira, S. Pombo, E. Vásquez-Justo, and C. Ribeiro. Executive Functions in Alcohol Use Disorder: The Positive Role of Neuropsychological Rehabilitation — Prospective Cohort Study. *Drug and Alcohol Review* **45(4): e70154**, 2026. `https://doi.org/10.1111/dar.70154`

[2] H. Bacelar-Nicolau, F. Nicolau, Á. Sousa, and L. Bacelar-Nicolau. Measuring Similarity of Complex and Heterogenous Data in Clustering of Large Data Sets. *Biocybernetics and Biomedical Engineering* **29(2)**, 9–18, 2009. `https://www.scopus.com/pages/publications/70450253231`

[3] A. P. Nascimento, A. Oliveira, B. M. Faria, R. Pimenta, M. Vieira, C. Prudêncio, and H. Bacelar-Nicolau. Affinity Coefficient for Clustering Autoregressive Moving Average Models. *Computational and Mathematical Methods* **5540143**, 13 pages, 2024. `https://doi.org/10.1155/2024/5540143`

Nonparametric Conditional Survival under Interval Censoring

Luis Meira-Machado[1,2]

[1]University of Minho, Portugal
[2]Centre of Mathematics, Portugal

E-mail addresses: *lmachado@math.uminho.pt*

Estimation of conditional survival functions is challenging when event times are interval censored and covariates are continuous or partially observed. We propose a fully nonparametric framework based on kernel-weighted Turnbull estimators, extending Beran's conditional Kaplan–Meier approach to interval-censored settings. The methodology also addresses conditional survival given an interval-censored intermediate event, avoiding ad hoc imputation strategies. Simulation results show improved performance over midpoint-based approaches, particularly under moderate to severe interval censoring. An application to breast cancer data illustrates the practical relevance of the proposed methods.

Keywords

Conditional survival, interval censoring, kernel smoothing, Turnbull estimator, nonparametric estimation.

Conditional survival analysis plays a central role in time-to-event studies, particularly in medical and epidemiological applications. While regression-based approaches, such as the Cox proportional hazards model [3], are widely used under right censoring, they rely on structural assumptions that may be restrictive in practice.

In many applications, event times are not exactly observed but are only known to lie within inspection intervals, leading to interval-censored data [5]. Although the Turnbull estimator provides a nonparametric solution for marginal survival estimation in this setting [2], extending such methods to conditional survival remains challenging, especially in the presence of continuous covariates or partially observed event times. Classical nonparametric approaches for conditional survival, such as Beran's estimator [4], are not directly applicable in this context.

In this work, we develop a unified nonparametric framework for conditional survival estimation under interval censoring. The proposed approach combines kernel smoothing with Turnbull-type estimators, yielding a flexible method that avoids parametric assumptions. Additionally, we address the problem of estimating survival probabilities conditional on the occurrence of an intermediate event within a time interval, a setting commonly encountered in longitudinal follow-up studies and multi-state processes [6, 7].

Acknowledgements: This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Program Contract of the Centre of Mathematics of the University of Minho (CMAT/UM), UID/00013/2025 (DOI: 10.54499/UID/00013/2025), and by the research project 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481, 1958.

[2] B. W. Turnbull. The empirical distribution function with arbitrarily censored data. *J. Roy. Statist. Soc. Ser. B* **38**, 290–295, 1976.

[3] D. R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187–220, 1972.

[4] R. Beran. Nonparametric regression with censored survival data. Technical Report, University of California, Berkeley, 1981.

[5] J. Sun. *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer, 2006.

[6] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer, 1993.

[7] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen. Multi-state models for the analysis of time to event data. *Stat. Methods Med. Res.* **18**, 195–222, 2009.

The impact of congested periods on high-speed distances in elite football - a statistical analysis of fixture congestion

Luísa Novais[1,2], Paulo Barreira[3], Pedro Antunes[3,4], Afonso Baptista[3,4], João Pedro Araújo[3] and Francisco Tavares[3]

[1]Department of Mathematics, University of Minho, Portugal
[2]Mathematics Center of the University of Porto, Portugal
[3]Medical and Performance Department, Sporting Clube de Portugal, Portugal
[4]Faculty of Human Kinetics, University of Lisbon, Portugal

E-mail address: *lnovais@math.uminho.pt*

Football calendars are becoming more demanding for football players, leading to less time to recover between matches. The physiological need for adequate recovery supports the study of the potential detrimental effects of congested periods on players' health and performance. The aim of this study is to compare the performance of high-speed distances during consecutive matches of congested and non-congested fixture periods in professional football players.

Keywords

data analysis, mixed models, physical performance, football.

Straight sprinting is the most frequent action in goal situations. Therefore, the capacity of the players to perform high-intense velocity efforts is of capital relevance. As such, adequate recovery is essential to maintain performance, particularly in high-speed actions critical to match outcomes. Despite the rationale associated with performance decrements during congested periods, the impact on high-velocity running efforts in football remains unclear.

This study analyses the effect of fixture congestion on external load performance using a longitudinal modelling framework. Data were collected from two competitive seasons of a portuguese professional football team, where the players were observed over sequences of 2 to 5 consecutive matches. Congested periods were defined as inter-match intervals of less than 5 days, while non-congested periods corresponded to intervals between 5 and 8 days. Only observations with a minimum of 70 minutes played were included to ensure comparability.

Inference focused on estimating the effect of congestion on performance trajectories across consecutive matches, accounting for within-player correlation and between-player variability. Model-based comparisons were conducted both within and between periods, enabling a formal assessment of performance differences under varying temporal constraints.

Acknowledgements: The author was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project with reference UID/00144/2025 and associated DOI given by `https://doi.org/10.54499/UID/00144/2025`.

References

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

[2] H. Wiig, T. Raastad, L.S. Luteberget, I. Ims, and M. Spencer. External load variables affect recovery markers up to 72 h after semi-professional football matches. *Frontiers in Physiology* **10: 689**, 2019.

[3] A. Gualtieri, E. Rampinini, R. Sassi, and M. Beato. Workload monitoring in top-level soccer players during congested fixture periods. *International Journal of Sports Medicine* **41: 677–681**, 2020.

Model-Based Clustering for Count Time Series: an Athlete Profiling Application

Luís Sousa[1,3] Magda Monteiro[2,3] and Isabel Pereira[1,3]

[1]DMat – Mathematics Department, University of Aveiro, Portugal
[2]ESTGA – Águeda School of Technology and Management, University of Aveiro, Portugal
[3]CIDMA – Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal

E-mail addresses: *luissousa2@ua.pt; msvm@ua.pt;isabel.pereira@ua.pt*

Clustering discrete-valued time series according to their temporal dynamics is challenging because distance-based methods are often difficult to apply and lack effectiveness in this context. We propose a flexible model-based clustering framework for count time series based on finite mixtures of INAR-type models, where each mixture component corresponds to a distinct data-generating mechanism and thus defines a cluster. Extending existing approaches, we allow for arbitrary autoregressive orders, multiple thinning operators, and a wide range of innovation distributions, resulting in a highly general class of models, allowing, for example, varying levels of dispersion and zero inflation. The finite mixture representation allows for likelihood-based inference and the use of an EM algorithm for clustering. The effectiveness of the methodology is demonstrated through a comprehensive simulation study and the athlete profiling of a long distance running.

Keywords

INAR models, EM algorithm, finite mixture, clustering of time series.

1 Introduction

Clustering time series characterized by small counts and distinct temporal dynamics, where continuous data models are ineffective can be challeging. By focusing on discrete observations like disease cases or daily purchases, the study emphasizes the need for specialized methods for non-negative integers. It introduces a framework designed to group these series based on their unique generative processes rather than simple distance metrics[1]. This work introduces a model-based clustering framework designed specifically for count time series [2], where data consist of non-negative integers, possible with zero inflation or over dispersion.

2 Methodology

Traditional distance-based methods often fail to capture the underlying generative processes, especially for discrete data. Hence, it is proposed a finite mixture model approach where each component represents a distinct time series model. While previous research has

focused on continuous data or limited INAR(1) structures, this work extends the methodology significantly. It allows for arbitrary lag orders (p), multiple thinning operators, and diverse innovation distributions. This generalization accommodates specific data traits like underdispersion and INARCH-type dynamics. By utilizing a flexible family of INAR-type models, the authors derive conditional distributions necessary for likelihood-based inference. This structure enables the implementation of a standard EM algorithm to identify optimal clusters. The resulting framework is generic, highly flexible, and applicable to various fields such as epidemiology and retail. Overall, it provides a robust statistical tool for profiling athletes or subjects based on temporal count patterns. The performance were assessed through a simulation study where three scenarios were defined based on the degree of similarity among the parameters of the INAR(p) processes.

3 Application

The methodology was applied to data from 186 ultramarathoners participating in the 2012 World Championship 24-hour race to identify performance patterns. The primary metric is the number of laps completed per hour, derived from original cumulative records. Only athletes demonstrating continuous effort were selected, as the analysis focuses strictly on modeling pacing dynamics regardless of demographic factors like age or gender.

The analysis evaluated a range of 2 to 6 clusters using INAR(p) models with orders $p \in \{3, 4, 5\}$. Two competitive specifications were tested: negative binomial thinning (nb-poi) and binomial thinning (bin-poi), both with Poisson innovations. The optimal configuration, identified by the lowest BIC and ICL values, was the three-cluster INAR(3) model with binomial thinning and Poisson innovations. This selection is consistent with the data's lack of overdispersion, which favors binomial thinning over the negative binomial alternative.

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

[2] C. H. Weiß. *An Introduction to Discrete-Valued Time Series*. John Wiley & Sons, Ltd., 2018.

On Parameter Estimation in Linear State-Space Models: A Double-Iterated GMM Framework

Marco Costa[1,2] and Magda Monteiro[1,2]

[1]ESTGA – Águeda School of Technology and Management, University of Aveiro, Portugal
[2]CIDMA – Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal

E-mail addresses: *marco@ua.pt; msvm@ua.pt*

State-space models are widely used to represent latent dynamic systems, but parameter estimation may become difficult under distributional misspecification or numerical instability. This work proposes a double-iterated Generalized Method of Moments (GMM) estimator for linear Gaussian state-space models based on Kalman prediction errors and iterative corrections. The approach aims to reduce the impact of parameter estimation errors on filtering and forecasting performance while relaxing strict Gaussian maximum likelihood assumptions. The methodology is illustrated using Portuguese unemployment data.

Keywords

forecasting accuracy, generalized method of moments, Kalman filter, parameter estimation state-space models.

4 Introduction

State-space models are widely used to analyse dynamic systems through latent state and observation equations. However, parameter estimation in these models is often a difficult task, particularly due to numerical instability, convergence problems and the possibility of obtaining estimates outside the admissible parameter space, [2]. These difficulties may compromise filtering and forecasting performance, especially in practical applications involving short samples or highly persistent dynamics, [1]. This work proposes a double-iterated Generalized Method of Moments estimator (GMM2i) for linear state-space models. The methodology combines moment conditions derived from Kalman filter prediction errors with an iterative correction mechanism designed to improve parameter estimation stability and forecasting accuracy.

5 Methodology

Consider a linear Gaussian state-space model composed of a state equation and an observation equation. The latent process evolves according to $\beta_t = \Phi_t\beta_{t-1}+\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$, while the observations satisfy $Y_t = H_t\beta_t + e_t, \quad e_t \sim \mathcal{N}(0, \Sigma_e)$. Here, β_t denotes the latent state vector, Y_t the observed variables, and Φ_t and H_t the transition and observation matrices, respectively. The disturbances are assumed mutually uncorrelated across time. In the stationary specification considered in this work, $\beta_t = \mu + \Phi(\beta_{t-1} - \mu) + \varepsilon_t$, where

μ is the stationary mean vector and stationarity is ensured when the eigenvalues of Φ satisfy $|\lambda_i| < 1$. The Kalman filter provides recursive predictions and filtered estimates of the latent states. Since these quantities depend on unknown parameters, estimation errors may propagate through the filtering recursions and deteriorate forecasting performance. The proposed methodology derives moment conditions from first-order approximations of Kalman prediction errors and incorporates them into a double-iterated Generalized Method of Moments framework (GMM2i). The estimation procedure iteratively updates the parameter vector using corrected moment conditions and weighting matrices until convergence. A hybrid estimator (h-ML) is also proposed, combining GMM corrections for mean and autoregressive parameters with Gaussian maximum likelihood estimation for variance components. A simulation study compared the proposed methods with standard maximum likelihood estimation regarding parameter estimation and forecasting performance.

6 Application

The methodology was applied to Portuguese unemployment data from 2001–2023, analysing the relationship between registered unemployed individuals and unemployment benefit recipients. Three estimation procedures were compared: Gaussian maximum likelihood (ML), the hybrid estimator (h-ML) and the proposed GMM2i estimator. Forecasting performance was evaluated using RMSE, MSSE and MAD measures. The results showed that the h-ML estimator achieved the lowest forecasting errors in most cases, while the proposed methods demonstrated good numerical stability and competitive predictive performance.

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] K. Newman, et al. State-space models for ecological time-series data: Practical model-fitting. *Methods Ecol Evol* **14: 26-42**, 2023.

[2] F. C. Pereira, M. Costa and A. M. Gonçalves. A bootstrap-enhanced fisher scoring algorithm for parameter estimation in state-space models. *Comput Stat* **41: 65**, 2026.

Calibrated Discrepancy Bands for Checking the Markov Property in Multi-State Models

Marta Azevedo[1], Luís Meira-Machado[1]

[1]Centre of Mathematics, Universidade do Minho, Braga, Portugal

E-mail addresses: *marta.vasconcelos4@gmail.com; lmachado@math.uminho.pt*

We propose a calibrated discrepancy framework for checking the Markov property in multi-state models. It compares Aalen–Johansen and landmark Aalen–Johansen estimators through a studentized discrepancy process, uses complementary KS- and L2-type functionals, and provides a bootstrap-calibrated simultaneous confidence band. The method supports both local and multi-landmark inference. Simulations and colon cancer data show good calibration, power, and interpretable time-localized departures.

Keywords

multi-state model, Markov property, landmarking, Aalen–Johansen estimator, simultaneous confidence band.

Markov multi-state models are widely used for prognosis and dynamic prediction, but the Markov assumption may fail when transition risks depend on duration in the current state or on earlier history. Existing checks include regression-based approaches and discrepancy-based comparisons between Markov and non-Markov transition probability estimators [1, 2]. We present a discrepancy-based diagnostic framework that turns the comparison between the Aalen–Johansen (AJ) and landmark Aalen–Johansen (LM-AJ) estimators into a calibrated inferential tool.

For a fixed landmark time and transition, we define the discrepancy process as the difference between LM-AJ and AJ transition probability estimates over time. Under the Markov assumption, both estimators target the same quantity, so the discrepancy should fluctuate around zero. To detect departures, we use two complementary studentized functionals: a Kolmogorov–Smirnov-type supremum statistic for localized deviations and an integrated L^2-type statistic for sustained differences. We also construct a simultaneous confidence band for the full discrepancy curve, so the graphical diagnostic supports formal inference and identifies the time regions where departures are most pronounced. This is especially useful in applications where scalar summaries may miss brief but relevant non-Markov behavior.

Calibration is performed by a subject-level martingale multiplier (wild) bootstrap [3]. The same multipliers are reused across AJ and LM-AJ perturbations to preserve their joint dependence. The framework also extends naturally to multiple landmark times through a global maximum statistic that aggregates evidence while controlling multiplicity by construction. To improve finite-sample behavior in late follow-up, we use a data-adaptive truncation rule that restricts inference to stable time windows when landmark risk sets become small and tail variability becomes dominant.

Simulation results in a progressive four-state model show near-nominal type-I error and good power under duration- and history-dependent alternatives. The KS statistic is more sensitive to sharp local departures, while the L^2 statistic captures broader sustained deviations. In a colon cancer application, the proposed band localizes time regions

where Markov modelling appears inadequate and yields conclusions consistent with established discrepancy-based diagnostics [4, 1]. Overall, the method provides a practical bridge between visual discrepancy plots and formally calibrated inference for Markov property assessment, while preserving direct interpretability for applied researchers.

Acknowledgements: This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Program Contract of the Centre of Mathematics of the University of Minho (CMAT/UM), UID/00013/2025 (DOI: 10.54499/UID/00013/2025), and by the research project 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] A. C. Titman and H. Putter. General tests of the Markov property in multi-state models. *Biostatistics* **23**(2):380–396, 2022.

[2] G. Soutinho and L. Meira-Machado. Methods for checking the Markov condition in multi-state survival data. *Computational Statistics* **37**(2):751–780, 2022.

[3] T. Bluhmki, C. Schmoor, D. Dobler, M. Pauly, J. Finke, M. Schumacher, and J. Beyersmann. A wild bootstrap approach for the Aalen–Johansen estimator. *Biometrics* **74**(3):977–985, 2018.

