PROGRAMA E RESUMOS

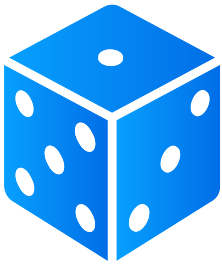
XXVII CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA

SPE 2025

FARO, 22 A 25 DE OUTUBRO DE 2025

EDIÇÕES SPE

SOCIEDADE PORTUGUESA DE ESTATÍSTICA



UNIVERSIDADE DO ALGARVE
Faro, Portugal
https://spe2025.mozellosite.com
✉: congressospe2025@gmail.com

Fundação
para a Ciência
e a Tecnologia

A organização do XXVII Congresso da Sociedade Portuguesa de Estatística (SPE 2025) foi parcialmente financiada por Fundos Portugueses através da FCT (Fundação para a Ciência e a Tecnologia) no âmbito dos Projetos UID/00006/2025 e UIDB/00006/2020, UID/4106/2025 e UID/PRR/4106/2025.

Ficha Técnica:

PROGRAMA E LIVRO DE RESUMOS

Carlos Sousa, Clara Cordeiro, Conceição Ribeiro, M. Helena Gonçalves, Nelson Antunes, Sílvia Pedro Rebouças

Editora: Sociedade Portuguesa de Estatística

Capa: Ludovico Silva (Gabinete de Comunicação da Universidade do Algarve)

Impressão: Instituto Nacional de Estatística

Tiragem: 200 exemplares

ISBN: 978-972-8890-52-0

Depósito Legal:

COMISSÃO ORGANIZADORA
(ORGANIZING COMMITTEE)

Carlos Sousa, UNIVERSIDADE *do* ALGARVE
Clara Cordeiro (Presidente CO), UNIVERSIDADE *do* ALGARVE
Conceição Ribeiro (Presidente CO), UNIVERSIDADE *do* ALGARVE
M. Helena Gonçalves, UNIVERSIDADE *do* ALGARVE
Nelson Antunes, UNIVERSIDADE *do* ALGARVE
Sílvia Pedro Rebouças, UNIVERSIDADE *do* ALGARVE

COMISSÃO CIENTÍFICA
(SCIENTIFIC COMMITTEE)

Luís Meira Machado (Presidente), UNIVERSIDADE *do* MINHO
Clara Cordeiro, UNIVERSIDADE *do* ALGARVE
Conceição Ribeiro, UNIVERSIDADE *do* ALGARVE
Isabel Natário, UNIVERSIDADE NOVA *de* LISBOA
Fernanda Otília Figueiredo, UNIVERSIDADE *do* PORTO
Magda Monteiro, UNIVERSIDADE *de* AVEIRO
Maria do Rosário Ramos, UNIVERSIDADE ABERTA
Maria João Polidoro, INSTITUTO POLITÉCNICO *do* PORTO
Marta Ferreira, UNIVERSIDADE *do* MINHO
Nuno Sepúlveda, WARSAW UNIVERSITY *of* TECHNOLOGY
Pedro Oliveira, UNIVERSIDADE *do* PORTO
Vanda Inácio, UNIVERSITY *of* EDINBURGH
Vanda M. Lourenço, UNIVERSIDADE NOVA *de* LISBOA

A todos os participantes, palestrantes e convidados,

É com enorme satisfação que a Sociedade Portuguesa de Estatística (SPE), em colaboração com a Universidade do Algarve (UAlg), acolhe o XXVII Congresso da Sociedade Portuguesa de Estatística, que de 22 a 25 de outubro de 2025, no Complexo Pedagógico do Campus da Penha, em Faro.

Este evento é um encontro aberto a todos os que partilham interesse pela Estatística e pela Ciência de Dados, sejam investigadores, profissionais ou estudantes. Reconhecido como uma das iniciativas mais marcantes da SPE, o congresso tem por missão divulgar, valorizar e impulsionar o estudo e a aplicação da Estatística e das suas áreas afins. O congresso inicia-se com o minicurso "The Hitchhiker's Guide to Responsible Machine Learning", ministrado pelo Professor Przemyslaw Biecek, da Warsaw University of Technology, Polónia. Este minicurso tem como objetivo introduzir os princípios e práticas de aprendizagem automática responsável.

O programa científico integra quatro sessões plenárias, conduzidas por investigadores de reconhecido mérito internacional:

A primeira, intitulada "Statistics and Artificial Intelligence in Medicine", de Ana Luísa Papoila, irá focar-se no papel crucial da estatística no desenvolvimento de modelos de IA. A segunda será apresentada por Przemyslaw Biecek, com o título "Explanatory Model Analysis". Durante a palestra, serão apresentados os recentes desenvolvimentos em técnicas de visualização e exploração de modelos preditivos. Na terceira, "The Unreasonable Effectiveness of Data Science", de Renato Assunção, será apresentada uma visão geral dos fatores responsáveis pela IA, com foco na contribuição da Estatística. Na quarta,"One Way to Estimate an Out-of-Sample Quantile of an Unknown Distribution Through Extreme Value Theory", de Cláudia Neves, serão abordadas várias definições de (extreme) out-of-sample quantile através da estrutura dos domínios de atração e a generalização da estimativa de um nível de retorno válido para ambos os casos de limite superior finito ou infinito para a distribuição real subjacente aos dados amostrados.

O congresso inclui ainda 10 sessões temáticas, organizadas por diferentes instituições e grupos de investigação, que abordam temas atuais e especializados da Estatística e suas aplicações. Na sessão temática organizada pelo Banco de Portugal será abordado o tema da Anonimização e Proteção de Microdados. Na sessão Estatística em Português – Estatística com África será apresentada a cooperação científica e ensino nos PALOP. Na sessão Paving the Way: Making Careers in Statistics Education Research Visible and Viable, serão exploradas questões relacionadas com carreiras na investigação em educação estatística, numa mesa redonda com os oradores online. A sessão Indústria

(PT-MATH.IN) foca-se nas aplicações da estatística à indústria. Na sessão organizada pela Comissão Especializada de Educação da SPE, CEE-SPE, as apresentações serão sobre Ensino e Divulgação da Estatística. Numa parceria entre a SPE e a Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións, será realizada online a sessão SPE–Biometria /SGAPEIO, organizada pela secção de Biometria da SPE e em simultâneo com o XVII Congreso Galego de Estatística e Investigación de Operacións. Na sessão organizada pelo Instituto Nacional de Estatística, INE, serão abordadas as Inovações em Estatísticas Oficiais. Numa parceria entre a SPE e a Associação Portuguesa de Classificação e Análise de Dados, será realizada a sessão SPE/CLAD. Numa parceria entre a SPE e a Associação Portuguesa de Investigação Operacional, será realizada a sessão SPE–APDIO sob o tema Otimização e Aplicações em Investigação Operacional. Na sessão SPE/IPQ/CT225 organizada pela Comissão Técnica do Instituto Português de Qualidade - Aplicação de Métodos Estatísticos será abordada a Standardização Nacional e Internacional em Métodos Estatísticos. Estas sessões promovem o intercâmbio de ideias e experiências entre investigadores, profissionais e estudantes, reforçando o papel da Estatística como ciência transversal e indispensável à compreensão e tomada de decisão em contextos multidisciplinares.

Para além destas sessões temáticas, contamos com mais 79 sessões orais e 48 pósteres, organizados em várias áreas, nomeadamente, Séries Temporais, Ciência de Dados, Aplicações em Ambiente, Clima Geociências e Agricultura, Estatística Multivariada, Estatística Computacional, Estatística Espacial, Métodos Não paramétricos, Análise de Sobrevivência, Extremos, Bioestatística e Epidemiologia, Aplicações em Econometria, Finanças e Gestão, Probabilidade e Processos Estocásticos e Métodos Bayesianos.

Este é o cenário ideal para homenagear pessoas que dedicaram a sua vida a esta nobre área. O Prémio Carreira representa não apenas o reconhecimento de um percurso de excelência, mas também um sincero tributo de gratidão por uma vida dedicada à ciência e ao ensino da Estatística. A Professora Eugénia Graça Martins, com a sua visão pedagógica e capacidade de motivar o pensamento crítico, contribuiu decisivamente para a valorização da Estatística como disciplina fundamental em tantas áreas do saber. O Professor Fernando Rosado, com o seu compromisso com a qualidade científica, elevou o padrão do trabalho estatístico, deixando um legado de rigor e ética.

Mantendo viva a valorização do talento e da inovação, há igualmente lugar para distinguir as comunicações mais recentes, com a atribuição do Prémio SPE 2024 e 2025 e também dos mais jovens com a atribuição do Prémio Estatístico Júnior.

Reunindo uma variedade de temas que refletem o dinamismo na área da Es-

tatística, o programa procura não só estimular novas parcerias e trajetórias profissionais, como também equilibrar o rigor académico com experiências sociais e culturais enriquecedoras.

"Abraçado pela marina e pela paisagem deslumbrante da Ria Formosa, o AP Eva Senses Hotel" é o local perfeito para a realização do Jantar do Congresso. Uma oportunidade privilegiada para fortalecer relações profissionais, complementada por uma visita guiada ao centro histórico ou um passeio de barco pela Ria Formosa, que acrescentam ao programa uma dimensão cultural e social, promovendo um equilíbrio harmonioso entre ciência e lazer.

Um agradecimento sincero aos nossos patrocinadores, voluntários e a todos aqueles que contribuíram com esforço e dedicação para tornar este evento uma realidade.

Desejamos a todos um excelente congresso!

A Comissão Organizadora

PROGRAM

—PROGRAMA—

4ªfeira - 22 de outubro

<table>
<tr><td>08:30</td><td colspan="4">Registo e entrega documentação. Hall do Auditório 1.5</td></tr>
<tr><td>09:00</td><td colspan="4">Minicurso: The Hitchhiker's Guide to Responsible Machine Learning
Przemyslaw Biecek
Auditório 0.4</td></tr>
<tr><td>10:30</td><td colspan="4">Pausa para café</td></tr>
<tr><td>11:00</td><td colspan="4">Minicurso (cont.)</td></tr>
<tr><td>12:30</td><td colspan="4">Pausa para almoço</td></tr>
<tr><td>13:30</td><td colspan="4">Minicurso (cont.)</td></tr>
<tr><td>15:30</td><td colspan="4">Pausa para café</td></tr>
<tr><td>16:00</td><td colspan="4">Sessão de abertura do congresso.
Auditório 1.5</td></tr>
<tr><td>16:15</td><td colspan="4">Sessão Plenária I - Statistics and Artificial Intelligence in Medicine
Ana Luísa Papoila
Auditório 1.5 Moderador: Luís Machado</td></tr>
<tr><td>17:15</td><td colspan="4">Comunicações Orais I</td></tr>
<tr><td></td><td>Sessão Temática - Banco de Portugal
Anonimização e Proteção de Microdados
Organizadora: Rita Sousa
Auditório 1.5</td><td>Séries Temporais I
Moderadora: Magda Monteiro
Auditório 1.4</td><td>Ciência de Dados I
Moderador: Fernando Rosado
Auditório 1.3</td><td>Aplicações em Ambiente, Clima, Geociências e Agricultura I
Moderadora: Irene Oliveira
Auditório 0.4</td></tr>
<tr><td></td><td>Data Anonymization Principles at Banco de Portugal
Francisco Fonseca, Mário Lourenço, Ricardo Marques and Ana Carvalho</td><td>Are Sequential Search Methods Effective for ARMA Model Selection?
Sónia Gouveia, Ana Martins and Manuel Scotto</td><td>The Influence of Primary Health Care Provision on Emergency Department Attendance
Loide Ascenso, Paulo Infante and Hugo Quintino</td><td>Avaliação do Potencial Energético na Zona Costeira Norte e Centro de Portugal
Ana Leonor Oliveira, Paula Milheiro-Oliveira e Paulo Avilez-Valente</td></tr>
<tr><td></td><td rowspan="2">Anonymization and Protection of Microdata
Rita Sousa, José Pedro Veiga and Susana Faria</td><td>Statistical Framework for Environmental Justice Using Models for Time Series of Counts
Adriano Gomes, Ana Martins and Sónia Gouveia</td><td>Topic Analysis and Classification on Simulated RNA-seq Datasets - a Comparative Study
João F. Carrilho, Susan P. Holmes and Marta B. Lopes</td><td>The Importance of Using Resolvable Row-Column Designs in Large Agricultural Experiments with Perennial Species
Elsa Gonçalves</td></tr>
<tr><td></td><td>Unemployment Nowcasts via Google Trends: Insights into Digital Divide in Brazil and Portugal
Maria Eduarda Silva and Eduardo Costa</td><td>Misturas Pseudo-Convexas de Funções Potência e suas Aplicações
Miguel Felgueiras, João Martins e Rui Santos</td><td>Insights from a Data-Driven Study in an Automotive Assembly Line: Defects' Analysis
Marco Silva, Ana Xambre and Helena Alvelos</td></tr>
<tr><td>18:15</td><td colspan="4">Deslocação para Museu de Faro</td></tr>
<tr><td>19:00</td><td colspan="4">Receção de boas vindas
Museu de faro</td></tr>
</table>

5ªfeira - 23 de outubro

<table>
<tr><td>08:30</td><td colspan="4">Registo e entrega documentação. Hall do Auditório 1.5</td></tr>
<tr><td>09:00</td><td colspan="4">Comunicações Orais II</td></tr>
<tr><td></td><td>Sessão Temática
Estatística em Português – Estatística com África
Organizadores: Giovanni Silva e Rita Gaio
Auditório 1.5</td><td>Séries Temporais II
Moderador: Isabel Pereira
Auditório 1.4</td><td>Sessão Temática
Paving the Way: Making Careers in Statistics Education Research Visible and Viable
Organizador: Bruno de Sousa
Auditório 1.3</td><td>Estatística Multivariada
Moderador: Paula Brito
Auditório 0.4</td></tr>
<tr><td></td><td>Cursos de Pós-Graduação em Estatística em Parceria com os PALOP no período 2021-2024
Rita Gaio, Margarida Brito, Joel Nuvunga, Paulino Fortes, Maria de Natividade, João de Almeida Pedro, Joana Chorincas e Maria Hermínia Cabral</td><td>Stability of Machine Learning Models for Time Series Forecasting
Mafalda Sá Ferreira and Regina Bispo</td><td rowspan="4">Paving the Way: Making Careers in Statistics Education Research Visible and Viable
Iddo Gal, Dani Ben-Zvi and Katie Makar</td><td>The 2024 WRC Points System and Its Dual Reward Effect
João Branco, Luis Margalho and Carla Fidalgo</td></tr>
<tr><td></td><td>O Ensino da Estatística em Moçambique
Joel Nuvunga</td><td>Clustering Zero-Inflated Time Series of Counts
Luís Sousa, Magda Monteiro and Isabel Pereira</td><td>Explorando a Perceção Face à Matemática de Alunos de 3.º e 4.º Anos de Escolaridade
Ana Felizardo Henriques, Adelaide Freitas, Fernando Sebastião e João Marôco</td></tr>
<tr><td></td><td>Estatística em Cabo Verde, uma Visão Global
Ivanilda Cabral</td><td>Signed Periodic INAR(1) Model: A Comparative Study
Cláudia Santos and Isabel Pereira</td><td>CDPCA: A New Starting Point
Guilherme Pereira, Mariline Costa and Adelaide Freitas</td></tr>
<tr><td></td><td>Discussão</td><td>Outliers in Dynamic Time Series Models: an Approach Using Robust Statistics and the Kalman Filter
Ana Ribeiro, Arminda Gonçalves and Marco Costa</td><td>Casewise and Cellwise Outliers in Panel Data: Challenges and Robust Estimation Strategies
Anabela Rocha, M. Cristina Miranda and Manuela Souto de Miranda</td></tr>
<tr><td>10:20</td><td colspan="4">Pausa para café e Sessão de Posters I</td></tr>
</table>

