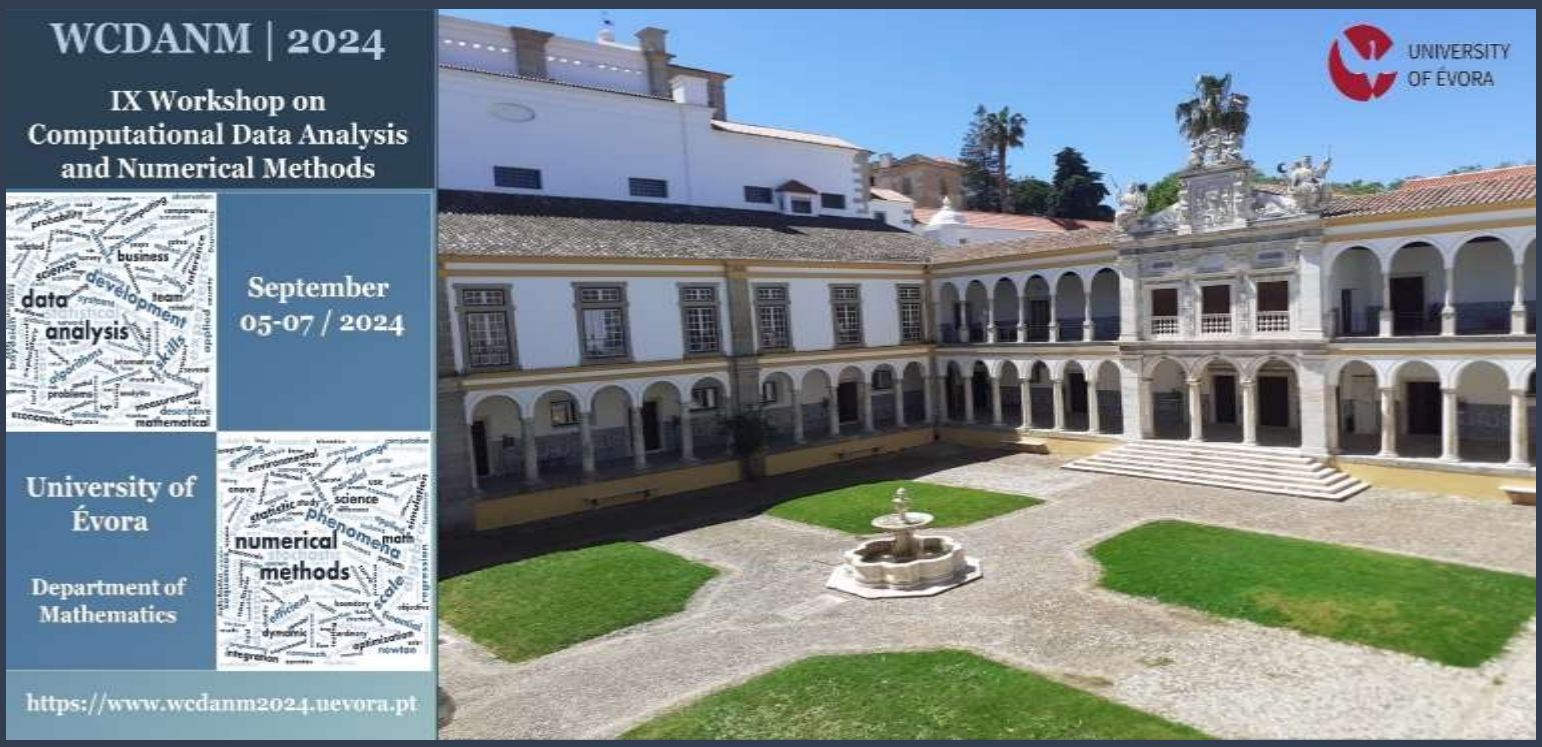

Book of Abstracts

Évora, Portugal

BOOK OF ABSTRACTS
IX WCDANM

University of Évora
Portugal
September 05–07, 2024

WELCOME TO THE IX WCDANM | 2024

Dear participants, colleagues and friends,

on behalf of the executive and organizing committees of the IX WCDANM (Workshop on Computational Data Analysis and Numerical Methods), we cordially welcome you to this workshop, hosted by the University of Évora. As a UNESCO World Heritage site (since 1986) and the upcoming European Capital of Culture (in 2027), Évora provides a unique setting for this event. This year, the event is also supported by research centers at several Portuguese universities: University of Minho, University of Aveiro, University of Lisbon and University of Évora. We aim to exceed the expectations of all participants, sponsors, and organizers.

A hybrid meeting, featuring both in-person and virtual participation, will be held. Renowned international scholars, including Ana Nieto (University of Salamanca, Spain), Carlos Ramos (University of Évora), Dharmendra Tripathi (National Institute of Technology, Uttarakhand, India), Ding-Geng Chen (Arizona State University, Phoenix, USA and University of Pretoria, South Africa), Padmanabhan Seshaiyer (George Mason University, USA), and Sotiris Bersimis (University of Piraeus, Greece), will deliver keynote addresses. The conference will feature over 50 presentations of scientific papers in various fields of research, fostering a dynamic exchange of ideas. The active engagement of the scientific community is instrumental in the success of this event.

The IX WCDANM offers two additional courses this year. Ding-Geng Chen from the Arizona State University, Phoenix, USA and University of Pretoria, South Africa, and Yiu-Fai Yung from SAS Institute Inc, USA, will lead a course on Structural Equation Modeling Using R and SAS, while Padhu Seshaiyer and Alonso Gabriel Ogueda from George Mason University will co-ordinate a course on Neural Computing. We extend our sincere gratitude to both teams for their timely acceptance of our invitation.

Our gratitude is also extended to the members of the Executive, Scientific, and Organizing Committees. In particular, we acknowledge the invaluable contributions of Anuj Mubayi (Illinois State University, USA) and Milan Stehlík (University of Applied Sciences Upper Austria & Universidad de Valparaiso, Chile) from the Executive Committee, as well as the tireless efforts of the local Organizing Committee members: Ana Nata (Polytechnic Institute of Tomar), Adelaide Freitas (University of Aveiro), A. Manuela Gonçalves (University of Minho), Dulce Gomes (University of Évora), Fernanda Figueiredo (University of Porto), Helena Luzia Grilo (Polytechnic

Institute of Tomar and Open University), and Pedro Marques (University of Évora).

The scientific outcomes of the meeting will be disseminated in special issues of the Journal of Applied Statistics and Research in Statistics (Taylor & Francis).

We anticipate that the IX WCDANM will serve as a catalyst for intellectual exchange, collaboration, and the dissemination of scientific research within the relevant community. We wish all participants a productive and enjoyable workshop.

University of Évora, September 05-07, 2024.

Chairman of the Executive Committee of IX 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

Luís Miguel Grilo, University of Évora, Portugal (Chairman of the Workshop)
Ana Nata, Polytechnic Institute of Tomar, Portugal
Adelaide Freitas, University of Aveiro, Portugal
A. Manuela Gonçalves, University of Minho, Portugal
Dulce Gomes, University of Évora, Portugal
Fernanda Figueiredo, University of Porto, Portugal
Helena Luzia Grilo, Polytechnic Institute of Tomar, Portugal
Pedro Marques, University of Évora, Portugal

Executive Committee

Luís M. Grilo, University of Évora, Portugal
Milan Stehlík, Univ. of Appl. Sciences Upper Austria & Universidad de Valparaíso, Chile
Anuj Mubayi, Illinois State University, USA

Scientific Committee

Anabela Afonso, University of Évora, Portugal
Ana Nata, Polytechnic Institute of Tomar, Portugal
Ana Isabel Borges, Polytechnic Institute of Porto, Portugal
Ana Rodrigues, University of Évora, Portugal
Anuj Mubayi, Illinois State University, 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
Catarina Marques, ISCTE-University Institute of Lisbon, Portugal
Catarina Nunes, University of Aberta, Portugal
Carlos Agra Coelho, NOVA University of Lisbon, Portugal
Carlos Braumann, University of Évora, Portugal
Carlos Ramos, University of Évora, Portugal
Célia Nunes, University of Beira Interior, Portugal
Clara Grácio, University of Évora, Portugal
Dharmendra Tripathi, National Institute of Technology, Uttarakhand, India
Ding-Geng Chen, Arizona State Univ., Phoenix, USA & Univ. of Pretoria, South Africa
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
Feliz Minhós, University of Évora, Portugal
Fernanda Figueiredo, University of Porto, Portugal
Fernando Carapau, University of Évora, Portugal
Filipe Marques, NOVA University of Lisbon, Portugal
Gonçalo Jacinto, University of Évora, Portugal
Joaquim M. C. Correia, University of Évora, Portugal
Lígia Henriques-Rodrigues, University of Évora, Portugal

Luís M. Grilo, University of Évora, Portugal
Malay Banerjee, Indian Institute of Technology Kampur, India
Manuel Branco, University of Évora, Portugal
Manuela Oliveira, University of Évora, Portugal
Maria Luísa Morgado, University of Trás-os-Montes e Alto Douro, Portugal
Maria do Rosário Ramos, University of Aberta, Portugal
Marília Pires, Universidade de Évora, Portugal
Milan Stehlík, Univ. of Appl. Sciences Upper Austria & Universidad de Valparaíso, Chile
Pablo Rogers, Federal University of Uberlândia, Brazil
Padmanabhan Seshaiyer, George Mason University, USA
Paulo Correia, University of Évora, Portugal
Paula Vicente, Lusófona University, Lisbon; Portugal
Pedro Marques, University of Évora, Portugal
Rui Albuquerque, University of Évora, Portugal
Russell Alpizar-Jara, University of Évora, Portugal
Sarada Ghosh, Department of Statistics, Kolkata, India
Sotiris Bersimis, University of Piraeus, Greece
Susana Faria, University of Minho, Portugal
Valter Vairinhos, Naval School, Portugal
Yiu-Fai Yung, SAS Institute Inc, USA

Sponsored by

Technical Specifications

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

Web-page
`https://www.wcdanm2024.uevora.pt/`

Editor
University of Évora, Portugal
Colégio Luís António Verney
Rua Romão Ramalho, 59
7000-671 Évora

Editors
Luís Miguel Grilo (University of Évora & CIMA & NOVA Math & Ci2, Portugal)
Ana Nata (Polytechnic Institute of Tomar & CMUC, Portugal)
Helena Luzia Grilo (Polytechnic Institute of Tomar & UAb, Portugal)
Dulce Gomes (University of Évora & CIMA, Portugal)
Anuj Mubayi (Illinois State University, USA)
Milan StehlíK (University of Applied Sciences Upper Austria & Universidad de Valparaiso, Chile)

Authors
Many authors.

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

ISBN: 978-972-778-417-2

Contents

Posters

Invited Speakers

Dynamic monitoring of streaming text data by integrating text visualization techniques and natural language processing

Sotirios Bersimis[1], Grigorios Papageorgiou[2] and Polychronis Economou[2]

[1]Department of Business Administration, University of Piraeus
[2]Department of Civil Engineering, University of Patras

E-mail addresses: *sbersim@unipi.gr; up1069953@upatras.gr; peconom@upatras.gr*

Managing unstructured text streams, such as business emails, presents a significant challenge for large organizations due to their inherent lack of structure, the prevalence of noise, and frequent use of abbreviations and acronyms. Effective monitoring of large volumes of emails is critical for making informed decisions and staying aware of market dynamics. To address this, a text visualization approach is proposed, which transforms word distances from their origin into a sequential data format. This sequential data format reveals key insights into temporal patterns and fluctuations within the text streams. The approach incorporates two powerful change point detection methods, Cumulative Sum (CUSUM) and Pruned Exact Linear Time (PELT), to monitor text streams and identify significant shifts or anomalies. The effectiveness of the proposed algorithm is assessed through simulations and demonstrated in a real-world context, inspired by a case involving the mail corpus received by a shipbroker agent in Greece. Additionally, the method successfully identified potential market changes in Asia, highlighting its value in detecting regional trends. In summary, the proposed method provides a robust solution for dynamically monitoring time-varying unstructured text streams, enabling better decision-making and improved market intelligence for large-scale organizations.

Keywords

Natural Language Processing, Process Monitoring of Textual Data, Text Data Streams.

Managing unstructured text streams, such as business emails, presents a critical challenge for large organizations due to their intrinsic lack of structure, the presence of excessive noise, and the frequent use of abbreviations and acronyms [1]. Efficient monitoring of these text streams is essential for informed decision-making and maintaining market awareness, but the complexity and volume of the data can make this task overwhelming. To address this, we propose a novel text visualization approach that transforms word distances from their origin into a sequential data format. This transformation enables the extraction of significant insights regarding temporal patterns and fluctuations in the data, which are often hidden within the unstructured text [2].

Our approach integrates two robust change point detection methods, Cumulative Sum (CUSUM) [1] and Pruned Exact Linear Time (PELT) [2], to monitor these sequential data formats effectively. CUSUM is adept at identifying shifts in the mean of a data sequence, making it suitable for detecting abrupt changes, while PELT offers an efficient and accurate method for identifying multiple change points across longer sequences. By leveraging these methods, our approach can detect significant shifts or anomalies in the

text streams, providing early warnings of potential changes in market conditions or other relevant areas.

The performance of the proposed algorithm is evaluated through a simulation approach to demonstrate its robustness and accuracy under various scenarios. Additionally, the method is applied in a real-world context inspired by a case study involving the mail corpus received by a shipbroker agent operating in Greece. This application showcases the algorithm's ability to identify potential market changes, specifically in the region of Asia, highlighting its utility in detecting emerging regional trends that could impact business decisions.

Overall, our proposed method offers a comprehensive solution for the dynamic monitoring of time-varying unstructured text streams. It enhances the ability of large-scale organizations to make informed decisions and gain deeper market intelligence by systematically unveiling hidden patterns and changes within the data. This approach not only improves the understanding and analysis of unstructured text streams but also provides a scalable and adaptable framework for integrating text stream monitoring into broader decision-making processes.

References

[1] Lipika Dey and SK Mirajul Haque. Opinion mining from noisy text data. *Proceedings of the second workshop on Analytics for noisy unstructured text data*. 2008.

[2] K. C. Garwood, C. Jones, N. Clements and V. Miori. Innovations to identifying the effects of clear information visualization: Reducing managers time in data interpretation. *Journal of Visual Literacy*, **37(1)** , 40–50, 2018.

[3] P. H. Ellaway. Cumulative sum technique and its application to the analysis of peristimulus time histograms. *Electroencephalography and clinical neurophysiology*, **45(2)**, 302–304, 1978.

[4] Rebecca Killick, Paul Fearnhead and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107(500)**, 1590–1598, 2012.

Big data and statistical meta-analysis

Ding-Geng Chen[1,2]

[1]Department of Statistics, University of Pretoria, South Africa
[2]College of Health Solutions, Arizona State University, Phoenix, USA

E-mail addresses: *dinchen@asu.edu; din.chen@up.ac.za*

Statistical meta-analysis (MA) is a common statistical approach in big data inference to combine meta-data from diverse studies to reach a more reliable and efficient conclusion. It can be performed by either synthesizing study-level summary statistics (MA-SS) or modeling individual participant-level data (MA-IPD), if available. In this talk, we review the classical statistical meta-analyses, and further discuss the relative efficiency between MA-SS and MA-IPD. We show theoretically that there is no gain of efficiency asymptotically by analyzing MA-IPD. Our findings are further confirmed by extensive Monte-Carlo simulation studies and real data analyses in [1].

Keywords

Big-Data, Meta-Analysis, Statistical efficiency.

References

[1] D.G. Chen, D. Liu, X. Min and H. Zhang. Relative efficiency of using summary and individual information in random-effects meta-analysis. *Biometrics* **76(4)** 119–1329, 2020. (https://doi.org/10.1111/biom.13238)

Contributions to supervised three-way data analysis

Ana B. Nieto-Librero[1,2] and Nerea González-García[1,2]

[1]Department of Statistics, University of Salamanca, Spain
[2]Centre for Human Rights and Public Policy Research, CIDH-Diversitas, Spain

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

Multiway covariates regression analysis is a technique that allows explaining a set of dependent variables from a set of independent variables that are measured at different points in time or in different situations and are stored in tensors. Despite its potential, it has not been widely used in practice due to its difficulty in interpretation. This paper presents a solution using disjoint components in a way that facilitates the process of interpreting the results.

Keywords

Regression, Tucker, Disjoint.

When the aim of our study is to analyse the behaviour of a dependent variable as a function of a set of independent variables, we use classification methods (Support Vector Machine [1], Decision Trees [2], Random Forest [3]...) if the variable is qualitative or regression methods if it is quantitative (Multivariate Regression, Principal Component Regression [4], Principal Covariates Regression [5]...). When these variables have been measured at different points in time or in different situations, techniques have also been developed to capture this temporal or spatial dimension (Sparse Tensor Discriminant [6], Higher Order Discriminant Analysis [7], Tensor Regression [8], U-PLS [9], Multiway Covariates Regression [10]...). However, they have not been widely used in practice due to the difficulty in interpretation. For this reason, a new technique is proposed to facilitate the interpretation of the current methods through the use of disjoint components. To this end, the MDCovT3R method is proposed, which uses multiway covariates regression analysis [10] and the method of constructing disjoint components [11,12] to provide a way of studying the influence of a regressor tensor on a dependent tensor so that the interpretations of the components and their interactions are facilitated by constructing components in which not all the variables under study are involved.

References

[1] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning* **20(3)**, 273–297, 1995.

[2] J. R. Quinlan. Induction of decision trees. *Machine Learning* **1**, 81–106, 1986.

[3] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

[4] H. Martens and T. Næs. Multivariate Calibration. *Chemometrics*, 147–156, 1984.

[5] M. Vervloet, H. A. L. Kiers, W. van den Noortgate and E. Ceulemans. PCovR: An R Package for Principal Covariates Regression. *Journal of Statistical Software* **65(8)**, 1–14, 2015.

[6] Z. Lai, Y. Xu,J. Yang, J. Tang and D. Zhang. Sparse tensor discriminant analysis. *IEEE Transactions on Image Processing*, **22(10)**, 3904–3915, 2013.

[7] A. H. Phan and A. Cichocki. Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear Theory and Its Applications, IEICE*, **1(1)**, 37–68, 2010.

[8] H. Zhou, L. Li and H. Zhu. Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association*, **108**, 540–552, 2013.

[9] S. Wold, P. Geladi, K. Esbensen and J. Öhman. Multi-way principal components-and PLS-analysis. *Journal of Chemometrics*, **1(1)**, 41–56, 1987.

[10] A. K. Smilde and H. A. L. Kiers. Multiway covariates regression models. *Journal of Chemometrics*, **13(1)**, 31–48, 1999.

[11] M. Vichi and G. Saporta. Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, **53**, 3194–3208, 2009.

[12] C. Ferrara, F. Martella and M. Vichi. *Dimensions of Well-Being and Their Statistical Measurements*. Alleva, G., Giommi, A. (eds) Topics in Theoretical and Applied Statistics. Studies in Theoretical and Applied Statistics. Springer, 2016.

Evolutionary dynamics and cellular automata

C. Correia Ramos[1,2]

[1]University of Évora, Portugal
[2]Research Center in Mathematics and Applications, Portugal

E-mail address: *ccr@uevora.pt*

We discuss evolutionary and genetic algorithms for cellular automata. We consider several processes, such as mutation, replication, recombination and assembly which transform cellular automata. These processes can be combined to define dynamical processes on cellular automata spaces. We analyse evolving populations of cellular automata and its main characteristics. Applications are discussed, namely pseudo random number generators, simulation of natural phenomena or image processing.

Keywords

Cellular automata, Evolutionary dynamics, Genetic algorithms.

Acknowledgements: This talk has been partially supported by FCT through the Research Center of Mathematics and Applications of University of Evora (CIMA), project CIMA-UIDB/04674/2020.

References

[1] Carlos Ramos and Marta Riera. Evolutionary dynamics and the generation of cellular automata. Iteration theory (ECIT '08), 219–236, Grazer Math. Ber., 354, Institut fur Mathematik, Karl-Franzens-Universitat Graz, 2009.

[2] C. Correia Ramos and Nada EL Bouziani. Mouhaydine Tlemçani and Sara Fernandes. Probabilistic simulation of fractures using cellular automata, submitted, 2024.

[3] C. Correia Ramos, Nada EL Bouziani, Mouhaydine Tlemçani and Sara Fernandes. Simulation of ideal material blocks using cellular automata. *Nonlinear Dynamics* **111**, 22381–22397, 2023.

[4] Carlos Ramos, Fernando Carapau and Paulo Correia. Cellular Automata describing non-equilibrium fluids with non-mixing substances, In: Carapau,F., Vaidya, A. (eds) Recent Advances in Mechanics and Fluid-Structure Interaction with Applications. *Advances in Mathematical Fluid Mechanics. Birkhauser, Cham.* 229–245, 2022.

[5] Luís Bandeira and C. Correia Ramos. Transition matrices characterizing a certain totally discontinuous map of the interval. *J. Math. Anal. Appl.* **444(2)**, 1274–1303, 2016.

Mathematical modeling, analysis and simulation for data-driven challenges in mathematical biology with applications

Padmanabhan Seshaiyer[1]

[1]Department of Mathematical Sciences, College of Science, George Mason University, Fairfax, USA

E-mail address: *pseshaiy@gmu.edu*

Transmission dynamics of infectious diseases such as COVID-19 has made us to re-envision how we model, analyze and simulate the spread of infectious diseases and evaluate the effectiveness of non-pharmaceutical control measures as important mechanisms for assessing the potential for sustained transmission. Incorporating human behavior into these models responding to a perceived increase of the infections in the local environment in real-time, adds another layer of complexity in these models. These epidemiological models are often modeled via compartmental models using system of nonlinear coupled ordinary differential equations which are often numerically solved to get insights into population behavior. Also, there have also been rapid developments in employing a Physics Informed Neural Networks (PINNs) approach to estimate the model parameters such as the transmission, infection, quarantine and recovery rate using real data sets. In this work, we present modeling, analysis and simulation through Disease Informed Neural Networks (DINNs) and its application to real data modeled using non-linear differential equations. We discuss how these approaches are capable of predicting the behavior of a disease described by modified compartmental models that include parameters and variables associated with the governing differential equations describing the dynamics of the disease. Through benchmark examples, we will show how DINNs can predict optimal parameters for given datasets for a variety of applications.

Mathematical analysis of tumour impacts on physiological flows modulated by electric and magnetic fields

Dharmendra Tripathi[1] and Ashvani Kumar[1]

[1]Department of Mathematics, National Institute of Technology Uttarakhand, Srinagar-246174, India

E-mail address: *dtripathi@nituk.ac.in*

Physiological flows like blood flow, urine flow, breathing, movement of chyme, sperm movement, etc. are very important mechanisms in the biological systems which are governed by very natural pumping process i.e., peristalsis, membrane pumping, heart pumping, compression and expansion of lungs, and rhythmic propagation of the muscles. However, the tumours in the vessels/parts of the body are challenging problem, creating obstruction in the fluids flow and many people are died due to infection and fast growth of the tumours in the body. This process introduces complexity in flow behaviour particularly at micro scale. To find out the mathematical solution at small context, a fluid flow model governed by the peristaltic pumping is developed in present of single tumour. An analysis for flow characteristics and influence of tumour shape and size in during the fluid flow in microchannel is simulated. Furthermore, how this obstruction due to tumour by applying the external electric field and magnetic field have been examined. For this biophysical model, governing equations based on mass conservation, momentum conservation and Maxwell equation for electro-magneto-hydrodynamics have been adopted. Low Reynolds number flow in microchannel is considered. MATLAB code is utilized for simulation of the results. The findings reveal that a larger tumor height enhances fluid flow by narrowing the microchannel and promoting abnormal fluid flow in microchannel however this tumour size may also stop the fluid flow if this size is very close to the diameter/width of the microchannel. Overall, this research provides insights into optimizing fluid dynamics for biomedical applications and gives recommendation for development of bio microfluidics devices.

Keywords

Tumour cell, Peristaltic transport, Electroosmosis, Magnetohydrodynamics, Zeta potential, Hartmann number.

Acknowledgements: I acknowledge to SERB, DST, Gov. of India (Ref. no. MTR/2023/000377), for providing the fund for this research work.

References

[1] Meijing Li and James G. Brasseur. Non-steady peristaltic transport in finite-length tubes. *Journal of Fluid Mechanics* **248** 129–151, 1993.

[2] Sanjay Kumar Pandey and Ankit Prajapati. An analytical and comparative study of swallowing in a tumor-infected oesophagus: a mathematical model. *Journal of Mathematical Biology* **88(3)**, 37, 2024.