[4] H. Putter and C. Spitoni. Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator. *Statistical Methods in Medical Research* **27**(7):2081–2092, 2018.

Efficient Clustering of Survival Curves: A k-Means and Log-Rank Approach

Marta Sestelo[1,2], Nora M. Villanueva[2] and Luís Meira-Machado[3]

[1]Galician Center for Mathematical Research and Technology (CITMAga), Santiago de Compostela, Spain
[2]Universidade de Vigo, Department of Statistics and O.R. & SiDOR Group, Vigo, Spain
[3]Centre of Mathematics, Universidade do Minho, Braga, Portugal

E-mail addresses: *sestelo@uvigo.gal; nmvillanueva@uvigo.gal; lmachado@math.uminho.pt*

Survival analysis is used for studying time-to-event data. Traditional bootstrap clustering of curves is effective but imposes heavy computational demands, limiting scalability in large datasets. We propose a novel method leveraging k-means and the log-rank test to efficiently identify and cluster multiple survival curves. Eliminating intensive resampling substantially reduces computation time while strictly preserving statistical validity. Simulations demonstrate performance matching bootstrap techniques efficiently, providing a practical and scalable model.

Keywords

survival analysis, clustering, hypothesis testing.

Survival analysis is a fundamental statistical approach used in medical, biological, and epidemiological research to study time-to-event data. A key aspect of survival analysis is the estimation of the survival function, particularly in the presence of right-censored data. The most commonly used non-parametric estimator in such cases is the Kaplan-Meier estimator [2], which provides an empirical estimate of the survival probability over time.

Several statistical tests have been developed to compare survival curves, with the log-rank test being the most widely used due to its efficiency in detecting differences in survival distributions [3]. The log-rank test evaluates whether survival functions differ significantly between groups under the null hypothesis that they are identical.

When the null hypothesis is rejected, indicating significant differences between survival curves, multiple pairwise comparisons may be required. In R, the `survminer` package provides the function `pairwise_survdiff`, which allows for two-by-two survival curve comparisons. However, as the number of groups increases, multiple testing becomes a challenge, necessitating corrections to control the family-wise error rate.

Commonly used multiple testing correction methods, including the Bonferroni approach [1] alongside others like Holm or Hochberg, are implemented in the `p.adjust` function of the R `stats` package. For example, when comparing survival curves among 15 groups, such as in a breast cancer dataset analyzed in our application, the number of pairwise comparisons reaches $\binom{15}{2} = 105$. Conducting multiple hypothesis tests at this scale makes interpretation complex and computationally intensive.

While multiple pairwise comparisons allow for survival curve differentiation, they do not inherently provide a clustering mechanism to group similar curves. The 2019 study by Villanueva et al. [4] introduced a methodology to address this issue by clustering survival curves based on resampling techniques. However, the method is computationally demanding due to the extensive use of bootstrap resampling. Although effective, this

method introduces substantial computational costs, making it impractical for large-scale applications.

A significant drawback of bootstrap-based clustering methods is the necessity of multiple resampling iterations, where each step involves estimating survival functions, computing test statistics, and evaluating clustering results. This process can be computationally prohibitive, especially when analyzing high-dimensional survival data. To address these limitations, our study proposes an alternative method that employs the log-rank test as a clustering criterion, thereby significantly reducing computational overhead while maintaining robust statistical properties. Unlike bootstrap-based approaches, it provides a direct and computationally efficient means of determining whether two or more survival curves differ significantly. Our method builds upon this principle, systematically forming clusters by iteratively merging survival curves with non-significant log-rank test results, ensuring an optimal partitioning of survival groups.

The main contributions of this study are as follows. We introduce a novel clustering approach based on the log-rank test, which removes the need for computationally intensive resampling techniques. Through simulation studies, we demonstrate that our method produces clustering results comparable to those obtained with bootstrap-based approaches, while significantly improving computational efficiency. Additionally, we provide a practical framework for applying this method in survival analysis, making it particularly suitable for large datasets commonly encountered in medical and epidemiological research.

Acknowledgements: The authors acknowledge financial support by the Spanish Ministry of Science, Innovation and Universities through project PID2023-148811NB-I00 (funded by (AEI/FEDER, UE) and by the Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013/2025: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association* **56(293)**, 52–64, 1961.

[2] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53(282)**, 457–481, 1958.

[3] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50(3)**, 163–170, 1966.

[4] N. M. Villanueva, M. Sestelo, and L. Meira-Machado. A Method for Determining Groups in Multiple Survival Curves. *Statistics in Medicine* **38**, 366–377, 2019.

SARIMA and STARMA modelling of Atlantic ocean temperature in regions of Portugal and Cape Verde

Arciolindo Pinheiro[1], M. Rosário Ramos[2,4], Elisabete Carolino[3,5],

[1] Universidade Aberta, Portugal
[2] LE@D, Universidade Aberta and CEG
[3] H&TRC – Health & Technology Research Center, ESSL-Escola Superior de Saúde, Instituto Politécnico de Lisboa, Portugal
[4] CEAUL, Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[5] ISAMB, Instituto de Saúde Ambiental da Faculdade de Medicina da Universidade de Lisboa, Portugal

E-mail addresses: *arciolindo@gmail.com; mariar.ramos@uab.pt; elisabete.carolino@essl.ipl.pt*

Modelling seawater enables the identification of temporal patterns and the forecasting of future trends, providing support for the development of adaptation strategies. This research analyses time series of monthly mean sea temperature in four Atlantic regions — Nazaré, Faro, São Miguel, and Praia — between 1968 and 2024, using the european EMODnet data at two depths. SARIMA models showed strong performance in capturing trends and seasonality and in forecasting. To account for spatial dependence, a STARMA model was applied. The results confirm the progressive warming of Atlantic waters and demonstrate the effectiveness of the applied models.

Keywords

Sea Temperature, Spatio-Temporal Modelling, Forecasting, Trend, Climate change,

The study of seawater temperature is of major importance in the context of climate change and ocean warming. Rising sea temperatures are associated with significant impacts on marine ecosystems, including biodiversity loss and shifts in species distribution. This study analyses the evolution of monthly mean sea temperature in the Atlantic Ocean between January 1968 and June 2024 across four regions: Nazaré, Faro and São Miguel (Portugal), and Praia (Cape Verde). It is also of interest to compare the evolution of sea temperature at different depths. Accordingly, time series at two distinct depths, namely 1 m (surface) and 5 m, were considered using data obtained from the European Marine Observation and Data Network (EMODnet).
Several time series methodologies were applied to model the univariate sea temperature series, including Holt–Winters and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Among these, SARIMA demonstrated the best performance, robustly capturing both trend and seasonality while producing low forecasting errors for the test period from July 2023 to June 2024. Morover the test for trend show a significant increase in sea temperature along time. Another approach incorporated the spatial dimension by accounting for correlations both over time and across geographical or networked locations. A Space-Time Autoregressive Moving Average (STARMA) model, which extends ARMA/ARIMA-type models, was fitted to the sea temperature time series at 1 m depth

(surface level), based on a spatial weight matrix derived from geodesic distances. The final model revealed dominant temporal dependence and moderate spatial influence, showing good performance and residuals close to white noise. STARMA forecasts preserved the thermal hierarchy among regions and reinforced the structural consistency of the univariate results. Overall, the study confirms the progressive warming of Atlantic waters and demonstrates that SARIMA and STARMA models are effective tools for understanding and predicting regional thermal variability. These findings provide relevant scientific support for ocean monitoring and climate adaptation strategies in vulnerable coastal and marine regions.

Acknowledgements: This work is partially funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, within the scope of the projects UID/4372/2025 and UID/PRR/04372/2025.

References

[1] Pfeifer, P. E., Deutsch, S. J. Seasonal space-time ARIMA modelling. *Geographical Analysis* **13**, 117–133, 1981.

[2] Bai, Z. Prediction of ocean temperature based on ARIMA model. *Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation*, 114–121, 2024. https://doi.org/10.54254/2753-8818/36/20240530

[3] Bilgili, M., Pınar, E., Durhasan, T. Global monthly sea surface temperature forecasting using the SARIMA, LSTM, and GRU models. *Earth Science Informatics*, **18**, Article 10.2025 https://doi.org/10.1007/s12145-024-01585-z

[4] Subba Rao, T., and Antunes, A. M. C. Spatio-temporal modelling of temperature time series: A comparative study. R. *Time series analysis and applications to geophysical systems*,Brillinger et al.(Eds.), Springer-Verlag. 2004

On the challenges of building a year-round climatic health indicator: methodological choices and trade-offs in the GATO-YR framework

Paulo Nogueira[1]

[1]Laboratório Associado TERRA, ISAMB – Instituto de Saúde Ambiental, Faculdade de Medicina, Universidade de Lisboa, Portugal

E-mail addresses: *pnogueira@medicina.ulisboa.pt*

Cumulative thermal-stress indices are central to environmental epidemiology, yet most are built for isolated extreme events in a single season. We present GATO-YR, a year-round, symmetric (heat and cold), hierarchical band-categorical extension of the Generalized Accumulated Thermal Overload, and show how five methodological decisions materially shape epidemiological inference, using 19 years of Portuguese hospital admissions and 14 years of birth records.

Keywords

cumulative thermal stress, GATO-YR, year-round indicator, categorical exposure, environmental epidemiology.

The Generalized Accumulated Thermal Overload (GATO) is a recursive heat-burden index that accumulates daily temperature exceedance over a percentile threshold while a run-length counter tracks the current heat episode. It extends the original Portuguese ÍCARO heat-warning methodology [1]; the English acronym GATO was first published internationally by Nogueira and Paixão [2], and its weekly-threshold form (GATO IV) was shown to outperform the Excess Heat Factor for Lisbon mortality [3]. GATO had, however, been restricted to the summer season, with thresholds built from summer-only data, and no symmetric cold-side accumulator existed. GATO-YR removes both limitations: weekly percentile thresholds (τ) are computed for all 53 ISO weeks and cyclically smoothed, separately for each of the 278 mainland Portuguese municipalities, and a symmetric cold-side index (CGATO) is defined on minimum temperature and low percentiles.

Building the year-round index exposed five methodological decisions, each of which measurably affects epidemiological inference: (i) the cyclic-smoothing scheme for τ, with two operationally equivalent variants; (ii) the operationalisation of exposure — a continuous smooth versus a hierarchical band-categorical cascade in which each day is classified into the most extreme percentile band it crosses; (iii) the climatological reference period, where the WMO current normal (1991–2020) empirically outperforms the classical 1961–1990 normal for contemporary data (heat F-statistic 15.66 vs 9.42), because historical thresholds applied to a warmed period dilute the "extreme" bands with merely-moderate days; (iv) seasonality control, where exhaustive factor(month) + factor(year) liberates about 2.6 times more genuine heat signal and removes some 67% of spurious cold signal relative to a cyclic spline; and (v) the operationalisation of cold exposure, where only the categorical cascade — not distributed-lag models on continuous temperature — detects a +10% preterm-birth risk at the most extreme cumulative-cold band.

The framework was developed and stress-tested on E-OBS gridded temperature [4] linked to 19 years of Portuguese hospital admissions (1.93 million municipality-day observations, 27 Major Diagnostic Categories) and 14 years of municipal birth records. To

our knowledge, GATO-YR is the first index to span heat and cold, acute and cumulative, across every season in one coherent construct — calibrated and validated for the Portuguese reality. This talk presents the framework, the five decisions, and the evidence for resolving each, drawing on three companion application papers.

References

[1] P. Nogueira, B. Nunes, C. M. Dias, and J. M. Falcão. Um sistema de vigilância e alerta de ondas de calor com efeitos na mortalidade: o índice Ícaro. *Rev. Port. Saúde Pública* **Vol. Temático 1**, 79–84, 1999.

[2] P. Nogueira and E. Paixão. Models for mortality associated with heatwaves: update of the Portuguese heat health warning system. *Int. J. Climatol.* **28(4)**, 545–562, 2008.

[3] L. Morais, A. Lopes, and P. Nogueira. Which heatwave measure has higher predictive power to prevent health risks related to heat: EHF or GATO IV? Evidence from modelling Lisbon mortality data from 1980 to 2016. *Weather Clim. Extrem.* **30**, 100287, 2020.

[4] R. C. Cornes, G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones. An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.* **123(17)**, 9391–9409, 2018.

Modeling Volatility in Count Time Series: The Zero-Inflated Generalized Poisson INGARCH model

Rafaela Rodrigues[1,2], Valdério Reisen[2,3] and Helena Mouriño[1,2]

[1]Faculdade de Ciências da Universidade de Lisboa, Portugal
[2]CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3]Universidade Federal do Espírito Santo, Vitória, Brazil

E-mail addresses: *rafaela.cvg.rodrigues@gmail.com; mhnunes@ciencias.ulisboa.pt; valderioanselmoreisen@gmail.com*

The ZIGP-INGARCH model is proposed to handle count time series with overdispersion and zero inflation, for which traditional models fall short. The asymptotic properties of its estimators are established and validated through Monte Carlo simulations. The results identified specific conditions required for stable parameter estimation, namely sample size and overdispersion levels. The model was applied to a real dataset of weekly Pseudo-nitzschia counts. The results showed that the model is flexible, but simpler options can sometimes perform better.

Keywords

Time series of counts, Overdispersion, Zero-inflation, ZIGP-INGARCH, *Pseudo-nitzschia*.

Acknowledgements: This work is partially funded by FCT/Mobility/1395440540/2024-25, 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` and by the European Union-NextGenerationEU through the project UID/PRR/00006/2025, DOI: `https://doi.org/10.54499/UID/PRR/00006/2025`.

References

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

[2] S. Palma, H. Mourino, A. Silva, M. I. Barão, and M. T. Moita. Can Pseudo-nitzschia blooms be modeled by coastal upwelling in Lisbon Bay?. *Harmful Algae* **9**, 294–303, 2010.

[3] C. W. S. Chen and S. Lee. Generalized Poisson autoregressive models for time series of counts. *Computational Statistics & Data Analysis* **99**, 51–67, 2016.

[4] S. Lee, Y. Lee, and C. W. S. Chen. Parameter change test for zero-inflated generalized Poisson autoregressive models. *Statistics* **50**, 540–557, 2016.

Optimal cutoff selection under scale mixtures of skew-normal distributions

Renato de Paula[1,2], Helena Mouriño[1,2] and Tiago Dias Domingues[1,2]

[1]Departamento de Ciências Matemáticas, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]Centro de Estatística e Aplicações (CEAUL), Faculdade de Ciências, Universidade de Lisboa, Portugal

E-mail addresses: *rrpaula@ciencias.ulisboa.pt; mhnunes@ciencias.ulisboa.pt; tmdomingues@ciencias.ulisboa.pt*

We develop a unified framework for ROC analysis and optimal cutoff selection under scale mixtures of skew-normal (SMSN) distributions. The optimal threshold minimises a weighted misclassification risk incorporating disease prevalence and asymmetric costs, characterised by a likelihood-ratio equation. Under monotone likelihood ratio, we establish existence, uniqueness, and global optimality of the cutoff, and derive consistency, asymptotic normality, and a plug-in variance estimator via the delta method. Monte Carlo simulations and an application to SARS-CoV-2 serological data illustrate the methodology.

Keywords

ROC curve, optimal cutoff, scale mixtures of skew-normal, decision theory, asymptotic inference.

Selecting a diagnostic threshold for a continuous biomarker is a routine but consequential problem. The widely used Youden index assumes equal misclassification costs and equal disease prevalence, assumptions that are often unrealistic. In addition, serological and immunological biomarkers frequently exhibit skewness and heavy tails, limiting the adequacy of Gaussian ROC models.

We develop a parametric framework for ROC analysis and optimal cutoff selection under the family of scale mixtures of skew-normal (SMSN) distributions, including the skew-normal and skew-t models. The ROC curve and AUC are estimated by plug-in maximum likelihood from separate-group fits.

The optimal cutoff is defined as the minimiser of a weighted misclassification risk that incorporates disease prevalence and asymmetric costs, leading to a likelihood-ratio equation that generalises the Youden criterion. Under a monotone likelihood ratio condition, we establish existence, uniqueness, and global optimality of the cutoff. We further derive consistency, asymptotic normality, and a closed-form plug-in variance estimator for the cutoff estimator, obtained through the implicit function theorem and the multivariate delta method. A key term in the variance is the local slope of the estimating equation at the optimum, which we interpret as a local identifiability diagnostic.

Monte Carlo simulations under skew-normal and skew-t scenarios confirm the accuracy of the asymptotic approximation and the nominal coverage of Wald confidence intervals. An application to SARS-CoV-2 IgG serological data shows that the proposed cutoff can differ substantially from the Youden threshold and may reduce estimated misclassification risk by up to 63% under asymmetric decision settings.