5ªfeira - 23 de outubro

10:20	**Pausa para café e Sessão de Posters I**
1	Adaptação e Validação da Escala Motives for Online Gaming Questionnaire (MOGQ) numa Amostra de Estudantes Universitários **Elisete Correia**, Ana Luisa Lopes e Ana Paula Monteiro
2	A Python-Based Guide to Modeling Interval-Censored Time-to-Event Data **Rui Alves**, Luis Machado and Carla Moreira
3	Analyzing Interval-Censored Survival Data: A Practical Guide Using R **Luís Filipe Meira Machado**, Carla Moreira, Rui Alves and Aurélio Sidumo
4	Survival Analysis of COVID-19 Symptom Resolution in a Portuguese Cohort Joana Costa, Leandro Duarte, Luís Machado, Ana Paula Amorim, Margarida Tavares, Paula Meireles and **Carla Moreira**
5	A Gap Time Model Based on a Defective Distribution with a Time-Varying Recurrence-Free Proportion **Ivo Sousa-Ferreira**, Ana Maria Abreu and Cristina Rocha
6	Forecasting Seasonal Influenza in Portugal Using Pharmacy Sales: A Logistic Growth Model Approach **João Brandão**, Rúben Pereira, Zilda Mendes and António Teixeira Rodrigues
7	Estimation of the Bivariate Distribution Function for Interval Censored Data **Gustavo Soutinho**, Luís Meira-Machado and Marta Azevedo
8	Early Identification of Eating Disorders: The Contribution of Statistics to Clinical Decision Support **Vânia Almeida**, Luís Machado and Andreia Gonçalves
9	A Comprehensive Exploratory Analysis on Pancreatic Adenocarcinoma **Isabel Fonseca**, Tiago Stoffel, Raquel Mugeiro Silva and Eunice Carrasquinha
10	Os Padrões Espaço-Temporais da Criminalidade de Rua e o Patrulhamento da Guarda Nacional Republicana Duarte Branco, Ana Romão e **Paula Simões**
11	Escore de Propensão: Uma Aplicação do Programa Bolsa Permanência **Rosemeire Fiaccone**, Italo Sá e Marcelo Taddeo
12	Linear Regression Analysis of Harmonized IgG Antibody Levels against the SARS-CoV-2 Spike Protein: A Cohort Study in Healthcare Workers **Ana Leonor Saraiva**, Vera Afreixo, Ausenda Machado and Vânia Gaio
13	Dor Musculosquelética e Sono em Profissionais de Saúde de Reabilitação: Aplicação do Modelo Beta-Binomial Inflacionado em Zero Rute Teixeira, **João Paulo Martins**, Simão Ferreira, Lucimere Bohn e Leonor Miranda
14	Avaliação da Fiabilidade e Eficiência de Métodos de Classificação Hierárquicos versus Não Hierárquicos **Rui Santos**, João Paulo Martins, Miguel Felgueiras e Susana Ferreira
15	Prevalence and Risk Factors of Subretinal Drusenoid Deposits in Age-related Macular Degeneration: The Coimbra Eye Study **Rita Coimbra**, Cláudia Farinha, Alina Humenyuk, Patrícia Barreto and Rufino Silva
16	Dietary Patterns in a Population-Based Cohort: The Coimbra Eye Study **Alina Humenyuk**, Patrícia Barreto, Cláudia Farinha, Rita Coimbra and Rufino Silva
17	Patient Engagement with a Digital Decision-Support Tool in Breast Cancer Surgery **Miguel Broes**, Giovani Silva and Marília Antunes
18	Classifying App Usage Data with Finite Mixture Models **Ana Sofia Barata**, Giovani Silva and Marília Antunes
19	The Impact of AI-Assisted Decision-Making on Patient Satisfaction: A BREAST-Q Study **Carolina Horta**, Giovani Silva and Marília Antunes
20	Intervalos de Referência por Métodos Indiretos **Lara Pereira**, Beatriz Saraiva, Henrique Reguengo, Ricardo Ribeiro, Margarida Brito e Rita Gaio
21	Functional Dependence at Admission as a Prognostic Factor in Palliative Care: A Survival Analysis **Nuno Domingues**, Joana Bragança, Ivo Sousa-Ferreira and Tiago Dias Domingues

5ªfeira - 23 de outubro

11:00	**Comunicações Orais III**			
	Sessão Temática Indústria (PT-MATH.IN) Organizadoras: Filipe Marques e Lígia Henriques- Auditório 1.5	**Aplicações em Ambiente, Clima, Geociências e Agricultura II** Moderador: Pedro Oliveira Auditório 1.4	**Estatística Computacional I** Moderadora: M.Helena Gonçalves Auditório 1.3	**Estatística Espacial I** Moderador: Isabel Natário Auditório 0.4
	Profit Optimization for Cattle Growth Using Stochastic Differential Equations **Gonçalo Jacinto**, Patrícia A. Filipe and Carlos A. Braumann	*Clustering, Time Series and Risk Analysis for Assessing Water Quality in a River Basin* **A. Manuela Gonçalves**, Irene Brito and Ana Pedra	*Modelo de Previsão: Uma Aplicação a uma Indústria de Calçado* Nadine Laranjeira e **Maria do Rosário Ramos**	*Supervised Statistical Learning Methods in the Presence of Spatial Correlation* **Beatriz Ferreira** and Raquel Menezes
	Estimação do AGB na Floresta Utilizando Dados de Satélite **Ricardo Coelho**, Isabel Natário e Silvia Fraile	*Boost Fisher Scoring: A Robust Approach to Parameter Estimation in State-Space Models* **F. Catarina Pereira**, Marco Costa and A. Manuela Gonçalves	*Group Lasso for Finite Mixtures of Linear Regression Models: A Simulation Study* **Ana Moreira** and Susana Faria	*Spatial Analysis of Ascending Thoracic Aortic Aneurysms* **Alda Carvalho**, Katalina Oviedo Rodríguez, Rodrigo Valente, José Xavier and António Tomás
	Our Journey to a More Sustainable Digital Infrastructure **Pedro Nobre**, Fábio Coelho and Mafalda Sá Ferreira	*Forecasting of Pollen Concentrations in Évora* **Ana Sapata**, Anabela Afonso, Célia Antunes and José Saias	*Analytical Properties of Kalman Predictor Derivatives in State-Space Models* **Marco Costa** and Magda Monteiro	*Modelling Wildfires In The UK Using Spatio-Temporal Point Processes* **Gordon Hannah**
12:00	**Sessão Plenária II - Explanatory Model Analysis** **Biecek Przemyslaw** Auditório 1.5 Moderadora: Clara Cordeiro			
13:00	**Pausa para almoço**			
14:00	**Sessão Plenária III - The Unreasonable Effectiveness of Data Science** **Renato Assunção** Auditório 1.5 Moderadora: Conceição Ribeiro			
15:30	**Passeio do Congresso**			

6ªfeira - 24 de outubro

08:30	**Registo e entrega documentação. Hall do Auditório**			
09:00	**Comunicações Orais IV**			
	Sessão Temática CEE-SPE ***Ensino e Divulgação da Estatística*** Organizadora: Adelaide Freitas Auditório 1.5	**Métodos Não Paramétricos** Moderadora: Sílvia Pedro Rebouças Auditório 1.4	**Análise de Sobrevivência** Moderador: Carlos Sousa Auditório 1.3	**Extremos** Moderador: Maria da Graça Temido Auditório 0.4
	Jornalismo de Dados: Uma Oportunidade para Educação e Divulgação Estatística **Cláudia Silvestre**	*Rank and Related Tests: A Randomization Procedure for Grouping Factor Levels in Cocoa Breeding Experiments* Kwaku Opoku-Ameyaw, **Célia Nunes** and Manuel L. Esquível	*Avanços em Modelos de Sobrevivência Multi-estado Não-Markovianos: Revisão Sistemática e Perspetivas Futuras* **Marta Azevedo**, Luís Meira-Machado e Carla Moreira	*Weibull Tail Coefficient Estimation via Linear Combinations* **Maria Ivette Gomes**, Frederico Caeiro, Fernanda Otília Figueiredo and Lígia Henriques-Rodrigues
		Métodos de Agregação: Contributo dos Estimadores Baseados em Entropia **Ana Helena Tavares**, Maria Costa e Pedro Macedo	*Evaluating the Linearity of a Covariate in Shared-Parameter Joint Models* **Xavier Piulachs**, Anouar El Ghouch and Ingrid Van Keilegom	*Testing the Domain of Attraction for Maxima* **Frederico Caeiro**, Ivette Gomes, Lígia Henriques-Rodrigues and Cláudia Neves
	A Importância da Estatística na Educação para a Cidadania e o Papel do Projeto ALEA **Pedro Campos**	*Métodos de Reamostragem na Estimação do Índice Extremal* **Dora Prata Gomes** and Manuela Neves	*Using A Joint Model for Longitudinal and Time-to-Event Data to Estimate the Causal Effect of Liver Transplantation on Survival in Hepatocellular Carcinoma Patients* **Pedro Miranda-Afonso**, Hao Liu, Michele Molinari and Dimitris Rizopoulos	*The Extremal Index P-Estimator: a Photovoltaic Energy Data Application* **M. Cristina Miranda**, Manuela Souto de Miranda, Conceição Amado and M. Ivette Gomes
		Simulation Study on Projection-based Goodness-of-fit Tests for Generalized Linear Models **Rui Costa-Miranda**, Rita Gaio and Wenceslao González-Manteiga	*A Bayesian Approach for Modeling Time-to-event Distal Outcomes* **Leila Denise A. F. Amorim**, Marcos Aurélio Eustorgio-Filho and Lilia Carolina C. Costa	*Deteção de Mudanças de Estruturas em Series Temporais* **Ludomilo Rebelo Almeida**, Dulce Gomes e Lígia Henriques-Rodrigues
	Educação Estatística 2025+ **Bruno de Sousa**	*A Delta Sequence Class of Density Derivative Estimators for Circular Data* **Carlos Tenreiro**	*Unveiling the Operational Reliability of Coastal Tide Gauges: A Comparative Survival Analysis Enhanced with Statistical and Machine Learning Approaches* **Dora Carinhas**, Paulo Infante and António Martinho	*A Novel Approach to Model Long Time Survival Times in Highly Censored Data* Eduardo Janotti Cavalcante, António Carlos Pedroso de Lima and **Lígia Henriques-Rodrigues**
10:40	**Pausa para café e Sessão de Posters II**			

6ªfeira - 24 de outubro

10:40	**Pausa para café e Sessão de Posters II**
22	Robustness Evaluation of Machine Learning Models in Genomic Prediction **Vanda Lourenço**, Joseph Ogutu and Hans-Peter Piepho
23	Optimizing Herbicide Use in Precision Agriculture through Classification Models **Alexandre Aparecido da Silva** and Luiz Fernando Carvalho
24	From Behaviour to Personas: A Machine Learning Approach to Understand Adaptive Thermal Comfort Strategies in Workspaces **Celina P. Leão**, Lumy Noda, Amanda V.P. Lima and Solange Leder
25	A Comparison Between Location Models and Regression Structures **Gabriel Macedo** and Luiz Carvalho
26	Applying Clustering Approaches to GAMLSS **Iago Macarini** and Luiz Carvalho
27	Multiplicative Algebra of Random Variables, Contraction and Expansion **Dinis Pestana**, Sandra Mendonça and Neto Pascoal
28	Transformação de Dados na Análise Estatística: Desenvolvimento de uma Framework de Apoio à Decisão **João Correia**, M. Rosário Ramos e Patricia Engrácia
29	Welch t and Power Means Sí**lvio Velosa** and Sandra Mendonça
30	Estandardização em Modelos Lineares: Quando é Útil, Quando é Neutra e Quando Atrapalha? **Dulce Pereira** e Anabela Afonso
31	Regressão Quantílica com Efeitos Fixos e Mistos: Comparação de Funções Disponíveis no R **Anabela Afonso** e Dulce Pereira
32	Pacing Strategies in 800m and 1500m Freestyle: A Data-Driven Analysis from the 2024 Olympics **Joana Pinto** and Carlos J. Costa
33	Relação entre o Microbioma Uterino e a Adesão à Dieta Mediterrânica: Metodologias Estatísticas **Laura Vieira**, Analuce Canha Gouveia e Délia Gouveia Reis
34	An Exploratory Comparison of Metaheuristic Algorithms for Threshold Selection in Extreme Value Analysis **Beatriz Leça Pereira**, Luiz Guerreiro Lopes and Délia Gouveia Reis
35	Misturas de Gaussianas e Cotação de Criptomoedas **Susana Ferreira**, Rui Santos e Miguel Felgueiras
36	Multivariate Empirical Bayes Analysis for Time-Resolved Omics: A Grapevine-Pathogen Infection Case Study Nuno Domingues, **Lisete Sousa**, Gonçalo Laureano, Vincent Carré, Jasmine Hertzog, Andreia Figueiredo and Marisa Maia
37	Maximum Likelihood Estimation for a Folded Directional Distribution **Adelaide Figueiredo** and Fernanda Otília Figueiredo
38	Predictor Variable Selection for Mathematics Achievement: A Study Using PISA Data **Susana Faria**
39	A Statistical Approach to Pandemic Impact Analysis: Clustering Time Series Features in the US Retail Sector José Canoso and **Joana Leite**
40	Bivariate INAR Models with Zero Inflated Innovations Sandra Dias, **Maria da Graça Temido** and Cristina Martins
41	Exploring Statistical Indices for Weekly Seasonality Analysis Joana Carvalheiro, Joana Leite and **Clara Viseu**
42	Improved Estimation of the Shape Parameter for the Shifted Log-Logistic Distribution: Theory and Applications **Ayana Mateus** and Frederico Caeiro

43	A Comparative Study of Some Parametric and Nonparametric Control Charts **Fernanda Otilia Figueiredo**, Adelaide Maria Figueiredo and Maria Ivette Gomes
44	Root Cause Analysis in Surface Mount Technology through Explainable AI **Ana Marinho** and Luís Araújo
45	Holistic Defect Detection in Surface-Mount Production Lines with a Machine Learning Approach **Gabriel Quartin** and Luís Araújo
46	Economic Inequality in Europe: Interplay Between Income and Wealth **Kamila Trzcińska** and Elżbieta Zalewska
47	Goodness-of-fit in multivariate ARMA-based models **Ana Martins** and Sónia Gouveia
48	Factors influencing student engagement in different forms Cristina Veríssimo, Joselina Barbosa, Milton Severo, Paula Mena Matos, **Pedro Oliveira** and Laura Ribeiro

6ªfeira - 24 de outubro

11:30 **Comunicações Orais V**

Sessão Temática - Biometria/SGAPEIO Organizador: Nuno Sepulveda SPE-SBIO Auditório 1.5	**Sessão Temática - INE** **Inovações em Estatística Oficiais** Organizador: Pedro Campos Auditório 1.4	**Ciência de Dados II** Moderador: Maria Eduarda Silva Auditório 1.3	**Estatística Espacial II** Moderadora: Raquel Meneses Auditório 0.4
Inferring Infectious Disease Risk in Real Time – Methodology and Public Health Implications **Ricardo Águas**	*Indicadores de Acessibilidade a Serviços de Interesse Geral* **Joana Malta**, Maria Aurindo, Carla Cardoso e Rita Santos	*Synthetic Integer-Valued Times Series Generation: an Experimental Study* I**sabel Silva**, Isabel Pereira and Maria Eduarda Silva	*Imputation for Net Income on Rotating Panel Data: na Approach with Conditional Autoregressive (CAR) Models on Portuguese Labor Survey* **Antonio Loría-García**, Lígia Henriques-Rodrigues and Pedro Campos
	A Inovação nas Estatísticas Oficiais em Portugal **Sofia Rodrigues,** Paulo Saraiva, Glória Carrilho e João Poças	*Modelação de Eventos Raros em Sinistralidade Rodoviária: Comparação de Técnicas de Reamostragem e Algoritmos de Classificação* **Lorena Santos**, Gonçalo Jacinto, Anabela Afonso e Paulo Infante	*Visual Spatial Learning: Single-Field Spatial Interpolation Using Convolutional Neural Networks* **Daniel Tinoco**, Raquel Menezes and Carlos Baquero
Nonparametric Model Check for Cure Rate Quantile Regression **Mercedes Conde Amboage,** Wenceslao González-Manteiga and César A. Sánchez-Sellero	*Not Just Another Black Box: AI in the Age of Trusted Statistics* **Sónia Quaresma**	*Statistical Techniques for Real-Time Digital Twins and the Industrial Metaverse* **Cecília Castro**	*A Zero-inflated Spatio-temporal Approach for Joint Modeling of Fishery-depedent and Fishery-independent Data to Understand Fish Distribution* **Daniela Silva**, Raquel Menezes, Gonçalo Araújo, Ana Teles-Machado, Renato Rosa, Ana Moreno, Alexandra Silva and Susana Garrido

12:30 **Sessão Plenária IV - One Way to Estimate an Out-of-Sample Quantile of an Unknown Distribution Through Extreme Value Theory**
Cláudia Neves
Auditório 1.5 Moderadora: Maria Ivette Gomes