Short Courses

Neural computing

Padmanabhan Seshaiyer[1] and Alonso Ogueda-Oliva[1]

[1]Department of Mathematical Sciences, College of Science, George Mason University, Fairfax, USA

E-mail address: *pseshaiy@gmu.edu; aogueda@gmu.edu*

In this short course, we will introduce foundations of computational problem solving for models and datasets associated with equations describing real world problems. We start the course with an introduction to computational thinking along with an overview of Python as a programming language. Following that, we will provide an overview of machine learning fundamentals and showcase some of the powerful state-of-the-art machine learning algorithms with applications. Finally, we will build on the tools and techniques to introduce neural computing as a platform for artificial intelligence that takes advantage of the architecture of neural networks and the structure of the physical-laws governing the equations describing real-world problems to make data-driven intelligent decisions.

Short course topics:

- Introduction to Computational Thinking
- Introduction to Python Basics
- Overview of Machine Learning Fundamentals
- Machine Learning Algorithms and Applications
- Introduction to Neural Computing
- Data-driven Neural Computing

Presenters

Prof. Padhu Seshaiyer is a Full Professor of Mathematical Sciences and works in the broad areas of Computational Mathematics, Data science, Numerical methods for differential equations, Mathematical Biology, Computational Biomechanics, Design Thinking and STEM Education. In particular, his research includes the development of new analytical techniques and efficient computational algorithms to obtain numerical solutions to mathematical models describing multi-physics interactions with applications to real-world problems. During the last two decades, he has initiated and directed a variety of educational programs including faculty development, post-graduate, graduate and undergraduate research, K-12 outreach, teacher professional development, and enrichment programs to foster the interest of students and teachers at all levels to apply well-developed research concepts, to fundamental applications arising in STEM disciplines. Over the years, he has won several prestigious awards and honors for his contributions to research, teaching and service. In 2019, he was one of the Plenary Speakers for the VI WCDANM conference. More details can be found at `https://math.gmu.edu/~pseshaiy/`

Mr. Alonso Gabriel Ogueda is currently a graduate student pursuing his doctoral studies with Dr. Padhu Seshaiyer on applications of Physics Informed Neural Networks. He holds a Master's degree in Mathematics from the Universidad Técnica Federico Santa María (2021) and a Mathematical Engineering degree from Universidad Técnica Federico Santa María (2019). He has worked on a variety of projects involving development of mathematical/statistical algorithms, data analysis, data science and engineering and Cloud computing. More details can be found at `https://aoguedao.github.io/cv/`

Structural equation modeling using R and SAS

Ding-Geng Chen[1] and Yiu-Fai Yung[2]

[1]Arizona State University
[2]SAS Institute Inc.

E-mail address: *Ding-Geng.Chen@asu.edu; yiu-fai.yung@sas.com*

Originated from social sciences, structural equation modeling (SEM) is becoming more popular in other fields such as education, health science, and medical sciences. This short course is aimed to provide an overview of SEM and to demonstrate its applications by using R and SAS software based on the newly published book: "Chen and Yung (2023). Structural Equation Modeling Using R/SAS: A Step-by-Step Approach with Real Data Analysis. Chapman and Hall/CRC". We will cover some main SEM topics, including path analysis, confirmatory factor analysis, structural relations with latent variables, and latent growth-curve modeling. Real-world application examples, most of which are based on our newly published SEM book (see reference), are compiled to demonstrate SEM in social, educational, behavioral, and marketing research. Mathematical and statistical foundations of SEM are discussed at a level suitable for general understanding. This course is designed for statisticians and data analysts who like to learn SEM techniques for their own research and applications. Both R package "lavaan" (latent variable analysis) and the CALIS procedure of SAS/STAT will be used to demonstrate model specifications, fitting, and result interpretations.

Attendees should have a basic understanding of regression analysis. Experience using R and SAS software is not required for understanding the general SEM techniques.

Outline (Two 2-hr Sessions)

Session 1 (2-hr) topics:

1. Overview of structural equation modeling
   - Historical background
   - Path analysis and SEM as an extension of regression analysis
   - Path diagram representations
   - Measurement errors in regressors
   - Confirmatory factor models for instrument validations
   - Combining measurement models and structural models for latent variables.
2. Statistical and mathematical backgrounds of structural equation modeling
   - Functional equations and matrix formulations
   - Estimation of parameters: Maximum-likelihood, generalized least squares, and asymptotically distribution-free method

Session 2 (2-hr) topics:

3. Latent growth-curve modeling
   - Latent growth-curve modeling with a single outcome
   - Latent growth-curve modeling with multiple outcomes
   - Latent growth-curve modeling with covariates
   - Illustrations using the Cancer Surgery data
4. Using SEM to assess direct, indirect, and total effects (if time permits)

- Definitions of direct, indirect, and total effects
- Why SEM is useful to study these effects
- Illustration: How mental abilities are affected remotely by social status

Presenters

Dr. (Din)Ding-Geng Chen is an elected fellow of American Statistical Association and an elected member of the International Statistical Institute. Currently he is the executive director and professor in biostatistics at the College of Health Solutions, Arizona State University. Dr. Chen is also an extraordinary professor and the SARChI research chair in biostatistics, at the Department of Statistics, University of Pretoria, South Africa, and an honorary professor at the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, South Africa. He was the Wallace H. Kuralt distinguished professor in Biostatistics in the University of North Carolina-Chapel Hill, a professor in Biostatistics at the University of Rochester, and the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. He is also a senior statistics consultant for biopharmaceuticals and government agencies with extensive expertise in Monte-Carlo simulations, clinical trial biostatistics and public health statistics. Dr. Chen has more than 200 referred professional publications, co-authored and co-edited 40 books on clinical trial methodology and analysis, meta-analysis, data sciences, causal inferences and public health applications. He has been invited nationally and internationally to give speeches on his research.

Dr. Yiu-Fai Yung is an analytic solution manager at the SAS Institute Inc. He has been developing commercial software for causal analysis, factor analysis, and structural equation modeling for more than 20 years. He has held several workshops and taught courses about causal analysis and structural equation modeling in conferences such as SAS Users' Group meetings, Joint Statistical Meetings, and International Meetings of Psychometric Society. Prior to joining SAS, he taught psychological and behavioral statistics at the University of North Carolina at Chapel Hill. He has published articles in Psychometrika, British Journal of Mathematical and Statistical Psychology, and Journal of Educational and Behavioral Statistics. His main research interests include latent variable modeling, mixture modeling, mediation analysis, and causal inferences.

Contributed Talks

Mathematical modeling, analysis and simulation of the spread of smoking in the United States using Optimal Control

Jonah Ascoli[1], Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]Mathematical Sciences Department, George Mason University, Fairfax, VA, United States

E-mail addresses: *jonah.ascoli@gmail.com; pseshaiy@gmu.edu; aogueda@gmu.edu*

We present a mathematical assessment of the dynamics of smoking and its public health impact. Using an infectious disease analogy, the model describes the spread of smoking through multiple subpopulations. We perform stability analysis, derive the basic reproduction number, and apply optimal control theory to minimize exposed and infected populations and maximize susceptible ones while also minimizing education costs. We show that moderate investment in education mitigates the spread of smoking.

Keywords

Optimal control, SEIR model, Basic reproduction number.

In this work, a compartmental model for smoking as an infectious disease is considered using a coupled system of ordinary differential equations. We introduce into this system the use of public education campaigns aimed at elucidating the health impacts of smoking. Specifically, we introduce a variable corresponding to education as a control [1] which causes a change in behavior resulting in two susceptible classes. Our numerical results show that education can be used as a regulatory mechanism to mitigate the spread of smoking and the basic reproduction number, a measure of the disease's transmissibility [2,?]. We hope that the model created can help provide insights to achieve maximum reduction in smoking prevalence while optimizing cost-effectiveness that will then allow policymakers to determine the optimal allocation of financial resources for such campaigns.

Acknowledgements: This work is supported in part by George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program (ASSIP), the GMU Department of Mathematical Sciences, and the National Science Foundation DMS 2230117.

References

[1] H. R. Joshi, S. Lenhart, S. Hota and F. Agusto. Optimal control of an SIR model with changing behavior through an education campaign. *Electronic Journal of Differential Equations* **2015:50**, 2015.

[2] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang and K. H. Jacobsen. Complexity of the Basic Reproduction Number (R_0). *Emerging Infectious Diseases* **25:1**, 2019.

[3] C. Ohajunwa, K. Kumar and P. Seshaiyer . Mathematical modeling, analysis, and simulation of the COVID-19 pandemic with explicit and implicit behavioral changes. Computational and Mathematical Biophysics **8(1)**, 216–232, 2020.

Topological insights into gait for Parkinsonism differentiation

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

[1]Centre of Mathematics, University of Minho, Portugal
[2]Neurology Department, Hospital da Senhora da Oliveira, Portugal
[3]School of Medicine, Life and Health Sciences Research Institute (ICVS), University of Minho, Portugal
[4]Algoritmi Centre, School of Engineering, University of Minho, Portugal

E-mail addresses: *jhonathanbarrios21@gmail.com; wolfram.erlhagen@math.uminho.pt; miguelgago@hospitaldeguimaraes.min-saude.pt; estela.bicho@dei.uminho.pt; fjferreira@math.uminho.pt*

Topological Data Analysis (TDA) has been used to generate features that distinguish gait patterns and improve machine learning classifiers in identifying gait-related pathologies. Due to the nonlinear relationships in human gait dynamics, TDA offers innovative topological and geometric tools to analyze time-series data. This work aims to implement TDA techniques to explore the topological properties of gait time series in patients with Parkinson's Disease. Gait data were collected from 34 healthy subjects and 29 Parkinson's patients using Physilog sensors. Topological descriptors were used for binary classification tasks, showing potential in distinguishing between healthy individuals and those with Parkinsonism, as well as between idiopathic and vascular Parkinsonism.

Keywords

Biomedical signal processing, Nonlinear dynamics, Time series analysis, Topological data analysis.

The analysis of time series data has seen significant advancements with new techniques that extract valuable predictive information from high-dimensional, complex datasets [1,2]. Human gait dynamics, characterized by nonlinear relationships among multiple inputs and outputs, exemplify such complexity. The incorporation of Algebraic Topology in biomedical data analysis has transformed the exploration of intricate datasets like human gait [1,3,4]. These methods reveal hidden patterns and relationships by leveraging the topological structure within gait data [1,4]. Topological Data Analysis (TDA) provides a robust framework for characterizing and classifying gait patterns, surpassing traditional linear methods. Furthermore, methodologies like phase space reconstruction have been pivotal in uncovering nonlinear dynamics within gait time series, aiding in disease assessment.

The aim of this work is to use the topological properties of gait time series to uncover new fundamental insights to help distinguish Parkinson's disease. Gait data for this work were collected using Physilog sensors (GaitUp®) to measure spatiotemporal variables in a 60-meter runner at self-selected speeds. Thirty-four healthy subjects (CO) and 29 patients with Parkinsonism (PD), including 15 patients diagnosed with idiopathic Parkinson's disease (IPD) and 14 with vascular parkinsonism (VaP), were studied. Parkinson's patients underwent evaluation in both the "Off" phase (after 24 hours without L-dopa) and the "On" phase (after a suprathreshold L-dopa test, equivalent to 150% of your morning dose).

Descriptors of Betti curves, persistence landscapes, and silhouette landscapes were extracted to analyze gait time series, serving as inputs for classification. Binary classification tasks, CO vs. PD and IPD vs. VaP (in both "Off" and "On" phases), were performed using leave-one-out cross-validation. A random forest classifier was employed in these classification tasks. The results demonstrated that these topological descriptors effectively capture the subtle differences in gait patterns associated with different conditions, offering a promising approach for distinguishing between healthy individuals and those with Parkinsonism, as well as between IPD and VaP. The use of topological features in conjunction with advanced machine learning techniques thus holds potential for enhancing diagnostic accuracy and understanding of gait dynamics in Parkinson's disease.

Acknowledgements: The authors thank the Fundação para a Ciência e a Tecnologia (FCT) for the financial support provided through the doctoral scholarship with reference 2023.02242.BDANA and the support of Portuguese funds through the Center of Mathematics of the University of Minho and FCT within the projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Y. Yan, O. O. Mumini, X. Yu-Cheng, L. Hui-Hui, L. Qiu-Hua, N. Ze-Dong, F. Jianping and W. Lei. Classification of neurodegenerative diseases via topological motion analysis-A comparison study for multiple gait fluctuations. *IEEE Access* **10.1109/ACCESS.2020.2996667**, 2020.

[2] C. Fernandes, F. Ferreira, R. Lopes, E. Bicho, W. Erlhagen, N. Sousa and M. F. Gago. Discrimination of idiopathic Parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect. *Journal of Biomechanics* **10.1016/j.jbiomech.2020.110214**, 2021.

[3] B. Safarbali and S. Hashemi Golpayegani. Nonlinear dynamic approaches to identify atrial fibrillation progression based on topological methods. *Biomedical Signal Processing and Control* **doi.org/10.1016/j.bspc.2019.101563**, 2019.

[4] Y. Yan, L. Yu-Shi, L. Cheng-Dong, W. Jia-Hong, M. Liang, X. Jing, Z. Xiu-Xu and W. Lei. Topological Descriptors of Gait Nonlinear Dynamics Toward Freezing-of-Gait Episodes Recognition in Parkinson's Disease. *IEEE Sensors Journal* **10.1109/JSEN.2022.3142750**, 2022.

Innovative approaches to breakpoint detection in electricity consumption patterns

Ana Borges[1] and Mariana Carvalho[1]

[1]CIICESI, ESTG, Polytechnic of Porto, Portugal

E-mail addresses: *aib@estg.ipp.pt; mrc@estg.ipp.pt*

An innovative strategy is proposed to detect breakpoints in monthly electricity consumption across Portuguese municipalities using methodologies adapted from hydrological data analysis on data from the E-Redes platform. Time series data is decomposed using Seasonal-Trend decomposition based on Loess (STL), followed by breakpoint analysis on the seasonally adjusted series. The Mann–Kendall test and Sen's slope estimator are then employed to analyze periods of statistical changes in consumption. Results demonstrate that this method effectively identifies significant breakpoints in energy consumption, which can be linked to policy changes, economic events, or technological advancements.

Keywords

Time Series, Breakpoint, Electricity Consumption.

Electricity consumption patterns are key indicators for understanding policy changes, economic events, and technological advancements. Detecting shifts in consumption is essential for economic, environmental, and operational reasons. Economically, it helps manage costs, forecast revenue, and plan resources. Environmentally, it supports sustainability and grid stability by facilitating demand-side management and reducing peak demand.
This study introduces a novel method for identifying monthly electricity consumption breakpoints in Portuguese municipalities using data from the E-Redes platform. It adapts techniques from hydrological data analysis, employing Seasonal-Trend decomposition based on Loess (STL), which has been successfully applied in various studies, such as those by [1] and [2]. Similarly, [5] used STL decomposition and breakpoint analysis to evaluate water meter performance, effectively identifying significant reductions in water consumption. The methodology here presented adapts and expands the approach proposed in [5], by detecting increases and decreases in monthly electricity consumption within Portuguese municipalities. Following STL decomposition, the study conducts breakpoint analysis on the seasonally adjusted time series. The Mann-Kendall test and Sen's slope estimator identify significant changes in electricity consumption, and the relative magnitude of change (RMC) quantifies the change.

To exemplify the procedure we present the results for three municipalities: Porto, Braga and Lisbon. In Porto, there were significant changes in trends in November 2020 and February 2022. The slope of the first segment was steep, indicating rapid growth, but it decelerated in November 2020 and experienced another significant shift in February 2022. The relative change suggests a decrease in the slope between segments (-0.64, 0.01). In Braga, a significant breakpoint occurred in June 2022. Prior to this, the slope was not significant, but after the breakpoint, there was a significant increase in the trend (1.62).

In Lisbon, significant changes in trends were observed in November 2020 and February 2022. The slope of the first segment was steep, but it decreased in November 2020 and

experienced another significant change in February 2022. The relative magnitude of change indicate a decrease in the slope between segments (-0.33, -0.47). The observed shifts in electricity consumption in Porto and Lisbon during November 2020 and February 2022 can be attributed to the impacts of the COVID-19 pandemic and subsequent economic recovery. In November 2020, the economic slowdown and remote work policies likely led to reduced industrial and commercial electricity usage, despite the onset of colder months. By February 2022, as economies recovered, increased industrial and commercial activities, coupled with a return to offices and potential changes in energy market dynamics, resulted in higher electricity demand.
In conclusion, the presented methodology is shown to successfully detect breakpoints in electricity consumption. As future work we intend to incorporate additional variables such as temperature, economic indicators, and policy changes to provide a complete analysis of factors influencing electricity consumption.

Acknowledgements: This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through project UIDB/04728/2020. The authors thank the hospital for providing the real data used in this study.

References

[1] C. Miller and F. Meggers. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. *Energy and Buildings* **156**, 360–373, 2017.

[2] A. E. Lafare, D. W. Peach and A. G. Hughes. Use of seasonal trend decomposition to understand groundwater behaviour in the Permo-Triassic sandstone aquifer, Eden Valley, UK. *Hydrogeology Journal* **24(1)**, 141–158, 2016.

[3] M. Gocic and S. Trajkovic. Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia. *Global and Planetary Change* **100**, 172–182, 2013.

[4] S. Sharma, D. A. Swayne and C. Obimbo. Trend analysis and change point techniques: A survey. *Energy, Ecology and Environment* **1(3)**, 123–130, 2016.

[5] C. Cordeiro, A. Borges and M. R. Ramos. A strategy to assess water meter performance: A case study. *Journal of Water Resources Planning and Management* **148(2)**, 05021027, 2021.

SDE harvesting models for populations in randomly varying environments: Impact of Allee effects

Carlos A. Braumann[1,2], Nuno M. Brites[3] and Clara Carlos[1,4]

[1]Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Portugal
[2]Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal
[3]ISEG/UL - Universidade de Lisboa, Department of Mathematics; REM - Research in Economics and Mathematics, CEMAPRE, Portugal
[4]Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal, Portugal

E-mail addresses: *braumann@uevora.pt; nbrites@iseg.ulisboa.pt; clara.carlos@estbarreiro.ips.pt*

General results on extinction and on the existence of a stochastic equilibrium are presented for general stochastic differential equation models for harvested populations living in a randomly varying environment, including populations with Allee effects. We then use stochastic optimal control theory to study profit optimization from the harvesting activity for the Pacific halibut data, using both the logistic model without Allee effects and the logistic-like model with Allee effects. The resulting optimal harvesting policy has severe shortcomings, so suboptimal policies were also considered. The comparison of the two models allows one to assess the impact of Allee effects and their intensity on the population, on the harvesting effort and profit of the different policies, and on the very design of an appropriate harvesting policy.

Keywords

Harvesting models, Stochastic differential equations, Allee effects, Profit optimization.

In a randomly varying environment, the dynamics of a harvested (say, a fish) population with size $X(t)$ can be described by a stochastic differential equation (SDE). [1] proves results on extinction and on the existence of a stochastic equilibrium that are robust w.r.t. the natural growth dynamics by using general models for population growth and harvesting. However, Allee effects were not considered.

Since some populations do show Allee effects, [4] extends the general model results to populations with Allee effects, but it did not consider harvesting. Here, we extend such results to harvested populations with Allee effects.

In [2], using data on the Pacific halibut and a specific model, the logistic model, we study the present value, i.e., the expected total discounted profit of the harvesting activity, and determine the optimal harvesting variable effort policy through stochastic optimal control theory. Unfortunately, this optimal policy is incompatible with the logistics of fisheries and therefore inapplicable. It also causes social problems, such as fishermen's unemployment during periods of no or low harvesting. Furthermore, it requires knowledge of the population size at each instant, and estimating population size is an inaccurate, lengthy, and expensive task. Therefore, [2] also studies suboptimal applicable policies, like the constant effort policy and stepwise effort policies, and compares them with the optimal policy.

For this first study of optimal and suboptimal policies, we have used the logistic model, assuming a population without Allee effects. What happens if the population has

Allee effects? Strong Allee effects were not worth studying since they led to population extinction even in the absence of harvesting. So, we consider weak Allee effects. To answer the above question, [3] and another paper of the same authors study the same harvesting policies for the same data but using a logistic-like model with weak Allee effects of different intensities and compare the results with the initial logistic model results. We present here such results and comparisons, allowing us to determine the impact of Allee effects and their intensity on population size, on harvesting efforts and profits, and on the very design of an appropriate harvesting policy. In other papers, we have also studied the (huge) impact of Allee effects on realistic extinction times.

Acknowledgements: C.A. Braumann and C. Carlos are members of the Centro de Investigação em Matemática e Aplicações, supported by Fundação para a Ciência e a Tecnologia - FCT (Portuguese Foundation for Science and Technology), Project UID/04674/2020, `https://doi.org/10.54499/UIDB/04674/2020`. N.M. Brites was partially funded by FCT, Project CEMAPRE/REM - UIDB/05069/2020, through national funds.

References

[1] C. A. Braumann. Variable effort fishing models in random environments: generalization to density-dependent noise intensities. *Mathematical Biosciences* **177 & 178**: 229–245, 2002.

[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 and C. A. Braumann. Profit optimization of stochastically fluctuating populations under harvesting: the effects of Allee effects, *Optimization*, **71(11)**: 3277–3293, 2022.

[4] C. Carlos and C. A. Braumann. General population growth models with Allee effects in a random environment. *Ecological Complexity* **30**: 26–33, 2017.

Statistical Modeling and Machine Learning: A Comparison

Ricardo Coelho[1], Isabel Natário[1,2] and Sílvia Fraile[3]

[1]Center for Mathematics and Applications (NOVA Math), Portugal
[2]Department of Mathematics, NOVA School of Science and Technology, Portugal
[3]GEOSAT, Portugal

E-mail addresses: *rpe.coelho@campus.fct.unl.pt; icn@fct.unl.pt; silvia.fraile@geosat.space*

Technological development is generating increasingly larger and structurally more complex data sets. To analyse these data set machine learning and statistical modeling can be used. Statistical modeling is centred on models for the data generation process and inference, having these models strong assumptions. On other hand, machine learning models do not depend on specific assumptions. This approach seeks to automatically discover patterns and establish relationships, where these methods to learn directly from data seeks a good predictive performance. Thus, within this context, this work aims to address the fundamentals of statistical and machine learning models, comparing these two approaches by highlighting their similarities and differences. Additionally, we will enumerate the advantages and disadvantages of these two approaches.

Keywords

Machine learning, Statistical modeling, Similarities, Differences.

In many applications of scientific areas such as health, economics, bioinformatics, environment, interest is to describe a certain phenomenon or predict a future value or state, [1]. For this, data analysis is necessary in order to establish relationships and make predictions, where lately the use of machine learning models has been favoured over traditional statistical models, [2]. Both approaches are powerful in this objective and despite may lead to very similar conclusions, are quite distinct in conception, assumptions and implementation which is not always obvious to users when applying these two approaches. The choice should be guided by the problem, by empirical evidence such as the size of the data set, the number of variables, assumptions or the lack thereof, and the expected outcomes such as prediction or causality [1].

Statistical modeling is centred on models for the data generation process and inference about the relationships between observed variables, based on strong assumptions about data distribution, the types of relationships to consider, and parsimony. These assumptions must be validated to ensure confidence in the results. Machine learning models aim to identify patterns and establish relationships automatically, using statistical and probabilistic methods to learn directly from the data, guided by the search of good predictive performance. These models are based on statistical learning theory. Statistical learning theory formalizes the model for making predictions based on data, while machine learning automates the modeling process. This approach does not rely on specific assumptions about data distribution, providing more flexibility compared to statistical modeling.

In statistical modeling various diagnostics on parameters are performed, like significance tests based on the p-value, while in machine learning models those are not generally considered, [3]. This is because machine learning models are based on less assumptions

when compared to statistical modeling, such as specific response distributions assumptions, collinearity, etc., that must be satisfied and verified before the results of fitting a statistical model can be trusted and used. Statistical modeling thus needs a greater insight of how data were obtained, the underlying distribution of the population, the statistical properties of the estimators. On the other hand, machine learning models are more suitable for predicting the outcome than to infer and establish causal relationship between the outcome and independent variables, which is is the domain of the statistical modeling. The reason for the machine learning models being less demanding on assumptions is related to their learning-from-the-data nature, the machine learning models are trained with part of the data (training set) to be able to do the better predict of the rest of the data (validation set).

Acknowledgements: This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects: UIDB/00297/2020 (`https://doi.org/10.54499/UIDB/00297/2020`) and UIDP/00297/2020 (`https://doi.org/10.54499/UIDP/00297/2020`) (Center for Mathematics and Applications). National funding by FCT, through the individual research grant 2023.01166.BDANA of Ricardo Coelho.