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

Finite volume solution of the incompressible Navier-Stokes equations with very high-order accuracy: a comparison between primitive-variable and streamfunction-vorticity formulations

Ricardo Costa[1], Stéphane Clain[2], Gaspar J. Machado[3], and João M. Nóbrega[1]

[1]Institute for Polymers and Composites, University of Minho, 4804-058 Guimarães, Portugal
[2]Centre of Mathematics, University of Coimbra, 3000-143 Coimbra, Portugal
[3]Centre of Mathematics, University of Minho, 4800-058 Guimarães, Portugal

E-mail addresses: *rcosta@dep.uminho.pt*

The numerical solution of the incompressible Navier-Stokes equations is commonly performed using classical primitive-variable formulations, in which the pressure-velocity coupling plays a central role in the overall complexity and efficiency of the numerical method. Alternatively, streamfunction-vorticity formulations eliminate the pressure variable and provide a natural framework for the analysis of vortex dynamics, although the accurate treatment of vorticity boundary conditions remains a significant challenge, particularly in curved geometries. This work presents a comparative study between primitive-variable and streamfunction-vorticity formulations within a very high-order accurate finite volume method for incompressible flows. Special attention is given to the mathematical derivation and numerical treatment of boundary conditions for the streamfunction and vorticity in arbitrary curved geometries. For both formulations, an appropriate boundary discretisation strategy on unstructured polygonal meshes is employed, allowing very accurate approximations on curved domains without requiring curved mesh elements. Several benchmark two-dimensional incompressible flow problems, including the semi-elliptical lid-driven cavity problem, are considered to assess the behaviour of both formulations in terms of numerical accuracy and computational performance. The results highlight the respective advantages and limitations of each approach, illustrating the potential of vorticity-based formulations as a competitive alternative for very high-order accurate simulations of incompressible flows in complex geometries.

Keywords

Incompressible fluid flows, Streamfunction-vorticity formulation, Finite volume method, Very high-order convergence, Arbitrary curved boundaries, Immersed boundaries.

References

[1] R. Costa, S. Clain, G.J. Machado, and J.M. Nóbrega. Very high-order accurate wall vorticity treatment on curved boundaries with polygonal meshes for incompressible vorticity equations. *Computers & Mathematics with Applications* **214**, 94-134, 2026. DOI: https://doi.org/10.1016/j.camwa.2026.04.004.

A partially linear model for the uterine artery pulsatility index: estimation and model checking

Rui Costa-Miranda[1] and Rita Gaio[1,2]

[1]Centre of Mathematics of the University of Porto, Portugal
[2]Department of Mathematics of the Faculty of Sciences of the University of Porto, Portugal

E-mail addresses: *up201804962@up.pt; argaio@fc.up.pt*

A partially linear model including a nonparametric effect of time and a linear effect of age group and parity, together with their interaction, is used to model uterine artery pulsatility index cross-sectional data. Goodness-of-fit is assessed by an integrated conditional moments test, implemented by a new estimation-robust approach. The results are compared with those from a fully parametric model and suggest the need for the nonparametric term.

Keywords

fast-bootstrap, gaussian processes, goodness-of-fit, partially linear model.

For $i = 1, \ldots, n$, consider the crude errors of a partially linear model setting, given by $\varepsilon(A_i, S_i, \beta, m) = Y_i - (A_i^T\beta + m(S_i))$, where 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) is equivalent to an infinite number of projected unconditional moment restrictions, so that the test can be stated as an integrated approach over residual-marked empirical Gaussian processes (GP) [3]. The test statistic is

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

where $R(X, \beta, m) := E\left[w(A, S)\varepsilon(A, S, \beta, m)\right]$, $X = (A, S)^t$, w is any measurable function, and K is the estimated covariance matrix of the GP. A refinement of this methodology was recently developed for (finite-dimensional) parametric models [1]. The key innovation is projecting X onto the orthocomplement of the conditional moment score subspace. The test statistic considers the estimated covariance-kernel matrix of the projected GP. This ensures robustness to the estimation, allowing for fast-bootstrap procedures and efficient computations. Using a local polynomial estimation-based approach, we extend this test procedure to accomodate the infinite-dimensional parameter $m(\cdot)$ of a PLM.

To illustrate, data regarding the uterine artery pulsatility index [2] are considered. For a better understanding of the cyclic changes in uterine perfusion, cross-sectional data were collected from 1668 women from different age groups (*1:* 18–26 y.o., *2:* 27–35 y.o., *3:* 36–50 y.o.), with measurements taken at various days of the menstrual cycle (from 1 to 34). Variables such as the day of the menstrual cycle (S), age group (A_j, binary, $j = 1, 2, 3$) and parity status (B, binary) were considered for modelling the log-transformed pulsatility index: $E[\log(Y_i) \mid A_i, B_i, S_i] = \beta_0 + \beta_{1j}B_i + \beta_{2j}(S/10) + \beta_{3j}(S/10)^2 + \beta_4(S/10)^3$, for age j [2]. For comparison, we propose the fit of a PLM of the form $E[\log(Y_i) \mid A_i, B_i, S_i] = m_{A_i,B_i}(S_i)$, where an interaction between variables A, B and the nonparametric curve is considered. The estimation of the curves is done separately for a common bandwith,

determined by the maximum of the bandwidths obtained for each interaction condition by leave-one-out cross-validation.

In Table 1, the results of 999 bootstrap iterations of different tests are presented, for each of the models: the usual GP based procedures are labelled by T_E, considering a Escanciano-type covariance kernel [3], and T_D, for a distance covariance kernel [1]; test statistics $T^{\perp}$ represent the proposed Neyman-orthogonal version of the latter. Results show that the novel orthogonal approach reduces time complexity in the PLM without compromising the conclusion of the test. All tests fail to reject the hypothesized LM, thus validating that the nonparametric effect of the day of the menstrual cycle offers a better fit to understand the behavior of the uterine artery pulsatility index.

Table 2. p-values and execution times obtained for each of the tests, in each model.

	LM				PLM			
Test statistic	T_D	T_E	$T_D^{\perp}$	$T_E^{\perp}$	T_D	T_E	$T_D^{\perp}$	$T_E^{\perp}$
p-value	< 0.001	< 0.001	< 0.001	< 0.001	0.967	0.994	0.865	0.571
time (seconds)	34.7	969.0	51.9	971.1	270.0	510.9	133.7	395.0

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

[2] L. Guedes-Martins, R. Gaio, J. Saraiva, S. Cerdeira, L. Matos, E. Silva, F. Macedo, and H. Almeida. Reference ranges for uterine artery pulsatility index during the menstrual cycle: a cross-sectional study. *PLoS ONE* **10(3)**, e0119103, 2015.

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

Diet Quality, Anemia Status, and Equity in Early-Life Nutrition Transitions in South and Southeast Asia

Sarada Ghosh[1,2]

[1] Department of Statistics, Gurudas College, Kolkata, India

E-mail addresses: *saradaghosha111@gmail.com*

Achieving global targets for reducing malnutrition and anemia depends on effective dietary and health-system interventions[1, 2]. This study examines how dietary diversity among children aged 6–23 months influences anemia outcomes in South and Southeast Asia using Demographic and Health Survey data from 75,619 mother–child dyads. Logistic regression models were applied to assess the relationship between minimum dietary diversity and childhood anemia while accounting for child, maternal, and household characteristics. Children consuming fewer than five of eight WHO-recommended food groups had significantly higher odds of anemia (OR: 1.12; 95% CI: 1.05–1.19) compared with those meeting minimum dietary diversity standards. Female children were slightly less likely to be anemic, whereas children aged 12–23 months showed greater vulnerability than those aged 6–11 months. Maternal anemia and undernutrition substantially increased childhood anemia risk, while higher maternal education and improved sanitation were associated with lower risk. The findings demonstrate that constrained diets and adverse social determinants continue to hinder equitable nutrition transitions in the region. The study emphasizes the need for spatially targeted interventions that improve dietary diversity, maternal nutrition, education, and sanitation to support more equitable child-health outcomes and sustainable nutrition transitions [3].

Keywords and phrases: Dietary diversity, Anemia, Odds ratio, Logistic regression.

References

[1] World Health Organization and United Nations Children's Fund. Indicators for assessing infant and young child feeding practices: Definitions and measurement methods. 2021.

[2] World Health Organization. Global Anaemia Estimates, 2021 Edition. 2021.

[3] B. M. Popkin, C. Corvalan, and L. M. Grummer-Strawn. Dynamics of the double burden of malnutrition and the changing nutrition reality. *The Lancet* **395(10217)**, 65–74, 2020.

Meshless Structural method

Stephane Clain[1], Jorge Figueiredo[2]

[1] Center of Mathematics of the FCTUC, Largo D. Dinis, 3000-143 Coimbra, Portugal.
[2] Center of Mathematics of the University of Minho, Campus de Azurém, 4080-058 Guimarães, Portugal.

E-mail addresses: *clain@mat.uc.pt*

We present a new very high-order compact method using a meshless discretization named structural method. The concept lies in providing linear implicit relations between the function and the derivatives of the approximation over stencils. After a brief presentation of the technique, we present some recent numerical results focusing on the order of convergence.

Keywords

Structural method, compact scheme, meshless.

The structural method has been recently introduced as a new way to provide very high-order compact schemes [1, 2, 3, 4]. In the talk we present the extension for a general 2D cloud of points using the classical convection-diffusion-reaction equation as an example:

$$\begin{aligned} &(1)\quad \kappa\Delta\phi(x) + u\cdot\nabla\phi(x) + \lambda\phi(x) = f(x),\\ &(2)\quad \nu_D\phi(x) + \nu_N\nabla\phi(x)\cdot n(x) = g(x), \end{aligned}$$

over a domain Ω and corresponding boundary $\partial\Omega$, respectively, where $x = (x_1, x_2)$, $u = (u_1, u_2)$ and $n = (n_1, n_2)$.

We denote by $\alpha = (\alpha_1, \alpha_2)$ the multi-index and, accordingly, $x^\beta = x_1^{\beta_1}x_2^{\beta_2}$ and $\phi^{(\alpha)}(x) = \partial^\alpha\phi(x)$. Furthermore, let $x_i = (x_{i,1}, x_{i,2})$ be a set of points with $\mathcal{I}$ as the nodes in Ω and $\mathcal{B}$ as the nodes on the boundary $\partial\Omega$. We introduce the approximation $\phi_i^\alpha \approx \phi^{(\alpha)}(x_i)$ and the physical equations read

$$\begin{aligned} &\texttt{PE1}\quad \kappa(\phi_i^{(2,0)} + \phi_i^{(0,2)}) + u_1\phi_i^{(1,0)} + u_2\phi_i^{(0,1)} + \lambda\phi_i^{(0,0)} = f(x_i), \quad i\in\mathcal{I},\\ &\texttt{PE2}\quad \nu_D\phi_i^{(0,0)} + \nu_N(n_1\phi_i^{(1,0)} + n_2\phi_i^{(0,1)}) = g(x_i), \quad i\in\mathcal{B}. \end{aligned}$$

Each node i supports the unknowns ϕ_i^α with $|\alpha|\le 2$ (*i.e.*, up to the second order derivative).

The key idea is to build a set of linear equations, named the structural equations $\texttt{SE}_\ell$, independent of the underlying physics of the problem, of the form

$$\sum_{j\in V_i}\sum_{|\alpha|\le 2} c_{ij}^{\alpha,\ell}\phi_j^\alpha = 0, \quad \ell = 1,\dots,5,$$

where V_i is a stencil centred on node i and $c_{ij}^{\alpha,\ell}$ are the coefficients of the ℓ-th structural equation.

We show the following properties:

- The structural relations are exact for all polynomial functions of total degree up to d.
- The coefficients are obtained from a basis of the null space of a local polynomial consistency matrix, which are very easy to set up.
- We select 5 linearly independent relations such that the total number of equations (Physical + Structural) is equal to the number of unknowns.

As a result, we obtain a sparse matrix of 6×6 blocks for which the solution provides a very accurate approximation for both the solution and the derivatives.

We present the results of some numerical experiments to give evidence of the efficiency of the method as well as its ability to handle compact stencils while providing very accurate solutions. Moreover, we are able to handle any type of boundary (non-polygonal boundary) which gives a strong advantage with respect to the traditional mesh-based method that requires a specific treatment of the boundary to preserve the accuracy.

Acknowledgements: S. Clain and J. Figueiredo acknowledge the financial support of the Portuguese Foundation for Science and Technology (FCT) through a national funding for projects IC&DT with the reference 2023.16854.ICDT. `https://doi.org/10.54499/2023.16854.ICDT`.
S. Clain 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.
The research of J. Figueiredo was financially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) under the scope of the project UID/00013/2025 `https://doi.org/10.54499/UID/00013/2025`, Center of Mathematics of the University of Minho.

References

[1] P. C. Chu and C. Fan. A three-point combined compact difference scheme. *J. Comput. Phys.* **140**, 1998.

[2] S. Clain, R. M. S. Pereira, P. A. Pereira, and D. Lopes. Structural schemes for one dimension stationary equations. *Appl. Math. Comput.* **457**, 2023. DOI: `10.1016/j.amc.2023.128207`

[3] S. Clain, G. J. Machado, and M. T. Malheiro. Compact schemes in time with applications to partial differential equations. *Comput. Math. Appl.* **140**, 2023. DOI: `10.1016/j.camwa.2023.03.011`

[4] S. Clain, M. T. Malheiro, G. J. Machado, and R. Costa. The structural method for Ordinary Differential Equations. `https://arxiv.org/abs/2508.02440`, 2025.

R-Block structural schemes for ordinary differential equations

M. Teresa Malheiro[1], Stéphane Clain[2], Gaspar J. Machado[1] and Ricardo Costa[3]

[1]Centre of Mathematics, University of Minho, Portugal
[2]Center of Mathematics, FCT-University of Coimbra, Portugal
[3]Inst. Polymers and Composites, University of Minho, Portugal

E-mail addresses: *mtm@math.uminho.pt*

We present a compact scheme whose core concept involves decomposing it into two subsystems of equations. The Physical Equations utilise the function and its K derivatives at a node by implementing physical relations. These equations operate locally, with no exchange of information with other nodes,as the physics involved are governed by local operators. The Structural Equations depend on linear relationships between the function and its derivatives across a stencil of R points, which we call a Rblock, establishing complete connections between a node and its neighbours. These relationships are independent of the physics involved since they are established regardless of the specific problem. In this presentation we address in particular the accuracy and stability of these methods.

Keywords

high order numerical methods.

Machine Learning Approaches for Benzene Price Forecasting

Tiago Fernandes[1,2], Sara Martins[1] and Eliana Costa e Silva[1,3]

[1]CIICESI, ESTG, Polytechnic of Porto, Portugal
[2]Supply Chain, Bondalti Chemicals, Estarreja, Portugal
[3]Algoritmi Center, University of Minho, Portugal

E-mail addresses: 8200497@estg.ipp.pt, ssbm@estg.ipp.pt, eos@estg.ipp.pt

Benzene is a key petrochemical feedstock used in plastics, resins, and fibers. Its price depends on oil, aromatics, and global shocks, yet dedicated studies are limited. This work addresses Benzene price forecasting through an integrated framework combining statistical analysis and machine learning models. Weekly data on oil, naphtha, and aromatics are used to capture temporal dependencies, including lag effects. Results show strong autoregressive behaviour and confirm the predictive value of energy-related variables.

Keywords

Benzene price, Time Series Forecasting, Artificial Neural Networks, Random Forest

Benzene is one of the most important feedstocks in the petrochemical industry. It serves as a base material for a wide range of downstream products such as plastics, resins and synthetic fibers [3, 6]. Despite its industrial relevance, dedicated forecasting studies focused specifically on benzene remain relatively underexplored when compared to other petrochemical and energy commodities (such as crude oil or naphtha) [4, 5]. Since benzene is mainly produced as a co-product of refining and cracking processes, its price is mainly influenced by multiple interconnected factors, including crude oil dynamics, aromatics markets, downstream derivatives and global economic disruptions [3, 2, 1].

This study focuses on benzene price forecasting through a data-driven framework combining statistical analysis and machine learning models. Weekly historical data from 2019 to 2025, including crude oil, naphtha, styrene, aromatics markets and international benzene prices, are used to capture temporal patterns and interdependencies.

Seasonality and external shocks, such as the COVID-19 pandemic and the Russia–Ukraine conflict, are also incorporated.

To support forecasting, several exploratory techniques were applied, including correlation analysis, Dynamic Time Warping (DTW), Cross-Correlation Functions (CCF), Autocorrelation Functions (ACF) and Partial Autocorrelation Functions (PACF), enabling the identification of relevant predictors and lag structures.