13:30 **Pausa para almoço**

6ªfeira - 24 de outubro

14:30	**Foto de Grupo**			
14:40	**Comunicações Orais VI**			
	Sessão Temática - SPE/CLAD Organizadoras: Conceição Amado Auditório 1.5	**Bioestatística e Epidemiologia I** Moderadora: Lisete Sousa Auditório 1.4	**Sessão Temática - APDIO** Organizador: Filipe Alvelos Auditório 1.3	**Estatística Espacial III** Moderadora: Arminda Manuela Gonçalves Auditório 0.4
	A Inclusão Racial na Publicidade de Cosméticos no Instagram: Representação e Impacto no Envolvimento dos Utilizadores **Catarina Marques**, Daniela Langaro e Mariana Cintra	*Integrating Supervised and Unsupervised Learning for Variant Post-Filtering in Whole-Genome Sequencing* **Vera Pinto**, Lisete Sousa and Carina Silva	*Modelização do Comportamento do Vento no Contexto da Propagação de Fogos Florestais* **Helena Alvelos**, Ana Raquel Xambre, Francisco Marques, Agostinho Agra e Filipe Alvelos	*Joint Modeling of Spatial Intensity, Detectability and Marks in Caribou Surveys using inlabru* **Iúri J. F. Correia**, Soraia A. Pereira, Tiago A. Marques, Christine Cuyler and Marta M. Rufino
	Avaliação Não-Destrutiva da Qualidade de Uvas com Imagiologia Hiperespectral **Irene Oliveira**	*Cutoff-Free Sero-Epidemiological Analysis of Infectious Diseases* **Nuno Sepulveda**	*Uma Abordagem de Otimização em Dois Níveis para a Agregação da Flexibilidade no Consumo de Energia Elétrica* **Carlos Henggeler Antunes**, Inês Soares, Ana Soares e Maria João Alves	*Modelling Street Crime in Almada, Portugal, Using Point Processes* **Inês Oliveira**, Paula Simões and Isabel Natário
	Aquaponics for Sustainable Production: some Statistical Approaches **Fernando Sebastião**, Judite Vieira, Luís Cotrim, Damariz Y. Ushina, Ounísia Santos, Vânia S. Ribeiro, Daniela C. Vaz and Raul Bernardino	*Comparative Study of the Most Used Growth Models Applied to Weight in Infants Aged 0 to 2 Years* **Marta Alves**, Marisol Garzón, Bruno Heleno, Ana Luisa Papoila and Carlos Geraldes	*Otimização do Layout de Parques Eólicos com Fatores de Carga Heterogéneos* **Adelaide Cerveira** e Agostinho Agra	*Sobre a Validação Cruzada em Dados Dependentes* **Isabel Natário** e Ricardo Coelho
		Stratification in ME/CFS: Association Between Domain-Specific Severity Profiles and Herpesvirus Antibody Responses **João Malato**, Luis Graca, Ji-Sook Lee, Jacqueline Cliff, Luis Nacul, Eliana Lacerda and Nuno Sepulveda	*Optimização Robusta com Distribuições para o Problema de Pré-posicionamento de Recursos em Incêndios Florestais* **Filipe Alvelos**, Agostinho Agra, Francisco Marques, Ana Raquel Xambre e Helena Alvelos	*Modeling the Spatial Distribution of Dinosaur Fossil Records* **Carolina S. Marques**, Soraia Pereira, Emmanuel Dufourq, Elisabete Malafaia, Pedro Mocho, Joana Órfão and Vanda F. Santos
16:00	**Pausa para café**			
16:30	***Sessão Temática - SPE IPQ/CT225*** Organizadora: Mafalda Costa Auditório 1.5 *(Inter)national Standardization: Ongoing Projects on Applications of Statistical Methods* **Maria João Polidoro**, **Natércia Durão**, Mafalda Costa, Fernanda Figueiredo, Sónia Quaresma and Pedro Campos Curation, Cleansing and Wrangling of Big and Large Datasets **Sónia Quaresma**			
17:30	**Prémios Carreira || Organizadores: Dinis Pestana e Manuela Neves**			
18:30	**Assembleia Geral**			
19:15	**Ida para o jantar**			
19:45	**Jantar do Congresso**			

sábado - 25 de outubro

09:00	**Comunicações Orais VII**		
	Aplicações em Econometria, Finanças e Gestão Moderador: Dulce Pereira Auditório 1.5	**Séries Temporais III** Moderadora: Maria do Rosário Auditório 1.4	**Ciência de Dados III** Moderadora: Anabela Afonso Auditório 1.3
	Stochastic Differential Equation Harvesting Models and Effects of Parameter Estimation Errors **Carlos A. Braumann** and Nuno M. Brites	*Forecasting Hotel Demand* **Clara Cordeiro**, Nuno António and Sara Galguinho	*The Work Climate Questionnaire (WCQ) for Volunteer Settings: Psychometric Properties in a Portuguese Sample* Ricardo Batista, **Conceição Ribeiro**, Rita dos Santos, Marta Brás, Maria Dulce Estevão, Cláudia Carmo, Saul Neves de Jesus, José Tomás da Silva and Cátia Martins
	Predicting Daily Euro-Dollar Exchange Rate with SARIMA, LSTM and Decomposition-based models **Vasco Carneiro**	*INAR Models with Structural Breaks: a CUSUM-Guided Maximum Likelihood Approach* **Magda Monteiro** and Isabel Pereira	*Neural Network Binary Predictions Explanation through Propensity Score Methodology* **Luis Garcez**, João Telhada and Eduardo Severino
	Distinguishing Repeat and First-Time Hotel Guests: A Comparative Analysis of Classification Models **Sílvia Pedro Rebouças**, Inês Gonçalves, Conceição Ribeiro, Tiago Candeias and Miguel Portugal	*Structural Breaks in Overdispersed INAR Models: A Bayesian Approach* **Isabel Pereira**, Magda Monteiro and Maniha Zafar	*Modelos Preditivos para Dados Longitudinais: Um Estudo de Simulação* **Elsa Soares** e Inês Sousa
	Aplicação de Redes Neuronais Artificiais à Previsão de Preços de Criptomoedas **José Cruz** e Tiago Marques	*Data-Driven Fragmented Autocorrelation for Improved Time Series Clustering* **Jorge Caiado** and Nuno Crato	*A Principal Component Analysis for Ordinal Data* **Hugo Alonso** and Adelaide Freitas
10:20	**Pausa para café**		

sábado - 25 de outubro

10:40	**Comunicações Orais VIII**			
	Probabilidade e Processos Estocásticos Moderador: Nelson Antunes Auditório 1.5	**Métodos Bayesianos** Moderadora: Regina Bispo Auditório 1.4	**Estatística Computacional II** Moderadora: Vanda Lourenço Auditório 1.3	**Bioestatística e Epidemiologia II** Moderadora: Marília Antunes Auditório 0.4
	Unravelling Causal Dependencies in Climate Indices Using Mutual Information Rate Decomposition **Helder Pinto**, Susana Barbosa, Maria Eduarda Silva and Ana Paula Rocha	*Modelling Mobility Data during COVID-19 in Portugal with R-INLA* **André Brito**, Ausenda Machado, Ana Paula Rodrigues, Paula Patrício and Regina Bispo	*Predicting Model Degradation under Noisy Conditions: A Robustness Study for Regression Problems* Catarina Fernandes, Fátima Rbaibi, Isabela Alves, Nelson Vieira and **Luís Silva**	*Unveiling the Power of the Aranda-Ordaz Link in GAMLSS: A Comparative Study for Binary Data* **Neto Pascoal**, Eunice Carrasquinha and Carlos Geraldes
	A New Approach to the Delta Approximation Method for Mixed Stochastic Differential Equation Models **Patrícia A. Filipe**, Gonçalo Jacinto and Carlos A. Braumann	*The Logistic-Normal distribution: a powerful prior on the simplex* **Rui Martins**	*Modelling Distributional Data as Matrix-valued Data* Marcus Mayrhofer, **Paula Brito**, A. Pedro Duarte Silva and Peter Filzmoser	*Analyzing Vaccination Risks Compared to Infection Risks: a Game Theory Perspective Considering Reinfection* **José Martins** and Alberto Pinto
	Longitudinal Count Data: A Simulaton Study using R **M. Helena Gonçalves** and M. Salomé Cabral	*Mapping Urban Fire Intensity in Portugal: A Bayesian Approach with INLA and SPDE* **Nádia Bachir**, Regina Bispo and Lígia Henriques-Rodrigues	*Maximum Likelihood Estimation of the Parameters of the Power-Normal Distribution* **Rui Gonçalves**	*Metascience In Ecology: The Role Of Hypothesis Testing In Ecology* **Gabriela Xavier Quintais**, Tiago André Marques and Daniel Lakens
11:40	**Prémio SPE 2024**			
	On the Theory of Spatio-Temporal Models for Time Series of Counts Dependence Models **Ana Martins**, Manuel Scotto, Christian Weiß and Sónia Gouveia			
12:00	**Prémio SPE 2025**			
	Neural Bayes Inference for Complex Bivariate Extremal Dependence Models **Lídia André**, Jennifer Wadsworth and Raphaël Huser			
12:20	**Prémio Estatístico Júnior**			
12:40	**Encerramento do Congresso** **Auditório 1.5**			
13:10	**Almoço**			

ABSTRACTS

—RESUMOS—

Table of Contents
(Índice)

Short Course

—Minicurso—

The Hitchhiker's Guide to Responsible Machine Learning

Przemyslaw Biecek [1,2] [0000-0001-8423-1823]
przemyslaw.biecek@pw.edu.pl

[1] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
[2] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland*

Abstract: The purpose of the workshop is to discuss and test methods for explaining predictive models such as LIME, SHAP, Partial Dependence, Variable Importance, etc.
We will discuss and practise these methods with the DALEX package (for R and Python), so in addition to the methodological part, part of the time will be devoted to a hands-on workshop on your own laptop.
As the backbone we will use is the monograph Explanatory Model Analysis - Explore, Explain, and Examine Predictive Models available online https://ema.drwhy.ai/.
After the workshop, participants will have a practical understanding and experience in techniques needed to compare and explain several predictive models.

Plenary Sessions

—Sessões Plenárias—

Statistics and Artificial Intelligence in Medicine

Ana Luísa Papoila [1,2] [0000-0002-2918-8364]
ana.papoila@nms.unl.pt

[1] *Gabinete de Estatística do Centro de Investigação do Centro Hospitalar Universitário de Lisboa Central, EPE, Nova Medical School, 1169-045 Lisbon, Portugal*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: The growing role of artificial intelligence (AI) in medicine is extremely promising, though it comes with its own challenges. Among many others, I emphasize the lack of transparency in AI and those more problematic issues related to clinical practice.
Despite impressive results in areas like disease diagnosis, image processing, and other medical applications, AI systems still struggle to replicate the depth of human knowledge, expertise and clinical sensitivity, necessary for better decision-making. To address the transparency issue, many researchers have turned their attention to Explainable AI (XAI) techniques, which aim to clarify both the algorithms and the results produced by AI systems. This talk will focus on the crucial role of statistics in the development of AI models and examine how machine learning tools, such as artificial neural networks, can improve both performance and interpretability through the integration of statistical methods.

Explanatory Model Analysis

Przemyslaw Biecek [1,2] [0000-0001-8423-1823]
przemyslaw.biecek@pw.edu.pl

[1] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
[2] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland*

Abstract: During the talk, I will present the recent development of techniques for visualization and exploration of predictive models, developed under the terms interpretable machine learning (IML) / explainable artificial intelligence (XAI). I will present the most popular approaches used for model exploration (attribution methods, variable profiles, variable importance, Rashomon set analysis) and demonstrate their usefulness using real medical data analysis as an example. I will conclude the lecture with a brief discussion of current challenges and open problems that we need to solve along the way to Responsible ML.

The Unreasonable Effectiveness of Data Science

Renato Assunção[1]
assuncao@dcc.ufmg.br

[1] *Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*

Abstract: There are three factors responsible for the revolution brought about by artificial intelligence: (1) the constant increase in computational capacity; (2) the accumulation of large amounts of data generating insights and enabling the creation of data-driven products; (3) the development of statistical learning theory and its algorithms. The alignment of these planets allowed great success in difficult tasks such as the development of virtual assistants and chatbots, self-driving car, the automatic translation between languages, and the early detection of unspecified anomalies in vital signs. In this talk, I will present an overview of these developments from a historical point of view focusing on the contribution brought by Statistics. I will illustrate this presentation with examples from my own research on epidemiological surveillance using social media data, space-time demographic forecasting, and the Bayesian spatial partitioning of space-time maps.

One Way to Estimate an Out-of-Sample Quantile of an Unknown Distribution Through Extreme Value Theory

Claúdia Neves [1,2] [0000-0003-1201-5720]
claudia.neves@kcl.ac.uk

[1] *Department of Mathematics, King's College London, London, UK*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: Within the general aim of extreme value statistics lies the estimation of an event that is so rare that might have never been witnessed in the past. While parametric estimation of an extreme quantile has now climbed to the lore of risk assessment, especially in terms of return levels by way of hazard curves, analogous non-parametric statistical methodology is far less explored. This creates an interesting topic, in part because there are distinct albeit equivalent ways to define an (extreme) out-of-sample quantile supported by different constructs rooted in the same foundational extreme value theorem.
In this talk, I will address two of these definitions through the domains of attraction framework and will explain how we succeeded in generalising the estimation of a return level that is valid for both cases of finite or infinite upper bound to the actual distribution underlying the sampled data.

Thematic Sessions

—Sessões Temáticas—

Anonymization and Protection of Microdata - Portuguese Central Bank

Anonimização e Proteção de Microdados - Banco de Portugal

Rita Sousa
Banco de Portugal, Centro de Matemática e Aplicações - FCT/UNL
rcsousa@bportugal.pt

Data Anonymization Principles at Banco de Portugal

Francisco Fonseca [1] [0009-0000-1744-364X], Ana Filipa Carvalho [1], Mário Lourenço [1] [0009-0009-6581-6450], Ricardo Marques [1]
ffonseca@bportugal.pt, afcarvalho@bportugal.pt, mfllourenco@bportugal.pt, rjlmarques@bportugal.pt

[1] *Banco de Portugal, Portugal*

Abstract: Central Banks collect increasingly larger sets of granular data which frequently contain personal data. The analytical potential of these datasets is greatly enhanced whenever they have shared identifiers, meaning they can be combined in order to maximize the value of the available information. However, given the sensitive nature of some of these datasets, particularly when referring to natural persons, the protection of personal data must always be a priority. We focus on the essential principles that must be observed when defining a robust anonymization model aimed at preserving the confidentiality of individual data, while still allowing for the use of different combinations of the available datasets for analytical purposes. We focus on pseudonymization mechanisms and the definition of different access profiles as cornerstones of a centralized anonymization policy for Banco de Portugal's databases.

Keywords: Anonymization · Data protection · Data utility · Personal data · Pseudonymization

References

[1] Moreno, M.C.: Data Governance: an orchestra of people, processes, and technology. In: Proceedings of the ISI IFC High Level Meeting on Data Governance. IFC Bulletin 54, pp.73-81, 2021.

[2] Gonçalves, A., Lourenço, M., Sousa, D., Verheij, T.: New strategy of data sharing and data access in statistics: the view from Banco de Portugal. In: Proceedings of the 3rd Irving Fisher Committee workshop on data science in central banking. IFC Bulletin 64, pp.471-487, 2025.

Anonymization and Protection of Microdata

Rita Sousa [1[0000-0003-3703-0476]],
José Pedro Veiga [2[0009-0008-6533-6986]],
Susana Faria [2,3[0000-0001-8014-9902]]
rcsousa@bportugal.pt, jpedroveiga07@gmail.com, sfaria@math.uminho.pt

[1] *Banco de Portugal, Centro de Matemática e Aplicações - FCT/UNL*
[2] *Universidade do Minho*
[3] *Centro de Matemática da Universidade do Minho*

Abstract: The exponential growth of data collection and analysis brings new opportunities and challenges in Research and Policy. However, it also raises critical concerns regarding privacy and confidentiality. *Microdata* - unit-level data about individuals or entities - are particularly sensitive and require robust protection mechanisms before being made available for analysis, in order to minimize the risk of re-identification [1]. The field of *Statistical Disclosure Control* (SDC) offers a set of methods designed to modify microdata while preserving their analytical value.
SDC techniques fall into three main categories: *Non-perturbative methods* - reduce detail without altering data values; *Perturbative methods* - modify data values to introduce uncertainty; *Synthetic data generation* - creates artificial datasets that preserve the statistical properties of the original [2]. Core concepts include *Identification Risk* and *Information Loss*, with the aim of balancing privacy protection and data utility. The distinction between *Pseudonymization* and *Anonymization* is particularly important in legal contexts: pseudonymized data can potentially be traced back to individuals while anonymization intends to be irreversible and compliant with regulations such as the General Data Protection Regulation (GDPR) [3].
This study evaluates risk and utility metrics using tools in R, such as `sdcMicro` for traditional methods and `diffpriv` for Differential Privacy (DP) [4], a mathematically rigorous approach that introduces randomness to ensure individual-level privacy. DP is tested through simulations and applied to a real dataset, offering practical insight into its effectiveness and limitations.
Studying different anonymization approaches is essential to ensure privacy, regulatory compliance, and utility in microdata disclosure.

Keywords: Anonymization · Differential privacy · Microdata

References

[1] Templ, M.: Statistical Disclosure Control for Microdata. Methods and Applications in R. Springer International Publishing (2017). DOI:http://dx.doi.org/10.1007/978-3-319-50272-4

[2] Hundepool, A. *et al.*: Statistical Disclosure Control. Wiley (2012). DOI:http://dx.doi.org/10.1002/9781118348239

[3] European Union. Regulation (EU) 2016/679, General Data Protection Regulation (2016).