References

[1] Michele Bennett et al. Evaluating Similarities and Differences between Machine Learning and Traditional Statistical Modeling in Healthcare Analytics. *In: Artificial Intelligence Annual* Volume 2022. Ed. by Marco Antonio Aceves Fernandez and Carlos M. Travieso-Gonzalez. Rijeka: IntechOpen, 2022. Chap. 2. doi: 10.5772/intechopen.105116. `https://doi.org/10.5772/intechopen.105116`.

[2] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. Points of Significance: Statistics versus machine learning. *In: Nature Methods* **15.4** 233–234, 2018, issn: 15487105. doi: 10.1038/nmeth.4642. `http://dx.doi.org/10.1038/nmeth.4642`.

[3] Pratap Dangeti. *Statistics for Machine Learning*. Packt Publishing, 2017. isbn: 9781788291224.

Grape maturation clustering and vineyard management in the vinhos verdes region: an analysis from 2007 to 2023

Alexander Cornejo[4], Mafalda Costa[3], Nelson Costa[2], Óscar Pereira[3] and Aldina Correia[1]

[1]CIICESI, ESTG, Instituto Politécnico do Porto, Portugal
[2]ISEP, Instituto Politécnico do Porto, Portugal
[3]Comissão de Coordenação e Desenvolvimento Regional do Norte, Portugal
[4]Comissão de Viticultura da Região ao dos Vinhos Verdes, Portugal

E-mail addresses: [1] *aic@estg.ipp.pt;* [2] *nfc@isep.ipp.pt; mafaldatcosta@gmail.com*

The quantification of physico-chemical parameters such as acidity, alcohol content, pH and weight of grapes during the final ripening stage is crucial for determining the optimal harvesting dates and thus positively influencing the quality of the wine produced. However, climate change is disrupting the biological cycles of vineyards, requiring growers to adapt to new production realities.
The analysis of data from the annual control cycle, together with climate data, provides valuable insights into future scenarios, allowing vineyard owners to make informed decisions and effectively adapt to climate change.
In the Vinhos Verdes region (northern Portugal), grape samples were collected during the ripening phase from 2007 to 2023. Using data from the Loureiro and Alvarinho grape varieties, the study analysed the variation of these characteristics over time, classified harvesting locations and created future scenarios. Based on these scenarios, vineyard owners will be able to select grape varieties that are more resistant to climate change, implement adaptive vineyard management practices that mitigate the effects of extreme weather events such as frost, heat waves and drought, and explore innovative winemaking techniques that improve wine quality and stability in response to the digitalisation of environmental conditions. This will improve the sustainability and profitability of their vineyards in the context of evolving climate patterns.

Keywords

Grape Maturation, Vineyard Management, Longitudinal clustering, Classification, Decision Support.

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

References

[1] G. D. Batty, T. C. Russ, E. Stamatakis and M. Kivimäk. Psychological distress in relation to site specific cancer mortality: pooling of unpublished data from 16 prospective cohort studies. *BMJ* **356: j108**, 2017.

[2] C. Genolini and B. Falissard . KmL: A package to cluster longitudinal data. *Computer methods and programs in biomedicine*, **104**(3), e112–e121, 2011.

[3] M. Ranalli and R. Rocci. Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, **26**, 529–547, 2016.

[4] M. Ranalli and R. Rocci. A model-based approach to simultaneous clustering and dimensional reduction of ordinal data. *Psychometrika*, **82**, 1007–1034, 2017.

Veracity evaluation of ChatGPT's mathematical solutions via m-polar fuzzy graphs

Mariana Durcheva[1] and Philip Slobodsky[2]

[1]Sami Shamoon College of Engineering, Israel
[2] Halomda Educational Software, Israel

E-mail addresses: *mariadu@sce.ac.il; halomda@netvision.net.il*

Research on the application of ChatGPT for educational purposes has recently gained significant attention. Students can trust ChatGPT to a certain extent for homework and exam preparation, but its correctness can vary. The explanations provided by ChatGPT are often clear and detailed, which can be beneficial for understanding concepts. It is observed that the correctness of its responses depends on the subject area: in Math and Exact Sciences, ChatGPT's performance can be diverse; it can solve standard problems correctly but might struggle with complex or non-standard problems. In Humanities and Social Sciences, as well as in Programming and Technical Fields, ChatGPT tends to perform well. Recent studies have assessed ChatGPT's mathematical performance, highlighting both its strengths and limitations. ChatGPT excels in areas like elementary arithmetic and logic problems, but its accuracy diminishes as problem complexity increases. Its performance varies across different mathematical topics and difficulty levels, showing proficiency in some areas while struggling in others. In this work, we leverage the abilities of m-polar fuzzy graphs to tackle decision-making problems. Our innovative approach is by using graph theory, particularly m-polar fuzzy graphs, to analyze ChatGPT's reliability. We suggest utilizing graph theoretical concepts to model the relationship between different subject areas and ChatGPT's performance. To this aim, we represented and analyzed the varying degrees of trust (as fuzzy values) students can place in ChatGPT across different subjects. In this model, we evaluate ChatGPT's responses to mathematics problems using m-polar fuzzy graphs. The nodes represent different fields selected to our research, and the edges represent the relationships between the responses given by ChatGPT in these fields. We considered eight fields of mathematics: Elementary Algebra, Linear Algebra, Elementary Geometry, Combinatorics, Differential Calculus, Integral Calculus, Differential Equations, and Complex Functions. In each field, we posed 5 problems to ChatGPT, making a total of 40 problems. We graded its responses from completely correct (1) to absolutely incorrect (0). The membership value of each node $(m_1, m_2, m_3, m_4, m_5)$ is determined based on the correctness and truthfulness of ChatGPT's responses to each posed problem. Membership values are represented as a 5-polar fuzzy subset to account for uncertainty. We constructed a 5-polar fuzzy evaluation graph corresponding to the evaluation of ChatGPT's responses. By investigating the d_2-degrees of edges, we aim to answer the critical question for both students and educators: which math fields can we rely on ChatGPT's responses, and where should we be more cautious?

Keywords

ChatGPT, m-polar fuzzy graph, Evaluation graph.

ChatGPT's performance varied across different mathematical competitions, excelling in some while struggling in others [1]; the probability of failure increased linearly with the number of addition and subtraction operations [2].

Classical graph theory, which is based on classical propositional logic, is not suitable for modeling different real-life problems. To address this, a concept of fuzzy graphs, and later, bipolar fuzzy graphs was introduced. Extending bipolar fuzzy sets and fuzzy graphs, the notion of m-polar fuzzy sets [3] and m-polar fuzzy graphs was presented [4].

Definition 1 [4] An m-polar fuzzy graph $G = (V, \sigma, \mu)$ is a triple consisting of a nonempty set V together with a pair of functions $\sigma : V \to [0,1]^m$ and $\mu : E = V \times V \to [0,1]^m$ where σ is an m-polar fuzzy set on the set of vertices V and μ is an m-polar fuzzy relation in V such that for all $x, y \in V$, $\mu(xy) \leq \sigma(x) \wedge \sigma(y)$ where $\wedge$ stands for minimum.

Definition 2 [5] The d_2–degree of a vertex $x \in V$ in m-polar fuzzy graph G is termed as: $d_2(x) = (d_2^{(1)}(x), d_2^{(2)}(x), \ldots, d_2^{(m)}(x))$, with $d_2^{(i)}(x) = \sum_i p_i \circ \mu^2(xy)$, where $p_i \circ \mu^2(xy) = \sup\{p_i \circ \mu^2(xz) \wedge p_i \circ \mu^2(zy)\}, i = 1, \ldots, m$. Here x, z, y is the shortest path of length 2 connecting vertices x and y.

References

[1] X. Dao and N. Le. Investigating the Effectiveness of ChatGPT in Mathematical Reasoning and Problem Solving: Evidence from the Vietnamese National High School Graduation Examination. *ArXiv, abs/2306.06331*, 2023.

[2] P. Shakarian, A. Koyyalamudi, N., Ngu and L. Mareedu. An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP) *ArXiv, abs/2302.13814*, 2023.

[3] J. Chen, S. Li, S. Ma and X. Wang, m-polar fuzzy sets: an extension of bipolar fuzzy sets *Hindwai Publishing Corporation, The Scientific World J.* vol. 2014. Article Id 416530.

[4] G. Ghorai and M. Pal. A Study on m-polar Fuzzy Graphs *J. Comp. Sci. Math.* 7(3):283-292, 2016.

[5] M. Akram. m–Polar Fuzzy Graphs. Theory, Methods & Applications. *Springer Nature Switzerland AG 2019* ISSN 1434–9922.

Finding the tail of a distribution: analysis of a method based on the coefficient of variation

Marta Ferreira[1,2] and Elisa Moreira[2]

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

E-mail addresses: *msferreira@math.uminho.pt; a103613@alunos.uminho.pt*

Estimation of extreme values concentrates on the tails of data, where observations are limited. A crucial aspect of making inferences on extreme values is determining where the tail begins. This is a well-explored subject in literature, with multiple approaches devised for this task. The present study seeks to examine one of these approaches through a computational simulation analysis. We assess the advantages and limitations of the method that could benefit practitioners. To conclude, we apply the method to real data and estimate parameters of rare events.

Keywords

Extreme value theory, Tail index, Threshold selection.

The main result in extreme value theory is the extremal types theorem which states the possible limiting laws of linear normalized maximum, commonly summarized by a generalized extreme values (GEV) distribution function (df) [3,4]. An equivalent approach is to consider the largest observations above a high threshold t, whose df can be well approximated by a generalized Pareto (GP) [1,7]. Both models share the same shape parameter γ denoted tail index, that rules the type of tail.

The main question is: from which threshold t we have X_t well approximated by a GP model? A common approach is to consider the mean excess plot based on the mean excess function of a GP model, which exists for $\gamma < 1$ and is a linear function of threshold t [2]. The mean excess plot is the respective empirical counterpart, and we choose threshold t from which one has linearity.

An alternative approach based on the coefficient of variation (CV) of a GP was proposed in [3], which is a constant function only depending on the tail index γ, thus being easier to apply than the mean excess function. Applications of the method on real data have been carried out in [3,4]. However, to the best of our knowledge, no simulation studies to evaluate the method's performance have yet been presented. Here is the objective of this work, along with formulation of practical rules to better guide users in applying the method. An application to financial data is presented where we infer tail measures like *Value-at-Risk* and *Expected Shortfall*.

Acknowledgements: The research at CMAT was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013 /2020 and UIDP/00013/2020.

References

[1] A. A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability* **2**, 792–804, 1974.

[2] A. C. Davison and R. L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)* **52**, 393–425, 1990.

[3] J. del Castillo and M. Padilla. Modeling extreme values by the residual coefficient of variation. *SORT-Statistics and Operations Research Transactions* **40(2)**, 303–320, 2016.

[4] J. del Castillo, D. M. Soler and I. Serra. ercv: Fitting Tails by the Empirical Residual Coefficient of Variation. *R package version 1.0.1.*, 2019.

[5] B. V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics* **44(6)**, 423–453, 1943.

[6] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24(2)**, 180–190, 1928.

[7] J. Pickands III. Statistical inference using extreme order statistics. *The Annals of Statistics* **3(1)**, 119–131, 1975.

Classification for a folded directional distribution

Adelaide Figueiredo[1] and Fernanda Figueiredo[2]

[1]University of Porto, School of Economics and Management and LIAAD-INESC TEC, Portugal
[2]University of Porto, School of Economics and Management and CEAUL, University of Lisbon, Portugal

E-mail addresses: *adelaide@fep.up.pt; otilia@fep.up.pt*

When the directional data fall on the positive orthant of the unit hypersphere, the data can be modeled with a folded directional distribution, such as the folded Watson distribution. In this study we consider the Bayes classification rules for this distribution and we analyse the performance of these rules in a simulation study. We also present an example using spherical data available in the literature.

Keywords

Classification rules, Directional data, Folded distribution, Watson distribution.

Directional data are unit vectors on the surface of the hypersphere, S^{p-1}. There are many applications of directional data in various fields, such as machine learning, text analysis, bioinformatics, genetics, neurology, etc (see, for example, [2] and [3]).

The Watson distribution is one of the most widely used distributions for modeling axial data (see, for example, [1] and [3]). This distribution has two parameters: a modal direction and a concentration parameter around the directional parameter.

The Bayes classification rules for the Watson distribution defined on the hypersphere were proposed in [1]. In some situations, the directional data fall on the positive orthant of the unit hypersphere. This occurs, for example, when compositional data are transformed into directional data using the square root transformation. When applying this transformation, we do not know the signal of the components of the vectors, so the resulting data can be modeled with a folded Watson distribution. If $\mathbf{X} = (X_1, X_2, ..., X_p)'$ has a Watson distribution on the unit hypersphere S^{p-1}, then $\mathbf{Y} = |\mathbf{X}| = (|X_1|, |X_2|, ..., |X_p|)'$ has a folded Watson distribution on the positive orthant of the unit hypersphere S_+^{p-1}.

In this study, we consider the Bayes classification rules for a folded Watson distribution. Then, due to the complexity of the folded Watson density, we analyze the performance of the Bayes classification rule by simulation. We also present an example using real spherical data from the literature.

Acknowledgements: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects LA/P/0063/2020 and UIDB/00006/2020, DOI 10.54499/LA/P/0063/2020 and DOI: 10.54499/UIDB/00006/2020, `https://doi.org/10.54499/LA/P/0063/2020`, `https://doi.org/10.54499/UIDB/00006/2020`.

References

[1] A. Figueiredo and P. Gomes. Discriminant analysis based on the Watson distribution defined on the hypersphere. *Statistics* **40:5**, 435–445, 2006.

[2] A. Banerjee, I. S. Dhillon, J. Ghosh and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* **6**, 1345–1382, 2005.

[3] J. L. Dortet-Bernadet and N. Wicker. Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics* **9:1**, 66–80, 2008.

[4] S. Sra and D. Karp. The multivariate Watson distribution: Maximum-likelihood estimation. *Journal of Multivariate Analysis* **114**, 256–269, 2013.

Multivariate analysis of some circular economy indicators

Fernanda Otília Figueiredo[1,2] and Adelaide Figueiredo[1,3]

[1]University of Porto, School of Economics and Management, Rua Dr Roberto Frias, 4200 s/n, Porto, Portugal
[2]CEAUL - Centre of Statistics and Applications, University of Lisbon, Portugal
[3]INESC TEC, University of Porto, Portugal

E-mail addresses: *otilia@fep.up.pt; adelaide@fep.up.pt*

This study aims to understand the progress made by EU countries in recent years in the following key areas of the circular economy: production and consumption, waste management, competitiveness and innovation, and global sustainability. Data for some indicators related to these areas have been collected from the Eurostat database for the period 2013-2021. The method of double principal component analysis has been used to analyze the data, which allows to have some insight on the evolution of countries and indicators over this period.

Keywords

Circular economy, Double principal component analysis, Eurostat indicators.

Our current economy is based on a model of extracting raw materials from nature, transforming them into products, and then discarding them as waste. The circular economy, on the other hand, aims to minimize waste and promote the sustainable use of natural resources by keeping products in use for longer through repair, recycling and redesign. This new economic model reduces the pollution and carbon emissions, protecting the environment, and contributes to a healthier living and to solve the challenges of climate change and biodiversity loss.

This study aims to understand the progress made by EU countries in recent years on some key areas of the circular economy. It focuses on production and consumption, waste management, competitiveness and innovation, and global sustainability, and we have collected data of the following indicators (variables), over the period 2013-2021, which will be analyzed using the double principal component analysis, a method introduced by Bouroche[1]:

- Resource productivity;
- Generation of municipal waste per capita;
- Recycling rate of municipal waste;
- Circular material use rate;
- Private investment related to circular economy sectors;
- Gross added value related to circular economy sectors;
- Persons employed in circular economy sectors;
- Greenhouse gases emissions from production activities.

Acknowledgements: This work is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the scope of the projects UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020) and LA/P/0063/2020 `https://doi.org/10.54499/LA/P/0063/2020`.

References

[1] J. M. Bouroche. *Analyse des donnés ternaire: la double analyse en composantes principales.* 3rd-cycle PhD thesis, Université de Paris VI, Paris, 1975.

An empirical study of the CDPCA on high-dimensional data sets

Adelaide Freitas[1,2] and Maurizio Vichi[3]

[1]University of Aveiro, Portugal
[2]Center for Research & Development in Mathematics and Applications, Portugal
[3]University of La Sapienza, Italy

E-mail addresses: *adelaide@ua.pt; maurizio.vichi@uniroma1.it*

Clustering and Disjoint Principal Component Analysis (CDPCA) is a constrained Principal Component Analysis that promotes sparsity in components while clustering objects. Using simulated and real gene expression datasets with more variables than objects, we evaluate the Alternating Least Square (ALS) algorithm's performance for CDPCA.

Keywords
Principal component analysis, Clustering of objects, Sparsity, Between cluster deviance.

CDPCA is a two-mode methodology originally proposed by [1] that is aimed at a simultaneous clustering of objects along a set of centroids and a partitioning of variables along a set of components in order to maximize the between cluster deviance of these components in the reduced space. In this way, the components are defined by disjoint classes of variables and then easier for interpretation purposes.

Based on simulated and real gene expression data sets where the number of variables is higher than the number of the objects, in [2] the performance of the Alternating Least Square (ALS) algorithm is compared with a procedure that use an approximation algorithmic framework based on a semidefinite programming approach. In this talk we focus on the evaluation of the ALS algorithm to perform CDPCA.

Our numerical tests show that ALS performs well and produces satisfactory results in terms of solution precision. In recovering the true object clusters, the complexity of the data structure (i.e., the error level of the CDPCA model on which the data was generated) seems to influence the ability of ALS when the sample size is lower. For lower sample size, ALS algorithm performs better when the error level is lower. The proportion of explained variance by each CDPCA component estimated by ALS is affected by the data structure complexity (higher error level, the lower explained variance).

Acknowledgements: This work has received funding from CIDMA – Center for Research and Development in Mathematics and Applications of the University of Aveiro, through the Portuguese Foundation: Fundação para a Ciência e a Tecnologia, I.P. (FCT, Funder ID = 50110000187) under the projects: `https://doi.org/10.54499/UIDB/04106/2020)` and `https://doi.org/10.54499/UIDP/04106/2020`.

References

[1] M. Vichi and G. Saporta. Clustering and disjoint principal component analysis, *Computational Statistics & Data Analysis*, **53**, 3194–3208, 2009.

[2] A. Freitas, E. Macedo and M. Vichi. An empirical comparison of two approaches for CDPCA in high-dimensional data, *Statistical Methods & Applications*, **30**, 1007–1031, 2021.

Prediction of key parameters and simulation in Asthma disease model Induced by pollution using Physics-Informed Neural Networks

Aadi Gannavaram[1], Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]Mathematical Sciences Department, George Mason University, Fairfax, VA

E-mail addresses: *aadigannavaram@gmail.com; pseshaiy@gmu.edu; aogueda@gmu.edu*

Asthma is a chronic lung condition in which the airways become inflamed and narrow, and overproduce mucus, making breathing difficult. [1] Asthma affects 262 million people worldwide, and as exposure to pollution is a key risk factor in developing conditions, increasing urbanization is often accompanied by an increase in asthma prevalence, particularly in lesser-developed regions. [2] To examine the relationship between asthma and pollution, we constructed a system of ordinary differential equations using a compartmental model with susceptible, exposed, and infected components, as well as a pollutants component to act as a pathogen. The parameters of the model that describe the relationship between components are difficult to measure and are currently unavailable. Thus, a physics-informed neural network (PINN) approach was taken to compute parameter values for a given dataset. [3] The network was trained on artificially generated time series data of each component. The error on estimated parameters was calculated and optimized based on known input parameters to the system. Once the error is reduced via hyperparameter tuning, the network can compute accurate parameters given real-world time series data to allow for realistic modeling of asthma-pollution epidemiology. This establishes PINNs as a method for prediction of key parameters in the model that allow for improved disease forecasting capability.

Keywords

Asthma, Compartmental models, Physics-Informed Neural Networks, Parameter estimation.

Acknowledgements: This work is supported in part by the George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program (ASSIP) and the GMU Department of Mathematical Sciences.

References

[1] S. Holgate, S. Wenzel, D. Postma et al. Asthma. Nat Rev Dis Primers 1, 15025, 2015. https://doi.org/10.1038/nrdp.2015.25

[2] V.J. Clemente-Suárez, J. Mielgo-Ayuso, D.J. Ramos-Campo, A.I. Beltran-Velasco, I. Martínez-Guardado, E. Navarro Jimenez, L. Redondo-Flórez, R. Yáñez-Sepúlveda and J.F. Tornero-Aguilera. Basis of preventive and non-pharmacological interventions in asthma. *Front Public Health.* 2023 Oct 18;11:1172391. https://doi.org/10.3389/fpubh.2023.1172391. PMID: 37920579; PMCID: PMC10619920.

[3] Long Nguyen, Maziar Raissi and Padmanabhan Seshaiyer. Modeling, Analysis and Physics Informed Neural Network approaches for studying the dynamics of COVID-19 involving human-human and human-pathogen interaction, *Computational and Mathematical Biophysics*, **10(1)**, 1–17, 2022. https://doi.org/10.1515/cmb-2022-0001

Effects of dietary diversity on growth outcomes of children aged 6-23 months in south and southeast Asian countries

Sarada Ghosh[1,2]

[1]Division of Nutritional Sciences, Cornell University, Ithaca, NY, USA
[2]Department of Statistics, Gurudas College, Kolkata, India

E-mail addresses: *saradaghosha111@gmail.com; sg2283@cornell.edu*

This study evaluates the impact of dietary diversity [1] on growth outcomes of children aged 6-23 months [2] in South and Southeast Asia using national health survey data, applying multivariable logistic regression while controlling for other covariates. Higher dietary diversity is expected to improve growth outcomes [3] and show regional variations. This research provides insights into dietary diversity's role in promoting healthy growth and informs targeted interventions for children in diverse settings.

Keywords

Dietary diversity, Stunting, Wasting, Underweight.

Acknowledgements: I would like to express our gratitude to the Demographic and Health Surveys for providing the valuable data used in this study.

References

[1] World Health Organization and United Nations Children's Fund (UNICEF). Indicators for assessing infant and young child feeding practices: Definitions and measurement methods. `https://apps.who.int/iris/handle/10665/340706` 2021.

[2] WHO. Malnutrition. 2021. Available online: `https://www.who.int/news-room/fact-sheets/detail/malnutrition` 2021 (accessed in 26.Oct.2022).

[3] A. Motbainor, A. Worku and A. Kumie. Stunting is associated with food diversity while wasting with food insecurity among under-five children in East and West Gojjam Zones of Amhara Region, Ethiopia. *PLoS One* **10:e0133542**, 2015.

More accurate extremal index estimation with Jackknife techniques

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

[1]NOVA School of Science and Technology, NOVA FCT, Portugal
[2]Instituto Superior de Agronomia (ISA), University of Lisbon, Portugal
[3]Center for Mathematics and Applications (NOVA Math), NOVA FCT, Portugal
[4]Centre of Statistics and its Applications (CEAUL), Portugal

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

In real life it is often observed that extremes cluster in time. Such clustering is accommodated by Extreme Value Theory via a parameter known as the *extremal index*. Therefore, the estimation of the *extremal index* is a topic of great interest. However, several challenges persist. One such challenge involves determining the appropriate number of upper order statistics to consider in semiparametric estimation. Overall, the concept of employing Jackknife Techniques for estimating the *extremal index* remains relatively underexplored, promising to offer novel insights.

Keywords

Extreme value theory, Extremal Index, Semiparametric Estimation, Jackknife Methodology.

Extreme value theory offers limiting distributions for rare events across a broad range of stationary time series. This theory deals not only with the magnitude of extremes but also with the frequency of their occurrence. Our attention in this context is directed towards understanding the clustering pattern of extremes. Frequently, it is observed that extremes, such as temperatures, water levels, wind speeds, or financial time series, exhibit clustering behavior over time. In other words, they do not occur randomly, as one would expect from a Poisson process. Extreme value theory can incorporate such clustering tendencies through the *extremal index*, while maintaining the unchanged shape of the Generalized Extreme Value (GEV) distribution. Therefore, the estimation of the *extremal index* is a topic of great interest. Its estimation has been addressed by numerous authors; however, several challenges persist. One such challenge involves determining the appropriate number of upper order statistics to consider in semiparametric estimation.

Most semiparametric estimators of this parameter show the same behavior: nice asymptotic properties but a high variance for small values of k, the number of upper order statistics used in the estimation and a high bias for large values of k. The Mean Square Error, a measure that encompasses bias and variance, usually shows a very sharp plot, needing an adequate choice of k. Using classical *extremal index* estimators considered in the literature, the emphasis is now directed to derive reduced bias estimators with more stable paths, obtained through Jackknife techniques, see [1] and [2]. An adaptive algorithm for estimating the level k for obtaining a reliable estimate of the *extremal index* is used. This algorithm has shown good results, but some improvements are still required. A simulation study will illustrate the properties of the estimators and the performance of the adaptive algorithm proposed, see [3].