Predictive models such as Random Forest and Artificial Neural Networks (ANN) were implemented to model nonlinear relationships and temporal dependencies. Multiple ANN architectures were tested to optimize predictive performance while avoiding overfitting. Model evaluation was conducted using walk-forward validation, ensuring robust assessment across different time periods and market conditions.

The analysis highlights the strong autoregressive nature of benzene prices, as well as the importance of energy-related and petrochemical variables in explaining market behavior. Variables associated with crude oil, aromatics markets and downstream petrochemical

products showed significant influence on benzene price dynamics, particularly when combined with appropriate temporal lag structures. The forecasting models also demonstrated that incorporating historical price dependencies together with external market indicators improves the ability to capture both short-term fluctuations and broader market trends across different market conditions.

The forecasting component is conceptually integrated with a procurement optimization model designed to support purchasing and inventory-related decisions in a specific industrial environment. The proposed optimization structure considers operational constraints such as storage capacity, shipment frequency, lead times and procurement costs, enabling forecasted benzene prices to inform procurement planning.

Overall, this work contributes to the application of predictive and prescriptive analytics in petrochemical markets by proposing a structured framework tailored to benzene price forecasting and procurement optimization.

Acknowledgements: This work was supported by national funds through FCT - Fundação para a Ciência e Tecnologia through projects UIDB/04728/2025 and UIDP/04728/2025 (https://doi.org/10.54499/UID/04728/2025). T. Fernandes was supported by Bondalti Chemicals through the provision of industrial and market-related data.

References

[1] J. F. Adolfsen, F. Kuik, E. M. Lis, and T. Schuler. The impact of the war in ukraine on euro area energy markets. *ECB Economic Bulletin* **(4)**, 2022.

[2] Argus Media. Higher us bz-to-crude ratio undermines sm unit margins, 2023.

[3] J. C. Gentry. Benzene production and economics: a review. *Asia-Pacific Journal of Chemical Engineering* **2(4)**, 272–277, 2007.

[4] H. Kwon, B. Lyu, K. Tak, J. Lee, J. H. Cho, and I. Moon. Optimization of naphtha purchase price using a price prediction model. *Computers & Chemical Engineering* **84**, 226–236, 2016.

[5] B. Lyu, H. Kwon, and I. Moon. A novel system dynamics model for forecasting naphtha price. *Korean Journal of Chemical Engineering* **35(4)**, 1033–1044, 2018.

[6] Z. Wang, M. Ke, Z. Song, J. Li, and J. Sun. Benzene reduction process simulation and optimization in catalytic cracking gasoline distillation. *Processes* **11(1)**, 151, 2023.

Posters

Outliers in dynamic time series models: a robust approach to parameter estimation and Kalman filter

A. Catarina Ribeiro[1,2] A. Manuela Gonçalves[1,2], and Marco Costa[3,4]

[1]Department of Mathematics (DMAT), University of Minho, Portugal
[2]Centre of Mathematics (CMAT), University of Minho, Portugal
[3] Águeda School of Technology and Management (ESTGA), University of Aveiro, Portugal
[4]Centre for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

E-mail addresses: *anaribeiro1421@outlook.pt; mneves@math.uminho.pt; marco@ua.pt*

In time series, the presence of outliers is common, resulting from natural phenomena or measurement errors. These observations compromise the effectiveness of classical estimation methods, such as the Kalman filter, reducing the accuracy of estimates and the reliability of forecasts [1]. The main objective of this work is to study and propose robust methodologies capable of adequately handling these observations, both in the prediction of states and in the estimation of model parameters within dynamic time series modelling. To this end, robust versions of the Kalman filter are proposed based on loss functions, which adjust the weights assigned to residuals, reducing the influence of these values [2], [3]. In parallel, robust likelihood estimation is explored through three distinct approaches: one based on the Huber function, a trimmed version of the classical likelihood that ignores a fraction of the most extreme observations, and a version based on the Cauchy loss function [4]. The performance of these approaches was evaluated through simulation studies, considering different combinations of parameters and sample sizes. Finally, the methods will be applied to real water quality data in a watershed, demonstrating their capabilities in real-world contexts [5].

Keywords

Kalman filter, outliers, robust estimation, state space models, time series.

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). Marco Costa was partially financed is supported by CIDMA (https://ror.org/05pm2mw36) under the FCT (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). Ana Catarina Ribeiro thanks CMAT for the research fellowship (BI) the support of CMAT through the grant UMINHO/BIM/2024/131.

References

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

[2] T. Cipra and R. Romera. Robust Kalman filter and its application in time series analysis. *Kybernetika* **27-6**, 1991.

[3] T. Cipra and R. Romera. Kalman filter with outliers and missing observations. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **6: 379–395**, 1997.

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

[5] SNIRH. Sistema Nacional de Informação de Recursos Hídricos. `https://snirh.apambiente.pt/` 2025 (accessed in 30 March 2025).

Forecasting and Interpretability of Bus Demand: An Application of SHAP in a Spatio-Temporal Context

Marina Estanislau[1], Ana Borges[1], Wellington Alves[1], Willian Machado[2] and Géremi Dranka[3]

[1]CIICESI, ESTG, Polytechnic of Porto, Portugal
[2]University of Campinas (UNICAMP), Brazil
[3]Production and Systems Engineering Program (PPGEPS), Federal University of Technology – Parana, Pato Branco, Brazil

E-mail address: *aib@estg.ipp.pt*

The growing complexity of urban mobility requires predictive models that balance performance and interpretability for public transportation planning. This study integrates SHapley Additive exPlanations (SHAP) into bus demand forecasting models using a high-resolution operational database of the city of Zurich, revealing the determinants of demand and exploring forecasting across planning-relevant time horizons.

Keywords

Public transportation demand; SHAP; Forecasting; Supply; Urban mobility.

The growing evolution and complexity of urban mobility require the use of predictive models that balance performance and interpretability, thereby effectively and efficiently supporting public transportation planning [1]. This study proposes an extension of bus demand analysis through the integration of Explainable Artificial Intelligence (XAI) techniques [2], specifically SHapley Additive exPlanations (SHAP) [3], into forecasting models based on a high-resolution operational database of the city of Zurich.

The SHAP framework [3] assigns each input feature an importance value for a given prediction using a game-theoretic approach rooted in cooperative game theory. This methodology provides both local and global interpretability, making it particularly well-suited for analysing complex spatiotemporal patterns in public transit demand [4].

The results reveal a strong dynamic component of demand, with high dependence on past values. The supply variable is identified as the main determinant, exhibiting a positive and consistent effect. In contrast, spatial variables are less directly relevant, exhibit high variability, and are not considered robust causal factors. These findings align with recent literature on stop-level modelling frameworks that combine machine learning algorithms with SHAP-based interpretability [4].

The introduction of an interaction term between a supply variable and dependence on public transportation allowed us to capture the alignment of the service with the population's needs, highlighting greater supply effectiveness in contexts of higher dependence. The SHAP analysis helped distinguish between predictive relevance and causal evidence [2], reinforcing the idea that demand is essentially determined by supply and its dynamics. Finally, different forecasting scenarios were explored across planning-relevant time horizons, contributing to a more comprehensive understanding of bus demand behaviour in urban settings.

Acknowledgements: 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) and National Council for Scientific and Technological Development – CNPq, under grant number 444556/2024-9.

References

[1] A. Nova, B. Bettoni, C. Biagioni, and D. Cossalter. Machine learning for public transportation demand prediction: a systematic literature review. *Engineering Applications of Artificial Intelligence*, **138**: 109359, 2024.

[2] G. Sariyer, S. K. Mangla, M. E. Sozen, G. Li, and Y. Kazancoglu. Leveraging explainable artificial intelligence in understanding public transportation usage rates for sustainable development. *Omega*, **128**: 103114, 2024.

[3] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, **30**, Curran Associates, Inc., 2017.

[4] A. Almeida and R. Pereira. Data-driven predictive modelling of stop-level public transit patterns. *Transportation*, 2025. https://doi.org/10.1007/s11116-025-10689-4

Spatial Analysis of Cancer Mortality in Portugal Using Autoregressive Random Forests

Anita Ferreira[1], Soraia Pereira[1,2,3] and Raquel Menezes[1,2]

[1]University of Minho, Portugal
[2]Centre of Mathematics, School of Sciences, University of Minho, Portugal
[3]Centre of Statistics and Applications, Faculty of Sciences, University of Lisbon, Portugal

E-mail addresses: *pg56093@alunos.uminho.pt; rmenezes@math.uminho.pt; soraia.pereira@math.uminho.pt*

Territorial inequalities in cancer mortality demand models able to capture complex determinants and residual spatial dependence. Focusing on lung and colon cancer in mainland Portugal, this study compares classical spatial models with hybrid spatial machine learning models, assessing their predictive ability, spatial autocorrelation capture, and contribution to epidemiological interpretation.

Keywords

cancer mortality, spatial epidemiology, hybrid spatial machine learning, Portugal.

Understanding the geographical distribution of cancer and its determinants is essential to support effective public policies, reduce health inequalities, and improve prevention. In Portugal, marked territorial differences persist, motivating the need for models that capture both complex covariate effects and spatial dependence. Classical spatial statistical models account for spatial autocorrelation but often rely on restrictive parametric assumptions. In contrast, machine learning methods such as Random Forests flexibly model nonlinear relationships, yet typically ignore spatial structure, leaving an important methodological gap. MacBride, Davies and Lee (2025) introduced the Spatial Autoregressive Random Forest (SPAR), which integrates Random Forests with a spatial autoregressive component to capture residual spatial dependence [1]. More recently, Lee and Davies (2026) proposed the Spatial Autoregressive Random Forest for Exposure-Response Functions (SPAR-Forest-ERF), a Bayesian extension enabling probabilistic inference [2]. We apply these methods to analyse lung and colon cancer mortality in mainland Portugal at the municipal level, incorporating sociodemographic, socioeconomic, environmental, general health, dietary propensity, anthropometric, and physical activity covariates. We compare classical spatial models, SPAR, and SPAR-Forest-ERF in terms of predictive performance, spatial autocorrelation capture, and interpretability. The results provide insight into the value of hybrid spatial machine learning models for epidemiological data and support the identification of modifiable risk factors, with implications for targeted prevention and resource allocation.

References

[1] C. MacBride, V. Davies, and D. Lee. A spatial autoregressive random forest algorithm for small-area spatial prediction. *Submitted to the Annals of Applied Statistics*, 2025.
[2] D. Lee and V. Davies. A spatial random forest algorithm for population-level epidemiological risk assessment. *arXiv preprint arXiv:2602.02277*, 2026.

Weighting for improved Stochastic Gradient Boosting in Genomic Prediction

Beatriz H. Comparado[1], João Lourenço[2], Vanda M. Lourenço[1]

[1] NOVA Math & Department of Mathematics, NOVA FCT, Portugal,
[2] NOVA LINCS & Department of Computer Science, NOVA FCT, Portugal

E-mail addresses: *b.comparado@campus.fct.unl.pt; joao.lourenco@fct.unl.pt; vmml@fct.unl.pt*

The presence of data contamination can compromise the performance of machine learning methods like Stochastic Gradient Boosting (SGB) in genomic prediction. This work addresses response contamination and evaluates SGB robustness via simulations on a synthetic animal breeding dataset. Our findings show that contamination reduces accuracy and that weighting strategies provide a straighforward way to improve SGB 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 (i) response transformation using weights derived from robust regression, (ii) observation weighting incorporated within the learning algorithm, and (iii) the combination of these approaches with alternative loss functions (default and L_1). Our findings clarify how contamination affects SGB and identify which adaptations most effectively improve robustness, providing practical guidance for genomic prediction in breeding studies where imperfect data are unavoidable.

Data contamination & Robust extensions of SGB
We consider a simulated animal dataset from the literature [1]. Data contamination is introduced through Huber's contamination model $(1-\varepsilon)\mathbf{F}+\varepsilon\mathbf{G}$, where $\mathbf{F} \sim N(\mu, \sigma^2)$ represents the distribution of the animal data, with μ and σ^2 estimated from the observed data. The contaminant distribution $\mathbf{G}$ is generated from the following Normal distributions: $N(\mu+k\sigma, \sigma^2)$ (`shift` contamination), $N(\mu, (s\sigma)^2)$ (`variance-inflated` contamination), and $N(\mu, (\sigma/\gamma)^2)$ (`variance-deflated` contamination). The contamination levels considered are $\varepsilon = 2\%, 5\%, 10\%$, with $k = 5, 7, 9$, $s = 5, 7$, and $\gamma = 1000, 10000$. Robust extensions of SGB are based on: (i) response transformation using weights derived from a robust regression fit, with $y_i^* = w_i y_i$ [2]; (ii) weighted loss optimization within the SGB fitting procedure, considering both the default and L_1 loss functions; and (iii) combinations of these two weighting strategies.

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* has limited impact on SGB performance; (ii) SGB performance degrades more markedly under *variance-inflated* contamination, followed by *shift* contamination; and (iii) the considered adaptations tend to improve performance relative to the classical SGB in the *shift* and *variance-inflated* scenarios, both in terms of predictive ability and prediction errors. These findings suggest that incorporating robustness into the SGB framework can mitigate the adverse effects of data contamination. In particular, approaches based on weighting and alternative loss functions appear to offer a promising and practically feasible direction for improving performance in genomic prediction settings where imperfect data are unavoidable.

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] V. M. Lourenço, J. Ogutu, R. Rodrigues, and H. P. Piepho. Genomic prediction using machine learning: a comparison of regularized regression, ensemble, instance-based and deep learning methods on synthetic and real data. *BMC Genomics* **25: 152**, 2024.

[2] V. M. Lourenço, J. O. Ogutu, and H.-P. Piepho. Robust Random Forests for Genomic Prediction: Challenges and Remedies. *Submitted*, `https://doi.org/10.64898/2026.03.30.715203` 2026.

Learning analytics for early detection of difficulties in Mathematics: a data-driven decision support system for teachers

Carla Martinho[1,2]

[1]ISCAL – Instituto Superior de Contabilidade e Administração de Lisboa, Instituto Politécnico de Lisboa, Portugal
[2]ICPOL – Centro de Investigação em Ciências Policiais, Lisboa, Portugal

E-mail address: *cmartinho@iscal.ipl.pt*

We present a learning analytics framework that integrates frequent low-stakes quiz data, item-level response patterns and longitudinal performance indicators to support timely pedagogical intervention in higher-education Mathematics. Aggregated indicators feed a teacher-facing decision dashboard, enabling early detection of struggling students. A pilot study at ISCAL (2025/26) illustrates the methodology and discusses statistical and computational challenges of small-cohort educational data.

Keywords

learning analytics, educational data mining, formative assessment, Mathematics education, decision support systems.

The growing availability of digital teaching platforms in higher education has made it feasible to collect fine-grained data on student learning that, when processed with appropriate computational and statistical methods, can support timely pedagogical decisions [1]. This work describes a learning analytics methodology designed to detect early signs of difficulty in undergraduate Mathematics courses and to translate raw assessment data into actionable indicators for teachers.

The methodological pipeline integrates three data sources collected throughout the semester: (i) short end-of-class quizzes administered with high frequency, (ii) item-level response data including time-on-task and error patterns, and (iii) longitudinal records of formative and summative assessments. Quiz items are tagged by curricular topic and cognitive demand, allowing the construction of multidimensional performance profiles at student, class and topic levels. Aggregated indicators – including topic-specific success rates, response-time distributions, intra-student variability and class-level dispersion measures – are computed and visualised in a teacher-facing decision dashboard. The dashboard is designed to surface deviations from expected trajectories early enough to enable targeted intervention, in line with the principles of formative assessment [3] and effective feedback [4].

A pilot study is being conducted in the second semester of 2025/26 at ISCAL with undergraduate students enrolled in a Mathematics course. Data are collected through the Mlearnix digital platform, which serves as the technical infrastructure for item delivery, response capture and indicator computation. Statistical analyses planned for the pilot include exploratory description of quiz-level distributions, clustering of student response

profiles and assessment of the predictive value of early-semester indicators for end-of-semester outcomes, following established educational data mining practice [2]. The contribution is twofold: a reproducible analytical pipeline transforming heterogeneous classroom data into pedagogically meaningful indicators, and a critical discussion of the statistical and computational challenges of working with small, sparse and longitudinal educational datasets typical of single-course studies.

Acknowledgements: The author thanks ISCAL – Instituto Politécnico de Lisboa for institutional support, and acknowledges the Mlearnix platform for providing the technical infrastructure used in data collection.

References

[1] G. Siemens. Learning analytics: the emergence of a discipline. *American Behavioral Scientist* **57(10)**, 1380–1400, 2013.

[2] C. Romero and S. Ventura. Educational data mining and learning analytics: an updated survey. *WIREs Data Mining and Knowledge Discovery* **10(3)**, e1355, 2020.

[3] P. Black and D. Wiliam. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* **21(1)**, 5–31, 2009.