[4] Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, **Vol.**(9), 211-407 (2014). DOI:https://doi.org/10.1561/0400000042

Statistics in Portuguese - Statistics with Africa

Estatística em Português - Estatística com África

Giovani Silva
Instituto Superior Técnico, Universidade de Lisboa e CEAUL
giovani.silva@tecnico.ulisboa.pt
Rita Gaio
Faculdade de Ciências, Universidade do Porto e CMUP
argaio@fc.up.pt

Estatística em Português – Estatística com África

Giovani Silva [1,2][0000-0002-7434-2383] e Rita Gaio [3,4][0000-0003-3906-0775]
giovani.silva@tecnico.ulisboa.pt, argaio@fc.up.pt

[1] *Dep. Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[2] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[3] *Dep. Matemática, Faculdade de Ciências, Universidade do Porto*
[4] *Centro de Matemática da Universidade do Porto (CMUP)*

Abstract: A sessão aborda o tema da cooperação com os Países Africanos de Língua Oficial Portuguesa (PALOP) e consistirá de 3 apresentações, feitas por Adilson Silva (Faculdade de Ciências e Tecnologia da Universidade de Cabo Verde), Joel Nuvunga (Faculdade de Ciências e Tecnologia da Universidade Joaquim Chissano, Moçambique) e Rita Gaio (Faculdade de Ciências da Universidade do Porto). Os 3 oradores prestarão testemunho da sua experiência de colaboração, quer científica, quer ao nível da implementação e docência de cursos de formação pós-graduada.
Pretende-se também discutir a possibilidade de criação de uma nova secção da Sociedade Portuguesa de Estatística (SPE) – por enquanto intitulada "Estatística em Português" - cujo objetivo principal será o reforço das colaborações científicas e de divulgação no âmbito da Estatística com a Comunidade dos Países de Língua Portuguesa (CPLP) ao nível da investigação, formação avançada ou divulgação. Pretende-se estabelecer, fomentar e consolidar parcerias científicas e pedagógicas, contribuindo para o desenvolvimento da literacia estatística nesses países.
De facto, são vários os estudos e organizações que reconhecem a importância do reforço do ensino da Estatística/Matemática, salientando o seu papel no desenvolvimento de competências ao nível do pensamento crítico e na resolução de problemas e contribuindo para o desenvolvimento socioeconómico. Melhores e mais fortes competências em Estatística e da Análise de Dados irão preparar docentes, profissionais e estudantes para melhor enfrentarem os desafios contemporâneos, permitindo uma participação ativa na economia global do conhecimento.

Keywords: CPLP · Estatística · PALOP

Paving the Way: Making Careers in Statistics Education Research Visible and Viable

Abrindo Caminho: Tornando Carreiras de Investigação em Estatística Educacional Visíveis e Viáveis

XXVII CONGRESSO SPE 25 SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Bruno de Sousa
Faculdade de Psicologia e de Ciências da Educação, CINEICC
Universidade de Coimbra
bruno.desousa@fpce.uc.pt

Paving the Way: Making Careers in Statistics Education Research Visible and Viable

Iddo Gal [1][0000-0001-9817-3150], **Dani Ben-Zvi** [2][0000-0002-9946-3456], **Katie Makar** [3][0000-0002-4707-8898], and **Bruno de Sousa** [4][0000-0001-9918-8100]

iddo@research.haifa.ac.il, dbenzvi@univ.haifa.ac.il, k.makar@uq.edu.au, bruno.desousa@fpce.uc.pt

[1] *Faculty of Social Welfare and Health Sciences, University of Haifa, Israel*
[2] *Faculty of Education, University of Haifa, Israel*
[3] *School of Education, The University of Queensland, Australia*
[4] *Faculdade de Psicologia e de Ciências da Educação, CINEICC, Universidade de Coimbra, Portugal*

Abstract: This roundtable chaired by Bruno de Sousa invites participants to explore the growing field of statistics education research by highlighting both career pathways and active research initiatives. Iddo Gal will discuss emerging opportunities in statistics education research, including academic, governmental, and interdisciplinary roles, with an emphasis on how to enter and sustain a career in this evolving field. Complementing this perspective, Dani Ben-Zvi and Katie Makar will share insights from their projects, illustrating the range and impact of research topics in the discipline. Together, the session aims to make careers in statistics education research more visible and viable by showcasing concrete examples, fostering dialogue, and encouraging participation from early-career scholars and educators interested in shaping the future of statistics learning and teaching.

Keywords: Educational pathways · Professional development · Research careers · Research projects · Statistics education

Industry (PT-MATH.IN)

Indústria (PT-MATH.IN)

Filipe Marques
NOVA School of Science and Technology and NOVA Math
fjm@fct.unl.pt
Lígia Henriques-Rodrigues
School of Science and Technology (ECT-UE) and CIMA
ligiahr@uevora.pt

Profit Optimization for Cattle Growth Using Stochastic Differential Equations

Gonçalo Jacinto [1[0000-0002-3292-2208]], Patrícia A. Filipe [2[0000-0003-3664-7239]], Carlos A. Braumann [1[0000-0003-2721-9750]]
gjcj@uevora.pt, patricia.filipe@iscte-iul.pt, braumann@uevora.pt

[1] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações (CIMA); Universidade de Évora, Escola de Ciências e Tecnologia, Portugal.*
[2] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações (CIMA); Iscte Business School, Iscte-Instituto Universitário de Lisboa; Business Research Unit (BRU-Iscte), Portugal*

Abstract: In previous studies we have used stochastic versions of classical growth models, formulated as stochastic differential equations (SDEs). We derive explicit expressions for the profit probability distribution, its first two moments, and other relevant quantities under a realistic market structure, where the price per kilogram depends on the animal's age and weight category. We apply these results to real weight data from Mertolengo cattle males, using the Gompertz SDE model. The analysis shows that animals are typically sold before reaching the optimal age, resulting in a lower average profit.
In this work we extend the model to a more complex SDE where key growth parameters, the asymptotic size and the growth rate, vary randomly across individuals. This mixed-effects SDE, takes into account the animal's variability, and using the new proposed Delta approximation method that approximates the integrals involved in the likelihood function, allows a greater efficiency and lower complexity.
Using these updated parameter estimates, we compare the expected profit and optimal selling age obtained from the mixed-effects model to those derived from the fixed-effects model. Finally, we incorporate the most recent data on breeding costs and updated market prices to provide revised estimates that reflect current economic conditions. Our framework offers practical guidance for farmers aiming to maximize profitability in cattle growth.

Keywords: Delta method · Mixed models · Profit optimization

Acknowledgements: The authors belong to the research center CIMA (Centro de Investigação em Matemática e Aplicações, Universidade de Évora), https://doi.org/10.54499/UIDB/04674/2020, supported by FCT (Fundação para a Ciência e a Tecnologia), project UID/MAT/04674/2020. We are grateful to ACBM and José Pais (ACBM head engineer) for providing the data.

References

[1] Jacinto, G., Filipe, P.A., Braumann, C.A.: Profit optimization of cattle growth with variable prices. Methodology and Computing in Applied Probability, **24**, 1917-1952 (2022). DOI:https://doi.org/10.1007/s11009-021-09889-z
[2] Jamba, N.T., Jacinto, G., Filipe, P.A., Braumann, C.A.: Estimation for stochastic differential equation mixed models using approximation methods. AIMS Mathematics, **9**4, 7866 – 7894 (2024). DOI:https://doi.org/10.3934/math.2024383

Estimação do AGB na Floresta Utilizando Dados de Satélite

Ricardo Coelho [1[0000-0002-8791-2840]], Isabel Natário [1,2[0000-0001-6020-9373]] e Silvia Fraile [3]

rpe.coelho@campus.fct.unl.pt, icn@fct.unl.pt, silvia.fraile@geosat.space

[1] *NOVA Math - Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal*
[2] *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa*
[3] *Geosat, Boecillo, Espanha*

Abstract:
Quantificar e monitorizar a quantidade de carbono presente na floresta é importante para o desenvolvimento de políticas globais, entendimento das alterações climáticas e apoio ao mercado emergente de créditos de carbono. Para tal, o AGB (*Above-Ground Biomass*) é comummente utilizado, pois está intimamente relacionado com a quantidade de carbono. Com o avanço da tecnologia, dados de satélite têm sido utilizados na estimação do AGB, complementando os tradicionais dados de inventário florestais.
Recentemente, vários métodos de modelação estatística e de *Machine Learning* (ML) têm sido aplicados na estimação do AGB. No entanto, muitos dos métodos de ML não quantificam a incerteza nem consideram a autocorrelação espacial, limitando a sua capacidade preditiva. Uma alternativa é considerar uma abordagem híbrida, combinando modelos de ML com modelos geoestatísticos, neste trabalho considerados no paradigma bayesiano. Nesta abordagem, os modelos de ML bayesianos são primeiro ajustados aos dados para captar relações complexas e não lineares. Em seguida, modela-se a estrutura espacial presente nos resíduos dos modelos ML através de um modelo geoestatístico bayesiano. A predição final resulta da soma dos valores preditos por ambos os modelos, e a incerteza é quantificada pela soma dos limites inferiores e superiores de ambos.
Neste trabalho, aplicamos esta estratégia de modelação para estimar o AGB na *Sierra de la Culebra*, Espanha, utilizando dados de satélite, como bandas de reflectância, índices de vegetação e variáveis de textura, provenientes do satélite GEOSAT-2. Comparamos ainda os resultados preditivos de diferentes modelos, incluindo o modelo geoestatístico bayesiano, diferentes ML bayesianos e diferentes modelos híbridos.

Keywords: AGB · Geoestatística · Machine learning bayesiano · Modelo geoestatístico bayesiano

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00297/2020, DOI: 10.54499/UIDB/00297/2020, e UIDP/00297/2020, DOI:10.54499/UIDP/00297/2020 (Centro de Matemática e Aplicações). Fundação para a Ciência e a Tecnologia pelo financiamento através de uma Bolsa de Doutoramento individual 2023.01166.BDANA de Ricardo Coelho.

Our Journey to a More Sustainable Digital Infrastructure

Pedro Nobre [1], Fabio Coelho [1], and Mafalda Sá Ferreira [2]
pan@startcampus.pt, fpc@startcampus.pt, msm.ferreira@campus.fct.unl.pt

[1] *Start Campus*
[2] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA FCT)*

Abstract: In this presentation, we share a joint perspective that bridges industrial innovation and academic research in the evolving landscape of sustainable digital infrastructure.
The first part of our presentation will highlight the work of Start Campus, which is developing one of Europe's largest hyperscale and AI-ready data center campuses in Sines, Portugal. With a capacity of 1.2 GW of IT load, and direct integration with global submarine cable networks, Start Campus is positioning Portugal as a key global data and connectivity hub. We will explore how infrastructure scalability, renewable energy integration, and innovative cooling solutions [1] — such as seawater systems — are enabling a new generation of low-carbon digital services.
Complementing this, Mafalda Sá Ferreira, PhD student at NOVA FCT, will present her research on advanced methods for energy forecasting and optimization in green data centers. Her recent literature review identifies the most effective machine learning models and optimization strategies for managing variable workloads and renewable energy supplies [2]. Her work addresses the growing need for 24/7 Carbon-Free Energy and outlines a framework to dynamically align energy demand with sustainable supply using predictive analytics and decision-support tools.
Together, our presentation demonstrates the powerful synergy between academic research and industrial practice in advancing sustainable and intelligent energy management for data centers.

Keywords: Data center · Energy management · Sustainability

References

[1] Liu, C., Hao, N., Zhang, T., Wang, D., Li, Z., Bian, W.: Optimization of data centre immersion cooling using liquid air energy storage. Journal of Energy Storage, **90**, 111806 (2024). https://doi.org/10.1016/j.est.2024.111806
[2] Kashyap, S., Singh, A.: Prediction-based scheduling techniques for cloud data center's workload: a systematic review. Cluster Computing, **26**, 3209-3235 (2023). https://doi.org/10.1007/s10586-023-04024-8

Teaching and Disclosure of Statistic - CEE-SPE

Ensino e Divulgação da Estatística - CEE-SPE

Adelaide Freitas
CIDMA and University of Aveiro
adelaide@ua.pt

Jornalismo de Dados: Uma Oportunidade para Educação e Divulgação Estatística

Cláudia Silvestre [1][0000-0002-8850-4304]
csilvestre@escs.ipl.pt

[1] *ESCS-IPL, CEAUL, LIACOM*

Abstract: A literacia estatística é hoje uma competência fundamental para o exercício de uma cidadania informada e ativa, sobretudo num mundo onde os dados orientam políticas, decisões e narrativas públicas. Por exemplo, a Agenda 2030 das Nações Unidas, através dos Objetivos de Desenvolvimento Sustentável (ODS), exige o acompanhamento de múltiplos indicadores quantitativos, o que torna ainda mais evidente a necessidade dos cidadãos desenvolverem competências ao nível da estatística.

Assim, no contexto atual, a literacia estatística assume um papel estruturante na educação, exigindo abordagens inovadoras que aproximem os cidadãos do conhecimento estatístico de forma acessível e rigorosa. Conscientes desta necessidade e do papel que desempenham no panorama educativo, a Sociedade Portuguesa de Estatística (SPE) em parceria com a Escola Superior de Comunicação Social do Politécnico de Lisboa (ESCS-IPL) criaram o Prémio Jornalismo de Dados.

Com este prémio, ao se reconhecer trabalhos que transformam, de forma rigorosa, dados complexos em histórias acessíveis e relevantes, realça-se o papel da estatística como instrumento de comunicação confiável e acessível. E também reforça a literacia estatística como um pilar essencial para uma sociedade mais informada e esclarecida.

Do ponto de vista pedagógico, este prémio também se revela particularmente importante na formação de futuros jornalistas e comunicadores, uma vez que evidencia a necessidade de aquisição de competências em análise e interpretação de dados, visualização gráfica e pensamento crítico. Por outro lado, ao criar pontes entre a educação, o jornalismo e a cidadania, esta iniciativa permite aproximar a ciência da sociedade.

Keywords: Educação estatística · Jornalismo de dados · Literacia estatística · Visualização de dados

References

[1] Cushion, S., Lewis, J., Callaghan, R.: Data journalism, impartiality and statistical claims: Towards more independent scrutiny in news reportin. Journalism practice, **11**(10), 1400–1419 (2017). DOI:https://doi.org/10.1080/17512786.2016.1256789

[2] Silvestre, C., Pina, H.: From data to stories: Statistics and creativity in data journalism. In: XXXII Meeting of CLAD, pp. 97–98, 2025.

[3] Spiegelhalter, D.: The art of statistics: Learning from data. Penguin (2019).

A Importância da Estatística na Educação para a Cidadania e o Papel do Projeto ALEA

Pedro Campos [1][0000-0001-5495-9434]
pedro.campos@ine.pt

[1] *Instituto Nacional de Estatística, Faculdade de Economia da Universidade do Porto e LIAAD INESC TEC*

Abstract: A literacia estatística é uma competência fundamental para a formação de cidadãos críticos, informados e capazes de participar ativamente na sociedade [1]. Vivemos rodeados de dados — em decisões políticas, notícias, redes sociais ou escolhas do quotidiano — e a capacidade de os compreender, questionar e utilizar de forma responsável é essencial para o exercício pleno da cidadania. A estatística permite analisar realidades complexas, reconhecer padrões, avaliar riscos e tomar decisões informadas, promovendo a autonomia, o pensamento crítico e a participação democrática.

Neste contexto, destaca-se o papel do projeto ALEA – Ação Local de Estatística Aplicada (www.alea.pt) [3], uma iniciativa conjunta do INE, da Escola Secundária Tomaz Pelayo e do Ministério da Educação. Trata-se de um espaço digital em constante atualização, que oferece recursos pedagógicos, propostas de atividades, jogos, gráficos e dados reais, facilitando o ensino e a aprendizagem da estatística em contexto escolar.

Mais recentemente, o ALEA criou uma área dedicada à Educação para a Cidadania, com sugestões organizadas por domínios como os Direitos Humanos, Desenvolvimento Sustentável, Literacia Financeira ou Participação Democrática. Estas propostas integram informação estatística produzida pelo INE, formatada com apoio de docentes, permitindo aos alunos explorar temas atuais com base em dados reais. Esta abordagem está alinhada com a ideia de Estatística Cívica, promovida pelo projeto ProCivicStat [2], que defende o uso da estatística para compreender questões sociais relevantes e fomentar a participação informal em sociedades democráticas.

Keywords: Educação Estatística · Educação para a Cidadania · Literacia Estatística

References

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

[2] Nicholson, J., Gal, I., Ridgway, J. (2018). *Understanding Civic Statistics: A Conceptual Framework and its Educational Applications.* A product of the ProCivicStat Project. Retrieved from http://IASE-web.org/ISLP/PCS

[3] ALEA – Ação Local de Estatística Aplicada. Disponível em: https://www.alea.pt. Acedido em [acedido em 15/06/2025].

Educação Estatística 2025+

Bruno de Sousa [1][0000-0001-9918-8100]
bruno.desousa@fpce.uc.pt

[1] *Faculdade de Psicologia e de Ciências da Educação, CINEICC, Universidade de Coimbra*

Abstract:
A educação estatística está a sofrer uma transformação rápida em resposta aos desafios globais e aos avanços tecnológicos. Há uma ênfase crescente na equidade, inclusão e diversidade cultural, exigindo práticas educativas que sejam globalmente relevantes, acessíveis e eticamente fundamentadas. A ascensão da inteligência artificial (IA) generativa, do *machine learning*, de algoritmos baseados em dados e da ciência de dados apresentam-se como desafios significativos e complexos. Os estudantes devem não só aprender a utilizar estas ferramentas, mas também a avaliá-las criticamente, o que reforça a necessidade urgente de currículos atualizados que promovam o pensamento crítico, a ética dos dados e uma compreensão clara das limitações dessas análises. À medida que os programas educativos se tornam cada vez mais internacionalizados, a inclusão deve alinhar-se não só com o modelo social das necessidades especiais, mas também com a diversidade cultural, os diferentes idiomas, a orientação sexual e as expressões de género, garantindo que a literacia de dados seja acessível a todos e que todas estas perspetivas sejam respeitadas e refletidas nas narrativas construídas a partir dos dados. Poderão estes desafios e avanços tecnológicos abrir caminho para a implementação do Desenho Universal de Aprendizagem (UDL - *Universal Design for Learning*)? Será imperativo olhar para a educação estatística através da IA?