Acknowledgements: This work has received funding from national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects

UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (Center for Mathematics and Applications) and under the scope of the project UIDB/00006/2020 (CEAUL) https://doi.org/10.54499/UIDB/00006/2020.

References

[1] M. I. Gomes, A. Hall and 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, M. J. Martins and M. M. Neves. Generalized Jackknife-Based Estimators for Univariate Extreme-Value Modeling. *Communications in Statistics - Theory and Methods* **42(7)**, 1227–1245, 2013.

[3] M. M. Neves, M. I. Gomes, F. Figueiredo and D. P. Gomes. Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation. *Journal of Statistical Theory and Practice* **9(1)**, 184–199, 2015.

Time series and risk analysis in the assessment of surface water quality in a river basin

A. Manuela Gonçalves[1,2], Irene Brito[1,2] and Ana Cristina Pedra[1]

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

E-mail addresses: *mneves@math.uminho.pt; ireneb@math.uminho.pt; pg46704@alunos.uminho.pt*

This study combines statistical methodologies (clustering, and time series modeling) and concepts from risk theory, to predict and analyze the surface water quality in a river basin. The aim is to propose an average risk index predictor for water pollution, and to investigate if the overall ranking result (obtained in the in-sample period) can serve as a risk index for future water pollution risk forecasts. The methodologies are illustrated using a data set of surface water quality variables in the Douro River basin (in Portugal).

Keywords

Clustering, Time series, Risk measures, Douro River basin, Surface water quality.

Water is a limited, irreplaceable and indispensable natural resource. In this sense, monitoring its quality is essential to avoid environmental and public health problems. Statistical methods are important tools for controlling and forecasting changes regarding surface water quality. The main objective of this work is to develop new methodologies by combining concepts from risk theory and times series approaches in order to predict and analyze surface water quality. The methodologies are illustrated using a data set regarding the Douro River basin (in Portugal) in terms of environmental water quality variables, measured monthly by 18 monitoring stations and recorded in the period from January 2002 to December 2013 [3].

A cluster analysis was carried out to group homogeneous stations in terms of surface water quality variables: dissolved oxygen (DO) and conductivity [1]. Since water quality depends on the flow variation, this approach differentiates and studies separately the time horizon from May to September (the dry period) and the time horizon from October to April (the wet period), after considering all the data. Several risk measures [2], such as value at risk, probability of excess, among others were determined for the considered clusters in order to assess the risk of water pollution for each one. Time series modeling approaches, such as SARIMA models and exponential smoothing methods, [4], were applied to the clusters to obtain predictions for the last 12 months. The results were compared with the classifications obtained through risk measurements. This identifies clusters (sampling stations) at the highest risk and concludes that pollution levels are higher in the dry period [3]. The cluster time series analysis reveals a clear seasonality pattern and suggest that the most appropriate approach to forecast values depend on the variable under study. Also, in order to analyze the risk of water pollution in the water station clusters, different risk measures, such as mean, and variance are calculated based on the DO and conductivity measurements in the in-sample period (from January 2002 to December 2012). The water station clusters are classified in terms of risk according to each risk measure, and a final ranking is established, considering the total, the dry and the wet periods. These results were compared with the ranking obtained from the average

predictions and forecasts of the time series models, in the in-sample period and in the out-of-sample period (between January 2013 to December 2013), to analyze the performance of the risk index predictor using the time series forecasts.

These methodologies can be easily applied to other river basins suffering from similar environmental problems.

Acknowledgements: A. Manuela Gonçalves and Irene Brito thank support from FCT through the projects UIDP/00013/2020 (`https://doi.org/10.54499/UIDP/00013/2020`) and UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020. Ana Pedra thanks CMAT for the research fellowship (BI) UMINHO/BIM/2022/100.

References

[1] SNIRH. Sistema Nacional de Informação de Recursos Hídricos. `https://snirh.apambiente.pt/` 2023 (accessed in 15 March 2023).

[2] J. Ganoulis. Clustering and forecasting of dissolved oxygen concentration on a river basin. *Risk Analysis of Water Pollution*. Wiley, Weinheim, 2009.

[3] H. W. Brachinger and M. Weber. Risk as a primitive: A survey of measures of perceived risk. *Operations-Research-Spektrum* **19**, 235–250, 1997.

[4] R. J. Hyndman, and G. Athanasopoulos. *Forecasting: principles and practice*. 2nd Ed., `https://otexts.com/fpp2/` 2018 (accessed in 15 March 2023).

[5] A. M. Gonçalves and M. Costa. Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*. **25**, 151–163, 2011.

Modeling the reflective higher-order construct 'student burnout' using the disjoint two-stage approach with PLS-SEM

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 this study, the structural equation modeling (SEM) was used to assess the predictors of the second-order construct 'student burnout', where 'emotional exhaustion', 'disbelief' and 'personal efficacy' were considered first-order constructs. Higher-order constructs have the main advantage of leading to more parsimonious models, as they allow reducing the number of relationships in the path model. The partial least squares (PLS) estimator has been used to investigate models at a higher level of abstraction. Therefore, a model was estimated using the reflective-reflective disjoint two-stage approach with the PLS-SEM, based on a sample collected through a survey carried out at a Portuguese Polytechnic Institute. As expected, the exogenous construct 'optimism' has a negative effect on the 'perceived stress' and also on the 'student burnout'. Moreover, the mediator construct 'perceived stress' has a positive direct effect on 'student burnout'. The estimated model provides some important information on the relationship between the variables involved and it can be used to analyze the global burnout score of the students' institution.

Keywords

Optimism, Second-order construct, Stress, Structural equation modeling.

As in some specialist literature [4], we specified burnout in college students as a higher-order reflective construct (a second-order construct in this case), with the three lower-order reflective constructs (first-order): 'emotional exhaustion' (EE), 'disbelief' (DB) and 'personal efficacy' (PE), which constitute the three-factor conceptualization of burnout in students, in the Maslach Burnout Inventory – Student Survey (MBI-SS) scale. The latent constructs 'optimism' and 'stress' have also been considered as predictors of burnout in some studies [2]. Therefore, to assess 'optimism', 'perceived stress', and 'student burnout', we developed a questionnaire incorporating the Revised Life Orientation Test (LOT-R), Perceived Stress Scale (PSS), and MBI-SS, respectively. An online survey at a Portuguese Polytechnic Institute was used to collect a nonrandom sample of 144 students (exactly 50% of each sex), being 57 (approximately 39.6%) from the Technology School and the remaining students from the Management School. Since the assumption of multivariate normality does not hold and the available sample is not large, we applied the disjoint two-stage approach (available in SmartPLS 4, [1]) to estimate a reflective-reflective endogenous second-order reflective construct, college student burnout, with the non-parametric

method partial least squares structural equation modeling (PLS-SEM) [3,5]. In the first stage, we estimated the model that connects all lower-order constructs (i.e., without the second-order construct 'student burnout') in the path model. Then, we evaluated the reflective measurement model of the lower-order constructs, namely reliability analysis, convergent and discriminant validity. In the second stage, we used the scores saved earlier (for the lower-order constructs: 'EE', 'DB' and 'PE') to measure the higher-order construct 'student burnout', with all other constructs in the path model estimated using their standard multi-item measures as in stage one [5]. The reflective measurement model of the higher-order construct was assessed and then the structural model was also evaluated, in particular the significance and relevance for the path coefficients, Q^2 and PLS predict. All predicted Q^2 values are greater than zero, which indicates that the model outperforms the most naïve benchmark. The comparison of PLS-SEM with a Linear Model (LM) the Mean Absolute Error (MAE) values indicate that the model has a good prediction quality.

Acknowledgements: This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/04674/2020, DOI 10.54499/UIDB/04674/2020 (https://doi.org/10.54499/UIDB/04674/2020).

References

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

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

[5] M. Sarstedt, J.F. Hair Jr and J.-H. Cheah et al., How to specify, estimate, and validate higher-order constructs in PLS-SEM, *AustralasianMarketingJournal*, 2019.

Control of Tuberculosis epidemic in South Africa using a Multi-stage Stochastic Recourse approach for resource allocation under various transmission rates.

Simran Gupta[1], Raina Saha[1] and Padmanabhan Seshaiyer[1]

[1]Department of Mathematical Sciences, George Mason University, Fairfax, VA

E-mail addresses: *simrangupta0824@gmail.com; rsaha3@gmu.edu; pseshaiy@gmu.edu*

Tuberculosis (TB), caused by bacteria Mycobacterium tuberculosis, is one of the leading infectious diseases globally. Every year, 10 million people fall ill with tuberculosis, and despite being a preventable and curable disease, TB kills 1.5 million people every year. South Africa, as of 2022, is on WHO's list of 30 countries with a high burden of tuberculosis and has an incidence rate of 615 per 100,000. The TB epidemic has proliferated in South Africa due in part to the HIV population. HIV is one of the most significant risk factors in TB spread. This is due to individuals with HIV having a greater rate of active TB, meaning a greater chance of spreading the disease and a greater need for health resources. While this connection between HIV and TB has significantly been researched, the implications on the budget for multiple locations have not. This work presents a comprehensive multistage stochastic recourse method applied to an epidemic compartmental model for TB dynamics. This model is analyzed for various TB transmission rates, taking into account various biological, environmental, and socioeconomic factors in South Africa. The mathematical model incorporates the progression from latent TB infection to active disease, accounting for variation in susceptibility, infectiousness, and treatment responses. We employ discretized differential equations to describe the interaction between susceptible, infected, and recovered populations, and incorporate stochastic transmission rates to capture the inherent randomness in disease spread and intervention impacts. Sensitivity analyses identify key parameters influencing disease dynamics, highlighting critical intervention points for effective TB control. Another contribution of this work involves the development of a Graphical User Interface to allow users to input their own values and determine the effect of different parameters on disease flow. Our results underscore the importance of early detection and targeted public health strategies. The model serves as a robust tool for policymakers to simulate various scenarios and optimize TB control measures, ultimately contributing to the global efforts in eradicating this enduring public health challenge.

Keywords

Tuberculosis, Multistage Stochastic Recourse Method, Compartmental Model, Optimization, Epidemiology.

Acknowledgements: This work is supported in part by George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program and the GMU Department of Mathematical Sciences.

References

[1] Centers for Disease Control and Prevention. HIV and Tuberculosis Overview: South Africa. Global HIV and TB, June 2024.

[2] J. Janson, Frederick Marais, Shaheen Mehtar and RMPM Baltussen. Costs and process of in-patient tuberculosis management at a central academic hospital, cape town, south africa. *Public health action*, **2(3)**, 61–65, 2012.

[3] Mmamapudi Kubjane, Muhammad Osman, Andrew Boulle, and Leigh F Johnson. The impact of hiv and tuberculosis interventions on south african adult tuberculosis trends, 1990-2019: a mathematical modeling analysis. *International Journal of Infectious Diseases*, **122**, 811–819, 2022.

[4] Xuecheng Yin and I Esra Büyüktahtakın. A multi-stage stochastic programming approach to epidemic resource allocation with equity considerations. *Health Care Management Science*, **24(3)**, 597–622, 2021.

Drivers of Bank and Trade Credit for SMEs in Portugal

Carla Henriques[1,2], Pedro Pinto[1,3] and Carolina Cardoso[1]

[1]School of Technology and Management of Viseu, Polytechnic Institute of Viseu, Portugal
[2]Centre for Mathematics of the University of Coimbra (CMUC), Portugal
[3]Research Centre in Digital Services (CISeD), Portugal

E-mail addresses: *carlahenriq@estgv.ipv.pt; spinto@estgv.ipv.pt; carolinaesteves96@outlook.com*

The growth and development of small and medium-sized firms (SME) is hindered by financial distress. Due to the limited size and poor liquidity of the Portuguese capital market, small and medium-sized enterprises (SMEs) rely significantly on bank credit and trade credit for their development. This study applies panel regression methodology to analyze drivers of the bank and trade credit, as well as to investigate the complementary or substitutive relationship between them.

Keywords

Regression modeling, Panel data.

Small companies in Portugal heavily rely on bank credit and trade credit for financing. Bank loans are the most common source of funding for businesses, however, due to the constraints encountered in the banking sector, trade credit emerges as a viable option to bank credit for numerous firms [1]. In order to examine the factors influencing bank credit and trade credit, regression models using panel data spanning a decade (2010 to 2019) were employed. As independent variables, the models take into account the firm's data on return on assets, collateral security, current ratio, turnover growth, and Altman's Z score for bankruptcy prediction [2]. The same conclusions were drawn using different estimating techniques. The findings reveal that return on assets has a negative effect on both bank and trade credit. Collateral security and liquidity (current ratio) demonstrated a positive significant relationship with bank credit, but a negative relationship with trade credit. Concerning the growth in turnover, the results suggest a significant and negative relationship with bank credit, while no significant relation was found with trade credit. Additionally, models indicate that firms in a "distress zone"—a region where bankruptcy is a possibility—resort more to bank credit and less to trade credit than other firms. Regression models were also used to investigate the complementary or substitutive relationship between trade and bank credit. The substitution hypothesis asserts that companies use more trade credit when facing difficulties in accessing bank financing. Another theory is that bank credit and trade credit typically rise or shrink together; in this sense trade and bank credit are complementary financial resources [1]. Estimation results in this study give evidence of the substitution hypothesis between trade and bank credit.

Acknowledgements: The authors would like to thank the CMUC for financial support.

References

[1] M. J. Palacín-Sánchez, F. J. Canto-Cuevas, and F. di-Pietro. Trade credit versus bank credit: a simultaneous analysis in European SMEs. *Small Bus Econ* **53: 1079–1096**, 2019.

[2] E. I. Altman. *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy.* Frontiers in Finance Series, ohn Wiley & Sons, 1983.

A comparative study of several classes of reduced-bias extreme value index estimators with applications

Lígia Henriques-Rodrigues[1,2,6], Frederico Caeiro[3,4] and M. Ivette Gomes[5,6]

[1]School of Sciences and Technology, University of Évora, Portugal
[2]Research Center in Mathematics and Applications (CIMA), Portugal
[3]NOVA School of Science and Technology (FCT NOVA), Nova University of Lisbon, Portugal
[4]Center for Mathematicas and Applications (NOVA MAth), Portugal
[5]DEIO, Faculty of Sciences, University of Lisbon, Portugal
[6]Center for Statistics and Applications of University of Lisbon (CEAUL), Portugal

E-mail addresses: *ligiahr@uevora.pt; fac@fct.unl.pt; migomes@ciencias.ulisboa.pt*

Extreme Value Statistics addresses the estimation of parameters of extreme events, being the extreme value index (EVI) the most important within this field. Under a semi-parametric approach, the Hill estimators are commonly used for EVI estimation, for models with an EVI > 0 (heavy tails). These estimators can be biased, which can make EVI estimates inaccurate. In this work we perform a comparative study of several classes of reduced bias EVI-estimators based on a particular class of generalized Hill estimators.

Keywords

Statistics of extremes, Reduced bias estimators, Extreme value index, Monte Carlo simulation.

The *extreme value index* (EVI), denoted by ξ, is the main parameter in *extreme value theory* (EVT), measuring the heaviness of the *right-tail function* (RTF), $\overline{F}(x) := 1 - F(x)$, and the heavier the right-tail, the larger ξ is. For Pareto-type models, with a positive EVI, the classical EVI-estimators are the Hill estimators [1], which are the averages of the log-excesses, $V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}$, $1 \le i \le k < n$, where $X_{i:n}$, $1 \le i \le n$, denotes the ascending *order statistics* (OSs) associated with a sample X_i, $1 \le i \le n$,

$$\widehat{\xi}^{\mathrm{H}}(k) := \frac{1}{k} \sum_{i=1}^{k} V_{ik}, 1 \le i \le k < n,$$

The Lehmer's mean-of-order-p ($\mathrm{L_p}$) EVI-estimators were introduced and studied, under a second order framework, in [2,3], being defined as

$$\mathrm{L}_p(k) := \frac{M_{k,n}^{(p)}}{p\, M_{k,n}^{(p-1)}}, \quad p > 0.5 \qquad [\mathrm{L}_1(k) \equiv \widehat{\xi}^{\mathrm{H}}(k)],$$

consistent for $\xi > 0$ and real $p > 0$, and where $M_{k,n}^{(p)}$ is the p-moment of the log-excesses, V_{ik}, i.e., $M_{k,n}^{(p)} := \frac{1}{k} \sum_{i=1}^{k} V_{ik}^{p}$, $p \ge 1$.

Based on the assymptotic behaviour of the L_p EVI-estimators a class of reduced-bias (RB) EVI-estimators was introduced in [4] with functional from:

$$\widehat{\xi}^{\mathrm{RBL}(p)}(k) = \widehat{\xi}^{\mathrm{L}(p)}(k) \left(1 - \frac{\widehat{\beta}}{(1-\widehat{\rho})^p} \left(\frac{n}{k}\right)^{\widehat{\rho}}\right), \quad p > 0.5, \quad 1 \le k < n,$$

with $(\widehat{\beta},\ \widehat{\rho})$ adequate estimators of the second order parameters (β, ρ). This class of RB estimators generalizes the well known minimum-variance RB (MVRB) corrected Hill (CH) estimators introduced in [5], and given by

$$\widehat{\xi}^{\mathrm{CH}}(k) \equiv \widehat{\xi}^{\mathrm{CH}}_{\widehat{\beta},\widehat{\rho}}(k) := \widehat{\xi}^{\mathrm{H}}(k) \left(1 - \frac{\widehat{\beta}}{1-\widehat{\rho}} \left(\frac{n}{k}\right)^{\widehat{\rho}}\right), \quad 1 \le k < n.$$

In this work we will introduce new classes of RB EVI-estimators based on the L_p EVI-estimators and perform a comparative study of their assymptotic properties. The performance for finite samples will be assessed through a small scale Monte-Carlo simulation study and with applications to real data sets.

Acknowledgements: Research partially supported by National Funds through **FCT** – Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/04674/2020 (CIMA) `https://doi.org/10.54499/UIDB/04674/2020`, UIDB/00297/2020 [`https://doi.org/10.54499/UIDB/00297/2020`] and UIDP/00297/2020 [`https://doi.org/10.54499/UIDP/00297/2020`] (Centro de Matemática e Aplicações), and UIDB/00006/2020 (CEAUL) [`https://doi.org/10.54499/UIDB/00006/2020`].

References

[1] B.M. Hill. A Simple General Approach to Inference About the Tail of a Distribution. *Annals of Statistics*, **3(5)**, 1163–1174, 1975.

[2] H. Penalva, M.I. Gomes, F. Caeiro and M.M. Neves. A couple of non reduced bias generalized means in extreme value theory: an asymptotic comparison. *Revstat - Statistical J.* **18(3)**, 281–298, 2020a.

[3] H. Penalva, M.I. Gomes, F. Caeiro and M.M. Neves. Lehmer's mean-of-order-p extreme value index estimation: a simulation study and applications. *J. Applied Statistics* **47(13–15)** (*Advances in Computational Data Analysis*), 2825–2845, 2020b.

[4] M.I. Gomes, H. Penalva, F. Caeiro and M.M. Neves. Non-reduced versus reduced-bias estimators of the extreme value index-efficiency and robustness. Em: *In Colubi*, A., Blanco A. and Gatu C. (eds.), Proceedings of COMPASTAT 2016: 22th International Conference on Computational Statistics, Oviedo, Spain, 279–290, 2016.

[5] F. Caeiro, M.I. Gomes and D.D. Pestana. Direct reduction of bias of the classical Hill estimator. *Revstat - Statistical J.* **3(2)**, 113–136, 2005.

Mathematical modeling and physics informed neural network approaches for studying the environmental impact of data centers on a county level

Sophie Hutter,[1] Alonso Ogueda-Oliva[1], Yehia Khalil[2] and Padmanabhan Seshaiyer[1]

[1]George Mason University, USA
[2]Yale University, USA

E-mail addresses: *sthutter7@gmail.com; aogueda@gmu.edu; Yehia.khalil@yale.edu; pseshaiy@gmu.edu*

Loudoun county in the state of Virginia in the United States is the world's data center hub, with around 115 data centers [1]. Rapid expansion of the internet of things (IoT) and AI are accelerating data center growth, energy and water use, and emissions, posing a challenge to the UN Sustainable Development Goals (SDG) of net-zero emissions by 2050 [2]. While estimates of global emissions from data centers exist, this would be the first study to estimate the direct and indirect environmental impact of data centers at the county level. Our study dynamically models the relationship between data center growth, population growth, loss of biomass, and increased CO2 emissions using a system of coupled ordinary differential equations. The mathematical model thus accounts for the broader implications of data center concentration, such as its role in stimulating further infrastructure and land development, and assesses the impact on human mortality. Physics Informed Neural Networks (PINNs) are used with input from real-time data in Loudoun County to quantify parameters. Findings identify key causes and impacts of emissions related to data center growth at a local level, and define quantitatively the problem that sustainable energy solutions must address.

Keywords

Mathematical model, Data center emissions, Atmospheric carbon dioxide, Sustainable development, Physics Informed Neural Networks.

Acknowledgements: This work is supported in part by George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program (ASSIP), and the GMU Department of Mathematical Sciences, as well the Yale Department of Chemical and Environmental Engineering.

References

[1] A. Olivoäk. Northern Va. is the heart of the internet. Not everyone is happy about that. *Washington Post*, 2023.

[2] United Nations. TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT. *A/RES/70/1.* 2015.

[3] M. Verma, A. K. Verma, and A. K. Misra. Mathematical modeling and optimal control of carbon dioxide emissions from energy sector *Environ. Dev. Sustain.* **23**, 13919–13944, 2021.

[4] P. Donald, M. Mayengo, and A. G. Lambura. Mathematical modeling of vehicle carbon dioxide emissions. *Heliyon.* **10:2**, 2024.

Stochastic differential equations mixed model for individual growth with inclusion of genetic values

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

[1]Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Évora, Portugal
[2]Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, Évora, Portugal
[3]Iscte - Instituto Universitário de Lisboa, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia, Lisboa, Portugal
[4]Business Research Unit - Iscte (BRU-Iscte), Iscte - Instituto Universitário de Lisboa, Lisboa, Portugal
[5]Instituto Superior de Ciências de Educação da Huíla, Lubango, Huíla, Angola and Instituto Superior Politécnico Sinodal da Huíla, Lubango, Huíla, Angola

E-mail addresses: *ntj01011980@gmail.com; patricia.filipe@iscte-iul.pt; gjcj@uevora.pt; braumann@uevora.pt*

Stochastic Differential Equations (SDE) models effectively describe individual growth dynamics in fluctuating environments. Applied to Mertolengo cattle, the model parameters are the average size at maturity (α), a growth parameter (β), and environmental fluctuation intensity (σ). These parameters can vary per animal in SDE mixed models. Here, we consider SDE mixed models with parameter α varying among animals and, in addition, we incorporate into α a dependence on the animal's genetic characteristics. Using maximum likelihood estimation, our study found a highly significant variability of α among animals and that some genetic values significantly contribute to explain such variability.

Keywords

Genetic traits, Individual growth, Mixed models, Stochastic differential equations.

Stochastic differential equations (SDE) models adequatly describe the dynamics of individual growth in a randomly fluctuating environment. We have applied them to model cattle weight evolution using real data from the Mertolengo cattle breed.

The model parameters are the average transformed weight at maturity α, a growth parameter β describing the rate of approach to maturity, and the intensity of the effect of environmental fluctuations σ. Considering M animals,

we use the following general SDE model

$$dY_i(t) = \beta\,(\alpha - Y_i(t))\;dt + \sigma\,dW_i(t), \quad Y_i(t_{i,0}) = y_{i,0}, \quad i = 1, \ldots, M,$$

where $Y_i(t) = h(X_i(t))$ is the modified weight by the transformation h, a monotonous continuously differentiable function (which we assume known) of the real size $X_i(t)$ at age t of the i^{th} individual ($i = 1, \ldots, M$). We have $Y_i(t_{i,0}) = y_{i,0} = h(x_{i,0})$, where $x_{i,0}$ is the weight observed at age $t_{i,0}$ (initial age) for individual i, and $\alpha = h(A)$, where A is the asymptotic weight or weight at maturity, and $W_i(t)$ ($i = 1, \ldots, M$) are independent standard Wiener processes. To incorporate individual characteristics of the animals, we have considered that the model parameters may vary randomly from animal to animal, resulting in SDE mixed models. Here we consider an SDE mixed model with random α and allow α to be a function of the genetic values of the animal. Let the genetic characterization of each animal be determined by K genetic values. For simplicity, we will consider incorporating one genetic value at a time ($K = 1$).