[4] J. Hattie and H. Timperley. The power of feedback. *Review of Educational Research* **77(1)**, 81–112, 2007.

On the Algebraic Structure and Efficiency of Stair Nesting

Carla Santos[1,3], Cristina Dias[2,3] and Célia Nunes[4]

[1]Polytechnic Institute of Beja, Portugal
[2]Polytechnic Institute of Portalegre, Portugal
[3]NOVAMath - Center of Mathematics and Applications, School of Science and Technology, NOVA University of Lisbon, Portugal
[4]Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Portugal

E-mail addresses: *carla.santos@ipbeja.pt; cpsd@ipportalegre.pt; celian@ubi.pt*

Balanced nested designs often require a very high number of observations and concentrate degrees of freedom at the last stage. To overcome these limitations, we consider stair nesting as an advantageous unbalanced alternative, offering substantial economy in the number of observations (which becomes the sum of the numbers of levels of the factors) and ensuring orthogonality. Building on the algebraic structure of stair nesting, we discuss how to estimate and test the variance components for the successive factors under the assumption of normality.

Keywords

Nested factors, normality, unbalanced design, variance components.

The simplicity of administration and statistical analysis of balanced nesting is a key factor underlying its popularity in hierarchical experimental settings, across various fields where nested designs are widely used. Nevertheless, traditional balanced nested designs often require a very high number of observations and lead to a concentration of degrees of freedom at the last stage, which can reduce efficiency and increase experimental costs. In response to these limitations, we consider stair nesting as an advantageous alternative to balanced nesting, offering substantial savings in the number of observations – which becomes the sum of the numbers of levels of the successive factors – while ensuring orthogonality [1]. Stair nesting is based on an unbalanced allocation of observations across nested factors, however, in contrast to other unbalanced approaches, such as the staggered nested design [2], stair nested designs preserve orthogonality. Building on the algebraic structure of stair nesting, we discuss the straightforward estimation of the variance components associated with the nested factors (see, e.g., [3]). Assuming normality, we construct F-tests for the hypothesis of nullity of these variance components (see, e.g., [4]). To illustrate and validate the efficiency of stair nesting in achieving reliable estimates with fewer observations, we present a simulation study comparing it with the usual balanced nesting.

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/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/-10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications – NOVA Math) and UID/00212/2025 (https://doi.org/10.54499/UID/00212/2025) (Center of Mathematics and Applications of University of Beira Interior).

References

[1] D. Cox and P. Solomon. *Components of Variance*. Chapman and Hall, New York, 2003.

[2] T.R. Bainbridge. Staggered Nested Designs for Estimating Variance Components. *Industrial Quality Control* **22:1**, 12–20, 1965.

[3] R. Bailey, C. Fernandes, and P. Ramos. Sparse designs for estimating variance components of nested factors with random effects. *Journal of Statistical Planning and Inference* **214**: 76–88, 2021.

[4] C. Nunes, I. Pinto, and J.T. Mexia. F and Selective F tests with balanced cross-nesting and associated models. *Discussiones Mathematicae Probability and Statistics* **26**(2): 193–205, 2006.

A Computational Framework for Extremal Index Estimation

Dora Prata Gomes[1,2] and M. Manuela Neves[3,4]

[1]NOVA School of Science and Technology (NOVA FCT), Portugal
[2]Center for Mathematics and Applications (NOVA Math), Portugal
[3]Instituto Superior de Agronomia, Universidade de Lisboa, Portugal
[4]Centro de Estatística e Aplicações (CEAUL), Portugal

E-mail addresses: *dsrp@fct.unl.pt; manela@isa.ulisboa.pt*

Extreme events often cluster over time, impacting risk evaluation. In Extreme Value Theory, the extremal index measures this clustering, but selecting the threshold, k, for the upper order statistics to consider in the estimation remains a challenge. We propose an automatic, adaptative computational procedure using the Generalized Jackknife to reduce the estimator bias. Simulations show high performance, and its practical use is proven through Tejo River (Portugal) discharge data (1974–2022), offering a robust tool for flood risk assessment.

Keywords

Adaptive algorithm, Extremal index, Generalized Jackknife.

Extreme value parameter estimation faces inherent data scarcity, requiring robust semiparametric techniques to evaluate tail properties. When dealing with stationary sequences, serial dependence often causes extreme observations to appear in clusters rather than in isolation. In Extreme Value Theory (EVT), this temporal clustering is formalized and measured by the extremal index (EI), denoted by $\theta \in (0,1]$, where smaller values indicate a higher degree of clustering.

Under weak dependence conditions—such as Leadbetter's $D(u_n)$ condition—the maxima of stationary sequences converge to the same extreme value distribution types as independent data, though dependence indicated by the parameter θ afect the limiting parameters. A classic method to estimate the EI is the upcrossing estimator, defined by:

$$\hat{\Theta}^{UC}(u_n) = \frac{\sum_{i=1}^{n-1} I(X_i \leq u_n < X_{i+1})}{\sum_{i=1}^{n} I(X_i > u_n)} \tag{3}$$

Using a deterministic threshold $u = X_{n-k:n}$, $\hat{\Theta}^{UC}(u_n)$ can be written as

$$\widehat{\Theta}^{\mathrm{UC}}(k) = \frac{1}{k} \sum_{i=1}^{n-1} I(X_i \leq X_{n-k:n} < X_{i+1}).$$

However, the bias expansion of this estimator shows a direct dependence on the threshold parameter k, posing a practical challenge for optimal selection.

To mitigate this systematic bias, a second-order Generalized Jackknife estimator $\hat{\Theta}^{GJ(\delta)}$ is employed, [1]. By linear combination of the upcrossing estimator evaluations at three flexible levels (k, $[\delta k]+1$, and $[\delta^2 k]+1$), governed by a tuning parameter $0 < \delta < 1$, the two leading bias terms are asymptotically eliminated:

$$\hat{\Theta}^{GJ(\delta)} = \frac{(\delta^2+1)\hat{\Theta}^{UC}([\delta k]+1) - \delta(\hat{\Theta}^{UC}([\delta^2 k]+1) + \hat{\Theta}^{UC}(k))}{(1-\delta)^2} \tag{4}$$

We implement a computational stability-oriented heuristic algorithm to jointly calibrate the tuning parameter δ and the threshold level k, adapting principles from heuristic algorithms widely compared in threshold estimation literature, [2, 4]. The procedure systematically evaluates a grid of δ values, rounds the respective estimates, and identifies the longest run of consecutive equal values to define a region of maximal stability. The optimal parameters (δ and k) are adaptively chosen based on these maximum run lengths, removing subjective guesswork, [3].

The practical utility of this methodology is demonstrated through the analysis of a real-world hydrological dataset consisting of daily mean river discharge recorded at the Tejo River (Portugal) from 1974 to 2022 ($n = 576$ monthly maxima observations). The stability-based algorithm automatically selected $\delta = 0.06$ and an optimal threshold of $k = 516$, yielding a highly stable extremal index estimate of $\hat{\theta}^{GJ(\delta)} = 0.514$. This underscores the methodology's capacity to deliver reproducible and reliable risk assessments for environmental hazards and flood mitigation planning.

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/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications).

References

[1] M. I. Gomes, A. Hall, M. C. Miranda. Subsampling techniques and the jackknife methodology in the estimation of the extremal index. *Computational statistics & data analysis* **52(4)**, 2022–2041, 2008.

[2] M. I. Gomes, L. Henriques-Rodrigues, M. I. Fraga Alves,B. G. Manjunath. Adaptive port–mvrb estimation: an empirical comparison of two heuristic algorithms. *Journal of Statistical Computation and Simulation* **83(6)**, 1129–1144, 2013.

[3] D. P. Gomes, M. M. Neves. Resampling Procedures for a More Reliable Extremal Index Estimation, chap. 6. John Wiley and Sons, Ltd 2020.

[4] M. M. Neves, M. I. Gomes, F. Figueiredo, D. P. Gomes. Modeling extreme events: an application to environmental data. In: Contributions to Statistics, 137–145. Springer, 2015.

A Comparative Study of Longitudinal Prediction Methods Using Simulated Data

Elsa Soares[1], Inês Sousa[1] and Pedro Miranda Afonso[2]

[1]Centre of Mathematics, School of Sciences, University of Minho, Portugal
[2]Department of Biostatistics, Erasmus University Medical Center, The Netherlands

E-mail addresses: *id10725@alunos.uminho.pt, isousa@math.uminho.pt, p.mirandaafonso@erasmusmc.nl*

In this study, we compare the predictive performance of mixed-effects models and machine learning methods through simulations conducted under linear and nonlinear scenarios across different prediction horizons. The results indicate that correctly specified mixed-effects models, as well as those subject to minor misspecification, remain competitive across prediction horizons. However, models that fail to account for nonlinear temporal structures show a loss of predictive accuracy when compared with longitudinal random forests, which provide flexible nonlinear modelling without requiring explicit model specification.

Keywords

Gaussian linear mixed-effect model, Longitudinal prediction, Machine learning, Random forests, Simulation study.

The analysis of longitudinal data is central to a wide range of fields, as it enables the modelling and prediction of both subject-specific and population-level trajectories while accounting for temporal dependence. In precision medicine, longitudinal modelling supports clinical decision-making tasks such as early diagnosis, risk prediction, individualised treatment strategies and prognosis estimation [1].

Generalised linear mixed models (GLMMs) [2] are widely used in longitudinal data analysis because they provide an interpretable parametric framework that separates population-level effects from subject-specific variability through random effects. When the mean structure is correctly specified, or only mildly misspecified, GLMMs can achieve competitive predictive performance. However, owing to their assumed functional form, their performance may deteriorate when the true trajectories involve nonlinear temporal patterns [3].

In contrast, machine learning methods offer a non-parametric, data-driven alternative that can flexibly learn nonlinearities and interactions without requiring explicit functional specification. In longitudinal settings, random-forest-based approaches include (S)MERF, (S)MERT, REEMtree and REEMforest [4].

The main objective of this study is to compare predictive performance within a simulation framework based on a known data-generating process. We simulate longitudinal trajectories for 100 individuals observed over 50 time points under both linear and nonlinear scenarios. The proportion of observed history used for training is varied in order to assess its impact on predictive accuracy. Model performance is evaluated using the root mean squared error (RMSE), allowing comparison between correctly specified and misspecified mixed-effects models and longitudinal random forests, with the true data-generating model serving as a benchmark. Figure 1 illustrates the evolution of RMSE over time for models trained with 10% of the available history.

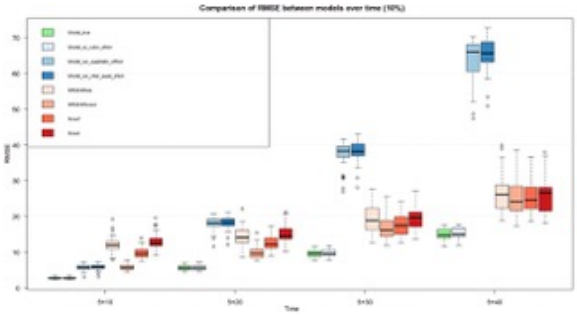


Fig. 1. RMSE evolution over horizons (10% observed history).

The results indicate that correctly specified and mildly misspecified mixed-effects models remain competitive across prediction horizons. By contrast, models that fail to capture the true temporal structure exhibit reduced long-term predictive performance when compared with longitudinal random forests.

Acknowledgements: This work was funded by Fundação para a Ciência e a Tecnologia (FCT) under the doctoral scholarship project UI/BD/154394/2023.

References

[1] J. Hu and S. Szymczak. A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics* **24**(2), 1–11, 2023.

[2] J. A. Nelder and R. W. Wedderburn. Generalized linear models. *J. R. Stat. Soc.* **135**, 370–384, 1972.

[3] A. Cascarano, J. Mur-Petit, J. Hernández-González, M. Camacho, N. de Toro Eadie, P. Gkontra, M. Chadeau-Hyam, J. Vitrià, and K. Lekadir. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review* **56**(Suppl 1), S1711–S1771, 2023.

[4] L. Capitaine, R. Genuer, and R. Thiébaut. Random forests for high-dimensional longitudinal data. arXiv:1901.11279, 2019.

Modeling Time to Syphilis Infection Using Interval-Censored Survival Data

Filipa Pinto[1], Rui Alves[1], Aurélio Sidumo[1], Carla Moreira[2], Luis Meira-Machado[1], Paula Meireles[3], Miguel Rocha[3,4] and Maria João Novais[3]

[1]Centre of Mathematics, University of Minho, Portugal
[2]RISE-Health, Faculty of Medicine, University of Porto, Porto, Portugal
[3]EPIUnit ITR, Instituto de Saúde Pública da Universidade do Porto, Universidade do Porto, Porto, Portugal
[4]GAT – Grupo de Ativistas em Tratamentos, Lisboa, Portugal

E-mail addresses: *filipa_pinto2003@hotmail.com; lmachado@math.uminho.pt*

Syphilis remains a major public health concern, with increasing incidence reported worldwide. In longitudinal screening programs, the exact time of infection is often unknown, resulting in interval-censored data. This study aimed to characterize a Portuguese cohort enrolled in a community-based syphilis screening program and identify factors associated with time to first reactive syphilis test.
Data were obtained from an observational longitudinal cohort of Men who have sex with Men (MSM) followed through periodic visits including rapid syphilis testing and structured questionnaires. Sociodemographic, behavioral, and clinical variables were analyzed, including age, education level, condom use, occasional sexual partners, substance use, knowledge of post-exposure prophylaxis (PEP), and country of birth. Time to infection was analyzed using survival methods adapted to interval-censored data, including nonparametric estimators, parametric accelerated failure time (AFT) models, and semiparametric proportional hazards models.
The cohort included more than 10,000 observations, with most participants being young adults. Participants reporting occasional sexual partners showed earlier occurrence of reactive syphilis tests, while higher educational level and knowledge of PEP were associated with longer time to infection. Individuals born in Brazil also presented higher risk of earlier infection. Machine learning methods for interval-censored data, including random survival forests, identified prevention knowledge, education, and sexual behavior as important predictors of risk profiles.

Keywords

syphilis, interval censoring, survival analysis, machine learning.

Acknowledgements: This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Program Contract of the Centre of Mathematics of the University of Minho (CMAT/UM), UID/00013/2025 (DOI: 10.54499/UID/00013/2025), and by the research project 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] B. W. Turnbull. The empirical distribution function with arbitrarily censored data. *J. Roy. Statist. Soc. Ser. B* **38**, 290–295, 1976.

[2] J. Sun. *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer, 2006.

A Topological Comparison of Uniform and Non-Uniform Embeddings for Weekly E-commerce Product Data

Patrícia Couto Neto[1,2], Susana Faria[2] and Flora Ferreira[3,4]

[1] Farfetch UK Limited, Porto, Portugal
[2] CMAT, Department of Mathematics, University of Minho, Portugal
[3] School of Economics and Management, University of Porto, Portugal
[4] CMUP, Faculty of Sciences, University of Porto, Portugal

E-mail addresses: *id12277@alunos.uminho.pt; sfaria@math.uminho.pt; flora.ferreira@fep.up.pt*

Topological Data Analysis is applied to weekly e-commerce product data through uniform and non-uniform delay embeddings. These representations generate point clouds in Euclidean space, from which Vietoris-Rips filtrations are constructed and persistent homology is computed. The results show that the induced topology is strongly representation-sensitive and that persistent homology summaries reveal structural heterogeneity beyond standard descriptive statistics.

Keywords

delay embeddings, e-commerce, persistent homology, time series, topological data analysis.

Topological Data Analysis is applied to a weekly e-commerce product panel through a representation-based pipeline built on delay embeddings and persistent homology [2, 3, 4]. This study aims to examine how alternative temporal embeddings of weekly product trajectories induce different geometric and topological structures, and whether these structures reveal heterogeneity not fully captured by standard descriptive statistics. For each product p and week t, a five-dimensional state vector

$$X_p(t) = \left(x_p^{\text{views}}(t), x_p^{\text{bags}}(t), x_p^{\text{orders}}(t), x_p^{\text{gtv}}(t), x_p^{\text{stock}}(t)\right) \in \mathbb{R}^5$$

is constructed using product views $\left(x_p^{\text{views}}(t)\right)$, add-to-bag activity $\left(x_p^{\text{bags}}(t)\right)$, sessions leading to orders $\left(x_p^{\text{orders}}(t)\right)$, gross transaction value $\left(x_p^{\text{gtv}}(t)\right)$, and stock availability $\left(x_p^{\text{stock}}(t)\right)$.