Keywords: Educação estatística · Inclusão e diversidade · Inteligência artificial (IA)

References

[1] Evmenova, A.S., Borup, J., Shin, J.K.: Harnessing the power of generative AI to support ALL Learners. TechTrends **68**, 820–831 (2024). https://doi.org/10.1007/s11528-024-00966-x
[2] de Sousa, B.: Universal design for inclusive education. In: R Helenius, E Falck (Eds.), Statistics Education in the Era of Data Science. Proceedings of the Online Satellite Conference of the International Association for Statistical Education (IASE), Aug-Sept 2021. DOI:https://doi.org/10.52041/iase.kxvpc
[3] Melo-López, V.-A., Basantes-Andrade, A., Gudiño-Mejía, C.-B., Hernández-Martínez, E.: The impact of artificial intelligence on inclusive education: A systematic review. Educ. Sci. **15**(5), 539 (2025). DOI:https://doi.org/10.3390/educsci15050539
[4] The Center for Universal Design (1989). Environments and products for all people. https://design.ncsu.edu/research/center-for-universal-design/

SPE/Biometrics & SGAPEIO

SPE/Biometria & SGAPEIO

Nuno Sepúlveda
Faculty of Mathematics & Information Science
Warsaw University of Technology, Poland and CEAUL
nuno.sepulveda@pw.edu.pl

Inferring Infectious Disease Risk in Real Time – Methodology and Public Health Implications

Ricardo Águas [1][0000-0002-6507-6597]
ricardo.aguas@ndm.ox.ac.uk

[1] *Nuffield Department of Medicine, University of Oxford, United Kingdom*

Abstract: Accounting for heterogeneity has been shown to be critical in models of infectious disease transmission. Traditional compartmental models, often defaulted to in outbreak scenarios, oversimplify disease dynamics by assuming homogeneous susceptibility within, and homogeneous mixing between, modelled populations. We consider an extension of the traditional susceptible-infected-recovered model that allows for heterogeneity in susceptibility and test our ability to statistically infer its parameters in real-time using simulated data with state-of-the-art software. A simple retroactive validation and real-time limit test across a set of reasonable parameter values demonstrate our ability to infer these parameters with a high degree of certainty. We quantify the value of detecting this heterogeneity in real-time and discuss the implications on the timing and scope of public health mitigation policies. An instantaneous measure of population-level heterogeneity in the context of a rapidly evolving infectious disease landscape with quantifiable uncertainty adds value to health policy decision-making. Milder interventions can be most effective for relatively high levels of heterogeneity due to the infection-by-selection phenomenon, which disproportionately decreases the mean level of susceptibility in the population as the disease dynamics evolve.

Keywords: Epidemiology · Infectious diseases · Mathematical modelling · Parameter estimation

Nonparametric model check for cure rate quantile regression

Mercedes Conde-Amboage[1,2], Wenceslao González-Manteiga[1,2] and César A. Sánchez-Sellero[1,2]
mercedes.amboage@usc.es, wenceslao.gonzalez@usc.es, cesar.sanchez@usc.es

[1] *Department of Statistics, Mathematical Analysis and Optimization. Faculty of Mathematics. Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain.*
[2] *Galician Center for Mathematical Research and Technology (CITMAga), Santiago de Compostela, Spain.*

Abstract: In classical survival analysis, a fundamental assumption is that all individuals will eventually experience the event of interest. However, it often occurs that a subset of subjects will never experience the event. These individuals are typically considered to have infinite survival times and are classified as "cured". To deal with this phenomenon, classical survival models have been extended to what is commonly referred to as cure models. A thorough review of this kind of models from the standpoint of classical mean regression can be found in [1].

On the other hand, in the context of quantile regression (which aim is to provide a more detailed description of the conditional distribution of the response variable) the problem of estimating a cure rate model has been scarcely addressed in the literature. Specifically, only the work of [3], and more recently that of [2], can be highlighted.

Throughout this talk, a new lack-of-fit test for cure models in the context of quantile regression is presented. This new proposal represents the first contribution in the literature to test the effect of a group of covariates on a survival time using empirical processes marked by residuals. The asymptotic behaviour of the test statistics will be derived. In addition, an extensive simulation study and a real data application will be presented to show the performance of the new proposal in practice.

Keywords: Cure models · Quantile regression · Lack-of-fit test

References

[1] Amico, M., and Van Keilegom, I.: Cure models in survival analysis. Annual Review of Statistics and Its Application, 5, 311–342 (2018). DOI:10.1146/annurev-statistics-031017-100101
[2] Narisetty, N., and Koenker, R:. Censored quantile regression survival models with a cure proportion. Journal of Econometrics, 226, 192–203 (2022). DOI:10.1016/j.jeconom.2020.12.005
[3] Wu, Y., and Yin, G.: Cure rate quantile regression for censored data with a survival fraction. Journal of the American Statistical Association, 108, 1517–1531 (2013). DOI:10.1080/01621459.2013.837368

Innovations in Official Statistics - Statistics Portugal

Inovações em Estatística Oficiais - Instituto Nacional de Estatística

Pedro Campos
Instituto Nacional de Estatística
Faculdade de Economia da Universidade do Porto e LIAAD INESC TEC
pedro.campos@ine.pt

Indicadores de Acessibilidade a Serviços de Interesse Geral

Joana Malta [1], Maria Aurindo [1], Carla Cardoso [1], Rita Santos [1]
(joana.malta, mjose.aurindo, carla.cardoso, rita.santos)@ine.pt

[1] *Instituto Nacional de Estatística*

Abstract: O acesso a equipamentos e serviços de interesse geral é uma dimensão central na avaliação da qualidade de vida das populações e da coesão dos territórios. Esta apresentação utiliza a aplicação Carta de Equipamentos e Serviços de Interesse Geral (CE-SIG), lançada pelo Instituto Nacional de Estatística (INE) em junho deste ano, para analisar os tempos de deslocação a pé e de automóvel para um conjunto de equipamentos de interesse geral.
Os resultados apresentados utilizam quartis como estatística de referência para analisar a diversidade dos tempos de acesso a equipamentos e serviços entre a população em diferentes territórios, procurando demonstrar o valor desta informação para as políticas públicas.
Estes novos indicadores inserem-se no projeto Indicadores de Assimetria ao Nível Local e Inter-regional (IAssLocal), que visa fornecer nova informação para caracterizar a diversidade socioeconómica dos territórios e tira partido do potencial da informação associada à Infraestrutura Nacional de Dados do INE.

Keywords: Equipamentos · Indicadores territoriais · Medidas de assimetria · Serviços de interesse geral · SIG

A Inovação nas Estatísticas Oficiais em Portugal

Sofia Rodrigues [1], Paulo Saraiva [1[0009-0002-1315-1396]], Glória Carrilho [1], João Poças[1]
(sofia.rodrigues, paulo.saraiva, gloria.carrilho, joao.pocas)@ine.pt

[1] *Instituto Nacional de Estatística*

Abstract: Num ecossistema de dados em rápida mudança, o Instituto Nacional de Estatística (INE) reconhece a inovação como um aspeto transversal à organização. Procura-se aumentar a capacidade de resposta à crescente procura por estatísticas mais relevantes, granulares e atempadas. As iniciativas do INE demonstram um compromisso estratégico e multidimensional. A Infraestrutura Nacional de Dados é um sistema que, visando tirar o máximo partido da cadeia produtiva do INE, integra novas abordagens – organizacionais, tecnológicas, metodológicas e de gestão - para a produção de estatísticas oficiais.
Entre variados exemplos de inovação no INE, podem destacar-se a integração de dados de diferentes fontes, crucial para projetos como o recenseamento da população e habitação com base em dados administrativos, a produção de novas estatísticas com base em fontes externas ou a Carta de Equipamentos e Serviços de Interesse Geral.
Para este artigo, foram selecionados o WebInq e a iniciativa Meia-hoRa. O WebInq, que celebra 20 anos, tem sido um impulsionador da modernização na recolha de dados por autopreenchimento, otimizando o processo e a relação com os respondentes através de funcionalidades como a Transmissão Automática de Dados. A iniciativa Meia-hoRa ilustra como a inovação se manifesta na cultura organizacional, promovendo a partilha de conhecimento e a colaboração através do uso do software R, cultivando um ambiente de aprendizagem contínua.
Em suma, a trajetória de inovação do INE reforça que a inovação não é uma opção, mas uma necessidade. Assenta numa estratégia holística, assegurando a adaptação do organismo aos novos desafios e reforçando o seu papel vital.

Keywords: Dados · Estratégia · Inovação

Not Just Another Black Box: AI in the Age of Trusted Statistics

Sónia Quaresma [1][0009-0004-1422-3712]
sonia.quaresma@ine.pt

[1] *Instituto Nacional de Estatística*

Abstract: This presentation explores the emerging role of Generative AI (GENAI) in the field of Official Statistics, positioning it both as a promising driver of innovation and a source of considerable challenges. GenAI offers powerful new capabilities that can enhance statistical workflows, from improving efficiency to enabling new analysis and communication.
Simultaneously, its application raises critical concerns regarding accuracy, reproducibility, transparency, and the preservation of public trust, elements that are foundational to the statistical offices. Addressing these opportunities and risks requires a framework of responsible use, in which technological adoption is guided by statistical principles and ethical considerations rather than by technological enthusiasm alone.
To ground this discussion in practice, the presentation examines prototyping initiatives currently being undertaken within the Portuguese Statistical Office. These initiatives not only demonstrate how GenAI can be applied in concrete statistical contexts but also provide a case study of controlled, scalable, and ethically sound integration, offering valuable insights for the broader community of statistical producers and users.

Keywords: Ethical integration · Generative AI · Official Statistics · Prototyping · Responsible innovation

References

[1] Grobelnik, M.: GenAI for official statistics - opportunities and dangers. In: Proceedings of the WIN Conference 2025, 2025. https://win2025.stat.gov.pl/Content/Presentations/03.M.Grobelnik.pdf
[2] Wirthmann, A.: Reusing MNO data for official statistics: a common methodological framework for the European Statistical System. In: Proceedings of the WIN Conference 2025, 2025. https://win2025.stat.gov.pl/Content/Presentations/02.%20A.Wirthmann.pdf
[3] AIML4OS WP12: Work in Progress First Deliverable. Eurostat (2025). https://github.com/AIML4OS/WP12/tree/main/Deliverables/D12.1

Sociedade Portuguesa de Estatística & Associação Portuguesa de Classificação de Dados - SPE/CLAD

Conceição Amado
CMAT and IST, University of Lisbon
conceicao.amado@tecnico.ulisboa.pt

A Inclusão Racial na Publicidade de Cosméticos no *Instagram*: Representação e Impacto no Envolvimento dos Utilizadores

Catarina Marques [1,2[0000-0003-2159-738X]], Daniela Langaro [1,2[0000-0002-1246-0720]] e Mariana Vicente Cintra [1]
catarina.marques@iscte-iul.pt, daniela.langaro@iscte-iul.pt, mvcae@iscte-iul.pt

[1] *ISCTE Instituto Universitário de Lisboa*
[2] *Business Research Unit (BRU-IUL), Portugal*

Abstract: A inclusão tornou-se um tema central no *marketing* e publicidade, sendo associada à justiça social e à identificação do consumidor com a marca. As redes sociais concentram grande parte do conteúdo de marca, sendo espaços privilegiados para estratégias inclusivas. No entanto, há poucos estudos que avaliam se essas intenções se traduzem em práticas eficazes, especialmente em relação à inclusão racial. Este estudo analisa a representação da diversidade de tons de pele em publicações de marcas de cosméticos no *Instagram*, bem como o impacto dessa inclusão no envolvimento dos utilizadores. Foi realizada uma análise de conteúdo a 750 publicações de 15 marcas de beleza, acompanhada de uma análise de sentimentos dos comentários. Utilizaram-se modelos de equações estruturais, nomeadamente análise de caminhos (*path analysis*), complementada por análise multigrupos para comparar marcas de grande consumo com marcas *premium*. Os resultados mostram que o tom de pele mais diverso conduz a menor envolvimento (menos *likes* e comentários), mas a maior positividade nos comentários. O menor envolvimento resulta do foco em minorias, que naturalmente atraem menos apoios comparativamente a outras publicações. Contudo, entre os que apoiam, o sentimento é claramente mais positivo e os comentários mais entusiastas. Nas marcas *premium*, a diversidade é mais bem aceite, não afetando negativamente o envolvimento. O estudo contribui para compreender as práticas reais de inclusão nas redes sociais e levanta questões importantes para a teoria e prática do *marketing* inclusivo. A inclusão deve ser uma estratégia autêntica e consistente, que reflita a diversidade da sociedade.

Keywords: Diversidade · Modelos de equações estruturais · Publicidade · Redes sociais

Avaliação Não-Destrutiva da Qualidade de Uvas com Imagiologia Hiperespectral

Irene Oliveira [1][0000-0002-9065-4336]
ioliveir@utad.pt

[1] *ECT-UTAD, DM, CEMAT (Centro de Matemática Computacional e Estocástica) e CITAB, (Centro de Investigação e de Tecnologias Agro-Ambientais e Biológicas)*

Abstract: A qualidade do vinho depende fortemente da colheita das uvas no ponto ótimo de maturação. Os métodos tradicionais de avaliação da maturação são destrutivos, demorados e dispendiosos. A espectroscopia hiperespectral em modo de reflectância (380–1028 nm) oferece uma abordagem não destrutiva para prever parâmetros enológicos como o pH, o teor de açúcares (° Brix) e o teor total de antocianinas. A elevada dimensionalidade e colinearidade dos dados hiperespectrais exigem técnicas robustas de redução de dimensionalidade [1]. A metodologia proposta inclui a seleção ótima de variáveis espectrais para melhorar o desempenho de modelos preditivos. Desenvolveu-se uma aplicação Shiny em R para a seleção interativa de subconjuntos informativos de comprimentos de onda. Analisaram-se 90 bagos de uva colhidos durante a maturação, cujos espectros foram adquiridos antes do congelamento para análises laboratoriais posteriores. Utilizando o pacote `subselect` do R [2, 3], que emprega critérios multivariados para seleção de variáveis, identificaram-se subconjuntos reduzidos de comprimentos de onda com alta capacidade preditiva para os parâmetros enológicos estudados. A aplicação desenvolvida permite a exploração visual de padrões espectrais, a seleção otimizada de variáveis e a construção de modelos de regressão com avaliação do seu desempenho. Os resultados indicam que é possível identificar subconjuntos reduzidos de comprimentos de onda com elevada capacidade preditiva. A aplicação facilita uma análise intuitiva e pode apoiar a tomada de decisão em tempo real na viticultura de precisão.

Keywords: Imagens hiperespectrais · Modelos preditivos · Redução da dimensionalidade

Acknowledgements: Agradecimento à Fundação para a Ciência e a Tecnologia (FCT) pelo apoio financeiro UID/MULTI/04621/2020 (CEMAT).

References

[1] Fernandes, A.M. et al.: Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks. Computers and Electronics in Agriculture **115.**(6), páginas 88-96 (2015). DOI: j.compag.2015.05.013

[2] Cerdeira, J.O. et al.: `subselect`: Selecting variable subsets. R package version 0.16.

[3] Duarte Silva, A.P.:Efficient variable screening for multivariate analysis. Journal of Multivariate Analysis, **76**, páginas 35–62, (2001).DOI:10.1006/jmva.2000.1920

Aquaponics for Sustainable Production: some Statistical Approaches

Fernando Sebastião [1,2,6[0000-0002-8792-4649]], Judite Vieira [1,2,6[0000-0001-7982-5686]], Luís Cotrim [1,2,6[0000-0001-9256-9349]], Damariz Y. Ushina [1,2,6], Ounísia Santos [1,2,6[0000-0001-5007-5252]], Vânia S. Ribeiro [1,5,6[0000-0001-6561-2065]], Daniela C. Vaz[1,5,6[0000-0001-7562-4676]], Raul Bernardino [1,3,4,6[0000-0002-7775-0614]]

fsebast@ipleiria.pt, judite.vieira@ipleiria.pt, luis.cotrim@ipleiria.pt, damariz.s.ushina@ipleiria.pt, ounisia.santos@ipleiria.pt, vania.ribeiro@ipleiria.pt, daniela.vaz@ipleiria.pt, raul.bernardino@ipleiria.pt

[1] *LSRE-LCM, Polytechnic Institute of Leiria, Portugal*
[2] *ESTG, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal*
[3] *ESTM, Polytechnic Institute of Leiria, 2520–614 Peniche, Portugal*
[4] *MARE, ESTM, Polytechnic Institute of Leiria, Portugal*
[5] *ESSLei, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal*
[6] *ALiCE, University of Porto, 4200-465 Porto, Portugal*

Abstract: The Polytechnic Institute of Leiria has been researching aquaponics as a sustainable alternative to traditional soilless cultivation methods since 2019. The system, consisting of a fish-rearing tank, drum filter connected to a sedimentation tank, biofilter, hydroponic unit, and sump tank, is monitored for physicochemical water parameters, environmental variables, plant growth and indicators of fish welfare. Several projects have been developed, including catfish-based aquaponic systems for leafy vegetables, yellow mealworms production for protein, papaya production analysis, and strawberry production comparisons. Across all these projects, various statistical approaches were employed, including experimental design and multivariate techniques, particularly in comparative analysis, assessing the existence or not of significant differences in the morphological characteristics of plant and fruit growth. These projects help mitigate climate change by cutting emissions, reducing chemical use, and promoting low-carbon and local food production.