Denoting by g_i the genetic value of animal $i, (i = 1, \ldots, M)$, let us consider α_i to be defined as a linear function of the genetic value in the form of

$$\alpha_i = c_0 + c_1 g_i + \delta_i,$$

where δ_i are independent and identically distributed with $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, ($i = 1, \ldots, M$). Then α_i follows a Gaussian distribution with mean $c_0 + c_1 g_i$ and variance σ_δ^2.

We present a comparison between the SDE non-mixed model with the SDE mixed model with random α and no genetic dependence that shows a highly significant variability of the parameter α among the animals. We also present, for each genetic value, a comparison between random α mixed models, repectively including and not including the dependence of α on the genetic value. Such comparisons show that some genetic values significantly contribute to explain the variability of α among animals, thus improving the estimation of the animal's growth curve [1].

Acknowledgements: The authors belong to the research center CIMA (Centro de Investigação em Matemática e Aplicações, Universidade de Évora), supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal), project UID/MAT/04674/2020, `https://doi.org/10.54499/UIDB/04674/2020`. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais - Productivity improvement in the system of bovine raising for meat, funded by PDR 2020. We are grateful to ACBM and José Pais (ACBM head engineer) for providing the data and for continuous support. We thank RuralBit for the help in extracting the data from the Genpro database.

References

[1] N. T. Jamba, P. A. Filipe, G. Jacinto and C. A. Braumann. Stochastic differential equations mixed model for individual growth with the inclusion of genetic characteristics. *Statistics, Optimization & Information Computing* **12(2)**: 298–309, 2024.

Improving infectious disease predictions through the use of metapopulation SIR modeling and graph convolutional neural networks

Petr Kisselev[1], Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]Department of Mathematical Sciences, George Mason University

E-mail addresses: *peter.kisselev@gmail.com; aogueda@gmu.edu; pseshaiy@gmu.edu*

Graph convolutional neural networks have shown tremendous promise in addressing data-intensive challenges in recent years. In particular, some attempts have been made to improve predictions of SIR models by incorporating human mobility between metapopulations and using graph approaches to estimate corresponding hyperparameters. In [1], researchers have found that a hybrid GCN-SIR approach outperformed existing methodologies when used on the data collected on a precinct level in Japan. However, deficiencies in capturing sudden outbreaks hint at the need for additional scrutiny when it comes to modeling assumptions. In our work, we extend this approach to data collected from the continental US, adjusting for the differing mobility patterns and varying policy responses. We study the effect of changing modeling assumptions with the goal of improving predictive power of the proposed hybrid model.

Keywords

Graph convolutional neural networks, SIR, Disease modeling.

Acknowledgements: This work is supported in part by George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program (ASSIP), the GMU Department of Mathematical Sciences, and the National Science Foundation DMS 2230117.

References

[1] Q. Cao, R. Jiang, C. Yang, Z. Fan, X. Song and R. Shibasaki. MepoGNN: Metapopulation epidemic forecasting with graph neural networks, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 453–468, 2022.

Identifying key skills for enhancing development, writing, and management of European Funded Projects: a multivariate analysis approach

Cristina Lopes[1], Cristina Torres[1], Kaisa Adair[2], Arina Ventelä[2], Kathrin Rath[3], Manuel da Silva[1] and Paula Carvalho[1]

[1]CEOS.PP, ISCAP, Polytechnic of Porto, Portugal
[2]Turun Ammattikorkeakoulu Oy, Finland
[3]Hochschule für Angewandte Wissenschaften Hamburg, Germany

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

ENCARE is a pioneering initiative aimed at improving the skills of researchers and science support staff across Universities of Applied Sciences to successfully develop, write and manage European funded projects. As part of this project, a survey was conducted to identify the skills gaps of UAS staff to successfully apply for funding, and their interest in enhancing their proficiency in pre-award tasks. Using multivariate statistical analysis, namely factor and cluster analysis, we identified different learner profiles, which was crucial for the design of a tailor-made capacity building program to develop the required skills.

Keywords

Research funding, Skills development, Cluster analysis, Profile identification.

ENCARE is an Erasmus+ Cooperation Partnership in higher education that aims at crafting a cross-university capacity-building program, designed to equip the staff of Universities of Applied Sciences (UASs) with the proficiency needed to navigate the intricacies of securing third-party funding for research (carpenetwork.org/encare/). The 3-year project (2023 - 2026) is led by Hochschule für Angewandte Wissenschaften Hamburg (Germany), in collaboration with Turun Ammattikorkeakoulu Oy, (Finland), Stichting Hogeschool Utrecht (The Netherlands), Instituto Politécnico do Porto (Portugal), Universitat Politecnica de Valencia (Spain), and University of the West of Scotland (Scotland). To design a good training course, it is necessary to analyse the needs and to identify profiles of learners [1,2].

The aim of the work presented here was to identify groups of learners among researchers, research managers, support staff, and managers of UASs, who want to acquire competences to write funding applications and manage research projects.

The needs analysis was based on a survey targeting the above-mentioned group. This survey included Likert scale questions to inquire respondents about the level of the skills already acquired in certain pre-award tasks - all work carried out before receiving funding, including finding opportunities and partners, writing proposals, planing research, and monitoring technical issues - as well as their respective interest in developing said skills.

The survey received 180 responses, mostly from researchers/academics or research support staff. Only a minority held managerial positions. However, 72% of the respondents had two or more roles. A few respondents showed 30 to 40 years of experience in pre-award activities, but, on average, participants in the survey had 8.4 years of experience, with a standard deviation of 8.0 years. Half of the participants had been engaged in pre-award activities for no longer than 5 years.

Using Factor Analysis (FA), we grouped the skills and interests questions into 4 factors, with a very good internal consistency (Cronbach's alpha = 0.948 and 0.951 respectively) and high total variance explained (74.415% and 74.891%, respectively). They were designated as follows: Factor 1 *Funding application writing*, Factor 2 *Management and research skills*, Factor 3 *Engaging Stakeholders*, and Factor 4 *Ethics and Regulations*. The scores obtained were used to cluster and classify participants into different profiles. Moreover, the hierarchical cluster analysis was performed using the Manhattan distance and the Ward method [3]. The analysis of the dendrogram revealed six clusters that were validated using silhouette width [4]. The identification of the six clusters led to the definition of the three most important learner profiles in UAS: **Profile 1** (Cluster 1) *Experienced research manager or researcher*; **Profile 2** (Cluster 3) *Eager to learn junior/intermediate level researcher or research manager*; **Profile 3** (Clusters 2/5/6) *Busy researchers with limited time to funding applications*. In Cluster 4 were people with no interest in developing any skills. Based on the results of this analysis, we are developing a three-strand training program - ICERpro - tailored to the identified profiles.

Acknowledgements: Project ENCARE is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

References

[1] L. A. Latif, T. T. Subramaniam, Z. A. Khatab. . Learner Profiling Towards Improving Learner Success. In: Li, K.C., Tsang, E.Y.M., Wong, B.T.M. (eds) *Innovating Education in Technology-Supported Environments*. Singapore: Springer, 2020. https://doi.org/10.1007/978-981-15-6591-5_14

[2] L. A. M. Zaina, G. Bressan, J. F. Rodrigues and M. A. C. A. Cardieri. Learning Profile Identification Based on the Analysis of the User's Context of Interaction. in *IEEE Latin America Transactions*, **9: 5**, 845–850, 2011. https://doi.org/10.1109/TLA.2011.6030999

[3] C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.). *Handbook of Cluster Analysis* 1st Ed., Chapman and Hall/CRC, 2015. https://doi.org/10.1201/b19706

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

Addressing data scarcity in classification of vertebrate footprints using transfer learning with CNNs and procedurally simulated footprints

Carolina S. Marques[1,2], Afonso Mota[3], Diego Castanera[4], Elisabete Malafaia[5], Soraia Pereira[1,2] , Vanda F. Santos[6,7] and Emmanuel Dufourq[8,9,10]

[1]Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, Portugal
[2]Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3]Faculdade de Ciências, Universidade do Porto, Portugal
[4]Fundación Conjunto Paleontológico de Teruel-Dinópolis/Museo Aragonés de Paleontología, Spain
[5]Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, Portugal
[6]Departamento de Geologia, Faculdade de Ciências, Universidade de Lisboa, Portugal
[7]Departamento de Geología y Geografía (Grupo PaleoIbérica), Universidad de Alcalá, Spain
[8]African Institute for Mathematical Sciences, South Africa
[9]African Institute for Mathematical Sciences Research and Innovation Centre, Rwanda
[10]Department of Mathematical Sciences, Stellenbosch University, South Africa

E-mail addresses: *csmarques@fc.ul.pt; afonsomm@gmail.com; castanera@fundaciondinopolis.org; efmalafaia@fc.ul.pt; sapereira@fc.ul.pt; vafsantos@fc.ul.pt; dufourq@aims.ac.za*

The study of vertebrate footprints is key for studying past and present wildlife. The lack of footprint photograph datasets available makes the task of creating automatic classifiers challenging. This study addresses this issue by generating a dataset with over 100,000 simulated footprints, which are used to train a CNN. The CNN is fine-tuned with real footprints, improving accuracy and robustness. These results highlight the importance of innovative data augmentation techniques for enhancing accuracy and reliability, especially when dealing with data scarcity.

Keywords

Transfer learning, CNN, Data scarcity, Vertebrate footprints, Simulated data.

Data augmentation has emerged as an essential technique in the field of machine learning, particularly for Convolutional Neural Networks (CNNs), which are extensively used for image classification tasks [1]. The robustness of CNNs depends on the availability of large, diverse, and labeled datasets [1]. However, in many real-world problems, acquiring such datasets is often impractical. Data augmentation addresses this problem by artificially expanding the training dataset [1].

Studying vertebrate footprints provides helpful insights into the distribution and movements of both past (e.g. dinosaurs) and present (e.g. mammals) fauna. Classifying vertebrate footprints automatically through photographs can be very challenging due to the variability among footprint images and the lack of available labeled datasets.

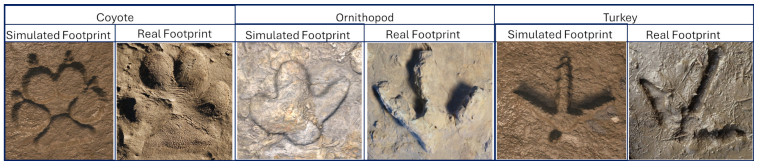

Fig. 1. Examples from simulated footprint photographs and real footprint photographs.

This study explores a novel approach to overcome the problems previously mentioned by employing transfer learning with CNNs trained on simulated vertebrate footprints. The simulated dataset was generated by an application build using Unity and contains around 100,000 footprints from different vertebrates, with variable light conditions, diverse footprint sizes and substrate types (Figure 1). The simulations are created using a set of different silhouettes from footprints found in published literature (e.g. [2,4]) and took less than 4 hours to be produced. The created dataset is used as the initial training set that enables the CNN to learn fundamental features of vertebrate footprints. The pre-trained model is then fine-tuned using a smaller dataset of real vertebrate footprints obtained from different papers (e.g. [3]), enhancing its robustness and adaptability. Our results show that this methodology significantly improves footprint recognition performance. This study highlights the importance of innovative data augmentation techniques in enhancing the accuracy and reliability of CNNs, especially when dealing with data scarcity.

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UIDB/00006/2020 (DOI: 10.54499/UIDB/00006/2020) and UI/BD/154258/2022.

References

[1] L. Alzubaidi et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **8**(1), 2021. doi:10.1186/s40537-021-00444-8

[2] D. Castanera et al.. Geometric morphometric analysis applied to theropod tracks from the Lower Cretaceous (Berriasian) of Spain. *Palaeontology* **58**, 183, 2015. doi:10.1111/pala.12132

[3] R. Shinoda and K. Shiohara. OpenAnimalTracks: A Dataset for Animal Track Recognition. *arXiv* **2406.09647**, 2024. Available at: `https://arxiv.org/abs/2406.09647`.

[4] T. Telander. *Animal Tracks: A Falcon Field Guide [tm] (Falcon Field Guide Series).* FalconGuides, 2012.

Reliability of a new clinical instrument: a case study

Ana Matos[1,2], Carla Henriques[1,3], Diogo Jesus[4] and Luís Inês[5]

[1]School of Technology and Management of Viseu, Polytechnic Institute of Viseu, Portugal
[2]Research Centre in Digital Services (CISeD), Portugal
[3]Centre for Mathematics of the University of Coimbra (CMUC), Portugal
[4]Leiria Hospital Center, Rheumatology Department and Faculty of Health Sciences, University Beira Interior, Covilhã, Portugal
[5]Faculty of Health Sciences, University Beira Interior and Coimbra Hospital and University Centre, Rheumatology Department, Portugal

E-mail addresses: *amatos@estgv.ipv.pt; carlahenriq@estgv.ipv.pt; jesus.p.diogo@gmail.com; luisines@gmail.com*

The purpose of this study is to evaluate the reliability of a recently developed instrument for measuring the activity of systemic lupus erythematosus disease. Inter-rater and intra-rater reliability were estimated using the intraclass correlation coefficient (ICC) calculated with a two-way random-effects model. Additionally, a comparison of this new instrument with classical methods was conducted using the Related-Samples Wilcoxon Signed Rank Test, Spearman's rho correlation, coefficient of variation, and minimal differences needed to be considered real.

Keywords

Intraclass Correlation Coefficient, Inter-rater reliability, Intra-rater reliability.

High-quality instruments are essential for both clinical and research applications. To establish an instrument's high quality, it is necessary to assess measurement properties such as reliability and validity using standardized criteria. Acceptance of any new method depends on a convincing demonstration that it is at least as good as, if not better than, an established method. This study evaluates the measurement properties of a recently developed instrument for assessing disease activity in Systemic Lupus Erythematosus, the SLE Disease Activity Score (SLE-DAS), and compares it with two other frequently used methods. For this purpose, 24 clinical vignettes were created, and 19 rheumatologists from multiple centers were randomly selected as raters. After undergoing training, the raters scored the vignettes using three SLE disease activity instruments in two rounds, spaced 7-14 days apart, with the vignettes presented in random order. Simultaneously, a highly experienced rheumatologist independently scored all the vignettes using the same three instruments. Inter-rater and intra-rater reliability were estimated using the intraclass correlation coefficient (ICC) with 95% confidence intervals, calculated through the kappaetc package by Daniel Klein [1]. To compare the reliability of the three instruments recorded on different scales, the comparison involved the use of the coefficient of variation and the "minimal differences needed to be considered real" [2].

Acknowledgements: This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Furthermore, we would like to thank the Research Centre in Digital Services (CISeD) and the Instituto Politécnico de Viseu for their support.

References

[1] Klein D. Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal* **18**, 871–901, 2018.

[2] Weir J. Quantifying Test-Retest Reliability Using The Intraclass Correlation Coefficient and the SEM. *Journal of Strength and Conditioning Research/ National Strength & Conditioning Association* **19**, 231–240, 2005.

Advances in photovoltaic parameter estimation using computational data analysis and numerical methods

Oumaima Mesbahi[1,2], Mouhaydine Tlemçani[1,2], Daruez Afonso[1,2], Fernando M. Janeiro[1,3] and Mourad Bouzzeghoud[2,4]

[1]Department of Mechatronics, University of Évora, Évora, Portugal
[2]Instrumentation and Control Laboratory, Institute of Earth Sciences, Évora, Portugal
[3]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[4]Departament of Physics, Universidade de Évora, Évora, Portugal

E-mail address: *omesbahi@uevora.pt*

Photovoltaic (PV) systems are a cornerstone of renewable energy technologies [1], necessitating precise parameter estimation for optimal performance and reliability [2] [3]. This study presents novel advancements in PV parameter estimation leveraging computational data analysis and numerical methods.

Our research integrates two key investigations. The first study employs total least squares and metaheuristic algorithms to enhance the accuracy of PV parameter estimation. These methodologies effectively address errors and uncertainties in measurement data (current and voltage), demonstrating significant improvements in the robustness and efficiency of PV systems. The application of these advanced computational techniques underscores their potential in improving the reliability of PV system performance, which is critical in enhancing its efficiency.

The second study explores the effect of measurement intervals on the accuracy of PV parameter estimation. By systematically varying measurement intervals, we illustrate the substantial impact on estimation precision. This investigation highlights the necessity of strategic measurement planning in PV system operation. Together, these studies provide a comprehensive framework for improving PV parameter estimation. The integration of total least squares, metaheuristic algorithms, and strategic measurement interval planning offers a robust approach to addressing challenges in PV systems. Our findings have wide-ranging implications for engineering, environmental sciences, and various interdisciplinary fields.

Keywords

Metaheuristic methods, Parameter estimation, Optimization.

References

[1] O. Mesbahi, D. Afonso, M. Tlemçani, A. Bouich and F. M. Janeiro. Measurement Interval Effect on Photovoltaic Parameters Estimation. *Energies*, **16(18)**, 6460, 2023.

[2] O. Mesbahi, M. Tlemçani, F. M. Janeiro, A. Hajjaji and K. Kandoussi. Recent development on photovoltaic parameters estimation: total least squares approach and metaheuristic algorithms. Energy Sources, Part A: Recovery, Utilization, and Environmental *Effects*, **44(2)**, 4546–4564, 2022.

[3] M. A. Navarro, D. Oliva, A. Ramos-Michel and E. H. Haro. An analysis on the performance of metaheuristic algorithms for the estimation of parameters in solar cell models. *Energy Conversion and Management*, **276**, 116523, 2023.

Neutrosophic odd generalized exponential family with applications

Mina Norouzirad[1] and Amin Roshani[2]

[1]Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal.
[2]Department of Statistics, Lorestan University, Khorramabad, Iran

E-mail addresses: *m.norouzirad@fct.unl.pt; roshani.amin@gmail.com*

The real world is full of ambiguity and uncertainty, where precise statistical values are often unattainable. Neutrosophic statistics addresses these situations. Lifetime distributions are crucial in fields like survival analysis and engineering. This paper introduces the neutrosophic odd generalized exponential family distribution, a flexible family with various hazard rate shapes, encompassing distributions like Odd exponential-G, Generalized exponential, and Exponentiated Weibull. We also provide simulations and applications.

Keywords

Odd generalized exponential family, Neutrosophic, Uncertainty.

The real world is full of ambiguity, unclear, and uncertain situations where precise values cannot always be assigned to statistical characteristics. Neutrosophic statistics addresses these imprecise conditions by incorporating degrees of truth, indeterminacy, and falsity. Lifetime distributions play a fundamental role in various fields such as survival analysis, biomedical sciences, engineering, and social sciences. Typically, lifetime refers to the duration of human life, the lifespan of a device before it fails, or the survival time of a patient from diagnosis to death.

Smarandache [1] introduced the concept of neutrosophy to accurately represent and model the inherent indeterminacy present in data. Classical probability distributions are applicable when samples are selected from populations with uncertain observations. Therefore, there is an essential need to introduce probability models under an indeterminacy environment. Various researchers have developed neutrosophic probability distributions in this case, producing better results compared to classical statistics. [?] introduced neutrosophic discrete random distributions such as the neutrosophic uniform discrete distribution, neutrosophic Bernoulli distribution, neutrosophic binomial distribution, neutrosophic geometric distribution, neutrosophic negative binomial distribution, neutrosophic hypergeometric distribution, and neutrosophic Poisson distribution. Additionally, [3] proposed the neutrosophic exponential distribution, neutrosophic gamma distribution, neutrosophic beta distribution, neutrosophic uniform distribution, and neutrosophic Poisson distribution.

In this paper, we introduce the neutrosophic odds generalized exponential (NOGE) distribution, a flexible family of distributions capable of modeling a wide range of hazard rate shapes, including increasing, decreasing, J-shaped, reversed-J, bathtub, and upside-down bathtub curves. This family of distributions [4] is versatile enough to encompass several well-known distributions, such as the Odd exponential-G family [5], the Generalized exponential [6], the Exponentiated Weibull [7], and the Generalized Gompertz [8].

We provide simulations to illustrate the performance and flexibility of the neutrosophic odds generalized distribution. Additionally, we present practical applications to demonstrate its utility in real-world scenarios, such as the analysis of nitrogen oxides emissions

and the COVID-19 pandemic. The maximum likelihood estimation approach is employed to estimate the parameters, and the performance of these estimators is evaluated. From the comparative analysis, the indeterminacy parameter is found to have a considerable impact on fitting quality.

Acknowledgements: The work of Mina Norouzirad is funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 (`https://doi.org/10.54499/UIDB/00297/2020`) and UIDP/00297/2020 (`https://doi.org/10.54499/UIDP/00297/2020`) (Center for Mathematics and Applications).

References

[1] F. Smarandache *Neutrosophy, neutrosophic probability, set and logic.* American Research Press, Rehoboth, DE, USA, 1998.

[2] C. Granados, New Notions On Neutrosophic Random Variables. *Neutrosophic Sets and Systems*, **47**, 286–297, 2022.

[3] C. Granados, A. K. Das, A. K., B. Das, Some Continuous Neutrosophic Distributions with Neutrosophic Parameters Based on Neutrosophic Random Variables. *Advances in the Theory of Nonlinear Analysis and its Applications*, **6(3)**, 380–389, 2022.

[4] M. H. Tahir, G. M. Cordeiro, M. Alizadeh, M. Mansoor, M. Zubair and G. G. Hamedani, The odd generalized exponential family of distributions with applications. *J. Stat. Distrib. App.*, **2**, 1–28 , 2015.

[5] M. Bourguignon, R. B. Silva, G. M. Cordeiro, The Weibull-G family of probability distributions. *J. Data Sci.*, **12**, 53–68, 2014.

[6] R. D. Gupta, D. Kundu, Generalized exponential distribution. *Aust. N. Z. J. Stat.*, **41**, 173–188, 1999.

[7] G. S. Mudholkar, D. K. Srivastava, Exponentiated Weibull family for analyzing bathtub failure data. *IEEE Trans. Reliab.*, **42**, 299–302, 1993.

[8] A. El-Gohary, A. Alshamrani, A. N. Al-Otaibi, The generalized Gompertz distribution. *Appl. Math. Model.*, **37**, 13–24 , 2013.

Tracing sea water parameters with spatial autocorrelation analysis

Christoper Ody[1,2], M. Rosário Ramos[1,2] and Elisabete Carolino[3]

[1]Universidade Aberta and CEG (Centro de Estudos Globais-UAb)
[2]CEAUL, Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3]H&TRC – Health & Technology Research Center, ESTeSL-Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal

E-mail addresses: *onleft@gmail.com, mariar.ramos@uab.pt; etcarolino@estesl.ipl.pt*

Studying the spatial distribution of environmental indicators makes it possible to monitor parameter values and their variation across a region. In this study, a statistical analysis is carried out on data collected in the southern region of the North Sea nearby Germany coast, a region of economic importance. The data is originally from the TRAM project (Tracing origin and distribution of geogenic and anthropogenic dissolved and particulate critical high-technology metals in the southern North Sea), collected over the course of a week by a scientific cruise ship. Physico-chemical variables were collected, such as temperature, salinity, N.NO3(Nitrate), DOCeq (Dissolved Organic Carbon) and Abs210 (absorption at wavelength), among others. Three maritime sub-areas were defined for the region, which in principle reflect different patterns in some parameters such as temperature and salinity. Data pre-processing involved screening for outliers, peak and gradient tests and data imputation methods. An exploratory spatial data analysis (ESDA) was carried to summarise its main characteristics. Spatial autocorrelation is a measure that tells us how a set of variables with associated geographic coordinates are related to each other in space. Spatial autocorrelation and can be positive or negative and can be evaluated at a global or local level. The Global measure quantifies the spatial pattern across the entire study area, while the local spatial association (LISA) identifies hotspots and coldspots of high or low values for each feature or variable. Here, the study of spatial autocorrelation used popular measures such as Moran's I statistic, both global and local, as well as Geary's c-statistic. Different profiles and significant spatial autocorrelation were identified in variables for the subregions and along the region.

Keywords

Environmental quality, Marine resources, Spatial autocorrelation.

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects: UIDP/05608/2020. DOI 10.54499/UIDP/05608/2020 (`https://doi.org/10.54499/UIDP/05608/2020`), UIDB/05608/2020. DOI 10.54499/UIDB/05608/2020 (`https://doi.org/10.54499/UIDB/05608/2020`) and UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (`https://doi.org/10.54499/UIDB/00006/2020`).

The authors thank to Dr. Andrea Koschin from Jacobs College in Germany, coordinator of the TRAM project for providing the data.

References

[1] C. Ody, E. Carolino and M. Rosário Ramos. Spatial and multivariate statistics in assessing water quality in the North Sea *Computational Science and Its Applications – ICCSA 2024 Workshops*, **LNCS 14816**, 1–17, 2024. https://link.springer.com/chapter/10.1007/978-3-031-65223-3_12

[2] M. Quante, and F. Colijns *North Sea Region Climate Change Assessment*. Springer International Publishing, 2016.

[3] Michael Schlundt *Continuous thermosalinograph oceanography along RV Meteor cruise M169-Data Processing Report*. GEOMAR - Helmholtz Centre for Ocean Research Kiel, 2021.