Two temporal embeddings are then defined, following the representation logic of delay-coordinate methods. The first is a uniform embedding

$$\Phi_p^{(U)}(r) = \left(X_p(\tau_{p,r}), X_p(\tau_{p,r-1}), X_p(\tau_{p,r-2}), X_p(\tau_{p,r-3})\right) \in \mathbb{R}^{20},$$

based on the current observed week and the previous three observed states. The second is a non-uniform embedding $\Phi_p^{(NU)}(r) \in \mathbb{R}^{30}$ using the delays $\{1, 2, 4, 13, 26\}$ in order to retain both short-range and longer-range temporal structure. In each case, the embedded observations form a point cloud $\mathcal{P} \subset \mathbb{R}^d$, with $d = 20$ or $d = 30$. Using the Euclidean metric, Vietoris-Rips complexes

$$VR_\varepsilon(\mathcal{P}) = \{\sigma \subseteq \mathcal{P} : \|x - y\|_2 \leq \varepsilon \text{ for all } x, y \in \sigma\}$$

are constructed across scales $\varepsilon \geq 0$, yielding a filtration from which persistent homology is computed [2, 3].

The analysis focuses on one-dimensional homology, using persistence diagrams to summarise the birth and death scales of H_1 features. These diagrams are summarised through the cardinality of H_1 persistence pairs, maximum persistence, mean persistence, persistence landscapes, silhouettes, and persistence images [1]. The results show that the induced topology is strongly representation-sensitive: the non-uniform embedding generates a broader distance structure and stronger H_1 summaries than the uniform embedding, indicating a richer multiscale organisation of weekly product behaviour. At the category level, category-specific persistence summaries produce different rankings across embeddings, showing that the topological structure is both heterogeneous and representation-dependent. This study therefore demonstrates that persistent homology provides a mathematically coherent exploratory framework for analysing structural heterogeneity in weekly e-commerce data.

Acknowledgments: The research of the first and second author 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`). The last author was partially supported by CMUP, member of LASI, which is financed by national funds through FCT under the project with reference `UID/00144/2025` and associated DOI given by (`https://doi.org/10.54499/UID/00144/2025`).

References

[1] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research* **18**(8), 1–35, 2017.

[2] G. Carlsson. Topology and Data. *Bulletin of the American Mathematical Society* **46**(2), 255–308, 2009.

[3] F. Chazal and B. Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence* **4**, 667963, 2021.

[4] L. Wasserman. Topological Data Analysis. *Annual Review of Statistics and Its Application* **5**, 501–532, 2018.

Ensuring High-Fidelity Data for LfD: The Importance of Shape-Based Anomaly Detection

Hugo Guimarães[1], Daniel Rodrigues[2], Luís Louro[2], André Cardoso[3], Ana Colim[3], Estela Bicho[2] and Eliana Costa e Silva[1,2]

[1]CIICESI, ESTG, Polytechnic of Porto, Portugal
[2]Algoritmi Center, University of Minho, Portugal
[3]Digital Transformation CoLab - DTx, Portugal

E-mail addresses: 8220337@estg.ipp.pt; id12193@alunos.uminho.pt; eos@estg.ipp.pt; luislouro@algoritmi.uminho.pt; andre.cardoso@dps.uminho.pt; ana.colim@dtx-colab.pt; estela.bicho@dei.uminho.pt

Manual programming for Human-Robot Collaboration (HRC) is time-consuming, and while Learning from Demonstration (LfD) offers a solution, it requires high-quality data. This study evaluates shape-based and feature-based outlier detection to clean kinematic trajectories. Results demonstrate that shape-based methods (DTW) are essential, successfully capturing spatial micro-hesitations that statistical summaries completely miss, ensuring high-fidelity data for LfD.

Keywords

Machine Learning, Time Series, Unsupervised

The success of Learning from Demonstration (LfD) in Human-Robot Collaboration (HRC) relies heavily on the quality of the training data [1]. Filtering out irregular human kinematics during preprocessing is a fundamental step to ensure robust robotic behaviour. Given the prohibitive cost and subjectivity of manually annotating video footage to find subtle kinematic errors, unsupervised anomaly detection has emerged as the most viable solution to establish baselines of normal execution [2].

In time-series anomaly detection, the literature contrasts two primary representation paradigms. The feature-based approach reduces high-dimensional trajectories into compact statistical descriptors to avoid the "curse of dimensionality" [4]. Conversely, the shape-based approach utilizes non-linear alignment methods like Dynamic Time Warping (DTW) to evaluate the elastic temporal morphology of the movement [3].

Despite extensive theoretical work on algorithms like LOF and DBSCAN, a significant gap remains in understanding how the choice of representation affects the detection of subtle spatial deviations in raw 3D kinematic trajectories. Accordingly, this study presents a comparative analysis of unsupervised anomaly detection models applied to raw 3D spatial coordinates from a simulated pick-and-place task, aiming to determine which paradigm is truly capable of capturing human spatial micro-hesitations and thereby ensuring the high-fidelity data required for LfD applications.

Motion data was collected from 19 participants performing a simulated window-assembly task using Xsens sensors (reconstructing the 3D spatial wrist coordinates: X, Y, Z). The movements were segmented into "Pick", "Move", and "Place" movements, in a total of 380, 380 and 95 respectively. To detect anomalies, we compared two data representation paradigms: a shape-based approach computing a pairwise Dynamic Time Warping (DTW) distance matrix, and a feature-based approach extracting statistical summaries (e.g., duration, spatial range). Unsupervised algorithms (LOF, DBSCAN, Isolation Forest, and

One-Class SVM) were applied to both representations. Performance was evaluated using the F2-Score against a visual ground truth, prioritizing the retrieval of known execution errors.

The evaluation revealed a clear distinction in the type of anomalies captured by each representation paradigm. The shape-based approach demonstrated superior sensitivity to trajectory morphology, with the DTW-LOF combination achieving the highest performance (77.46% F2-Score) during “Pick" movements. Qualitative visual inspection showed that DTW successfully flagged spatial micro-hesitations and subtle trajectory tremors that were completely missed by both human annotators and feature-based algorithms. By reducing kinematic curves to static global summaries, feature-based models (like DBSCAN) only reacted to extreme spatial deviations, ignoring internal procedural nuances.

Statistical feature extraction methods systematically miss critical spatial micro-hesitations in kinematic trajectories. This study therefore demonstrates that shape-based anomaly detection, specifically utilizing DTW, is highly recommended for ensuring high-fidelity data in LfD systems.

Acknowledgements: H. Guimarães and E. Costa e Silva were supported by national funds through FCT - Fundação para a Ciência e Tecnologia through projects UIDB/04728/2025 and UIDP/04728/2025 (https://doi.org/10.54499/UID/04728/2025). This work has been supported by Intelligent Robotic Coworker Assistant for Industrial Tasks with an Ergonomics Rationale (ref. PTDC/EEI-ROB/3488/2021).

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* **57**, 469–483, 2009.

[2] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Comput. Surv.* **54**, 2021.

[3] H. A. Dau, D. F. Silva, F. Petitjean, G. Forestier, A. Bagnall, A. Mueen, and E. Keogh. Optimizing dynamic time warping's window width for time series data mining applications. *Data Mining and Knowledge Discovery* **32**, 1074–1120, 2018.

[4] B. D. Fulcher and N. S. Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering* **26**, 3026–3037, 2014.

Nonparametric methods for functional data homogenization

Rubén Fernández-Casal[1], Manuel Oviedo de la Fuente[1] and Miguel Flores[2]

[1]University of A Coruña, Spain
[2]Escuela Politécnica Nacional, Ecuador

E-mail addresses: *ruben.fcasal@udc.es; manuel.oviedo@udc.es; miguel.flores@epn.edu.ec*

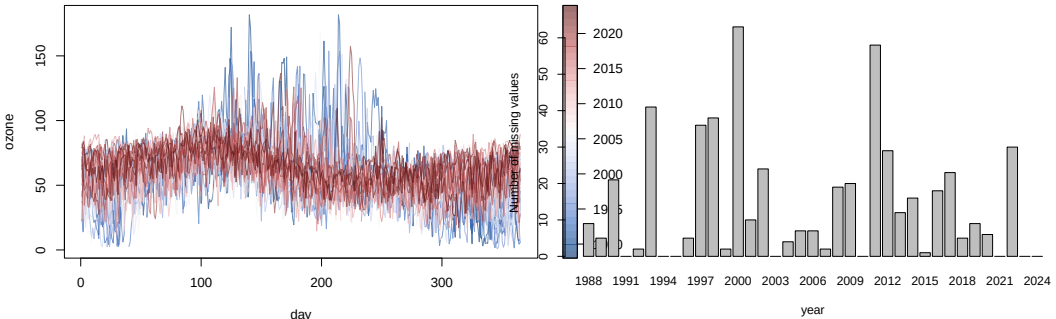
Functional data may contain point-wise anomalies, missing observations and outlying trajectories. We propose a fully nonparametric homogenization procedure that estimates trend, variance and dependence, winsorizes anomalous observations by leave-one-out kriging, reconstructs incomplete curves and detects functional outliers. The method is implemented in the R package `npfda` and illustrated with ozone data.

Keywords

functional data analysis, homogenization, kriging, depth measures, bootstrap.

Functional observations from environmental monitoring are rarely clean. We consider the heteroscedastic functional model $Y(t) = \mu(t) + \sigma(t)\varepsilon(t)$, observed on a finite grid with possible missing values. Here $\mu(t)$ is the trend, $\sigma^2(t)$ is the conditional variance and the standardized error process has correlogram $\rho(|t - t'|)$. The aim is to obtain a homogenized sample before applying FDA methods.

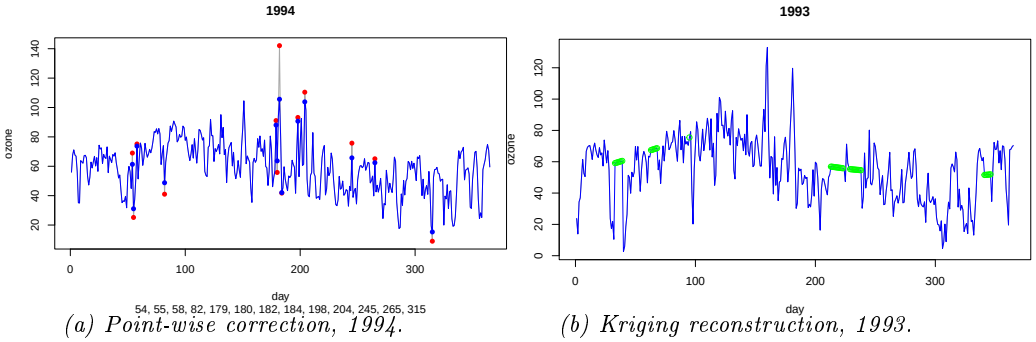
The procedure is illustrated with ground-level ozone data from the Yarner Wood monitoring site in the United Kingdom, recorded from 1988 to 2024. The annual curves display seasonal structure, changing variability and abnormal patterns; missingness is irregular, with some curves containing more than 50 unobserved daily values (Figure 2).



(a) Annual ozone curves. *(b) Number of missing observations.*

Figure 2. Ozone dataset from 1988 to 2024.

The proposed workflow has four stages. First, $\mu(t)$, $\sigma^2(t)$ and the semivariogram $\gamma_\varepsilon(u) = 1 - \rho(u)$ are estimated nonparametrically, following [2], and combined into a covariance model for residual kriging. Second, leave-one-out kriging intervals are computed at the observed grid points and observations outside their interval are winsorized to the nearest bound (Figure 3(a)). Third, missing positions are reconstructed by kriging from the corrected observations (Figure 3(b)).



(a) Point-wise correction, 1994. *(b) Kriging reconstruction, 1993.*

Figure 3. Homogenization examples.

Finally, completed curves are ranked by functional depth. The deepest $(1-\alpha)$ fraction defines the reference set, from which heteroscedastic bootstrap samples [2] provide lower depth quantiles. Their bootstrap median is used as cutoff, following the depth-based trimming idea of [1]. Curves below this cutoff are flagged as functional outliers (Figure 4).

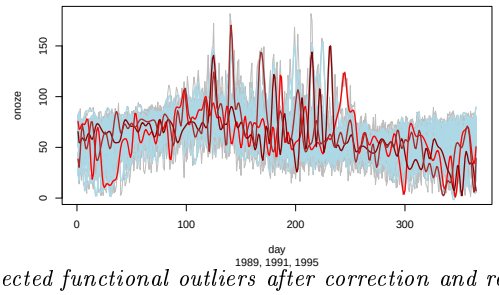


Figure 4. Detected functional outliers after correction and reconstruction.

The computational implementation can reuse precomputed pairwise distances in the bootstrap depth step, whenever resampling does not alter the curves, and can parallelize bootstrap replicates. This reduces the cost of the final outlier detection stage. The procedure jointly addresses point-wise contamination, incomplete curves and functional outliers, and is implemented in `npfda` [3].

Acknowledgements: This work has been supported by MCIN/AEI grant PID2020-113578RB-I00.

References

[1] M. Febrero, P. Galeano, and W. González-Manteiga. Outlier detection in functional data by depth measures. *Environmetrics* **19**, 331–345, 2008.

[2] R. Fernćasal, S. Castillo, and M. Flores. A nonparametric bootstrap method for heteroscedastic functional data. *JABES* **29**, 169–184, 2024.

[3] R. Fernćasal, S. Castillo, M. Flores, and M. Oviedo. *Npfda: Nonparametric functional data analysis*. R package version 0.2-1, 2025.

Quadratic filter for multi-rate systems with missing measurements and packet dropouts

Raquel Caballero-Águila[1], María Pilar Frías[1] and Antonia Oya-Lechuga[1]

[1]University of Jaén, Spain

E-mail addresses: *raguila@ujaen.es; mpfrias@ujaen.es; aoya@ujaen.es*

The least-squares quadratic filtering problem is studied for multi-rate systems with missing measurements. Additionally, random packet dropouts in transmission are considered and a hold-input compensation strategy is introduced. The proposed filter uses an augmented model with second-order Kronecker products and a covariance-based framework, without requiring the explicit identification of the signal evolution model. The filter performance is validated via a numerical simulation example.

Keywords

Multi-rate stochastic systems, missing measurements, packet dropouts, hold-input compensation strategy, quadratic filtering.

Multi-rate systems play a crucial role in modern engineering, where different sampling rates are employed to accommodate the distinct physical characteristics of system components, thereby enhancing both performance and resource utilization. Since estimation methods developed for single-rate systems cannot be directly applied, traditional approaches typically rely on transformation techniques, such as lifting or iterating the state equation, to unify sampling rates [1]. These methods require accurate knowledge of the state-space model, which may not always be available in practice. This limitation has motivated the development of covariance-based approaches that enable signal estimation without requiring explicit identification of the signal evolution equation [2].

Furthermore, when these multi-rate systems operate over networks, additional uncertainties arise. In particular, observation reliability is commonly degraded by missing measurements, often caused by imperfect measurement devices, sensor saturation or limited sensing capability. This phenomenon is typically modeled using Bernoulli-distributed random variables, which distinguish successful acquisitions from only-noise observations [3]. More general formulations based on random parameter matrices have also been proposed, providing a flexible framework to accommodate different random phenomena, including missing measurements [2, 4].

Additionally, information transmission is usually affected by packet dropouts, which reduce the availability of real-time data and may degrade estimation accuracy. Different compensation strategies, including hold-input mechanisms that replace lost packets with the most recently received data, have been proposed in the literature to mitigate these effects [2].

In this context, the presence of heterogeneous uncertainties significantly degrade the performance of conventional least-squares linear estimators, whose accuracy deteriorates in the presence of network-induced random phenomena, such as missing measurements or packet dropouts. To overcome these limitations, quadratic estimation has emerged as a robust alternative with moderate computational complexity, especially effective when

dealing with non-Gaussian uncertainties. In particular, the approach in [4] exploits second-order Kronecker powers of the observations to capture higher-order statistical information without requiring a full state-space model. This yields a recursive quadratic filter that outperforms conventional linear filters by better exploiting the data structure in scenarios involving random parameter matrices and deception attacks.

The current work addresses the recursive least-squares quadratic filtering problem for discrete-time stochastic multi-rate systems. The research specifically accounts for missing measurements, characterized by random parameter matrices defined as the product of a Bernoulli-distributed random sequence and a deterministic matrix. Furthermore, the system is subject to random packet dropouts during data transmission. To mitigate the impact of these losses, a hold-input compensation strategy is adopted. The proposed method is based on an augmented observation model using second-order Kronecker products, enabling covariance-based estimation without explicit signal evolution modeling. Finally, the effectiveness of the proposed filter is validated through numerical simulations.

Acknowledgements: This work is partially supported by the University of Jaén under the Research and Knowledge Transfer Plan 2025 (grant 2025/00345/001).

References

[1] Y. Shen, Z. Wang, H. Dong, and H. Liu. Multi-sensor multi-rate fusion estimation for networked systems: Advances and perspectives. *Inf. Fusion* **82**, 19-27, 2022.