Keywords: Aquaponics · Integrated multitrophic systems · Statistical analysis · Sustainable production

Acknowledgements: This work is partially financed by national funds through FCT/MCTES (PIDDAC): LSRE-LCM, UIDB/50020/2020 (DOI: 10.54499/UIDB/50020/2020), UIDP/50020/2020 (DOI:10.54499/UIDP/50020/2020), and ALiCE, LA/P/0045/2020 (DOI: 10.54499/LA/P/0045/2020).

References

[1] Sebastião, F. et al: Nutrient-efficient catfish-based aquaponics for producing lamb's lettuce at two light intensities. Journal of the Science of Food and Agriculture, **104. 11.**, 6541–6552 (2024). DOI:https://doi.org/10.1002/jsfa.13478

Associação Portuguesa de Investigação Operacional - APDIO

Filipe Alvelos
Centro de Investigação ALGORITMI e LASI, Universidade do Minho
falvelos@dps.uminho.pt

Modelização do Comportamento do Vento no Contexto da Propagação de Fogos Florestais

Helena Alvelos [1,3][0000-0002-6450-5521], Ana Raquel Xambre [1,3][0000-0001-8615-3443], Francisco Marques [1,2] Agostinho Agra [2,3][0000-0002-4672-6099]
Filipe Alvelos[4][0000-0002-1851-4339]
helena.alvelos@ua.pt, raquelx@ua.pt, franciscocmarques@ua.pt, aagra@ua.pt, falvelos@dps.uminho.pt

[1] *Departamento de Economia, Gestão, Engenharia Industrial e Turismo da Universidade de Aveiro, Portugal*
[2] *Departamento de Matemática da Universidade de Aveiro, Portugal*
[3] *Centro de Investigação em Matemática e Aplicações, Universidade de Aveiro, Portugal*
[4] *Centro de Investigação ALGORITMI e LASI, Universidade do Minho, Braga, Portugal*

Abstract: Os fogos florestais representam ameaças significativas à segurança pública, aos bens e aos recursos florestais e são cada vez mais frequentes, para os quais têm contribuído, também, as alterações climáticas. Prevenir a sua ocorrência ou reduzir os seus efeitos são questões cruciais, o que tem originado uma crescente atenção a este tópico. Este trabalho enquadra-se no desenvolvimento de uma framework que utiliza modelos de otimização para o pré-posicionamento de recursos e para o movimento dos recursos durante a fase de supressão do fogo. Esses problemas envolvem várias fontes de incerteza, sendo uma delas o comportamento do vento (direção e velocidade). O objetivo deste trabalho é o de modelizar os dados históricos do vento recolhidos numa estação meteorológica do norte de Portugal. Inicialmente, foram utilizadas várias técnicas estatísticas (como estatísticas descritivas, análise de correlação e testes de qualidade de ajuste) para compreender as variáveis em causa. Dado que existe uma grande quantidade de dados disponíveis, decidiu-se utilizar, nos modelos de otimização, a distribuição empírica conjunta da velocidade e da direção do vento, em alternativa ao ajuste de distribuições teóricas. Para tal, os dados foram divididos em três épocas de risco de incêndio: baixo, médio e alto. Utilizando os dados da época de alto risco, os valores da velocidade e da direção do vento e as respetivas probabilidades conjuntas estimadas foram usados para criar cenários que foram incorporados nos referidos modelos de otimização.

Keywords: Análise estatística · Comportamento do vento · Fogos florestais

Acknowledgements: Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia (FCT) no âmbito dos projetos PCIF/GRF/0141/2019 "O3F - An Optimization Framework to Reduce Forest Fire", DOI 10.54499/PCIF/GRF/0141/2019, UID/00319/Centro ALGORITMI (ALGORITMI/UM) e do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), no âmbito do projeto UID/04106.

Uma Abordagem de Otimização em Dois Níveis para a Agregação da Flexibilidade no Consumo de Energia Elétrica

Carlos Henggeler Antunes [1,2][0000-0003-4754-2168],
Inês Soares [1,2][0000-0001-7146-4273], Ana Soares [1,2][0000-0002-5431-7041] e
Maria João Alves [2,3][0000-0002-2268-0110]
ch@deec.uc.pt, inesgsoares@deec.uc.pt, ana.soares@inescc.pt, mjalves@fe.uc.pt

[1] *Departamento de Eng. Eletrotécnica e de Computadores, Universidade de Coimbra*
[2] *INESC Coimbra*
[3] *Faculdade de Economia, Universidade de Coimbra e CeBER*

Abstract: A flexibilidade é fundamental para a gestão de sistemas de energia, tendo como objetivo incentivar a alteração dos padrões de consumo de eletricidade face à crescente produção renovável variável. Os agregadores desempenham um papel importante, recolhendo a flexibilidade dos consumidores/produtores, que pode depois ser transacionada em mercados, criando benefícios económicos e operacionais para todas as partes interessadas. A interação agregador-consumidores pode ser modelada como um problema de otimização em dois níveis com múltiplos seguidores. O agregador, no nível superior, estabelece incentivos financeiros, enquanto os consumidores, no nível inferior, otimizam a utilização de energia em resposta a estes incentivos e preços da eletricidade, tendo em conta preferências de conforto. Para resolver este problema, propomos uma abordagem híbrida que combina otimização por enxame de partículas para o problema de nível superior com um *solver* exato para o problema de nível inferior de programação inteira-mista. Apresentam-se experiências computacionais para um conjunto de consumidores residenciais típicos, considerando diferentes eletrodomésticos, veículo elétrico, baterias e microgeração.

Keywords: Energia · Otimização em dois níveis · Programação inteira-mista

Otimização do *Layout* de Parques Eólicos com Fatores de Carga Heterogéneos

Adelaide Cerveira [1][0000-0002-7494-6566] e Agostinho Agra [2][0000-0002-4672-6099]
cerveira@utad.pt, aagra@ua.pt

[1] *Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro, Vila Real e INESCTEC, Portugal*
[2] *Departamento de Matemática e CIDMA, Universidade de Aveiro, Portugal*

Abstract: A crescente aposta nas energias renováveis torna fundamental a otimização dos custos associados aos parques eólicos, especialmente no que diz respeito ao layout da rede elétrica que liga as turbinas à subestação. Este trabalho aborda o problema do desenho de parques eólicos terrestres, assumindo posições fixas para as turbinas e subestação, com o objetivo de minimizar os custos de infraestrutura e as perdas de energia ao longo da vida útil do parque.
Ao contrário da maioria dos estudos existentes, que assumem fatores de carga iguais para todas as turbinas, propomos uma abordagem mais realista que considera fatores de carga heterogéneos. Essa variante reflete diferenças na exposição ao vento causadas pela orografia e orientação das turbinas. Como as perdas de energia são uma função quadrática da corrente elétrica, apresentamos duas formulações com funções objetivo quadráticas e, a partir destas, desenvolvemos formulações lineares inteiras mistas. Propomos um algoritmo de geração de colunas para resolução eficientemente estas formulações que apresentam um número exponencial de restrições. Introduzimos também vários melhoramentos, tais como eliminação de variáveis e introdução de desigualdades válidas e duas matheurísticas para tratar instâncias de grande dimensão. Os métodos propostos foram validados com dados reais e demonstram vantagens significativas em termos de qualidade de solução e tempo de processamento.

Keywords: Desenho de parques eólicos · Fator de carga · Linearização · Otimização linear inteira mista

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UID/50014/2023 (https://doi.org/10.54499/UID/50014/2023).

Optimização Robusta com Distribuições para o Problema de Pré-posicionamento de Recursos em Incêndios Florestais

Filipe Alvelos[1][0000-0002-1851-4339], Agostinho Agra [2,3,4][0000-0002-4672-6099], Francisco Marques [2,5], Ana Raquel Xambre [4,5][0000-0001-8615-3443] e Helena Alvelos [4,5][0000-0002-6450-5521]

falvelos@dps.uminho.pt, amagra@fc.ul.pt, franciscocmarques@ua.pt, raquelx@ua.pt, helena.alvelos@ua.pt

[1] *Departamento de Produção e Sistemas / Centro ALGORITMI / LASI, Escola de Engenharia, Universidade do Minho, Braga, Portugal*

[2] *Departamento de Matemática, Universidade de Aveiro, Portugal*

[3] *CEMS.UL – Center for Mathematical Studies, Universidade de Lisboa, Portugal*

[4] *Centro de Investigação em Matemática e Aplicações, Universidade de Aveiro, Portugal*

[5] *Departamento de Economia, Gestão, Engenharia Industrial e Turismo da Universidade de Aveiro, Portugal*

Abstract: Os incêndios florestais têm impactos significativos nas populações, ecossistemas e economia. A redução destes seus impactos negativos passa necessariamente pela contribuição de diversas disciplinas, entre elas a Optimização e a Estatística.
Nesta apresentação considera-se o problema do pré-posicionamento de recursos (meios terrestres ou aéreos) de combate a incêndios. Dada a incerteza inerente aos incêndios florestais, propõe-se um modelo de optimização sob incerteza em duas fases. Na primeira fase, as decisões de pré-posicionamento dos recursos são tomadas em antecipação de possíveis ignições e características do vento; na segunda fase, conhecida(s) a(s) ignição(ções) e características do vento, são decididas as posições de ataque. A propagação do incêndio é modelada utilizando o princípio do tempo mínimo de transmissão do fogo e incorporada num modelo de Programação Inteira Mista que também integra as decisões relativas ao posicionamento e movimentação dos recursos.
Para capturar a incerteza, propõe-se um modelo de Optimização Robusta com Distribuições (ORD, *Distributionally Robust Optimization*). Considera-se um conjunto discreto de cenários, cada um caracterizado por uma ou mais ignições, e direcções e intensidades do vento. Na ORD assume-se que a distribuição de probabilidade dos parâmetros incertos pertence a um conjunto de ambiguidade. Uma solução óptima minimiza o valor esperado da área ardida considerando a pior distribuição de probabilidade desse conjunto.
Um método de decomposição é desenvolvido para resolver eficientemente o modelo de grande dimensão de Programação Inteira Mista central em que se baseia a abordagem. Apresentam-se resultados de experiências computacionais realizadas em instâncias derivadas de uma paisagem real.

Keywords: Incêndios florestais · Optimização com incerteza · Programação Inteira Mista

Acknowledgements: Este trabalho é parcialmente financiado pelo projecto Firesys - AI - empowered Decision Support System for Wildfire Management e pela Fundação para a Ciência e a Tecnologia através do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), no âmbito do projeto UID/04106.

(Inter)national Standardization: Ongoing Projects on Applications of Statistical Methods - SPE IPQ/CT225

Normalização (inter)nacional: Projetos em curso sobre aplicações de métodos estatísticos - SPE IPQ/CT225

XXVII CONGRESSO SPE 25 SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Mafalda Costa
Divisão de Programas e Avaliação, CCDR-Norte, Portugal
anamafalda.costa@ccdr-n.pt

(Inter)national Standardization: Ongoing Projects on Applications of Statistical Methods

Maria J. Polidoro [1,6[0000-0002-2220-4077]], Mafalda Costa [2[0009-0007-6887-6307]], **Natércia Durão** [3[0000-0002-0845-263X]], Fernanda Figueiredo [4,6[0000-0003-0255-4106]], Sónia Quaresma [5[0009-0004-1422-3712]], and Pedro Campos [4,5[0000-0001-5495-9434]]
mjp@estg.ipp.pt, anamafalda.costa@ccdr-n.pt, natercia@upt.pt, otilia@fep.up.pt, sonia.quaresma@ine.pt, pedro.campos@ine.pt

[1] *Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, Felgueiras, Portugal*
[2] *Divisão de Programas e Avaliação, CCDR-Norte, Portugal*
[3] *Departamento de Ciências e Tecnologia e REMIT da Universidade Portucalense, Porto, Portugal*
[4] *Faculdade de Economia da Universidade do Porto, Portugal*
[5] *Instituto Nacional de Estatística, Lisboa, Portugal*
[6] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: (Inter)national standardization on applications of statistical methods involves the development and establishment of globally recognized guidelines, protocols, and best practices for using statistical techniques across various fields. This process aims to ensure consistency, accuracy, and reliability in statistical analyses worldwide, facilitating better communication and collaboration among researchers, industries, and regulatory bodies. The Portuguese Quality Institute and the Portuguese Technical Commission 225 for standardization on Applications of Statistical Methods follow the developments of the International Organization for Standardization - Technical Commission 69 (ISO/TC 69: Applications of Statistical Methods) to disseminate this knowledge at the national level. Through (Inter)national organizations, these standards help harmonize methodologies, improve data quality, and support informed decision-making in areas such as manufacturing, healthcare, environmental monitoring, and more. Ongoing projects focus on creating, updating, and implementing these standards to keep pace with technological advancements and emerging applications of statistical methods. This session will cover four ongoing developments: Statistics - Vocabulary and Symbols, Control Charts, Curation, Cleansing, and Wrangling of Big and Large Datasets, and Sampling Methods.

Keywords: Big and large datasets · Control charts · Sampling methods · Standardization · Vocabulary and symbols

Acknowledgements: This work is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the projects UID/00006/2025 and UIDB/00006/2020 (DOI:10.54499/UIDB/00006/2020).

Curation, Cleansing and Wrangling of Big and Large Datasets

Sónia Quaresma [1][0009-0004-1422-3712]
sonia.quaresma@ine.pt

[1] *Instituto Nacional de Estatística, Lisboa, Portugal*

Abstract: The Data Curation, Cleansing, Wrangling, and Quality Assurance (DCCWQA) framework provides a structured approach for enhancing data quality and ensuring the reproducibility of source data throughout the preparation lifecycle. It is built upon five guiding principles that inform each phase of the process: Data Type Awareness, Contextual Integrity, Fit-for-Purpose Processing, Documentation and Traceability, and Iterative Quality Assurance. These principles are interpreted in a differentiated yet complementary manner depending on the nature of the data source. For survey data, the emphasis is on understanding the design and behavior of the respondent. For administrative data, alignment with statistical definitions and documentation of transformations is key. For machine-generated data, such as Mobile Network Operator (MNO) records or sensor outputs, priority is given to platform-specific patterns, metadata quality, and the traceability of automated processing pipelines. In all cases, the principles guide the way data are interpreted, transformed, and validated. By consistently applying the DCCWQA framework across diverse data types, statistical institutions can manage data in a manner that is technically robust and analytically transparent. The framework promotes standardization of practices, supports adaptability, and promotes coherence between heterogeneous sources. This ensures that the resulting statistics maintain the core quality principles of relevance, accuracy, timeliness, accessibility, comparability, and coherence, even as data ecosystems evolve.

Keywords: DCCWQA framework · Machine-generated data · Survey data

Oral Sessions

—Comunicações Orais—

Séries Temporais I

Are Sequential Search Methods Effective for ARMA Model Selection?

Sónia Gouveia [1,2] [0000-0002-0375-7610], Ana Martins [1,2] [0000-0003-4860-7795], and Manuel Scotto [3] [0000-0001-8427-2684]

sonia.gouveia@ua.pt, a.r.matins@ua.pt, manuel.scotto@tecnico.ulisboa.pt

[1] *Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and DETI, University of Aveiro, Aveiro, Portugal*
[2] *Intelligent Systems Associate Laboratory (LASI), Portugal*
[3] *Center for Computational and Stochastic Mathematics (CEMAT) and Department of Mathematics, IST, Lisbon, Portugal*

Abstract: Univariate analysis using autoregression moving average (ARMA) models remains foundational for modeling time series. However, the identification of the model order (p, q) continues to pose a substantial challenge.
The seminal work of Box and Jenkins [1] marked a turning point in ARMA model selection, by introducing a systematic iterative approach involving model identification, parameter estimation, and diagnostic checking. In practice, model building relies on interpreting autocorrelation plots and comparing candidates using an Information Criterion (IC). Exhaustive search methods are common but computationally demanding, leading to the development of automated procedures that limit the search space or systematically explore feasible model orders. Sequential search methods like the Hyndman-Khandakar (HK) algorithm [2] provide an efficient alternative by starting from initial models, selecting the one minimizing an IC and iteratively exploring neighboring models to find a local optimum. Notwithstanding its widespread use, the HK algorithm has not undergone a comprehensive validation. This study evaluates performance of the HK algorithm in selecting ARMA model orders, building upon preliminary results [3]. The ability of the algorithm to recover the true model order is assessed through a systematic simulation study under varying conditions (different ARMA structures, parameter values and time series length), ultimately revisiting the provocative question: Are Sequential Search Methods Effective for ARMA Model Selection?