[4] P. Moraga *patial Statistics for Data Science: Theory and Practice with R*. Chapman & Hall/CRC Data Science Series, 2023.

A technique to improve General Linear Methods when integrating linear initial boundary value problems

Nuria Reguera[1]

[1]University of Burgos, Spain

E-mail address: *nreguera@ubu.es*

In this work we present a study of the order reduction exhibited by General Linear Methods (GLMs) when used for the numerical integration of linear initial boundary value problems with time-dependent boundary conditions. In addition, we propose a technique that improves the implementation of GLMs by increasing the practical order by one unit in the general case in which the phenomenon of order reduction is present.

Keywords

General Linear Methods (GLMs), order reduction, numerical methods for ODEs, method ol lines.

General linear methods (GLMs) are a class of multistep and multistage methods used for time integration of systems of ordinary differential equations that have been widely used in the literature since Burrage and Butcher introduced them in [1]. In fact, several of the most commonly used time integrators for numerically solving ordinary differential equations, such as Runge–Kutta methods and linear multistep methods, are particular cases of GLMs.

Despite the importance of these methods, GLMs have a practical drawback when used for the time discretization of rigid systems of ordinary differential equations, since in these cases the order observed in practice may be lower than the classical GLM order. This phenomenon of order reduction is a major practical drawback for all GLMs except those of high stage order.

An important case of very rigid systems of ordinary differential equations appears after the discretization in space, with the intention of using the method of lines, of a time-evolving partial differential equation. In this work we present a study [2] on the order reduction that GLMs present in these cases. In addition, we propose a technique with which it is possible to recover a unit of order and that, therefore, presents a practical improvement for the implementation of GLMs methods. This technique is based on the appropriate choice of the boundary values of the internal stages of the method.

Acknowledgements: This work has received funding from the University of Burgos under the project Y054GI "Grupo de Física Matemática FISMAT-UBU".

References

[1] K. Burrage and J.C. Butcher. Nonlinear stability of a general class of differential equation methods. *BIT Numer. Math.* **20**, 185–203, 1980.

[2] I. Alonso-Mallo and N. Reguera (2024). Avoiding order reduction phenomenon for general linear methods when integrating linear problems with time dependent boundary values. *Journal of Computational and Applied Mathematics* **439** 15629, 2024.

Solving steady-state heat conduction in irregular domains using physics-informed neural networks and fictitious domain method

José A. Rodrigues[1]

[1]CIMA and Department of Mathematics of ISEL Lisbon, Portugal

E-mail address: *jose.rodrigues@isel.pt*

This presentation introduces a novel approach for solving heat conduction in irregular domains using the fictitious domain method and Physics-Informed Neural Networks (PINNs). By embedding complex geometries within a regular grid and integrating PINNs, we efficiently address boundary conditions and internal heat sources. The method, demonstrated with NURBS curves and a central heat source, offers improved accuracy and computational efficiency for complex thermal problems.

Keywords

Physics-Informed Neural Networks (PINNs), Fictitious Domain Method, Irregular Domains, Steady-State Heat Conduction, NURBS Curves.

Introduction

The study of heat transfer in irregular domains presents a significant challenge due to the complex geometries often encountered in practical applications. Traditional numerical methods, such as finite element and finite difference methods, struggle with these complexities, especially when dealing with boundary conditions and the inclusion of internal sources. This work explores the application of Physics-Informed Neural Networks (PINNs) to solve the steady-state heat conduction problem in an irregular domain using a fictitious domain approach. PINNs, which leverage the power of neural networks to incorporate physical laws into the learning process, provide a versatile and efficient alternative to conventional methods.

Fictitious Domain Method

The fictitious domain method [1] is a powerful technique for handling irregular domains. By embedding the irregular domain within a larger, regular computational domain, the method simplifies the application of boundary conditions and the formulation of the governing equations. This approach allows the use of simple Cartesian grids while accurately capturing the effects of complex boundaries. In this study, the irregular domain is defined by a NURBS (Non-Uniform Rational B-Splines) curve, which offers a flexible and precise representation of complex geometries. The fictitious domain approach, combined with PINNs, enables the seamless handling of the irregular boundary without the need for mesh generation or complex geometric transformations.

Physics-Informed Neural Networks (PINNs)

PINNs represent a significant advancement in solving partial differential equations (PDEs)[2] due to their ability to integrate data and physical laws into the learning process. By embedding the governing PDEs and boundary conditions directly into the neural network's loss function, PINNs provide a robust framework for solving complex physical problems [3] and [4]. In this study, a fully connected neural network is employed to

approximate the temperature distribution within the domain. The network is trained to minimize a composite loss function that includes the PDE residual, boundary conditions, and an additional term for the fictitious domain.

Examples and Results

We demonstrate the effectiveness of this approach through several examples involving irregular domains defined by NURBS curves. These examples illustrate the flexibility of PINNs in handling complex boundary geometries and the fictitious domain method's ability to streamline the problem formulation. Each example involves the specification of a central heat source and zero boundary conditions, showcasing how the combined approach effectively captures the temperature distribution within the domain.

The results reveal that PINNs, when used in conjunction with the fictitious domain method, provide accurate and computationally efficient solutions. The temperature fields predicted by the model align well with expected physical behavior, demonstrating the method's robustness in managing complex boundary conditions and internal heat sources. The fictitious domain approach simplifies the computational implementation while maintaining high fidelity in the representation of the irregular domain.

Acknowledgements: This research was partially sponsored with national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT, under projects UIDB/04674/2020 (CIMA). DOI: `https://doi.org/10.54499/UIDB/04674/2020`.

References

[1] R. Glowinski, T. W. Pan, J. Periaux. A fictitious domain method for Dirichlet problem and applications. *Computer Methods in Applied Mechanics and Engineering*, **111(3-4)**, 283–303, 1994.

[2] M. Raissi, P. Perdikaris and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *Journal of Computational Physics*, **378**, 686–707, 2019.

[3] J.A. Rodrigues. Using Physics-Informed Neural Networks (PINNs) for Tumor Cell Growth Modeling, *Mathematics*, **12**, 1195, 2024.

[4] J.A. Rodrigues. Exploring Linear Elasticity: Unveiling the Power of Physics-Informed Neural Networks (PINNs).Oral communication at the 2nd International Workshop on Mathematics and Physical Sciences (MatPhys), Évora, Portugal, 11–12 July 2024.

Modeling, analysis and prediction of COVID-19 dynamics with interacting subpopulations and implicit behavior using Physics-Informed Neural Networks

Naima Aubry-Romero[1], Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]Department of Mathematical Sciences, George Mason University, Fairfax, USA

E-mail addresses: *naima.aubry.usa@gmail.com; aogueda@gmu.edu; pseshaiy@gmu.edu*

The COVID-19 pandemic has underlined the importance of research in epidemiological modeling concerning adaptive mathematical models, governed by nonlinear ordinary differential equations, that account for evolving behavioral responses to understand and predict the spread of infectious diseases [1]. In this work, we consider an extended SEIR compartmental model [2] that incorporates two interacting subpopulations representing young and old age groups, allowing for cross-group transmission dynamics. The basic reproduction number, the average number of secondary cases of infection produced by a single primary case, is derived using the Next Generation Matrix method for our model. Furthermore, we incorporated implicit behavioral changes into our extended model which allows us to determine its influence on the effectiveness of public health interventions. We solve the governing differential equation system numerically as well as estimate useful parameters in the model using Physics-Informed Neural Networks (PINNs) [3]. Our results point to how the PINNs approach offers an effective framework to predict the unique parameters of our model, forecast disease progression, and determine the impact of behavioral modifications on the reproduction number and transmission dynamics. Finally, an interactive dashboard is also created that will allow users to make important data-driven decisions by manipulating parameters, analyzing the resulting graphs, and understanding the impacts of the reproductive number.

Keywords

COVID-19, Compartmental models, Human behavior, Epidemiology, Physics Informed Neural Networks, Deep learning.

Acknowledgements: This work is supported in part by George Mason University (GMU) College of Science Aspiring Scientists Summer Internship Program (ASSIP), the GMU Department of Mathematical Sciences, and the National Science Foundation DMS 2230117.

References

[1] Comfort Ohajunwa, Kirthi Kumar and Padmanabhan Seshaiyer. Mathematical modeling, analysis, and simulation of the covid-19 pandemic with explicit and implicit behavioral changes. *Computational and Mathematical Biophysics*, **8(1)**, 216–232, 2020.

[2] Fred Brauer, Carlos Castillo-Chavez and Carlos Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 2. Springer, 2012.

[3] Maziar Raissi, Niloofar Ramezani and Padmanabhan Seshaiyer. On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. *Letters in biomathematics*, **6(2)**, 1–26, 2019.

A simulation approach to an optimal electric and diesel bus fleet design

Raina Saha[1,2], Madeline Haas[1] and Katherine McCrum[1]

[1]Department of Systems Engineering and Operations Research, George Mason University, Virginia, USA
[2]Department of Mathematics, George Mason University, Virginia, USA

E-mail addresses: *rsaha3@gmu.edu; mhaas5@gmu.edu; kmccrum@gmu.edu*

Due to electric buses having zero tailpipe emissions, less greenhouse gasses, and reduced maintenance costs and noise pollution, there is an ongoing national push towards electric vehicles [2]. This work investigates the feasibility of implementing a hybrid electric/diesel bus fleet in shuttle transit networks. The primary goal of the simulation is to sustain shuttle service over normal operating schedules while maintaining each bus's charge. It considers several constraints, including limited access to charging, existing bus schedules, and battery behavior. The simulation consists of three primary parts: a battery simulation modeling internal battery behavior, an electric bus simulation modeling the behavior of a single electric bus, and a fleet simulator modeling the behavior of a hybrid electric/diesel bus fleet. The electric bus simulation models the battery behavior on a route to determine the minimum charge we can send an electric bus on a route. The fleet simulator models the behavior of an electric/diesel bus fleet over a week to determine the ratio of electric buses to diesel buses to ensure that all routes of interest can be run. Our findings indicate that bus placement is very sensitive to the initial charge and the ability to charge on-route. These results underscore the importance of taking into account characteristics of each bus route when creating bus fleets. The model serves as a robust tool for transportation planners to simulate various scenarios and optimize electric bus placement. Overall, this work contributes to the ongoing efforts to reduce reliance on fossil fuels, offering hope for a more sustainable future.

Keywords

Electric Vehicle, Lithium-Ion Battery, Bus Transit, Agent-based Simulation, Battery Modeling.

This work investigates a simulation for determining the optimal composition of small transportation fleets consisting of a mixture of electric and diesel-powered buses. It will determine via simulation the number of diesel buses to be included in the hybrid fleet to minimize the probability of a complete discharge event on the electrical buses on their assigned routes.

A model-based battery simulation is used to model the battery voltage. This strategy was chosen from the initial alternatives assessed due to its robust performance associated with its closed-loop feedback mechanism, enabling the reevaluation of parameter estimates and ease of implementation [1,3].

Acknowledgements: This work has received support from the George Mason University Department of Systems Engineering and Operations Research as part of master's

program coursework. The simulation work done in support of this project is planned for incorporation in an four-year multi-department project commencing Fall 2024. This effort collaborates with multiple local government transit entities and school districts, and focuses on facilitating their transitions to hybrid or zero-emission bus fleets.

References

[1] Sandeep Dhameja. *Electric vehicle battery systems.* Elsevier, 2001.

[2] Aisling Doyle, Mohan Lal Kolhe, and T Muneer. *Electric Vehicles: Prospects and Challenges.* Elsevier, 2017.

[3] Gregory L Plett. *Battery management systems, Volume I: Battery modeling.* Artech House, 2015.

Assessing and enhancing data quality in data streams

Eliana Costa e Silva[1], Óscar Oliveira[1] and Bruno Oliveira[1]

[1]CIICESI, Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Porto

E-mail addresses: *eos@estg.ipp.pt; oao@estg.ipp.pt; bmo@estg.ipp.pt*

Streaming data scenarios face critical quality issues like incompleteness, inconsistency, and inaccuracy. This work applies previously established data quality scores to a real-world streaming pipeline, demonstrating their effectiveness in monitoring and ensuring data integrity. The study showcases how these metrics can provide real-time insights and immediate detection of anomalies, validating their practical utility in maintaining high data quality standards.

Keywords

Data Quality Scores, Monitoring, Evolving Datasets, Data Streaming, Online Analytical Processing, Weighted Sum Method.

In the era of big data, the integrity and reliability of data streams have become pivotal for decision-making processes and operational efficiency. Streaming data, characterized by its continuous and real-time nature, presents unique challenges that are not typically encountered with batch data processing.

Streaming data is prone to numerous quality issues that can significantly impact downstream applications and analytics. To address these challenges, data quality metrics are crucial for quantifying and monitoring the integrity of streaming data. These metrics offer objective measures to evaluate various dimensions of data quality, such as completeness (proportion of non-missing values in the stream), consistency (degree to which data adheres to predefined formats or standards), and accuracy (degree to which data reflects true values). For readers interested in a comprehensive overview of data quality in Industry 4.0, particularly in the context of IoT and Cyber-Physical Systems (CPS), [1] provides an in-depth systematic literature review of data quality techniques and issues in these advanced industrial applications.

In [3], three data quality metrics were proposed: the Weighted Quality Score (WQS), the Longitudinal Weighted Quality Score (LWQS), and the Quality Score Delta (QSD). The WQS evaluates data adherence to predefined quality rules of different dimensions, giving a snapshot of data quality for a particular data block. While, the LWQS measures data quality over time, prioritizing newer data, whenever required. The QSD, calculates the difference between WQS and LWQS, and offers valuable insights for monitoring and improving data quality. QSD helps organizations assess changes in data quality performance, evaluate the impact of quality interventions, set improvement goals, compare performance across units, and mitigate risks associated with declining data quality. Both WQS and LWQS are determined using a weighted sum method. For each block of data, collected at a given instance, the WQS is computed as the weighted sum of the proportion of rows that follow a given quality rule of a corresponding dimension, which are weighted across all dimensions. The LWQS is defined similarly to WQS, but it considers the ratio between the number of row in accordance to each rule and the total number of rows for every block of data prior to the present one.

In [2], a comprehensive framework for managing data quality is introduced, providing a flexible and extensible approach for consuming, analyzing, and handling data prior to generating business insights. To evaluate the effectiveness of these data quality scores, we integrated them into a high-demand data streaming pipeline based on the framework proposed in [2], which processes industry sensor data from multiple sources. This pipeline was selected for its performance requirements and the critical nature of the data it processes.

The integration and evaluation of data quality scores in the streaming pipeline yielded several valuable advantages, e.g., enhanced monitoring (enabling immediate responses to data quality issues), improved data reliability (continuous quality checks ensured that only high-integrity data was processed and analyzed), operational efficiency (automated alerts and anomaly detection), data-driven insights (historical analysis of metrics can provide deeper insights into recurring quality issues and their root causes).

The evaluation of data quality scores in the streaming data pipeline validated their effectiveness in maintaining high standards of data integrity and reliability. The scenarios tested demonstrated that these metrics are not only critical for real-time monitoring but also for gaining actionable insights into the overall health of the data stream. By addressing and mitigating quality issues promptly, organizations can ensure that their streaming data remains a robust foundation for analytics and decision-making processes.

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

References

[1] A. Goknil, P. Nguyen, S. Sen, D. Politaki, H. Niavis, K. J. Pedersen, A. Suyuthi, A. Anand and A. Ziegenbein. A Systematic Review of Data Quality in CPS and IoT for Industry 4.0. ACM Computing Surveys, **55(14s)**, 1–38, 2023.

[2] Ó. Oliveira and B. Oliveira. An Extensible Framework for Data Reliability Assessment. International Conference on Enterprise Information Systems, ICEIS - Proceedings, 2022.

[3] E. Costa e Silva, Ó. Oliveira and B. Oliveira. Enhancing Real-Time Analytics: Streaming Data Quality Metrics for Continuous Monitoring, In Proceedings of the 2024 7th International Conference on Mathematics and Statistics (ICoMS 2024), 2024 (in press).

Modeling the dynamics of the opioid epidemic using efficient computational approaches

Ryan Singh[1], Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]Department of Mathematical Sciences, George Mason University, Fairfax, USA

E-mail addresses: *ryjs78@gmail.com; aogueda@gmu.edu; pseshaiy@gmu.edu*

The opioid epidemic is one of the fastest growing public health issues in the United States. As a disease that spreads through complex behavior and interaction as opposed to a shared medium, addiction poses a challenge in terms of modeling as it cannot be analyzed through traditional infectious disease frameworks [1]. In this work, the nature of the spread of opioid dynamics is modeled using coupled non-linear Ordinary Differential Equations (ODEs). These equations employ a compartmental model composed of seven subpopulations including Susceptible, Prescribed, Exposed, Addicted, Deceased, Treated, and Recovered [2]. Specifically, the model considers the impact of human behavior and interaction of the subpopulations by incorporating both prescription-based addictions and social exposure from addicted individuals. The governing equations are solved using numerical approaches and the parameters in the model are estimated using Physics-Informed Neural Networks (PINNs) [3]. Additionally, we derive the basic reproduction number to provide further information on the spread of this epidemic. By employing a data-driven interaction-based model, the proposed framework helps us to understand the spread of opioid addiction and implement improved policy [4] to further mitigate its effects.

Keywords

Opioid Addiction, Compartmental Model, Physics-Informed Neural Networks.

Acknowledgements: This work is supported by the George Mason University Aspiring Scientists Summer Internship Program (ASSIP) and the GMU Department of Mathematical Sciences.

References

[1] Carlos Blanco, Melanie M. Wall and Mark Olfson. Data needs and models for the opioid epidemic. *Molecular psychiatry*, **27(2)**, 787–792, 2022.

[2] Nicholas A. Battista, Leigh B. Pearcy, and W. Christopher Strickland. Modeling the prescription opioid epidemic. *Bulletin of mathematical biology*, **81**, 2258–2289, 2019.

[3] Maziar Raissi, Paris Perdikaris and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, **378**, 686–707, 2019.

[4] Jeromie Ballreich, Omar Mansour, Ellen Hu, Francine Chingcuanco, Harold A. Pollack, David W. Dowdy and G. Caleb Alexander. Modeling mitigation strategies to reduce opioidrelated morbidity and mortality in the us. *JAMA network open*, **3(11)**, e2023677–e2023677, 2020.

DExPSO: a double exponential particle swarm optimization with non-uniform variates as stochastic tuning and guaranteed convergence to a global optimum

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*

Nature-inspired metaheuristic algorithms, like Particle Swarm Optimization (PSO), are powerful general-purpose optimization tools, but they invariably do not come with rigorous theoretical justifications and can fail to find a global optimum. By treating PSO as a random search optimization process and repairing the famous Global Search Convergence Theorem by imposing an additional condition in the proof, in [1], we created a novel theory-based algorithm called the Double Exponential Particle Swarm Optimization algorithm (DExPSO) that converges to a global optimum. In particular, we show that the common practice of using uniform variates as stochastic components in PSO and related algorithms does not satisfy the sufficient conditions in DExPSO and hence may provide a reason why PSO and other nature-inspired algorithms like QPSO, LcRiPSO, and CSO can fail to converge. Additionally, in more complicated design problems, we show that DExPSO tends to converge to the support points of the optimal design more frequently and faster than PSO and its variants do. Moreover, there is a possibility to modify other PSO variants to DExPSO variants, and such hybridization offers promising improvement in the quality of the global search. Our applications include finding designs that minimize the integrated mean squared prediction error and locally D-optimal exact designs for a 68-compartmental model to assess radioactive particles retained in the human lung after exposure, see [2]. Because PSO, and more generally, metaheuristics are used across disciplines, including ecology, pharmacokinetics and pharmacodynamics studies, agriculture, engineering, and computer science, there are potentially broad and deep implications of our results.

Keywords

DExPSO, PSO, Random search, Optimization, Correlated design.

References

[1] M. Stehlík, W. K. Wong,P. Y. Chen, J. Kiselak. 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*, 2024, `https://doi.org/10.1016/j.asoc.2024.111913`.

[2] M. Stehlík, J. M. Rodríguez-Díaz, W. G. Müller and J. López-Fidalgo . Optimal allocation of bioassays in the case of parametrized covariance functions: an application in lung's retention of radioactive particles. *TEST* **17**, 56–68, 2008.

Mathematical analysis of tumour impacts on physiological flows modulated by electric and magnetic fields

Dharmendra Tripathi[1] and Ashvani Kumar[1]

[1]Department of Mathematics, National Institute of Technology Uttarakhand, Srinagar-246174, India

E-mail address: *dtripathi@nituk.ac.in*

Physiological flows like blood flow, urine flow, breathing, movement of chyme, sperm movement, etc. are very important mechanisms in the biological systems which are governed by very natural pumping process i.e., peristalsis, membrane pumping, heart pumping, compression and expansion of lungs, and rhythmic propagation of the muscles. However, the tumours in the vessels/parts of the body are challenging problem, creating obstruction in the fluids flow and many people are died due to infection and fast growth of the tumours in the body. This process introduces complexity in flow behaviour particularly at micro scale. To find out the mathematical solution at small context, a fluid flow model governed by the peristaltic pumping is developed in present of single tumour. An analysis for flow characteristics and influence of tumour shape and size in during the fluid flow in microchannel is simulated. Furthermore, how this obstruction due to tumour by applying the external electric field and magnetic field have been examined. For this biophysical model, governing equations based on mass conservation, momentum conservation and Maxwell equation for electro-magneto-hydrodynamics have been adopted. Low Reynolds number flow in microchannel is considered. MATLAB code is utilized for simulation of the results. The findings reveal that a larger tumor height enhances fluid flow by narrowing the microchannel and promoting abnormal fluid flow in microchannel however this tumour size may also stop the fluid flow if this size is very close to the diameter/width of the microchannel. Overall, this research provides insights into optimizing fluid dynamics for biomedical applications and gives recommendation for development of bio microfluidics devices.

Keywords

Tumour cell; Peristaltic transport, Electroosmosis, Magnetohydrodynamics, Zeta potential, Hartmann number.

Acknowledgements: I acknowledge to SERB, DST, Gov. of India (Ref. no. MTR/2023/000377), for providing the fund for this research work.

References

[1] Meijing Li and James G. Brasseur. Non-steady peristaltic transport in finite-length tubes. *Journal of Fluid Mechanics*, **248** 129–151, 1993.

[2] Sanjay Kumar Pandey and Ankit Prajapati. An analytical and comparative study of swallowing in a tumor-infected oesophagus: a mathematical model. *Journal of Mathematical Biology*, **88, no. 3**, 37, 2024.

The effect of model misspecification on fit measures when there is a planned pattern of missingness

Paula C.R. Vicente[1]

[1]Lusófona University, Lisbon

E-mail address: *p951@ulusofona.pt*

This research aims to assess the sensitivity of the typical fit measures, specifically the root mean square error of approximation, standardized root mean square residual, comparative fit index, and Tucker-Lewis index, in the adjustment of misspecified structural equation models when there is a planned pattern of missingness according a two-method design. The existence of this type of omissions when considering misspecified models can result in unacceptable values for all the examined fit measures, especially in small sample sizes.

Keywords

Fit Measures, Misspecified Models, Two-Method Design, Structural Equation Model.

Missing or incomplete data represent a persistent problem in several studies in different fields, such as education, psychology or marketing, and remove the omissions from the data modelling is a common procedure. The absence of certain observations can be attributed to the extensive time needed to answer numerous questions in a lengthy survey, the effort required from each participant and the costs associated with obtaining specific responses. Employing a planned missing design could provide a potential resolution to this issue. Furthermore, full information maximum likelihood or multiple imputation approaches enables researchers to analyse and fit models without excluding incomplete cases, but instead incorporating omissions in the study design [1].

Two different planned missing designs are the three-form design and the two-method design. In a two-method design there are two distinct measures of a construct, one more expensive or time consuming and other, that is inexpensive or time saving. The least expensive measure will be observed for all participants, but only a small proportion of participants will be randomly selected to receive the expensive measure. This procedure allows to have more participants than in a complete data design using only the expensive measure, given the same budget [2].

Confirmatory factor analysis (CFA) and structural equation models (SEM) are statistical techniques for analysing multivariate data in social sciences. The quantification of the adjustment of this models is very important and can be performed using some different indices, which are based in distinct criteria to identify the best model. The most used with this type of modelling are Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI). In fact, most researchers concur that utilizing more than one fit index is necessary to ensure confidence in the model's fit [3].

The objective of this research is to evaluate the consequences of misspecification within a structural equation model, particularly in scenarios where there are omissions stemming from a two-method design, through a simulation study. Furthermore, this study considers the effects of different sample sizes and types of models.

References

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

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

[3] Shi et al. Understanding the model size effect on sem fit indices. *Educational and Psychological Measurement*, **79(2)**, 310–334, 2019.