[2] R. Caballero-Águila, M. P. Frías, and A. Oya-Lechuga. Least-squares linear estimation for multirate uncertain systems subject to DoS attacks. *Int. J. Netw. Dyn. Intell.* **4(2)**, 100014, 2025.

[3] Z. Yang, X. Zhang, W. Xiang, and X. Lin. A Novel Particle Filter Based on One-Step Smoothing for Nonlinear Systems with Random One-Step Delay and Missing Measurements. *Sensors* **25(2)**, 318, 2025.

[4] R. Caballero-Águila, and J. Linares-Pérez. Quadratic estimation for stochastic systems in the presence of random parameter matrices, time-correlated additive noise and deception attacks. *J. Frankl. Inst.* **360(15)**, 11141-11164, 2023.

Evaluation of Estimation Methods for the Residual Tail Dependence Parameter

Marta Ferreira[1,2]

[1]Centro de Matemática, Universidade do Minho
[2]Departamento de Matemática, Universidade do Minho

E-mail addresses: *msferreira@math.uminho.pt*

This study addresses the critical issue of threshold selection in the Peaks-Over-Threshold (POT) framework for estimating the residual tail dependence coefficient η, as introduced by Ledford and Tawn. Five automated threshold selection methods are explored and compared through a Monte Carlo simulation study. The methods are evaluated across several copula-based dependence models to assess their accuracy and robustness in estimating η. Finally, the approaches are applied to real environmental data involving atmospheric pollutant concentrations ($PM_{2.5}$, PM_{10}, and NO_2) from Lisbon, as well as NO_2 measurements from selected European capital pairs. The findings contribute to improving inference in multivariate extreme value analysis, particularly in the context of risk assessment.

Keywords

extreme value theory, coefficient of Ledford and Tawn, POT approach, threshold selection, risk assessment.

In multivariate extreme value analysis, it is essential to study the dependence structure between variables in their joint tails. Standard measures of dependence, such as correlation, may fail to capture extremal behavior. In particular, pairs of variables may exhibit asymptotic independence — that is, the probability of simultaneous extremes decays faster than the marginal probabilities — even though there is still a significant chance of large joint values.

To model such cases, Ledford and Tawn [2] introduced the residual tail dependence coefficient $\eta \in (0, 1]$, which quantifies the degree of dependence in the joint tail of a bivariate distribution.

Asymptotic independence has important applications in environmental and climate risk assessment (e.g., joint modeling of extreme rainfall, temperature, and flood drivers), financial risk management and systemic contagion analysis (where asset returns may be asymptotically independent yet exhibit substantial co-movement during crises), infrastructure reliability and engineering (simultaneous component failures), and more broadly in data science tasks involving rare-event classification and anomaly detection in multivariate settings. See, e.g., Beirlant et al. [1], Wadsworth and Tawn [4], Legrand et al. [3], and references therein.

Estimation of η typically requires a transformation of the marginal distributions to a common scale. This is necessary to ensure that the dependence structure in the joint tails is appropriately captured, independent of the marginal behavior. A common approach consists of transforming the marginal distributions to unit Fréchet or standard Pareto form, so that η corresponds to the tail index of the minimum of the standardized variables.

As in the univariate setting, the estimation of η depends crucially on the choice of a high threshold. In this work, we apply four threshold selection methods to the estimation of η, comparing their performance and robustness in this specific context.

We also investigate residual tail dependence in an air-pollution setting where compound extremes are of direct practical interest.

Acknowledgements: This research 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] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

[2] A. W. Ledford and J. A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169–187, 1996.

[3] J. Legrand, P. Naveau, and M. Oesting. Evaluation of Binary Classifiers for Asymptotically Dependent and Independent Extremes. *Journal of the American Statistical Association* **120**, 1558–1568, 2025.

[4] J. L. Wadsworth and J. A. Tawn. Dependence modelling for spatial extremes. *Biometrika* **99**, 253–272, 2012b.

Tail Dependence in Extremes: Empirical TDC Sample Paths, Eye-ball Thresholding, and Evidence from European Banks

Marta Ferreira[1,2]

[1]Centro de Matemática, Universidade do Minho
[2]Departamento de Matemática, Universidade do Minho

E-mail addresses: *msferreira@math.uminho.pt*

The tail dependence coefficient (TDC) introduced in Sibuya [1] and later formalized in Joe [2] is a key measure for quantifying extremal dependence between random variables. Unlike linear correlation, TDC captures the probability of joint extreme events and is therefore particularly relevant in risk management and financial stability analysis (Embrechts et al. [4]). Its theoretical foundation lies in Extreme Value Theory (EVT), which provides the asymptotic framework for modeling tail behavior and joint exceedances.

This study focuses on a simulation analysis of a heuristic method proposed by Danielsson et al. [3], commonly referred to as the "eye-ball" method, for detecting a stability region in the trajectory plot of successive TDC estimates. The procedure aims to identify an appropriate threshold by inspecting where the estimates stabilize, thus addressing the classical bias-variance trade-off inherent in tail estimation. Choosing too low a threshold induces bias, while too high a threshold increases variance due to limited extreme observations.

The performance of the heuristic approach is assessed through Monte Carlo simulations, evaluating its ability to balance bias and variance in finite samples. The methodology is then applied to financial data to analyze extremal dependence among equities of major European banks. Both upper-tail dependence (large joint gains) and lower-tail dependence (large joint losses) are examined, providing insights into asymmetries in extreme co-movements and implications for systemic risk.

Keywords

extreme value theory, tail dependence, heuristic threshold choice.

Acknowledgements: This research 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. Sibuya. Bivariate extreme statistics, I. *Annals of the Institute of Statistical Mathematics* **11(2)**, 195-210, 1959.

[2] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Taylor & Francis, 1997.

[3] J. Danielsson, L. M. Ergun, L. de Haan, and C. G. de Vries. Tail index estimation: quantile driven threshold selection. Systemic Risk Centre Discussion Papers, No. 58, 2016.

[4] P. Embrechts, A. J. McNeil, and D. Straumann. Correlation and Dependence in Risk Management: Properties and Pitfalls. In: M. A. H. Dempster (Ed.), *Risk Management: Value at Risk and Beyond*, Cambridge University Press, pp. 176–223, 2002.

Statistical Assessment of Forced Exercise Effects under Demyelination

Miguel L. Grilo[1,2,3], Sara Inteiro-Olveira[4], Tiago Costa-Coelho[4], Sandra H. Vaz[4], Sara A. Xapelli[4], Ana M. Sebastião[4]

[1]Instituto Superior Técnico, University of Lisbon, Portugal
[2]Department of Science and Technology, Universidade Aberta, Lisbon, Portugal
[3]Institute for Research and Advanced Training, University of Évora, Évora, Portugal
[4]Institute of Pharmacology and Neurosciences, University of Lisbon, Lisbon, Portugal

E-mail address: *miguel.l.grilo@uevora.pt*

This study investigated the therapeutic potential of a treadmill-based physical exercise protocol initiated following cuprizone-induced demyelination in mice. Cuprizone-fed C57BL/6J mice underwent forced exercise encompassing behavioural assessments and hippocampal electrophysiological recordings. Statistical results demonstrated that physical exercise enhanced long-term recognition memory and synaptic resilience under long-term potentiation saturation, with effects on synaptic plasticity most evident under repeated induction, in the cuprizone model.

Keywords

ANOVA, Bootstrap, Cuprizone Model, Behavioral Tests, Eletrophysiological Recordings

Multiple Sclerosis (MS) is a chronic inflammatory demyelinating disorder of the central nervous system in which limited remyelination impairs functional recovery, resulting in substantial neurological disability [1, 2, 3, 4], affecting ∼3.000.000 individuals worldwide [5]. Physical exercise (PE) has been proposed as a non-pharmacological intervention capable of modulating neuroprotective mechanisms and neuroinflammatory responses [6, 7]; however, its effects during active demyelination remain incompletely understood [6, 8]. This study employed a Cuprizone (CPZ)-induced demyelination model to investigate whether a treadmill-based PE protocol, initiated after the onset of demyelination, exerts a therapeutic effect on motor performance, anxiety-like behaviour, cognition, and hippocampal synaptic plasticity. The experimental paradigm, comprising four experimental groups (CTRL, PE, CPZ, CPZ&PE), facilitated the identification of either one independente variable (Group, with four categories, $n = 5$) or two independent variables (Treatment and Model, with two categories each, $n = 10$), while considering a significance level of 5%. Additionally, appropriate metrics were quantified for subsequent analysis as dependent variables: exploratory locomotor activity, anxiety-like behaviour, short-term spatial memory, long-term recognition memory, motor coordination and balance, hippocampal synaptic plasticity. Standard Analyses of Variance (ANOVA) tested main effects and interactions, with effect sizes reported as η^2. *Post hoc* pairwise comparisons were conducted via uncorrected Fisher's LSD, given the exploratory nature of the study and the small sample size ($n \leq 5$), with parametric significant results validated through non-parametric 95% bootstrap confidence intervals (percentile method; 5000 subsamples). Paired and unpaired (non)parametric t-tests were applied to within-subject and independent-sample comparisons, respectively. PE significantly reduced exploratory locomotor activity, likely reflecting

post-exercise relaxation rather than motor deficit, while simultaneously enhancing long-term recognition memory. The CPZ&PE exhibited superior motor coordination relative to PE, possibly reflecting conditioned anxiety induced by aversive treadmill-associated stimuli. No group differences emerged after a single long-term potentiation induction, whereas significant differences arose under synaptic saturation, suggesting PE-enhanced synaptic resilience. The results support that PE modulates synaptic plasticity and long-term recognition memory in the CPZ model, with effects most evident under repeated induction and saturation.

References

[1] R. H. Miller, S. Fyffe-Maricich, and A. C. Caprariello. Animal Models for the Study of Multiple Sclerosis. In: *Animal Models for the Study of Human Disease*, 967–988, 2017. DOI: https://doi.org/10.1016/b978-0-12-809468-6.00037-1

[2] M. Filippi, A. Bar-Or, F. Piehl, et al. Multiple sclerosis. *Nature Reviews Disease Primers* **4(1)**, 2018. DOI: https://doi.org/10.1038/s41572-018-0041-4

[3] D. E. Harlow, J. M. Honce, and A. A. Miravalle. Remyelination Therapy in Multiple Sclerosis. *Frontiers in Neurology* **6**, 2015. DOI: https://doi.org/10.3389/fneur.2015.00257

[4] G. D. F. Nunes, L. A. Osso, J. A. Haynes, et al. Incomplete remyelination via therapeutically enhanced oligodendrogenesis is sufficient to recover visual cortical function. *Nature Communications* **16(1)**, 2025. DOI: https://doi.org/10.1038/s41467-025-56092-6

[5] MS International Federation (MSIF). Atlas of MS, 3rd Edition. *MSIF*, 2020. Retrieved from https://www.msif.org/wp-content/uploads/2020/12/Atlas-3rd-Edition-Epidemiology-report-EN-updated-30-9-20.pdf

[6] M. Cefis, R. Chaney, J. Wirtz, et al. Molecular mechanisms underlying physical exercise-induced brain BDNF overproduction. *Frontiers in Molecular Neuroscience* **16**, 2023. DOI: https://doi.org/10.3389/fnmol.2023.1275924

[7] D. Hwang, J. Kim, S. Kyun, et al. Exogenous lactate augments exercise-induced improvement in memory but not in hippocampal neurogenesis. *Scientific Reports* **13(1)**, 2023. DOI: https://doi.org/10.1038/s41598-023-33017-1

[8] P. G. Nagappan, H. Chen, and D. Y. Wang. Neuroregeneration and plasticity: a review of the physiological mechanisms for achieving functional recovery postinjury. *Military Medical Research* **7(1)**, 2020. DOI: https://doi.org/10.1186/s40779-020-00259-3

A Real Options Model for Harvesting

Nuno M. Brites[1], João Brazão[1] and Miguel Reis[1,2]

[1]ISEG Research, ISEG Lisbon School of Economics & Management, Universidade de Lisboa, Lisbon, Portugal
[2]CMVM - Comissão do Mercado de Valores Mobiliários, Portugal

E-mail addresses: *nbrites@iseg.ulisboa.pt; l60438@aln.iseg.ulisboa.pt; mreis@iseg.ulisboa.pt*

Harvesting fisheries requires balancing economic incentives with ecological sustainability. We model the harvesting decision as a real option under uncertainty, leading to a stochastic control problem governed by a Hamilton–Jacobi–Bellman equation. Numerical solutions are used to analyse optimal harvesting policies and to illustrate the role of managerial flexibility in sustainable resource exploitation and economic efficiency.

Keywords

Fisheries management, Stochastic control, Real options.

Harvesting natural resources such as fisheries involves balancing economic incentives with ecological sustainability. We formulate the harvesting decision as a real option under uncertainty, leading to a stochastic control problem governed by a Hamilton–Jacobi–Bellman equation. Following recent numerical approaches, we analyse optimal harvesting policies and illustrate how managerial flexibility contributes to sustainable resource exploitation and economic efficiency.

Harvesting natural resources, particularly fisheries, has played a central role in supporting human society, both as a source of food and economic activity. Decisions related to harvesting are influenced not only by biological and environmental conditions but also by economic incentives. Understanding how to manage these resources sustainably is essential for balancing short-term gains with long-term viability.

Despite its importance, the fishing industry faces significant challenges. When fishing efforts exceed ecological limits, fish stocks can collapse due to direct human impact. In contexts where access to fishing grounds is open or poorly regulated, excessive effort may be exerted simultaneously by many agents, placing unsustainable pressure on the resource.

To better manage this uncertainty and the irreversibility of investment, the decision to harvest can be framed as a real option (see [1]). This approach treats the opportunity to fish as a right rather than an obligation, allowing fishermen to delay harvesting until conditions become favourable.

We formulate the optimal harvesting policy as a stochastic control problem, leading to a Hamilton–Jacobi–Bellman (HJB) partial differential equation (PDE), as in [3, 4, 2]. Following [5], we solve the HJB equation numerically, allowing us to simulate and analyse optimal policies under various scenarios. The results contribute to a better understanding of sustainable exploitation and highlight the economic value of flexibility in fisheries management.

Acknowledgements: Funded by national funds through FCT- Fundação para a Ciência e a Tecnologia, I.P., in the framework of the project UID/06522/2025.

References

[1] A. Murillas and J. M. Chamorro. Valuation and management of fishing resources under price uncertainty. *Environmental & Resource Economics* **33**(1), 39–71, 2006.

[2] N. M. Brites and C. A. Braumann. Fisheries management in random environments: comparison of harvesting policies for the logistic model. *Fisheries Research* **195**, 238–246, 2017.

[3] N. M. Brites. *Stochastic Differential Equation Harvesting Models: Sustainable Policies and Profit Optimization*. PhD Thesis, Universidade de Évora, Portugal, 2017.

[4] M. Reis, N. M. Brites, C. Santos, and C. Dias. Comparison of optimal harvesting policies with general logistic growth and a general harvesting function. *Mathematical Methods in the Applied Sciences* **47**(10), 8076–8088, 2024.

[5] M. Reis and N. M. Brites. Stochastic differential equations harvesting optimization with stochastic prices: formulation and numerical solution. *Results in Applied Mathematics* **25**, Article 100533, 2025. https://doi.org/10.1016/j.rinam.2024.100533

A Spectral Approach to Structured Designs

Cristina Dias[1,3], Carla Santos[2,3] and Nuno M. Brites[4]

[1] Polytechnic Institute of Portalegre,Portalegre,Portugal
[2] Polytechnic Institute of Beja, Beja, Portugal
[3] NOVAMath – Center for Mathematics and Applications, SST, NOVA University of Lisbon, Caparica, Portugal
[4] ISEG Research, ISEG Lisbon School of Economics & Management, Universidade de Lisboa, Lisbon, Portugal

E-mail addresses: *cpsd@ipportalegre.pt; carla.santos@ipbeja.pt; nbrites@iseg.ulisboa.pt*

We present a general formulation for information condensation in symmetric matrices whose spectral structure is governed by a dominant eigenvalue, allowing the structure to be reduced to its principal vector, to simplify analysis while retaining essential information. This applies to singular matrices as well as to structured families of matrices arising from the treatment structure of a base design with fixed effects. We focus on the action of base design factors on the structure vectors associated with the matrices. This framework provides a basis for further developments.