Keywords: ARMA model · Box-Jenkins model selection · Hyndman-Khandakar

Acknowledgements: This work was funded by the Fundação para a Ciência e Tecnologia, Portugal (FCT, https://www.fct.pt) under the research unit 00127-IEETA (https://www.ieeta.pt).

References

[1] Box, G. E. P., Jenkins, G. M.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1970).

[2] Hyndman, R. J., Khandakar, Y.: Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, **26**(3), 1–22 (2008).

[3] Gouveia, S. On the performance evaluation of algorithms for the identification of ARMA models. In: Proceedings of JOCLAD 2025 Porto, Portugal, 53–54, 2025.

Statistical Framework for Environmental Justice Using Models for Time Series of Counts

Adriano Gomes [1] [0000-0002-9384-627X], Ana Martins [1,2] [0000-0003-4860-7795], and Sónia Gouveia [1,2] [0000-0002-0375-7610]

aog@ua.pt, a.r.matins@ua.pt, sonia.gouveia@ua.pt

[1] *Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and DETI, University of Aveiro, Aveiro, Portugal*
[2] *Intelligent Systems Associate Laboratory (LASI), Portugal*

Abstract: While environmental justice (EJ) promotes equal protection from environmental risks, objective metrics to quantify it remain nonexistent. This work proposes a statistical framework to quantify EJ indicators at the country level. The framework begins by estimating the parameters of an INGARCH model (INteger-valued Generalized AutoRegressive Conditional Heteroskedasticity, [1]) at the district level, where the daily number of hospital admissions is modeled as a function of its past values, past conditional means and environmental covariates [1, 2]. Then, a proxy of EJ for a given socioeconomic group (e.g., age group 0-14, 15-64 or ≥ 65, education level such as basic, high school or university, etc.) is computed by multiplying the effects of a covariate on hospital admissions – captured by the corresponding normalized INGARCH coefficients – by weights based on the distribution of that group across districts.

Two important aspects are relevant to investigate: the first is whether the impact of a given covariate is statistically significant for a given socioeconomic group, and the second is to identify which groups may experience different impacts. This requires studying the distributional properties of the proposed EJ estimator, examined under the hypothesis of asymptotic normality for the estimators of the INGARCH parameters related to the covariates and, for comparison, via a parametric bootstrap approach. The EJ framework is applied to anonymized daily respiratory hospital admissions (ACSS, 2013–2017) aggregated by district for 18 mainland Portugal districts, alongside meteorological and air quality data (ERA5-Land and CAMS services) and socioeconomic variables (Census) aggregated at the district level.

Keywords: INGARCH · Time series of counts · Weighted estimators

Acknowledgements: This work was funded by the Fundação para a Ciência e Tecnologia, Portugal (FCT, www.fct.pt) under R&D unit 00127-IEETA (www.ieeta.pt). AG acknowledges the grant from the FCT project ALICE [2022.04351.PTDC].

References

[1] Ferland, R, Latour, A, Oraichi, D. Integer-valued GARCH process. J. Time Ser. Anal. **27**(6): 923–942 (2006).

[2] Liboschik, T, Fokianos, K, Fried, R. tscount: An R package for analysis of count time series following generalized linear models. J. Stat. Softw. **82**(5): 1–51 (2017).

Unemployment Nowcasts via Google Trends: Insights into Digital Divide in Brazil and Portugal

Eduardo André Costa [1[0000-0002-0636-8821]], and
Maria Eduarda Silva [1,2[0000-0003-2972-2050]]
up201800115@edu.fep.up.pt, mesilva@fep.up.pt

[1] *Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-464 Porto, Portugal*
[2] *LIADD-INESC TEC, Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-464 Porto, Portugal*

Abstract: Data on online behaviour — particularly Google Trends (GT) search indices — offer *quasi*-real-time information on search activity and allow for fresh perspectives on forecasting. These data have been employed as macroeconomic predictors, promising more timely estimates than traditional, time-lagged sources, particularly for unemployment indicators. However, GT's sampling bias can skew forecasts since search activity varies with Internet access, digital literacy, and socio-demographic attributes.

This study assesses how the effectiveness of GT as an unemployment predictor varies across socio-demographic groups. To that end, we use high-frequency GT time series as predictors for the lower-frequency unemployment data in Brazil (a developing economy) and Portugal (developed), disaggregating by sex, age and education in a total of 48 unemployment time series. We resort to Mixed Data Sampling (MIDAS) models and enforce coherence across demographic and educational hierarchies using forecast reconciliation methods.

Results for the period 2021 Q1–2023 Q4 show that social inequalities shape how valuable GT data is for forecasting unemployment levels. Overall, incorporating GT data improves forecast accuracy, especially for women and people with higher education levels in Brazil and Portugal. Yet the digital divide remains a serious barrier: groups with lower digital literacy or limited Internet access are under-represented in search data, leading to less reliable predictions for these populations. These results underline the promise of GT for more timely, detailed labour-market insights but also its limitations when large segments of the population fall outside the digital mainstream.

Keywords: Digital divide · Google Trends · Hierarchical reconciliation · Mixed data sampling · Nowcasting

Acknowledgements: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia and ESF (European Social Fund) under the reference 2021.07583.BD and within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020. The authors gratefully acknowledge support from CEF.UP.

Ciência de Dados I

The Influence of Primary Health Care Provision on Emergency Department Attendance

Loide Ascenso [1[0000-0002-9341-6149]], Hugo Quintino [2[0000-0002-1317-5606]], and Paulo Infante [3[0000-0002-1644-9502]]

loide.ascenso@ipleiria.pt, hquintino@ulsac.min-saude.pt, pinfante@uevora.pt

[1] *ESTM - Instituto Politécnico de Leiria e PDMAT/IIFA, Universidade de Évora, Portugal*
[2] *Unidade Local de Saúde do Alentejo Central, EPE*
[3] *DMAT/ECT e CIMA/IIFA, Universidade de Évora*

Abstract: The organisation and accessibility of Primary Health Care (PHC) play a crucial role in the efficient use of Emergency Services (ES). Ideally, PHC should serve as the entry point for users into the healthcare system, with the family doctor responsible for managing health needs and referring patients to hospital care only when strictly necessary. Therefore, it is essential to understand how the availability of PHC influences users' decisions to seek hospital emergency services directly.
This study aims to assess the impact of family doctor coverage on visits to the Emergency Department of Hospital do Espírito Santo, EPE, also considering the general availability of PHC. The analysis will be based on the number of ES admissions in 2024 at the largest and most specialised hospital in the Alentejo region.
The study will begin with a correlation analysis to evaluate preliminary relationships between doctor coverage and frequency of ES use. Subsequently, a linear mixed-effects model will be applied to examine the relationship between family doctor coverage and the number of ES admissions, including fixed effects such as distance to hospital, time of admission (day/night), PHC availability, triage category, and patient diagnosis. A multinomial logistic regression model will also be used to analyse the influence of family doctor coverage on the probability of a patient being classified under a specific triage colour. Potential interactions between variables, such as distance to the hospital and doctor coverage, will also be assessed to understand combined effects on patient decision-making.
Additionally, machine learning models such as Random Forest and XGBoost will be explored to compare their predictive performance with classical statistical approaches. These models will help identify more complex patterns in the data and assess the robustness of the results. The comparison will be based on metrics such as accuracy, precision, sensitivity, specificity, AUC, mean squared error, and generalisation capability.
The results aim to support strategic policies for optimising the distribution of human resources in healthcare, reducing pressure on emergency services, strengthening the effectiveness of PHC, and promoting a more balanced, efficient, and citizen-centred healthcare response.

Keywords: Emergency Department Attendance· Generalized Linear Mixed Models · Machine Learning

Topic Analysis and Classification on Simulated RNA-seq Datasets - a Comparative Study

João F. Carrilho [1,2][0009-0002-2265-8922], Susan P. Holmes [3][0000-0002-2208-8168], and Marta B. Lopes [1,2,4][0000-0002-4135-1857]
jf.carrilho@campus.fct.unl.pt, susan@stat.stanford.edu, marta.lopes@fct.unl.pt

[1] *Center for Mathematics and Applications (NOVA Math), NOVA FCT, Caparica, Portugal*
[2] *Department of Mathematics, NOVA FCT, Caparica, Portugal*
[3] *Department of Statistics, Stanford University, Stanford, California, USA*
[4] *Research and Development Unit for Mechanical and Industrial Engineering (UNIDEMI), NOVA FCT, Caparica, Portugal*

Abstract: RNA sequencing (RNA-seq) datasets are characterized by their high dimensionality and correlation structure due to the biological interactions among genes. When samples come from different classes we can search for subsets of biomarkers explaining the classes. We ran a simulation study in R based on glioma RNA-seq data from The Cancer Genome Atlas, which has 3 classes, to assess the behavior of latent Dirichlet allocation (LDA) followed by random forest (RF) on datasets obtained using 3 different generative algorithms: LDA; linear combination of genuine signal, systematic noise and random noise using the `RUVcorr` package; semi-parametric estimation, based on the counts from a real dataset, of gene-specific distributions via a log-linear model-base density estimation approach, and the pairwise correlation structure via Gaussian copulas using the `SPsimSeq` package.
The inferred topics on the dataset generated through the LDA model were coherent with the simulated topics, as well as the classification accuracy with the induced separability of classes. Regarding `RUVcorr`, a one-class scenario resulted in topics with low proportions. On a 3-class simulation design, we obtained even lower proportions within the topics and achieved very high class prediction accuracy. With `SPsimSeq`, by simulating one dataset with 3 classes we observed high intersection of highlighted genes among topics and very low class prediction accuracy, while the row binding of 3 simulated datasets, one per class, led to higher classification accuracy and very low proportions within the topics.

Keywords: Classification · Latent Dirichlet allocation · RNA-seq · Simulation

References

[1] Sankaran, K., Holmes, S.P.: Latent variable modeling for the microbiome. Biostatistics, **20**(4), 599-614 (2019). DOI:10.1093/biostatistics/kxy018
[2] Freytag, S., Gagnon-Bartsch, J., Speed, T.P., Mahlo, M.: Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinformatics, **16**(309) (2015). DOI:10.1186/s12859-015-0745-3
[3] Assefa, A.T., Vandesompele, J., Thas, O.: SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. Bioinformatics, **36**(10), 3276-3278 (2020). DOI:10.1093/bioinformatics/btaa105

Misturas Pseudo-Convexas de Funções Potência e suas Aplicações

Miguel Felgueiras [1,2,3][0000-0001-5450-7374], João P. Martins [2,4,5][0000-0002-0474-1397] e Rui Santos [1,2][0000-0002-7371-363X]
mfelg@ipleiria.pt, jom@ess.ipp.pt, rui.santos@ipleiria.pt

[1] *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Portugal*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal*
[4] *Escola Superior de Saúde, Instituto Politécnico do Porto, Portugal*
[5] *Center for Health Studies and Research, University of Coimbra, Portugal*

Abstract: Neste trabalho, exploramos misturas pseudo-convexas de funções potência. Estas misturas distinguem-se por permitirem obter funções de sobrevivência côncavas, cuja relevância advém da sua aplicabilidade em contextos reais. O estudo de funções de sobrevivência côncavas revela-se pertinente em áreas tão diversas como a medicina, a economia ou a fiabilidade. Como exemplo, basta notar que normalmente uma componente tem maior probabilidade de falhar no início do seu ciclo de vida, depois esta probabilidade diminui e permanece razoavelmente constante, voltando a aumentar no final do ciclo de vida da componente. Neste contexto, é investigada a estimação de parâmetros, a moda e a função de sobrevivência destas misturas.
Concluímos que as misturas propostas têm interesse no estudo das funções de sobrevivência côncavas, pois não só a estimação dos parâmetros envolvidos é razoavelmente precisa, como a sua aplicação a dados reais permite obter modelos bem ajustados, sem aumentar o número de parâmetros envolvidos, em comparação com outros modelos comummente utilizados para modelar dados com função de sobrevivência côncava.

Keywords: Função potência · Função de sobrevivência · Mistura pseudo-convexa

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/00006/2025 e UIDB/00006/ 2020. https://doi.org/10.54499/UIDB/00006/2020.

References

[1] Santos, R., Martins, J., Felgueiras, M.: Pseudo-convex Mixtures Generated by Shape-extended Stable Distributions for Extremes. J Stat Theory Pract **10**(2), 357–374 (2016). https://doi.org/10.1080/15598608.2016.1146929.
[2] Liu, B., Ananda, M.: A New Insight into Reliability Data Modeling with an Exponentiated Composite Exponential-Pareto Model. Applied Sciences **13**(1), 645 (2023). https://doi.org/10.3390/app13010645.

Aplicações em Ambiente, Clima, Geociências e Agricultura I

Avaliação do Potencial Energético na Zona Costeira Norte e Centro de Portugal

Ana Leonor Oliveira [1], **Paula Milheiro-Oliveira** [1,2[0000-0002-4685-1615]] e Paulo Avilez-Valente [2,3[0000-0002-2562-6603]]

up201907697@edu.fc.up.pt, poliv@fe.up.pt, pvalente@fe.up.pt

[1] *Centro de Matemática da Universidade do Porto, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal*
[2] *Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal*
[3] *Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzerios do Porto de Leixões, Av. General Norton de Matos s/n, 4450-208 Matosinhos, Portugal*

Abstract: Este trabalho tem como objetivo avaliar o potencial energético das ondas na zona costeira portuguesa do Norte e Centro, assim como comparar a eficácia de diferentes dispositivos conversores de energia (WEC). Para isso os autores recorrem a dados sobre agitação marítima disponíveis nas bases de dados ERA5 e IBIS. O estudo centra-se em variáveis como a altura significativa e o período das ondas ao longo de uma larga faixa da região costeira portuguesa, bem como em pontos selecionados da costa ocidental ibérica. Foi também realizada uma avaliação da variabilidade temporal da potência disponível, de modo a aprofundar essa análise. Foi ainda realizada uma comparação entre as bases de dados ERA5 e IBIS. O estudo conduziu à identificação dos locais de maior potencial energético na costa portuguesa e permitui avaliar a viabilidade e eficiência de dispositivos conversores de energia, como o Pelamis, o Archimedes e o WaveDragon. O presente estudo vem contribuir para a identificação de estratégias eficazes na exploração da energia das ondas em Portugal.

Keywords: Distribuição da altura significativa · Distribuição do período de energia · Energia das ondas · Potencial energético · WEC

Acknowledgements: Esta investigação foi parcialmente apoiada pelo Financiamento Estratégico (UIDB/00144/2020 e UIDP/00144/2020) através de fundos nacionais disponibilizados pela FCT — Fundação para a Ciência e Tecnologia — e pelo Fundo Europeu de Desenvolvimento Regional (FEDER) ao abrigo do acordo de parceria Portugal 2020. O primeiro autor realizou grande parte da investigação enquanto estudante da Faculdade de Ciências da Universidade do Porto.

The Importance of Using Resolvable Row-Column Designs in Large Agricultural Experiments with Perennial Species

Elsa Gonçalves [1][0000-0003-0216-436X]
elsagoncalves@isa.ulisboa.pt

[1] *LEAF - Linking Landscape, Environment, Agriculture and Food Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Portugal*

Abstract: The experimental design adopted in agricultural experiments with a perennial species is a key issue. For instance, a field trial for grapevine selection comprises hundreds of different clones (treatments), which remain in the field for a minimum period of 25–30 years. Consequently, a field trial must address all potential challenges that may arise in experimental vineyards over their lifetime. Resolvable row-column designs have been shown to be an effective solution for addressing challenges in grapevine trials when yield is being evaluated. The objective of this work is to demonstrate the usefulness of these experimental designs in the prediction of genotypic effects for several important traits (yield, quality traits of the berries, and traits used to measure abiotic and biotic stress tolerance). Real data from several years obtained in selection grapevine field trials installed according to resolvable row-column designs were utilised. The efficiency of the effects of the design (resolvable replicate, and rows and columns within resolvable replicate) in the prediction of genotypic effects was evaluated by applying the principles of mixed models theory and by obtaining quantitative genetic indicators from the estimated covariance parameters of the model. The effectiveness of each effect associated with the experimental design varied depending on the year and trait evaluated, showing a higher influence on the efficiency of the prediction of genotypic effects of yield and traits related to abiotic and biotic stress tolerance. The adequacy and importance of using resolvable row-column designs in field trials with a perennial species was demonstrated.

Keywords: Experimental designs · Genetic selection · Grapevine field trials · Mixed models

Acknowledgements: The author would like to thank their colleagues at the National Network for Grapevine Selection and the Portuguese Association for Grapevine Diversity (PORVID) for their contribution to the management of the field experiments and data collection. This research was supported through funding of the projects "Save the intra-varietal diversity of autochthonous grapevine varieties" (PRR-C05-i03-000016) and "BioGrapeSustain" (C644866286-011, PRR – Agendas Mobilizadoras, B6.1).