Posters

An application of Box-Jenkins methodology to model the series of currency in circulation in Mozambique

Samuel G. Arone[1,2], Catarina S. Nunes[2,4] and Luís M. Grilo[3,5,6]

[1]ISMU - Instituto Superior Mutasa, Maputo, Moçambique
[2]DCeT- Departamento de Ciências e Tecnologia, Secção de Matemática, Universidade Aberta, Portugal
[3]Departamento de Matemática, Universidade de Évora, Portugal
[4]LAETA / INEGI, Porto, Portugal
[5]CIMA, Universidade de Évora, Portugal
[6]NOVA Math, Universidade NOVA de Lisboa, Portugal

E-mail addresses: *sboboti@gmail.com; 1500887@estudante.uab.pt; CatarinaS.Nunes@uab.pt; luis.grilo@uevora.pt*

The use of banknotes and coins depends exclusively on the preference of the public for liquidity, and specifically, on the functionalities of various denominations put into circulation, making it difficult to determine exactly the quantity of means that the monetary authorities should put into circulation in the economy. Considering its relevance, monetary authorities must adopt models that allow describing and forecast currency in circulation. Thus, we use Box–Jenkins methodology to estimate and forecasting the series of currency in circulation in Mozambique. The main result reveals that SARIMA$(1,1,0)(0,1,1)_{12}$ model, is the most suitable to estimate and forecast this series.

Keywords

Box-Cox transformation, Forecast, Non-stationary, Seasonality, Trend.

Currency in circulation represents banknotes and coins issued by a monetary authority and placed into the economy to carry out day-to-day transactions. It is a more liquid and indispensable monetary aggregate for any economy. With the growing trend of digitalization, several alternative payment methods to cash are available, but the use of banknotes and coins continues to be universally accepted in many countries, including the most developed ones [3]. In the case of Mozambique, data available in the statistical database published by the Banco de Moçambique, in its website *www.bancomoc.mz*, indicate that in the last five years, the average annual growth of currency in circulation stood at around 9%. However, based on the periodic variation, from 2018 to 2023, currency in circulation grew up by 50%, a reduction compared to 76% recorded in the previous five years, a period associated with average annual growth of 12% [4]. These figures suggest the growth of currency in circulation at a decreasing rate.

Statistical analyzes carried out on the monthly series of currency in circulation in Mozambique, covering a period from January 2011 to December 2023 allowed us to conclude that this series has an increasing trend, is seasonal and is autocorrelated. Due to these specific characteristics Box–Jenkins methodology was used for its estimation and forecasting.

Box-Jenkins methodology is based on a class of stochastic processes called ARIMA (Autoregressive Integrated Moving Average) or SARIMA (Seasonal Autoregressive Integrated Moving Average) models if it includes a seasonal component. The general notation

of this class of models is SARIMA (p,d,q)(P,D,Q)$_s$, where p and P represent, respectively, non-seasonal and seasonal autoregressive parameters, q and Q represent non-seasonal and seasonal moving average parameters, d and D represent the integration levels of the series in a non-seasonal and seasonal components respectively, and s represents the seasonal amplitude. Therefore, identifying a model means discovering the integers p, d, q, P, D and Q that allow a better description of the process [1,2,5].

Based on the analysis of the correlogram of the series (autocorrelation and partial autocorrelation functions), two processes: SARIMA$(1,1,0)(0,1,1)_{12}$ and SARIMA$(1,1,0)(1,1,0)_{12}$ were identified as those that faithfully describe the series of currency in circulation in Mozambique. After identification, we moved on to estimation of models. Subsequently, using the t-ratio test, we found that the parameters estimates of the models are all statistically significant (p-values < 0.001). Regarding the diagnosis of the residuals, in both models, the Box-Pierce and Ljung-Box tests were unanimous in not rejecting the null hypothesis of absence of autocorrelation in the residuals. However, the Kolmogorov normality test did not support the hypothesis of normality of residuals at the 5% level, but it was relatively close. Therefore, the adjustments of the two models can perhaps be considered acceptable.

Once the models were adequately adjusted, and there was a need to choose the best one, models selection criteria that take into account the residuals of the estimated models were considered, specifically: Akaike Information Criterion, Bayesian Schwart Criterion and Hannan-Quinn Information Criterion, as well as the criteria that take into account the analysis of forecast errors namely: Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error and U-Theil coefficient [1, 2, 5]. All models selection criteria were unanimous in suggesting that SARIMA$(1,1,0)(0,1,1)_{12}$ model is the most appropriate to estimate and forecast the currency in circulation in Mozambique.

References

[1] J. Caiado. *Métodos de Previsão em Gestão com Aplicações em Excel*. Edições Sílabo, Lisboa, 2011.

[2] C. Chatfield. *Time Series Forecasting*. Chapman Hall/CRC. New York, 2000.

[3] FIS WorldPay. *The Global Payments Report. Payment insights that drive growth*. 8th edition, 2023.

[4] `https://www.bancomoc.mz/pt/areas-de-actuacao/estatisticas/`.

[5] W. Wei, Time Series Analysis: Univariate and Multivariate Methods. Philadelphia, 2006.

Analysis of economic and innovative variables in product innovation in SMEs of the 27 EU member states

Marta Azevedo[1], Aldina Correia[1] and Ana Borges[1]

[1]CIICESI, ESTG, Instituto Polit´ecnico do Porto, Portugal

E-mail addresses: *8230570@estg.ipp.pt; aic@estg.ipp.pt, aib@estg.ipp.pt*

The aim of this work is to understand which factors influence the introduction of product innovation in small and medium-sized companies (SMEs).
Initially, a literature review is carried out on the concept of innovation, in the context of small and medium-sized companies. What product innovation is, highlighting the impact that innovation indicators have on the introduction of product innovation in SMEs.
Subsequently, using the European Innovation Scoreboard 2023 (EIS 2023) database, a statistical analysis (descriptive statistics and statistical inference) of the data is carried out, and using two multivariate techniques, namely a multiple linear regression model and cluster analysis.
Finally, the results obtained indicate that innovation expenses not related to R& D, collaboration between innovative SMEs, average annual GDP growth and the Eco-Innovation Index are factors that influence the introduction of product innovation in SMEs.
Regarding the cluster analysis, it was established the creation of 5 clusters, and it can be concluded that the countries in cluster 5 are those with the greatest propensity to introduce product innovation in their SMEs.

Keywords

Innovation, Product Innovation, SMEs, Multiple Linear Regression, Clustering.

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

References

[1] O. A. Acar, A. Tuncdogan, D. van Knippenberg and K. R. Lakhani. Collective creativity and innovation: An interdisciplinary review, integration, and research agenda. *Journal of Management*, **50(6)**, 2119–2151, 2024.

[2] W. Jun, W. Ali, M. Y. Bhutto, H. Hussain and N. A. Khan. Examining the determinants of green innovation adoption in SMEs: a PLS-SEM approach. *European Journal of Innovation Management*, **24(1)**, 67–87, 2021.

[3] Organisation for Economic Co-operation and Development & Statistical Office of the European Communities. Oslo manual: guidelines for collecting and interpreting technological innovation data, 2005. Paris and Luxembourg: OECD/Euro-stat.

Evaluation of economic and innovative factors in process innovation among SMEs in the 27 EU member states

Mariana Azevedo[1], Aldina Correia[1] and Ana Borges[1]

[1]CIICESI, ESTG, Instituto Politécnico do Porto, Portugal

E-mail addresses: *8200414@estg.ipp.pt; aic@estg.ipp.pt, aib@estg.ipp.pt*

The present study aims to identify and explain the influence of small and mediumsized companies, of several factors on the introduction of process innovation in their businesses. For this purpose, the European Innovation Scoreboard 2023 (EIS 2023) database was used.
Initially, a literature review is carried out regarding the concept of innovation in a broad sense, that is, what process innovation is and the impact it causes on companies and finally, it is investigated what the literature refers to in relation to some innovation factors and what is their impact on process innovation.
In addition to the literature review, data analysis is also carried out using descriptive statistics and statistical inference. Finally, a multiple linear regression model is also considered with the aim of identifying which variables under study are statistically significant for the number of SMEs that introduce process innovation in their businesses. A cluster analysis was grouping similar European countries with regard to process innovation, innovation spending, collaboration between SMEs, GDP, and eco-innovation index.
The study results suggest that innovation expenditure other than R&D, collaboration between innovative companies, average GDP growth and the Eco- Innovation Index are significant for process innovation in small and mediumsized enterprises and that it can be defined 3 clusters of countries.

Keywords

Innovation; Innovative SMEs; Process Innovation, Multiple Linear Regression, Clustering.

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

References

[1] E. M. O. Coutinho and M. Au-Yong-Oliveira. Factors Influencing Innovation Performance in Portugal: A Cross-Country Comparative Analysis Based on the Global Innovation Index and on the European Innovation Scoreboard. *Sustainability.* **15(13)**, Page 10446, 2023.

[2] D. Kariv, N. Krueger, G. Kashy and L. Cisneros. Process innovation is technology transfer too! How entrepreneurial businesses manage product and process innovation. *Journal of Technology Transfer*, 1–25, 2024.

[3] S. Oduro. Eco-innovation and SMEs' sustainable performance: a meta-analysis. *European Journal of Innovation Management*, **27(9)**, 248–279, 2024.

[4] G. Y. Oukawa, P. Krecl and A. C. Targino. Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. *Science of The Total Environment*, **815**, 152836, 2022.

[5] H. Zhang. Non-R&D innovation in SMEs: is there complementarity or substitutability between internal and external innovation sourcing strategies? *Technology Analysis and Strategic Management*, **36(5)**, 916–930, 2024.

Pandemic preparedness: first steps on the use of non-traditional data for estimating mobility-incidence links of the COVID-19 pandemic in Portugal

André Brito[1,2,3], Ausenda Machado[3,4], Ana Paula Rodrigues[2], Paula Patrício[1,2] and Regina Bispo[1,2]

[1]Center for Mathematics and Applications (NOVAMath) Portugal, Lisbon, Portugal
[2]Department of Mathematics, NOVA School of Science and Technology, NOVA University of Lisbon, Lisbon, Portugal
[3]Department of Epidemiology, National Institute of Health Doctor Ricardo Jorge, Lisbon, Portugal
[4]Comprehensive Health Research Center (CHRC), NOVA University of Lisbon, Lisbon, Portugal

E-mail address: *anm.brito@campus.fct.unl.pt*

Mobility data can be used to approximate the likelihood of risky in-person interactions that lead to increased infection rates. Several studies carried out during the COVID-19 pandemic have shown that decreased mobility was associated with decreased in reported case incidence. In this review we look specifically at the Google Mobility Reports for Portugal and test dimensionality reduction techniques, such as PCA, to reach a mobility indicator that could best link mobility to transmission across several regions and pandemic or epdimeic stages. We highlight the challenges of working with non-traditional data.

Keywords

COVID-19, Viral Respiratory Infections, Mobility, Statistical Modelling.

This abstract provides a critical foundation for the first steps of my PhD project entitled *Spatio-Temporal Dependencies of Viral Respiratory Infections driven by Non-Traditional Data*. The objective of the project is to underscore the importance of understanding how viral respiratory infections spread over time and space, with COVID-19 as a case study. These insights provide a valuable starting point for exploring the spatio-temporal dynamics of viral infections, ultimately contributing to the broader understanding, management and control of infectious diseases. In this abstract we focus on how to deal with mobility data and how it can link to disease incidence.

The COVID-19 pandemic was an unprecedented epidemiological event, and the onset of the pandemic concurrent with digital transition allowed for an exceptional amount of data to be gathered that was then made available to public health officers and researchers [2]. Mobility data in particular as been used by researchers to evaluate its importance on disease incidence [3]. Principal Component Analysis (PCA) has been used to summarise the 6 mobility measures provided in the Google Mobility Reports [1] where the resulting indicator showed that reduced mobility was associated with decreased disease incidence during the first year of the pandemic [4,5]. In Portugal, investigations have highlighted the contribution of Non-Pharmaceutical Interventions (NPIs) such as lockdowns, that directly affect mobility, on lowering the effective reproduction number below 1 [6]. Mobility in

retail and grocery spaces in Portugal was found to have had a higher correlation with new COVID-19 cases compared to mobility in parks. The correlation was weaker in less densely populated regions, underscoring the need for differentiated confinement strategies depending on spatial-specific social characteristics. Mobility was also shown to be a good predictor to forecasting disease incidence [7,8].

These studies were conducted at the onset of the pandemic trough the initial part of 2021. Also, if we are to develop more complex models including more predictors while restricting collinearity, mobility data needs to be summarised so as to be included in a model. The contribution of this work focuses on providing an holistic and retrospective review of the whole pandemic making use of the full mobility report timeline.

References

[1] Google LLC. Google COVID-19 Community Mobility Reports. 2022 `https://www.google.com/covid19/mobility/` Accessed: 14/07/2024.

[2] K. Bolt, D. Gil-González and N. Oliver. Unconventional data, unprecedented insights: leveraging non-traditional data during a pandemic. *Frontiers In Public Health*, **12** 2024.

[3] K. Lee and J. Eom. Systematic literature review on impacts of COVID-19 pandemic and corresponding measures on mobility. *Transportation*, 2023.

[4] S. Jewell, J. Futoma, L. Hannah, A. Miller, N. Foti and E. Fox. It's complicated: characterizing the time-varying relationship between cell phone mobility and COVID-19 spread in the US. *Npj Digital Medicine*, **4**, 2021.

[5] J. Elarde, J. Kim, H. Kavak, A. Züfle and T. Anderson. Change of human mobility during COVID-19: A United States case study. *PLoS ONE*. **16**, 2021.

[6] C. Caetano, M. Morgado, P. Patrício, J. Pereira and B. Nunes. Mathematical Modelling of the Impact of Non-Pharmacological Strategies to Control the COVID-19 Epidemic in Portugal. *Mathematics*, **9**, 2021.

[7] A. Nova, P. Ferreira, D. Almeida, A. Dionísio and D. Quintino. Are mobility and covid-19 related? A dynamic analysis for portuguese districts. *Entropy*, **23** 2021.

[8] N. Mileu, N. Costa, E. Costa and A. Alves. Mobility and Dissemination of COVID-19 in Portugal: Correlations and Estimates from Google's Mobility Data. *Data*, **7**, 2022.

A generalized jacknife estimator of a negative extreme value index

Frederico Caeiro[1] and M. Ivette Gomes[2]

[1]NOVA School of Science and Technology (NOVA FCT) and CMA, NOVA University of Lisbon, Campus de Caparica, Portugal
[2]Faculdade de Ciências, Universidade de Lisboa (UL), and Centro de Estatística e Aplicações/UL (CEA/UL), Portugal

E-mail addresses: *fac@fct.unl.pt; migomes@ciencias.ulisboa.pt*

Let $X_1, X_2, \ldots, X_n$ be independent random variables with a common distribution function F in the max-domain of attraction of a non-degenerate distribution function G. Then G is the extreme value distribution with a shape parameter ξ (Gnedenko, [3]). Let us assume that the shape parameter ξ, also know as the extreme value index, is negative ($\xi < 0$). Under a semi parametric framework, inference is usually based on the upper $k+1$ order statistics of the sample of size n. Classical estimators of the extreme value index such as the moment estimator (Dekkers et al. [2]) or Pickand's [4] estimator have high variance for small values of k and a high bias for large values of k. This leads to the usual bias-variance trade-off problem when we choose the threshold k. In this work, following the lines of Caeiro and Gomes [1], we use the generalized jackknife methodology to obtain a new asymptotically unbiased estimator of a negative extreme value index. Under a second order condition on the tail function $1 - F$ we derive the non degenerate asymptotic behaviour of the new estimator. Using Monte-Carlo simulation techniques, we validate the asymptotic results for finite samples.

Keywords

Extreme value index, Generalized jackknife methodology, Moment estimator, Semi-parametric estimation.

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/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020) (CEAUL), and projects UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (Center for Mathematics and Applications).

References

[1] F. Caeiro and M.I. Gomes. Bias reduction for light tail models—the generalized jackknife methodology. In Gomes, M.I., de Haan, L., Pestana, D., Canto e Castro, L. and Fraga Alves, M.I. (eds.). *Extremes, Risk and Resampling Techniques*: 7–12, Edições CEAUL, 2003.

[2] A.L.M. Dekkers, J.H.J. Einmahl and L. de Haan. A moment estimator for the index of an extreme-value distribution, *The Annals of Statistics* **17(4)**, 1833–1855, 1989.

[3] B. V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire, *The Annals of Mathematics* **44**(3): 423–453, 1943.

[4] J. Pickands III. Statistical inference using extreme order statistics. *The Annals of Statistics* **3(1)**, 119–131, 1975.

Weighted biplot models and statis methodology: a comparative study

Cristina Dias[1,2] and Carla Santos[2,3]

[1]Polytechnic Institute of Portalegre, Portugal
[2]NOVAMath-SST- New University of Lisbon, Portugal
[3]Polytechnic Institute of Beja, Portugal

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

Today there is significant interest in analysing several tables of data together, often referred to as multi-block or multi-way analysis. Many of these methods extend the principles of principal component analysis. This paper briefly reviews the Statis and Metabiplot methods. In both procedures, matrices containing information on a set of variables measured by a set of individuals can be studied. Thus, the aim is to simultaneously explore several data matrices where each collects information on J variables in I individuals on T occasions or experimental situations, seeking a common structure for all studies. We apply both methods to a real data set and compare these techniques.

Keywords
Principal Component Analysis, Statis methodology, Weighted Biplot, Matrices.

One of the most common situations in multiple data analysis is when a set of matrices results from studying a set of variables on a set of individuals at different times (temporal data) or corresponding to different experimental situations (different studies) (see [3-4]). When working with this type of data, the aim is generally to carry out a simultaneous analysis that makes it possible to obtain a common structure or compromise (see [1-3]). We briefly review the Statis and Biplot methods for studying matrices containing information on a set of variables measured on a set of individuals. In this way, several data matrices are explored simultaneously, each collecting information on J variables in I individuals on T occasions or experimental situations, looking for a common structure in all the studies.

Acknowledgements: This work was partially supported by national founds of FCT-Foundation for Science and Technology under UIDB/00297/2020 and UIDP/00297/2020

References

[1] C. Chaya; C. Perez-Hugalde; L. Judez; C. S. WEE; J. X. Guinard. Use of the STATIS method to analyze time-intensity profiling data. *Food quality and preference*, **15(1)**, 312–320, 2004.

[2] C. Dias; C. Santos; J. T. Mexia. Isolated and structured families of models for stochastic symmetric matrices. *Comp and Math Methods*, **2(8)**, e1152, 2021.

[3] C. Lavit. *Analyse conjointe de tableaux quantitatifs*. Editions Masson, 1988.

[4] H. Abdi; L. J. Williams; D. Valentin; M. Bennani-Dosse. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *WIRES Computational Statistics*, **4(2)**, 124–167, 2012.

Variable selection methods in the context of mixtures of linear regression models

Susana Faria[1] and Ana Moreira[1]

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

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

Selecting variables is a crucial step in building a regression model, as it determines which covariates will be used to explain or predict the response variable. In this study, we explore the issue of variable selection in mixtures of linear regression models using penalized maximum likelihood estimation, employing both the Expectation-Maximization and Classification Expectation-Maximization algorithms. We perform a simulation study to compare the performance of various variable selection methods.

Keywords

Alasso, Lasso, Rlasso, Mixtures of linear regression models, Simulation study.

Finite Mixture Regression (FMR) models provide a flexible tool for modeling data that arise from a heterogeneous population, where the relationship between the dependent variable and the explanatory variables varies among the various subpopulations ([1]). In the applications of these models, a large number of explanatory variables is often considered and their contributions to explaining or predicting the dependent variable vary from component to component in the mixture model. For this reason, variable selection assumes great importance for mixture models. However, since all subset selection methods are computationally intensive, to overcome this problem, more efficient methodologies were developed such as, for example, methods based on penalty functions. The Least Absolute Shrinkage and Selection Operator (LASSO) method ([3]), the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) method ([3]) and the Relaxed Least Absolute Shrinkage and Selection Operator (RLASSO) method [2]] are some examples of these methods.

In this work, we analyze the problem of variable selection in mixtures of linear regression models with a large number of explanatory variables. We compare the performance of LASSO, ALASSO, and RLASSO in selecting explanatory variables, using the Expectation-Maximization (EM) and Classification Expectation-Maximization (CEM) algorithms for parameter estimation.

The study reveals how different scenarios affect the performance of these algorithms and penalization functions in selecting explanatory variables. Nevertheless, the ALASSO variable selection method consistently shows superior performance overall. For this reason, we highly recommend its use. Also, we recommend considering the ALASSO method combined with the CEM algorithm for estimation, because it is substantially less computationally demanding than the EM algorithm.

Acknowledgements: The research at CMAT was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020. Special thanks to FCT - Fundação para a Ciência e a Tecnologia, for financing this research through the PhD scholarship, reference number 2022.12256.BD.

References

[1] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2002.

[2] N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, **52(1)**, 374–393. 10.1016/j.csda.2006.12.019, 2007.

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

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

Generalized linear mixed models: an application to road traffic accident

Ana Maia[1], Susana Faria[2] and Elisabete Freitas[3]

[1]Department of Mathematics (DMAT), University of Minho, Guimarães, Portugal
[2]Centre of Mathematics (CMAT) and Department of Mathematics (DMAT), University of Minho, Guimarães, Portugal
[3] Civil Engineering Department, University of Minho, Guimarães, Portugal

E-mail addresses: *pg49144@alunos.uminho.pt; sfaria@math.uminho.pt; efreitas@civil.uminho.pt*

Although the number of road traffic accidents has decreased in developed countries over the past decades, understanding the factors that determine these accidents, aiming to reduce the severity of the damage caused to victims, remains incomplete. The objective of this work is to develop statistical models for count data that allow for the identification of factors (human, infrastructural, and environmental) associated with the occurrence of road accidents.

Keywords

Generalized linear mixed models, Count data, Road safety.

The World Health Organization (WHO) states that road traffic accidents are one of the leading causes of death worldwide among the younger age group. Companies and institutions involved in road safety bear significant responsibility for implementing measures aimed at improving road safety ([2]). These measures can lead to a significant reduction in the frequency of accidents as well as the severity of injuries sustained.

In this work, we pretend to develop statistical models for count data that allow the identification of the most key factors (such as traffic volume, the number of lanes, and road characteristics, among others) associated with the occurrence of road traffic accidents. ([1]) This is a problem of high importance in the context of road safety, as it enables the responsible companies to adopt appropriate measures to improve road safety.

In road accident data, it is common for the same experimental unit to be observed over time, resulting in correlated observations and violating the independence assumption. Generalized Linear Mixed Models (GLMM), an extension of Generalized Linear Models (GLM), address this issue by accounting for the correlation within repeated measurements, thus allowing for more accurate modeling of this type of data ([3]).

The data used in this study were collected at various points along the A4 and A41 highway, which is part of the Ascendi highway network, over the period from 2014 to 2021.

The results of the analyses show that friction and the geometric conditions of the roads contributed the most to the increase in the number of accidents.

Acknowledgements: The research at CMAT was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] H. Huang, H.C. Chin, Modeling road traffic crashes with zero-inflation and site-specific random effects. *Stat Methods Appl*, **19**, 445–462. https://doi.org/10.1007/s10260-010-0136-x, 2010.

[2] G. Yang, K. Wang, Q. Li, Y. Zhan and C. Wang. Panel data analysis of surface skid resistance for various pavement preventive maintenance treatments using long term pavement performance (ltpp) data. *Canadian Journal of Civil Engineering*, **44(5)**, 358–366, 2017.

[3] W. W. Stroup, *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, 2016.

Power function mixtures in reliability

Miguel Felgueiras[1,2,3], João Martins[2,4] and Rui Santos[1,2]

[1]ESTG, Polytechnic Institute of Leiria, Portugal
[2]Centro de Estatística e Aplicações, Universidade de Lisboa, Portugal
[3]Center for Research and Development in Mathematics and Applications, Aveiro, Portugal
[4]ESS, Polytechnic Institute of Porto, Portugal

E-mail addresses: *mfelg@ipleiria.pt; rui.santos@ipleiria.pt; jom@ess.ipp.pt*

In previous work, [1,4] introduced pseudo-convex mixtures generated by shape-extended stable distributions for extremes. This set of distributions and their main properties were derived.
The power function emerges as a particularly interesting distribution function since its pseudo-convex mixture can be used to fit bathtub-shaped hazard rate data. In reliability theory, a bathtub-shaped hazard rate curve is suitable for systems where the hazard rate is initially decreasing (burn-in phase), then more or less constant (random phase), and finally increasing (wear-out phase) [3]. In medicine, it is adequate to model human lifetime. Many statistical distributions have been used to address this kind of hazard rate function, mainly through generalizations or transformations of the Weibull distribution.
This work presents a pseudo-convex mixture from the power function to fit data with a bathtub-shaped hazard rate. This model is quite different from those based on the Weibull distribution, providing new insights into this subject. After a theoretical background, a simulation study and a real data example concerning the failure times of electronic devices [2] are presented to evaluate the goodness of fit of the new proposed model.