Keywords

dominant eigenvalue, symmetric matrices, structure vectors, fixed effects

The analysis of complex data structures often relies on the ability to extract essential patterns from high-dimensional matrices. Spectral decomposition provides a powerful tool for this purpose, particularly when a significant portion of the matrix information is concentrated in its primary components [1]. In this work we present a general formulation for information condensation in symmetric matrices whose spectral structure is governed by a dominant eigenvalue. This approach generalizes previous frameworks where families of matrices were mapped to the treatments of a base design, allowing for a systematic study of factor actions on structure vectors [2]. The proposed methodology applies to singular matrices as well as to structured families arising from the treatment structure of a base design with fixed effects. Particular attention is given to the action of the factors of the base design on the structure vectors associated with these matrices. This inferential structure, based on spectral decomposition, has proven robust even in the presence of incomplete datasets and missing observations [3]. When the leading eigenvalue is dominant, the representation of the matrix structure can be reduced to its principal structure vector, thereby simplifying the analysis while retaining the essential information encoded in the matrix. This framework provides a versatile basis for broad applications (see, e.g.[4]), and for further developments in spectral methods for structured designs, fixed-effect models, and matrix-based data representations.

Acknowledgements: This work is funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDB/MAT/00212/2020.
Nuno M. Brites was partially supported by project CEMAPRE/REM-UIDB/05069/2020 funded by FCT/MCTES through national funds.

References

[1] J.R. Schott *Matrix Analysis for Statistics*, 3rd ed., John Wiley & Sons, New York, 2016.

[2] C. Dias, C. Santos, and J.T. Mexia. Isolated and structured families of models for stochastic symmetric matrices. *Comp and Math Methods* **3:6**, e1152, 2021.

[3] C. Dias, C. Santos, N.M. Brites, and J.T. Mexia. Inferential Techniques in Structured Families of Models: Application to Durum Wheat Breeding Data. *Mathematical Methods in the Applied Sciences* **48**, 10386–10399, 2025.

[4] C. Santos, C. Nunes, C. Dias, and J.T. Mexia. Joining models with commutative orthogonal block structure. *Linear Algebra and its Applications* **517**, 235–245, 2017.

From salivary modulation to biological absence: A systems approach to tick-induced susceptibility

Sara Zúquete[1,2,3], Ludovina Padre[1], Clara Grácio[4], Luís Lopes[5]

[1] MED Mediterranean Institute for Agriculture, Environment and Development & CHANGE Global Change and Sustainability Institute, Instituto de Investigação e Formação Avançada, Universidade de Évora, Portugal;
[2]FMV – Faculdade de Medicina Veterinária, Universidade Lusófona—Centro Universitário de Lisboa, 1749-024 Lisboa, Portugal;
[3]CIISA – Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon;
[4]Universidade de Évora, CIMA, Rua Romão Ramalho, 59, Évora, Portugal;
[5]ISEL, IPL Lisboa, CIMA, R. Conselheiro Emídio Navarro 1, Lisboa, Portugal

E-mail addresses: *siortz@uevora.pt; lpadre@uevora.pt; mgracio@uevora.pt; luismariolopes@gmail.com*

We are aiming to a bottom-up, dynamical and information-theoretic framework that treats absence as an organising constraint with measurable signatures in time series, network structure and sequence distributions. At the immune level, we are moving beyond qualitative notions of "tolerance" towards a specified transient system in which activator and repressor fields interact across spatial diffusion scales and delays. This framework is intended to support a rational programme of fine-tuned immunomodulation, in tick-systems.

Keywords

Dynamical Systems, SIR model, Ticks, Nullomers

Ticks are embedded in multi-host, multi-pathogen systems representing a burden on veterinary and Public Health. Susceptibility to ticks must be treated as a dynamically system, a maintained property of hosts whose immune systems, once primed by an initial infestation, become either resistant or sensitive, driven by tissue damage, inflammatory activation, spatial diffusion and delayed salivary repression. Our central hypothesis is that current tick immunobiology remains conceptually local, focused on modulation at the feeding area, not really exploring on how system-level constraints and biological absence may jointly shape transient failures or successes in mounting resistance. We, therefore, argue that antigen-based approaches, seem to fail to provide a theoretically explicit treatment of how dynamical laws, informational constraints and structured absences influence tick systems. At the molecular level, nullomer-derived peptides and related constructs (including other rare or systematically excluded sequence motifs) can be used as tools for interrogating the biophysical and evolutionary conditions under which particular motifs are excluded from genomes and proteomes.

Acknowledgements: This work has received funding from MED (Universidade de Évora) under the project FCT - UID/05183/2025.

Latent Heterogeneity in Health Care Utilization Counts Using Negative Binomial Mixtures

Katy Freitas[1,2], Susana Faria[3], and Thiago Pavin[2]

[1] Department of Mathematics, University of Minho, Portugal
[2] Independent Researcher, Health Management and Data Science, Brazil
[3] CMAT, Department of Mathematics, University of Minho, Portugal

E-mail addresses: *katygdf@gmail.com; sfaria@math.uminho.pt; thiagopavin@gmail.com*

We assess whether finite Negative Binomial mixture models recover latent utilization groups more accurately than clustering in overdispersed count data. We generated a synthetic dataset with five health-care utilization variables, structured into four latent groups and calibrated to Brazilian regulatory benchmarks. Model parameters were estimated by maximum likelihood using Expectation-Maximization algorithm. The results suggest that finite Negative Binomial mixture models provide a more suitable approach to identifying latent utilization profiles in overdispersed health-care count data.

Keywords

finite mixture models, negative binomial distribution, health-care utilization .

These features challenge the adequacy of purely geometric clustering approaches, such as K-means, particularly when the underlying population comprises latent subgroups. Finite mixture models provide a probabilistic framework for representing unobserved heterogeneity by assuming that the observed population arises from a combination of latent components [1, 2].

In this work, we investigate whether a finite mixture model based on Negative Binomial components can recover latent utilization profiles more accurately than K-means. We generated synthetic to emulate health-care utilization in Brazil, following the simulation approach of Deb and Trivedi [3] for modelling health-care demand via finite mixtures. The data consists of 5,000 beneficiaries, structured into four latent groups and described by five count variables: *medical consultations*, *emergency visits*, *diagnostic exams*, *hospitalizations* and *therapies*. The marginal distributions of the simulated variables were calibrated using aggregate indicators from the Brazilian National Regulatory Agency for Private Health Insurance and Plans. The performance of each clustering method was evaluated using the Adjusted Rand Index (ARI) [7] and Classification accuracy [8].

The proposed model assumes that each latent component follows a product of independent Negative Binomial distributions [4], one for each utilization variable. Hospitalization and therapy variables exhibit structural excess of zeros [5], accommodated within each component. Parameter estimation was conducted using maximum likelihood via the Expectation-Maximization algorithm [6], with the number of components selected according to the Bayesian Information Criterion (BIC). For comparison, K-means clustering was applied to log-transformed and standardized count data, with the number of clusters selected using the Silhouette criterion. These preprocessing steps are necessary because

K-means relies on Euclidean distances, which are sensitive to scale differences and distributional skewness.

The best performance was obtained by the Negative Binomial mixture model when all five variables were included. In this specification, the BIC selected four components, matching the true number of latent groups. The results indicate that the mixture model outperformed K-means, achieving ARI=0.729, NMI=0.594 and accuracy =0.877, compared with ARI=0.673 and accuracy =0.815 for K-means. The inclusion of rare but informative variables, particularly therapies, improved the recovery of the latent structure.

These results provide evidence that, in overdispersed count data with latent heterogeneity, Negative Binomial mixture models constitute a more suitable framework than traditional clustering methods for recovering hidden utilization profiles. This is a methodological study and does not aim to classify real beneficiaries or detect operational anomalies.

Acknowledgements. The research of K. Freitas is supported by the Auditoria Inteligente 360 project under a technological co-development partnership agreement. The research of S. Faria was partially financed by Portuguese Funds through Fundação para a Ciência e a Tecnologia within the Project UID/00013/2025 (`https://doi.org/10.54499/UID/00013/2025`).

References

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

[2] S.-K. Ng, L. Xiang, and K. K. W. Yau. *Mixture Modelling for Medical and Health Sciences*. Chapman and Hall/CRC, 2019.

[3] P. Deb and P. K. Trivedi. Demand for medical care by the elderly: a finite mixture approach. *J. Appl. Econom.* **12(3)**, 313–336, 1997.

[4] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. 2nd ed., Cambridge University Press, 2013.

[5] D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34(1)**, 1–14, 1992.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39(1)**, 1–38, 1977.

[7] L. Hubert and P. Arabie. Comparing partitions. *J. Classification* **2(1)**, 193–218, 1985.

[8] J. Munkres. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* **5(1)**, 32–38, 1957.

Portuguese Fourth-Grade Students' Views of Scientists: A Statistical Analysis

Sara Michels[1], Susana Faria[2], and Glória Carvalho[3]

[1] Native Scientists, Portugal
[2] CMAT, Department of Mathematics, University of Minho, Portugal
[3] Department of Mathematics, University of Minho, Portugal

E-mail addresses: *sara.michels@nativescientists.org; sfaria@math.uminho.pt; gloria.carvalho@math.uminho.pt*

An increasing number of studies have examined students' representations of scientists using the Draw-A-Scientist Test. This study explored Portuguese fourth-grade students' images of scientists, with a particular focus on gender differences. The results showed that single-scientist drawings more frequently reproduced a male image of science, whereas group drawings appeared to allow greater space for gender diversity, although some stereotypical characteristics remained predominantly associated with male figures.

Keywords

draw-a-scientist test, stereotypical images, children's drawings.

The Draw-A-Scientist Test (DAST) has been used for more than 40 years to identify children's perceptions of scientists through drawings [1]. This method has consistently revealed stereotypical representations of scientists, typically depicting a male scientist working in a chemistry laboratory, wearing a lab coat and safety eyeglasses [2].

The present study attempts to further clarify Portuguese fourth-grade students' images of scientists through their drawings. Specifically, it examined whether students depicted no scientist, a single scientist, or multiple scientists, and how these representations varied according to the gender of the figures portrayed. The study also analysed the presence of stereotypical visual characteristics commonly associated with scientists, including lab coats, eyeglasses, unkempt hair, and older age.

The participants were 500 fourth-grade students from different regions of Portugal. The vast majority of the students drew at least one scientist (77.0%), indicating that children's representations of science remain strongly associated with the figure of the individual scientist. Nonetheless, 74 drawings represented two or more scientists, corresponding to 14.8% of the total sample and 19.2% of the drawings that included scientists. With respect to stereotypical characteristics, the most frequently observed were lab coats (48.8%), eyeglasses or goggles (30.6%), and unkempt hair (17.4%). Representations of middle-aged or elderly scientists were uncommon, appearing in only 3.4% of the drawings.

Chi-square tests were performed to examine whether the presence of stereotypical characteristics differed according to the number of scientists represented, comparing drawings depicting a single scientist with those depicting two or more scientists. No statistically significant differences were found for eyeglasses, lab coats, unkempt hair, or age category. These results suggest that the visual stereotype of the scientist was relatively stable and did not vary substantially as a function of whether children represented one scientist or multiple scientists.

Gender distributions were also compared between drawings depicting a single scientist and those depicting two or more scientists. Among drawings representing a single scientist, 54.0% depicted a male scientist, whereas 37.6% depicted a female scientist. In contrast, among drawings representing two or more scientists, the proportion of exclusively male groups decreased to 31.2%. Accordingly, approximately 70% of group representations were not exclusively male, including mixed-gender and exclusively female groups. These findings suggest that when children represent a scientist as an individual figure, they are more likely to depict a man. However, when scientists are represented collectively, children's drawings appear to show greater gender diversity.

The association between gender and stereotypical characteristics was also statistically significant, indicating that children did not apply the scientist stereotype uniformly to male and female figures. Unkempt hair was attributed almost exclusively to male scientists, appearing in 52 male representations compared with only 6 female representations. Eyeglasses showed a similar pattern, being more frequently associated with male than with female scientists, with 67 and 33 representations, respectively. The lab coat was the only stereotypical feature that appeared to be relatively evenly distributed across genders. These results indicate that some of the most salient visual markers of the scientist stereotype, particularly those associated with eccentricity or unconventional appearance, were predominantly linked to male figures. Female scientists, by contrast, tended to be represented in a more visually neutral manner.

Overall, these findings indicate that children's representations of scientists were shaped by both gender and the number of scientists depicted. Single-scientist drawings more frequently reflected a male-dominated representation of science, whereas group drawings appeared to support more diverse gender representation.

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

References

[1] D. W. Chambers. Stereotypic images of the scientist: The draw-a-scientist test. *Science Education* **67(2)**, 255–265, 1983.

[2] P. Bozzato, M. A. Fabris, and C. Longobardi. Gender, stereotypes and grade level in the draw-a-scientist test in Italian schoolchildren. *International Journal of Science Education* **43(16)**, 2640–2662, 2021.

Integrating Logistic Regression and Machine Learning Algorithms to Predict Postpartum Hemorrhage

Raquel Mugeiro Silva[1], Muriel Lérias Cambeiro[2,3], Anabela Rodrigues[4], António Vaz Carneiro[5], Filipa Lança[2,3] and Tiago Dias Domingues[1]

[1]Centre of Statistics and Its Applications, Faculty of Sciences, University of Lisbon, Portugal
[2]Faculty of Medicine, University of Lisbon, Portugal
[3]Department of Anesthesiology, Santa Maria University Hospital, Portugal
[4]Department of Transfusion Medicine, Santa Maria University Hospital, Portugal
[5] Institute for Evidence Based Healthcare (ISBE), University of Lisbon, Portugal

E-mail addresses: *fc57945@alunos.ciencias.ulisboa.pt; muriel@medicina.ulisboa.pt; anabela.rodrigues@ulssm.min-saude.pt; avc@isbe.research.ulisboa.pt; filipa.rodrigues@ulssm.min-saude.pt; tmdomingues@ciencias.ulisboa.pt*

Postpartum haemorrhage (PPH) is a major cause of maternal morbidity and mortality. This study evaluates logistic regression and machine learning approaches for predicting PPH using obstetric and clinical data. Classical logistic regression identified routinely collected predictors, while machine learning models were able to improve predictive performance. The findings support combining statistical and AI-based approaches for early risk stratification and improve clinical decision-making.

Keywords

unsupervised clustering, penalised regression, machine learning algorithms, postpartum haemorrhage.

Postpartum hemorrhage (PPH) remains one of the leading causes of maternal morbidity and mortality worldwide, underscoring the importance of identifying patients at risk as early as possible [1]. In this context, the present study proposes and evaluates a data analysis pipeline that integrates exploratory techniques with classification models to predict PPH based on demographic, obstetric, and laboratory variables.

To address missing values, Multiple Imputation by Chained Equations (MICE) was applied under the assumption that the data were missing at random (MAR) [2]. As an initial exploratory step, K-means clustering was used to uncover underlying patient profiles. The resulting clusters were visualized using Principal Component Analysis (PCA), revealing three distinct subgroups, one of which was characterized by a comparatively lower propensity for PPH.

For the predictive analysis, a two-stage modeling strategy was employed. First, stepwise Logistic Regression was used to identify six key clinical predictors commonly available in routine practice. Subsequently, machine learning models including Logistic Regression, Ridge Logistic Regression and Random Forest, were trained and evaluated using cross-validation [4]. Decision thresholds were optimized using Youden's index, and random oversampling was also employed to address class imbalance [3].

The findings demonstrated that Ridge Logistic Regression provided the most favorable predictive performance (Accuracy = 0.871, AUC = 0.907), which may be attributable to the stabilizing effect of L2 regularization on model coefficients. Although oversampling enhanced sensitivity, it was at times associated with a decrease in overall model performance metrics [3].

Collectively, this framework demonstrates that while conventional statistical models provide physicians with an interpretable tool by identifying easily obtainable, routinely collected predictors, the integration of contemporary machine learning algorithms can significantly improve predictive performance while enabling the inclusion of a broader set of covariates for PPH risk prediction.

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`, and by the European Union – NextGenerationEU through the project UID/PRR/00006/2025, DOI: `https://doi.org/10.54499/UID/PRR/00006/2025`.

The authors gratefully acknowledge the dedication of the Maternity Ward team, Mafalda Teodoro, Paula Mendes, and Werfen Portugal.

References

[1] L. Say, D. Chou, A. Gemmill, O. Tunçalp, A. B. Moller, J. Daniels, A. M. Gülmezoglu, M. Temmerman, and L. Alkema. Global causes of maternal death: A WHO systematic analysis. *Lancet Global Health* **2**, e323–e333, 2014.

[2] L. O. Joel, W. Doorsamy, and B. S. Paul. A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics* **20**, 6357–6373, 2025.

[3] M. Lérias-Cambeiro, R. Mugeiro-Silva, A. Rodrigues, T. Dias-Domingues, F. Lança, and A. Vaz Carneiro. Enhancing postpartum haemorrhage prediction through the integration of classical logistic regression and machine learning algorithms. *Mathematics* **13**, 3376, 2025.

[4] S. Çelik, B. Doğanlı, M. Ü. Şaşmaz, and U. Akkucu. Accuracy comparison of machine learning algorithms on World Happiness Index data. *Mathematics* **13**, 1176, 2025.

Index of authors