Keywords
Power function, Bathtub shaped hazard rate, Failure times.

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

References

[1] M. Felgueiras, J. Martins and R. Santos. Pseudo-convex mixtures. *AIP Conf. Proc.* **1479**, 1125–1128, 2012.

[2] R. Jiang. Two Methods to Estimate the Change Points of a Bathtub Curve. *IJPE* **9(5)**, 569–579, 2013.

[3] Q. Liao, Z. Ahmad, E. Mahmoudi and G. Hamedani. A New Flexible Bathtub-Shaped Modification of the Weibull Model: Properties and Applications. *Math. Probl. Eng.* **3206257**, 2020.

[4] R. Santos, J. Martins and M. Felgueiras. Pseudo-convex Mixtures Generated by Shape-extended Stable Distributions for Extremes. *Journal of Statistical Theory and Practice* **10(2)**, 357–374, 2016.

Tail (in)dependence on extreme value models

Marta Ferreira[1,2]

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

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

Extreme value theory has application in multiple areas, with particular interest in risk analysis. This often involves inferring the extent to which the occurrence of an extreme value can trigger others. Extremal models prove to be suitable for modeling tails. This work focuses on the tail dependence of bivariate extreme value models, addressing measures for this purpose. A simulation study will be presented, comparing different estimators. We finish with an application to financial data.

Keywords

Extreme value theory, Asymptotic independence, Extremal dependence.

The extremal types theorem [3,4] states the possible limiting models of linear normalized maximum distribution function (df). This result can be extended to any dimension. In the bivariate case, let $\{(X_i, Y_i)\}$ be a sequence of independent and identically distributed (iid) random pairs with df $\mathbf{F} = (F_X, F_Y)$. If there exist sequences of real constants $\{a_n, c_n > 0\}$, $\{b_n, d_n\}$, such that

$$P\left(\frac{\max(X_1,...,X_n) - b_n}{a_n} \leq x, \frac{\max(Y_1,...,Y_n) - d_n}{c_n} \leq y\right) = \mathbf{F}^n(a_n x + b_n, c_n y + d_n) \to \mathbf{G}(x, y),$$

as $n \to \infty$, with non degenerate df $\mathbf{G}(x, y) = (G_X(x), G_Y(y))$, then $\mathbf{F}$ belongs to the max-domain of attraction of $\mathbf{G}$ (notation $\mathbf{F} \in \mathcal{D}(\mathbf{G})$) where $\mathbf{G}$ is a bivariate extreme value (BEV) distribution, given by

$$\mathbf{G}(x, y) = \exp(-V(-1/\log G_X(x), -1/\log G_Y(y)))\,. \tag{1}$$

In particular, we have $F_i \in \mathcal{D}(G_i)$, $i \in \{X, Y\}$. Function V expresses the dependence and satisfies some properties like homogeneity of order -1, and $V(1, 1)$ ranging between 1 (complete dependence) and 2 (exact independence). Another formulation of BEV $\mathbf{G}$ is based on Pickands dependence convex function A [7], as follows

$$\mathbf{G}(x, y) = \exp\left[\log(G_X(x)G_Y(y))A\left(\frac{\log G_Y(y)}{\log(G_X(x)G_Y(y))}\right)\right], \tag{2}$$

where $\max(1 - t, t) \leq A(t) \leq 1$, for $t \in [0, 1]$. Both dependence functions are related, in particular, $V(1, 1) = 2A(1/2)$.

The Ledford and Tawn coefficient (LTC) $\eta \in (0, 1]$ describes the type of tail dependence within a random pair (X, Y) [6]. A unit η means that X and Y are tail dependent while $0 < \eta < 1$ means asymptotic tail independence that gradually vanishes as approaching the limit. If $\eta = 1/2$ then X and Y are (almost) independent.

BEV models present, in the upper tail, an LTC $\eta = 1$ corresponding to tail dependence, while the lower tail exhibits asymptotic independence with $\eta = (V(1, 1)^{-1} = (2A(1/2))^{-1})$, given $V(1, 1) < 1$ ($A(1/2) < 1/2$).

We address the estimation of η by considering estimators of V [2] and A [1,7]. We conduct a simulation study to evaluate performances and we illustrate with an application to real data.

Acknowledgements: The research was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] P. Capéraà, A. L. Fougères, and C. Genest. A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika* **84(3)**, 567–577, 1997.

[2] H. Ferreira, and M. Ferreira. On extremal dependence of block vectors. *Kybernetika* **48(5)**, 988–1006, 2012.

[3] R. A. Fisher, and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24(2)**, 180–190, 1928.

[4] B. V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics* **44(6)**, 423-453, 1943.

[5] J. E. Heffernan. A directory of coefficients of tail dependence. *Extremes* **3**, 279–290, 2000.

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

[7] J. Pickands. Multivariate extreme value distributions. In *Bulletin of the International Statistical Institute, Proceedings of the 43rd Session*, Buenos Aires, pp 859–878, 1981.

Salary estimation using random forest based on economic indicators

Catarina Monteiro[1], Ana Borges[1], José M. Soares[2], Pedro Pacheco[2] and Flora Ferreira[2,3]

[1]CIICESI, ESTG, Polytechnic of Porto, 4610-156 Felgueiras, Portugal
[2]BERD - Bridge Engineering Research and Design, Matosinhos, Portugal
[3]Centre of Mathematics, University of Minho, Guimarães, Portugal

E-mail addresses: *8160570@estg.ipp.pt, aib@estg.ipp.pt, jose.soares@berd.eu, pedro.pacheco@berd.eu, fjferreira@math.uminho.pt*

Salary forecasting is critical for effective management, salary negotiations, and talent retention. This study utilized Random Forest (RF) models to estimate salaries based on economic indicators. Factors included average wage, Consumer Price Index (CPI), Producer Price Index (PPI), unemployment rate, Gross Domestic Product (GDP) per capita, labor force, and inflation rate. Data from the International Labour Organization (ILO) on annual salaries by occupation from 2017 to 2022 across 14 countries was analyzed. RF models were trained with data from 2017 to 2021 and tested with 2022 data. Evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), showed that models using a broad set of economic variables offered better accuracy. SHapley Additive exPlanations (SHAP) identified average wage and GDP per capita as key factors.

Keywords

Economic Indicators, Random Forest (RF), Salary Estimation.

Salary forecasting is crucial in strategic decision-making for both employees and employers, impacting management strategies, salary negotiations, and talent retention policies. Machine learning (ML) methodologies, particularly Random Forest (RF), have shown promise in forecasting precision by uncovering valuable insights into salary trends and influencing factors [1,2]. Salaries are intricately influenced by various factors including the economic situation of a country. The objective in this study is to develop a model using RF for estimating salaries, based on economic indicators. The selection of factors was based on their accessibility and likelihood of impacting salary levels. Chosen factors include the average wage and economic variables such as the Consumer Price Index (CPI), Producer Price Index (PPI), unemployment rate, Gross Domestic Product (GDP) per capita, labor force, and inflation rate. For this study, a dataset with the annual salary by job occupation from 2017 to 2022 across 14 countries was used. This dataset is a subset of a larger dataset available from the International Labour Organization (ILO) [3]. Different RF models were evaluated with different sets of inputs. The data from 2022 was used for testing, and the data from 2017 to 2021 was used for training. Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared score (R^2) were used to evaluate the performance of the models. Feature importance scores were evaluated using SHapley Additive exPlanations (SHAP) method.

The RF models demonstrated varying degrees of accuracy in salary prediction, with some models significantly outperforming others based on the selected input factors. The

evaluation metrics indicated that models incorporating all chosen economic variables achieved the lowest MAE and MSE values, along with higher R^2 scores, suggesting better prediction accuracy and fit.

The SHAP analysis provided insights into the contribution of each economic indicator to the salary predictions. The average wage and GDP per capita emerged as the most significant factors.

The use of Random Forest for salary estimation proved effective in this study, highlighting the importance of economic indicators in predicting salary trends. By understanding the key factors influencing salary trends, organizations can make more informed decisions regarding compensation strategies and workforce planning.

Acknowledgements: This work has received supported from Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within the projects UIDB/00013/2020, UIDP/00013/2020 and UIDB/04728/2020.

References

[1] J. Chen, S. Mao and Q. Yuan. Salary prediction using random forest with fundamental features. In Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021) (Vol. 12167, pp. 491-498). SPIE, March 2022.

[2] F. Eichinger and M. Mayer. Predicting salaries with random-forest regression. In Machine Learning and Data Analytics for Solving Business Problems: Methods, Applications, and Case Studies (pp. 1-21). Cham: Springer International Publishing, 2022.

[3] International Labour Organization, *Average monthly earnings of employees by sex and occupation - annual*, `https://rshiny.ilo.org/dataexplorer16/?lang=en&id=EAR_4MTH_SEX_OCU_CUR_NB_A`, Accessed: 2024-03-01.

Application of machine learning to predict dynamics of epidemiological models that incorporate human behavior

Alonso Ogueda-Oliva[1] and Padmanabhan Seshaiyer[1]

[1]George Mason University, USA

E-mail addresses: *aogueda@gmu.edu; pseshaiy@gmu.edu*

In this work, we present a mathematical epidemiological model which incorporates human behavior. We discuss an approach based in Physics-Informed Neural Network (PINNs) that is capable of estimating the parameters of the epidemiological model. Numerical experiments were performed in order to validate this approach.

Keywords

Epidemiology, Physics-Informed Neural Networks, Data-driven scientific computing.

In this work, we present modeling, analysis and simulation of a mathematical epidemiological model which incorporates human social, behavioral, and economic interactions. In particular a COVID-19 pandemic model with explicit and implicit behavioral changes [3].

We discuss an approach based in Physics-Informed Neural Network [1], a data-driven framework for partial differential equations. PINNs can predict the dynamics of a disease described by modified compartmental models that include parameters, and variables associated with the governing differential equations [2]. Our work employs an approach capable of learning how diseases spread and finding their unique parameters with the help of prior knowledge of the subject. We validate our approach with synthetic data used as benchmark.

Acknowledgements: This work is partially supported by the National Science Foundation DMS-2230117.

References

[1] M. Raissi, P. Perdikaris, G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *Journal of Computational Physics*, **378**, 2019.

[2] S. Shaier, M. Raissi. and P. Seshaiyer. Data-driven approaches for predicting spread of infectious diseases through DINNs: Disease Informed Neural Networks. *Letters In Biomathematics*, **9**, 2022.

[3] C. Ohajunwa, K. Kumar and P. Seshaiyer. Mathematical modeling, analysis, and simulation of the COVID-19 pandemic with explicit and implicit behavioral changes. *Computational and Mathematical Biophysics*, **8(1)**, 2020.

Generalized linear models and Quantile Regression Models for *Pinus pinea* Pine Nuts and Kernels Characteristics

Dulce G. Pereira[1], Anabela Afonso[1] and Ana Cristina Gonçalves[2]

[1]CIMA – Centro de Investigação em Matemática e Aplicações/IIFA, Departamento de Matemática/ECT, Universidade de Évora, Portugal
[2]MED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, Departamento de Engenharia Rural, Escola de Ciências e Tecnologia, Universidade de Évora, Pólo da Mitra, Ap. 94, 7002-554 Évora, Portugal

E-mail addresses: *dgsp@uevora.pt; aafonso@uevora.pt; acag@uevora.pt*

This study aims to model the relationship between the weight of pine nuts and kernels and the characteristics of pine cones and trees of *Pinus pinea*. Given the presence of outliers, the asymmetry of the response variable, and the violation of homoscedasticity assumptions, both generalized linear models and quantile regression models were employed. The findings will contribute to developing management tools for optimizing kernel efficiency, particularly in Portugal, the largest producer of pine nuts.

Keywords

Parametric quantile regression, Generalized linear model, *Pinus pinea*, Pine nuts, Kernels.

Pinus pinea stands are valued for their fruit, timber, and resin, with fruit production gaining increased interest due to the high market value and nutritional content of the kernels. Spain, Portugal, and Italy are the leading producers of pine cones. Pine nuts, rich in protein, fats (especially oleic and linoleic acids), vitamins B1 and B3, and minerals, are considered a delicacy. Industrial profitability is influenced by the weight of individual kernels and the number of kernels per cone. While cone production has been extensively studied, there is limited research on the relationship between pine nut/kernel weight per cone and tree dimensions/stand structure, e.g.[1,2].

This study utilizes a dataset comprising over 16,000 observations from a three-year survey (2003, 2004 and 2005). Data collection involved selecting three cones from each of 480 trees across four plots (120 trees per plot). Five pine nuts from each cone were randomly selected and weighted individually. The pine nuts of each cone were broken manually and the kernels weighted. A random sample of five kernels per cone was taken and each kernel was weighted individually. A precise milligram scale was used for all measurements.

Generalized linear models are a versatile extension of traditional linear regression models. They offer a comprehensive framework for modeling different types of response variables by defining a suitable distribution, a linear predictor, and a link function. This adaptability enables generalized linear models to effectively handle a broad spectrum of data types and practical applications, making them a powerful tool in statistical analysis [3]. Quantile regression models [4] can characterize the entire distribution of the dependent variable by analyzing different quantiles. Generalized linear models and quantile regression models were fitted to the data. The quantile regression models were fitted to the data to account for outliers and response variable asymmetry. Different formats for the response variable,

such as individual pine nut weights and summary measures of pine nuts per cone, were evaluated using the same initial set of covariates. Model comparisons were made using AIC criteria and EQM, with statistical analyses performed in R Project, version 4.4.1. Quantile regression is more resistant to outliers and can handle data with diverse distributions, offering a more comprehensive perspective than classic regression.
Quantile regression offers a robust approach for modeling pine nut and kernel characteristics, accommodating the variability and outliers present in the data. This approach provides valuable insights for optimizing kernel efficiency in *Pinus pinea* stands.

Acknowledgements: This work was supported by PROGRAMA AGRO 200 (Project AGRO/2000/2001: "Colheita mecânica da pinha (*Pinus pinea* L.)") and funded by FCT - Foundation for Science and Technology under Projects `https://doi.org/10.54499/UIDB/05183/2020` (MED) and `https://doi.org/10.54499/UIDB/04674/2020` (CIMA).

References

[1] A. Afonso, A. C. Gonçalves and D. G. Pereira. Pinus pinea (L.) nut and kernel productivity in relation to cone, tree, and stand characteristics. *Agroforest Syst.* **94**, 2065–2079, 2020. `https://doi.org/10.1007/s10457-020-00523-4`

[2] A. C. Gonçalves, A. Afonso, D. G. Pereira and A. Pinheiro. Influence of umbrella pine (Pinus pinea L.) stand type and tree characteristics on cone production. *Agroforest Syst.* **91**, 1019–1030, 2017. `https://doi.org/10.1007/s10457-016-9975-2`

[3] P. McCullagh and J. A. Nelder. *Generalized Linear Models* . 2nd Ed., Chapman and Hall/CRC, 1989.

[4] R. Koenker and B. Basset. Regression Quantiles. *Econometrica* **46(1)**, 33–50, 1978.

A comparative analysis of inequality measures

Carla Santos[1,2] and Cristina Dias[2,3]

[1]Polytechnic Institute of Beja, Portugal
[2]NOVAMath-SST- New University of Lisbon, Portugal
[3]Polytechnic Institute of Portalegre, Portugal

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

In the literature on inequality there is a great emphasis on income inequality, and the measure that immediately emerges as a candidate for evaluating this inequality is the Gini index. Considering different fields of society, inequality appears associated with economic, social and spatial dimensions, occurring in circumstances in which different measures appear as viable alternatives to measure this inequality. Two commonly used measures of inequality are the coefficient of variation and the Variance-to-Mean Ratio. In this work we carry out a comparative analysis of the coefficient of variation and the Variance-to-Mean Ratio regarding their compliance with the four basic criteria for inequality measures, adopting a formal and axiomatic approach.

Keywords

Basic criteria for inequality measures, Coefficient of variation, Dispersion, Variance-to-mean ratio.

When the distribution of resources, opportunities or other attributes is not uniform, among the elements of a set, there is inequality. To measure this inequality, there are different statistical measures.The present study, focused on the properties of the coefficient of variation and the Variance-to-Mean ratio, was triggered by our interest in evaluating measures of inequality that could constitute alternatives to the Gini index, or complement the information provided by this index. Going beyond measuring income inequality, alternative measures to the Gini index acquire special relevance in addressing inequality in other areas of society, considering economic, social and spatial dimensions. The literature on the topic of inequality measures is very vast, however there are few studies that formally address the evaluation of measures in terms of the principles that establish their validity as inequality measures. With this work, we hope to be able to contribute to the enrichment of the literature on inequality measures, carrying out a comparative analysis of the coefficient of variation and the Variance-to-Mean Ratio regarding their compliance with the four basic criteria for inequality measures, adopting a formal and axiomatic approach.

Acknowledgements This work was partially supported by national founds of FCT-Foundation for Science and Technology under UIDB/00297/2020 and UIDP/00297/2020

References

[1] P. D. Allison. Measures of inequality. *American sociological review*, 865–880, 1978.

[2] S. R. Chakravarty. *Measuring Inequality: The Axiomatic Approach.* In: Silber, J. (eds) Handbook of Income Inequality Measurement. Recent Economic Thought Series, **71**. Springer, Dordrecht, 1999.

[3] R. Costa, S. Pérez-Duarte et al. *Not all inequality measures were created equal-the measurement of wealth inequality, its decompositions, and an application to european household wealth.* European Central Bank, Tech. Rep., 2019.

[4] C. Santos and C. Dias. An assessment of the true Gini coefficient regarding the fulfilment of the basic criteria for inequality measures. *Acta Scientiarum. Technology*, **46(1)**, e64563, 2023.

Inference for coefficient of variation and noncentrality parameters

Célia Nunes[1], Carla Santos[2,3], Manuela Oliveira[4], Isaac Akoto[5] and João Tiago Mexia[3]

[1]Department of Mathematics, and Center of Mathematics and Applications - University of Beira Interior, Portugal
[2]Polytechnic Institute of Beja, Portugal
[3]NOVAMath-SST- New University of Lisbon, Portugal
[4]Department of Mathematics, and CIMA - Center for Research on Mathematics and its Applications - University of Évora, Portugal
[5]Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana

E-mail address: *carla.santos@ipbeja.pt*

Limit distributions rising from the increase of sample non-centrality and not from sample size have been recently obtained, and may be useful for inference.This work addresses inference on coefficients of variation and non-centrality parameters, considering the intrinsic relation between them.

Keywords

ANOVA, inference, Non-centrality parameters, Coefficient of variation.

The coefficient of variation establishes the degree of variability of a data set in relation to the mean. Taking advantage of its dimensionless characteristic, the coefficient of variation is used to quantify the variability of data in several fields. Going beyond point estimation, it is often important to know information about the value of the population coefficient of variation, through confidence intervals or hypothesis tests. Limit distributions rising from the increase of sample non-centrality and not from sample size have been recently obtained, and may be useful for inference. In this paper we consider inference for coefficient of variation and non-centrality parameters, obtaining, for the normal case, confidence intervals and testing hypothesis through duality. The results for coefficient of variation are applied to normal samples, linear regression and products of variables. The results on the non-centrality parameters were applied to one-way and multi-way ANOVA.

Acknowledgements This work was partially supported by national founds of FCT-Foundation for Science and Technology under UIDB/00297/2020, UIDP/00297/2020 and UIDB/00212/2020

References

[1] J. T. Mexia and M.M. Oliveira. Asymptotic linearity and limit distributions, approximations. J. Statist. Plann. Inference, **140**, 353–357, 2010.

[2] C. Nunes, M.M. Oliveira and J. T. Mexia.Application domains for the Delta method. Statistics, **47**, 317–328, 2013.

[3] M.G. Vangel. Confidence Intervals for a Normal Coefficient of Variation. Amer. Statist **15**, 21–26, 1996.

[4] C. Nunes and J.T. and Mexia. Non-central generalized F distributions. Discuss. Math. Probab. Stat., **26(1)**, 47–61, 2006.

[5] A. T. McKay. Distribution of the coefficient of variation and the extended t distribution. J. Roy. Statist. Soc, **9(5)**, 695–698, 1932.

[6] G.E. Miller. Asymptotic test statistics for coefficients of variation. Communications in Statistics - Theory and Methods, **20(10)**, 335–3363, 1991.

ecpdist: an R package for the extended Chen-Poisson lifetime distribution

Ana Maria Abreu[1,2] and Ivo Sousa-Ferreira[1,3]

[1]Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal
[2]CIMA – Centro de Investigação em Matemática e Aplicações, Portugal
[3]CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal

E-mail addresses: *abreu@staff.uma.pt; ivo.ferreira@staff.uma.pt*

Recently, the extended Chen-Poisson (ECP) lifetime distribution [2] was proposed by compounding the Chen and zero-truncated Poisson distributions. This distribution can take a rich variety of flexible hazard shapes, being a suitable parametric alternative for modelling time-to-event data.
To enable practitioners with a user-friendly implementation of the ECP distribution, an `R` statistical package was developed, named `ecpdist` [1]. Currently, this new package computes the values of the cumulative distribution, survival, density, hazard, cumulative hazard, quantile and mean residual life functions of the ECP distribution. It will be shown that these functions can be easily plotted using a generic `R` function programmed by us. Moreover, the `ecpdist` package allows to generate pseudo-random samples and to compute some statistics, such as the measures of shape (Bowley skewness and Moors kurtosis), k-th raw moments and k-th conditional moments. The package's functionalities will be presented using several examples, where the `R` codes will be concisely explained.

Keywords

`ecpdist` package, Extended Chen-Poisson distribution, `R` statistical software.

Acknowledgements: This work is partially financed by national funds through FCT–Fundação para a Ciência e a Tecnologia, under the projects UIDB/04674/2020 (`https://doi.org/10.54499/UIDB/04674/2020`) and UIDB/00006/2020 (`https://doi.org/10.54499/UIDB/00006/2020`).

References

[1] A. M. Abreu and I. Sousa-Ferreira. *ecpdist: an R package for the extended Chen-Poisson lifetime distribution*. R package version 0.2.0, 2024. URL: `https://github.com/abreu-uma/ecpdist`.

[2] I. Sousa-Ferreira, A. M. Abreu and C. Rocha. The extended Chen-Poisson lifetime distribution. *Revstat Stat. J.* **21**(2), 173–196, 2023. DOI: `10.57805/revstat.v21i2.405`.

The impact of innovation indicators on employment in innovative companies: a European perspective using EIS

Ana Teixeira[1,2,3], Aldina Correia[1,2,3] and Ana Borges[1,2,3]

[1]CIICESI, ESTG, Instituto Polit´ecnico do Porto, Portugal

E-mail addresses: *8190090@estg.ipp.pt; aic@estg.ipp.pt; aib@estg.ipp.pt*

This study examines the relationship between several indicators derived from the "European Innovation Scoreboard 2023" on employment in innovative companies. Innovation data were considered, such as innovation expenditures not related to R&D (Research and Development), SMEs (Small and Medium Enterprises) with product innovations, collaboration between SMEs and worker qualifications. Statistical analyses, including ANOVA and multiple linear regression, were performed to demonstrate the significance of relationships between variables, along with a k-means data mining classification technique of EU Countries.

The analysis demonstrated that non-R&D innovation expenditure, the introduction of SMEs with product innovations, and collaboration between SMEs have a positive and statistically significant impact on employment in innovative companies. In contrast, worker qualifications were not found to be statistically significant. Using the k-means data mining classification technique, four clusters were identified based on the same concepts: one cluster comprises the most developed countries, another includes the least developed countries, and the remaining two clusters represent countries with intermediate levels of development.

This study contributes to the literature by providing insights into how different aspects of innovation affect employment in innovative companies and by offering a refined mathematical model based on experimental data.

Keywords

Innovation, Employment, Multiple Linear Regression, Data mining classification.

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

References

[1] L. Aldieri and C. P. Vinci. Innovation effects on employment in high-tech and low-tech industries: Evidence from large international firms within the triad. *Eurasian Business Review*, **8**, 229–243, 2018.

[2] F. Bogliacino, M. Piva and M. Vivarelli. R&D and employment: An application of the LSDVC estimator using European microdata. *Economics Letters*, **116(1)**, 56–59, 2012.

[3] J. P. Bustamante Izquierdo. Complementarities between product and process innovation and their effects on employment: A firm-level analysis of manufacturing firms in Colombia. *International Review of Applied Economics*, **38(1-2)**, 129–154, 2024.

Index of authors