Insights from a Data-Driven Study in an Automotive Assembly Line: Defects' Analysis

Marco Silva [1[0009-0000-8877-3689]], Ana Raquel Xambre [1[0000-0001-8615-3443]], Helena Alvelos [1[0000-0002-6450-5521]] Carlos Rodrigues [2[0009-0006-8984-9458]]
mass@ua.pt, raquelx@ua.pt, helena.alvelos@ua.pt , cmor@ua.pt

[1] *Center for Research & Development in Mathematics and Applications and Department of Economics, Management, Industrial Engineeing and Tourism, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*
[2] *University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*

Abstract: In the context of Industry 4.0, where large volumes of data are generated continuously, statistical techniques enable organizations to transform raw operational data into actionable insights, representing a critical tool in modern industrial organizations to assist continuous improvement. This study explores the critical human-related factors contributing to quality issues in the assembly sector of an automotive manufacturing company, with particular attention to the Trimming, Chassis, and Final Assembly lines. The investigation was motivated by a persistently high defect rate attributed to human error. A statistical methodology based on logistic regression was applied to a dataset comprising 240 monthly observations collected between September 2024 and April 2025. The variables considered include company seniority, assembly line type, gender, age, multiskill training, production volume, and overtime hours. The analysis aimed to determine which factors significantly influence the performance. The results reveal that higher seniority, assignment to the Chassis line, and a specific gender profile are associated with a higher probability of good performance. Conversely, age and cumulative production volume showed slight negative associations, suggesting potential links to fatigue or cognitive overload. These findings offer insights into workload management within industrial settings. The study shows the value of statistical modeling in transforming raw operational data into strategic knowledge, supporting evidence-based decision-making. Future research should build upon this foundation by integrating multicriteria decision-making (MCDM) algorithms. These tools can enable more comprehensive evaluation frameworks for assigning personnel to critical roles. Such integration aligns with the principles of Quality 4.0.

Keywords: Defect analysis · Quality 4.0 · Statistical analysis

Acknowledgements: This work is supported by CIDMA under the FCT (Portuguese Foundation for Science and Technology) Multi-Annual Financing Program for R&D Units, within project reference UID/04106., and also by the Portuguese Foundation for Science and Technology (FCT, Portugal) through the PhD grant 2024.04472.BDANA.

Séries Temporais II

Stability of Machine Learning Models for Time Series Forecasting

Mafalda Sá Ferreira [1] and Regina Bispo [1,2]
msm.ferreira@campus.fct.unl.pt

[1] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA FCT)*
[2] *School of Mathematics and Statistics and Centre for Research into Ecological and Environmental Modelling (CREEM), University of St Andrews, Scotland*

Abstract: Machine Learning models are powerful tools for time series forecasting [1]. However, their utility depends heavily on the accuracy and stability of their predictions, which are essential for informed decision making [2]. Among the most commonly used methods, Support Vector Machines and Random Forests offer distinct approaches to handling time series data [3].
This study focuses on comparing the stability of these two models, with the ultimate goal of providing analytical guidelines for Machine Learning researchers. Using bootstrap resampling, we assess how small variations in training data impact model performance in a simulated time series scenario. Our results indicate that Support Vector Machines demonstrates greater stability compared to Random Forests.

Keywords: Machine Learning · Stability · Time Series

Acknowledgements: This work was partially supported by Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects UIDB/00297/2020 (https://doi.org/10.54499/ UIDB/00297/2020), UIDP/00297/2020 (https://doi.org/10.54499/ UIDP/00297/2020) (Center for Mathematics and Applications) and 2023.02525.BDANA (https://doi.org/10.54499/2023.02525.BDANA).

References

[1] Masini, R.P., Medeiros, M.C., Mendes, E.F.: Machine learning advances for time series forecasting. Journal of Economic Surveys, **7**(1), 76-111 (2023). https://doi.org/10.1111/joes.12429

[2] Philipp, M., Rushc, T., Hornik, K., Strobl, C.: Measuring the stability of results from supervised statistical learning. Journal of Computational and Graphical Statistics, **27**(4), 685-700 (2018). https://doi.org/10.1080/10618600.2018.1473779

[3] Yu, P.S., Yang, T.C., Chen, S.Y., Kuo, C.M., Tseng, H.W.: Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. Journal of Hydrology, **552**, 92-104 (2017). https://doi.org/10.1016/j.jhydrol.2017.06.020

Clustering Zero-Inflated Time Series of Counts

Luís Sousa [1,2][0009-0000-2934-6193], Magda Monteiro [1,3][0000-0001-8585-4440] e Isabel Pereira [1,2][0000-0002-5152-546X]

luissousa2@ua.pt, msvm@ua.pt, isabel.pereira@ua.pt

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

Abstract: The clustering of time series has proven to be of interest in various fields, ranging from economics and finance to environment and medicine, among others. Specifically in the context of logistics, clustering time series may help to outline strategies for better decision-making, particularly in maritime ports. Some of the problems that may arise within this context range from clustering of ships based on the types of materials they are transporting to clustering cranes based on their productivity levels.
Much of the work developed over the recent decades has been conducted within the framework of continuous-valued time series, with few studies on clustering for count time series. The aim is to establish and apply model-based clustering to appropriately define discrete-valued time series, particularly those that allow for differing levels of dispersion and/or zero inflation. The idea is to use a finite mixture model that accommodates the mentioned characteristics, and several existing techniques, such as the selection of the number of clusters, estimation using expectation-maximization and model selection, are applicable. The methodology proposed employs a mixture of count models to cluster discrete-valued time series, in which each time series is allocated to a specific process. The processes considered follow an INAR(p) recursion with a zero-inflated innovation distribution. A simulation study is then carried out, and an illustration with a real data set is made as well.

Keywords: Clustering · Time Series of Counts · Zero-Inflation

Acknowledgements: This work is partially supported by CIDMA under the Portuguese Foundation for Science and Technology, reference UID/04106/2025. This content is also produced within the scope of the Agenda "NEXUS - Pacto de Inovação – Transição Verde e Digital para Transportes, Logística e Mobilidade", financed by the Portuguese Recovery and Resilience Plan (PRR), with no. C645112083-00000059 (investment project no.º 53).

Signed Periodic INAR(1) Model: A Comparative Study

Cláudia Santos [1,2][0000-0001-6792-2382] e Isabel Pereira [1][0000-0002-5152-546X]
csps@ua.pt, isabel.pereira@ua.pt

[1] *CIDMA - Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal*
[2] *CERNAS - Research Center for Natural Resources, Environment and Society, Polytechnic Institute of Coimbra, Portugal*

Abstract: Over the past few decades, discrete-valued time series have gained growing significance in both research and practical applications, leading to extensive creation of new models and methods. Among these, the integer-valued autoregressive (INAR) model—built using the binomial thinning operator—have emerged as a widely used approach for modeling count data. Nevertheless, this class of models is restricted to non-negative count data.
An extension of the INAR(1) model involves the use of alternative thinning operators to handle $\mathbb{Z}$-valued time series, including negative values. [1] proposed an INAR(1) model using the relative binomial thinning operator. In these signed models, innovations must follow a $\mathbb{Z}$-supported distribution to ensure that the discrete nature of the process is defined on the set of integers.
In this work, we consider two signed periodic INAR(1) models with different innovation distributions, namely the Skellam distribution and the extended Poisson distribution, both with support in $\mathbb{Z}$. Some properties of the periodic models are presented. The estimation of the parameters is addressed via two methods. The proposed models are applied to a real environmental data set.

Keywords: Extended Poisson distribution · INAR(1) model · Signed thinning operator · Skellam distribution

Acknowledgements: This work is partially supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT), reference UID/04106/2025.

References

[1] Kachour, M., Bakouch, H. S., Mohammadi, Z.: A New INAR (1) Model for Z-Valued Time Series Using the Relative Binomial Thinning Operator. Jahrbücher für Nationalökonomie und Statistik, **243**(2), 125-152 (2023). https://doi.org/10.1515/jbnst-2022-0059

Outliers in Dynamic Time Series Models: an Approach Using Robust Statistics and the Kalman Filter

A. Catarina Ribeiro [1], A. Manuela Gonçalves[1][0000-0001-8491-6048], and Marco Costa [2][0000-0001-7686-2430]

pg52209@alunos.uminho.pt, mneves@math.uminho.pt, marco@ua.pt

[1] *Department of Mathematics, Centre of Mathematics, University of Minho, Portugal*
[2] *Águeda School of Technology and Management, Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal*

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

Keywords: Kalman filter · Outliers · Robust estimation · State-space models · Time series

Acknowledgements: Marco Costa and A. Manuela Gonçalves were partially financed by Portuguese Funds through FCT within CIDMA under projects ref. UID/ 04106 and UID/00013: CMAT/UM, respectively. A. Catarina Ribeiro gratefully acknowledges the support of CMAT through the grant UMINHO/BIM/2024/131.

References

[1] Harvey, A. C.: Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press (2009). DOI:10.1017/CBO9781107049994
[2] Cipra, T., Romera, R.: Kalman filter with outliers and missing observations. TEST, pp. 379-395, 1997. DOI:10.1007/BF02564705
[3] Crevits, R., Croux, C.: Robust estimation of linear state space models. Communications in Statistics - Simulation and Computation, pp. 1694–1705, 2019. 10.1080/03610918.2017.1422752

Estatística Multivariada

The 2024 WRC Points System and Its Dual Reward Effect

J.R. Branco [1,2][0000-0003-2258-1902], L. Margalho [1,3][0000-0001-7259-3747] and C. Fidalgo [1][0000-0001-8646-4380]

jrbranco@isec.pt, lmelo@isec.pt, cfidalgo@isec.pt

[1] *Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo ,3045-093 Coimbra, Portugal*
[2] *CMUC, Department of Mathematics, University of Coimbra, Largo D. Dinis, 3000-143 Coimbra, Portugal*
[3] *RCM2+, Research Centre in Asset Management and System Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal*

Abstract: This study examines the controversial changes made to the 2024 World Rally Championship (WRC) scoring system, which introduced points at three different moments: *Saturday*, *Super Sunday* and *Power Stage*. The aim of these changes was to encourage more aggressive driving from the teams until the end of the event, but they faced strong criticism from drivers.
By analyzing the statistical relationship between the *Super Sunday* and*Power Stage* classifications, our study reveals a high correlation between the two, indicating that they essentially reward the same performance. These results are significant as they suggest that this structure may not accurately reflect the drivers' performance in a way that differentiates between the stages. The relevance of this work lies in its potential to inform future regulatory changes, ensuring that the WRC scoring system is both fair and meaningful, enhancing the competition's integrity.

Keywords: Point System · Power Stage · Rally · Super Sunday · World Rally Championship

Acknowledgements: This work was partially supported by the Centre for Mathematics of the University of Coimbra (UID/00324 - Centro de Matemática da Universidade de Coimbra).

References

[1] Dirtfish (2024). https://dirtfish.com/rally/wrc/drivers-call-for-evolution-of-points-system/. Accessed 4 June 2024;

[2] eWRC (2024). https://www.ewrc-results.com/season/. Accessed 4 June 2024 & 2 December 2024;

[3] WRC FIA World Rally Championship (2024). https://www.wrc.com. Accessed 4 June & 2 December 2024;

[4] What's new for the 2025 WRC? https://www.wrc.com/en/news/whats-new-for-the-2025-wrc. Accessed 21 January 2025.

Explorando a perceção face à Matemática de alunos de 3.º e 4.º anos de escolaridade

Ana Felizardo Henriques [1,2][0009-0003-0683-4604], Adelaide Freitas [1,3][0000-0002-4685-1615], Fernando Sebastião [2,4][0000-0002-8792-4649], and João Marôco [5][0000-0001-9214-5378]
anac@ua.pt, adelaide@ua.pt, fsebast@ipleiria.pt, jpmaroco@gmail.com

[1] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*
[2] *School of Technology and Management, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal*
[3] *Department of Mathematics, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal*
[4] *Laboratory of Separation and Reaction Engineering-Laboratory of Catalysis and Materials (LSRE-LCM), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal. ALiCE – Associate Laboratory in Chemical Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal.*
[5] *Intrepid Lab, ECEO, Lusófona University & CETRAD, University of Trás-os-Montes e Alto Douro*

Abstract: Pretende-se percecionar como os alunos de 3.º e 4.º anos de escolaridade se relacionam com a Matemática. Para tal foi desenhado um processo de amostragem por conveniência, aplicado nos primeiros meses do ano de 2025 em várias escolas do 1.º ciclo do Ensino Básico da zona Centro de Portugal. Como instrumento de recolha de dados optou-se por usar um questionário, devidamente autorizado por entidades institucionais e agentes educativos intervenientes, contendo questões sobre "Gosto por aprender Matemática" (9 itens), "Clareza na lecionação das aulas de Matemática" (6 itens), "Comportamento desordenado durante as aulas de Matemática" (6 itens) e "Confiança em Matemática" (9 itens) da versão portuguesa do "Student Questionnaire - Grade 4" do Trends in International Mathematics and Science Study (TIMSS 2019). Foram inquiridos 775 alunos do 1.º ciclo do Ensino Básino (377 do 3.º ano e 398 do 4.º ano), pertencentes a 15 escolas distribuídas por três agrupamentos de escolas.
Este trabalho apresenta uma análise estatística preliminar univariada e multivariada dos dados recolhidos. Em particular, é investigada a consistência interna das respostas obtidas e realizada uma Análise de Componentes Principais cujos resultados serão comparados com os obtidos nas questões correspondentes no TIMSS 2019, em Portugal. Para além disso, pretende-se usar o pacote `ClustOfVar` do software R para explorar e interpretar agrupamentos detetados para os itens (variáveis).

Keywords: Análise de componentes principais · Clustering · TIMSS

Acknowledgements: Este trabalho é suportado pelo CIDMA ao abrigo do Programa de Financiamento Plurianual de Unidades de I&D da Fundação para a Ciência e a Tecnologia (FCT, https://ror.org/00snfqn58).

CDPCA: A new starting point

Guilherme Pereira [1,4][0009-0006-1686-8279], Mariline Costa [1][0009-0005-0355-9947], and **Adelaide Freitas** [2,3][0000-0002-4685-1615]
g.pereira@ua.pt, marilinemcosta@ua.pt, adelaide@ua.pt

[1] *Physics Department, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal*
[2] *Department of Mathematics, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal*
[3] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*
[4] *CICECO – Aveiro Institute of Materials, Physics Department, University of Aveiro, Aveiro 3810-193, Portugal*

Abstract: Clustering and Disjoint Principal Component Analysis (CDPCA) is a constrained principal component analysis method which simultaneously detects P clusters of objects and identifies Q sparse and disjoint components, such that the between-cluster deviance in the reduced space of those components is maximized [1, 2]. Analogous to k-means, a prior selection of the P and Q parameters is required for any application of the CDPCA method.
The Calinski and Harabasz Index (CHI) is an internal evaluation metric for clustering algorithms which has been proposed to simultaneously select the number of object clusters and the number of components in two-mode methodologies. CHI is defined as the ratio of between-cluster deviance to within-cluster deviance, normalized by their respective degrees of freedom. By adapting CHI to measure deviances in the reduced space of disjoint components, a similar index has been proposed to estimate the (P, Q) values to be used in CDPCA. Nevertheless, in practical contexts, computing this index may require considerable processing time.
In this work, we introduce a new approach to estimate the optimal pair (P, Q) for CDPCA given a dataset. Using simulated datasets, results show that this novel method can be up to three times more efficient than the previously reported CHI-based algorithm. Applications on real datasets using the R function `CDpca` (from the `biplotboot` package in the CRAN) are also presented.

Keywords: Clustering · Kmeans · PCA

Acknowledgements: This work is supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58) Multi-Annual Financing Program for R&D Units. Project ref. UID/04106.

References

[1] Freitas, A., Macedo, E., M. Vichi. An empirical comparison of two approaches for CDPCA in high-dimensional data. Statistical Methods & Applications, **30**, 1007–1031 (2021).
[2] Vichi. M., Saporta, G.: Clustering and Disjoint Principal Component Analysis. Computational Statistics & Data Analysis, **53**, 3194–3208 (2009).

Casewise and Cellwise Outliers in Panel Data: Challenges and Robust Estimation Strategies

Anabela Rocha [1,2][0000-0002-3918-2476], M. Cristina Miranda [1,2,3][0000-0002-4642-5683], and Manuela Souto de Miranda [2][0000-0002-5703-5082]
anabela.rocha@ua.pt, cristina.miranda@ua.pt, manuela.souto@ua.pt

[1] *ISCA, University of Aveiro, Portugal*
[2] *CIDMA, University of Aveiro, Portugal*
[3] *CEAUL, University of Lisbon, Portugal*

Abstract: Robust methods are important for analyzing real data because they are less affected by violations of model assumptions. Real datasets often contain outliers—observations that deviate substantially from the majority of the data. Identifying these outliers is essential, as they can bias estimates and lead to incorrect conclusions when using traditional methods such as least squares. Robust methods fit models using the majority of the data and detect outliers based on deviations from the fitted model. This work focuses on panel data, a data structure common in fields such as health, biology, environmental science, economics, and finance. Panel data models enable analysis of both unit-specific and time-specific effects, offering advantages over cross-sectional or time-series data alone. We adapt recent techniques for detecting cellwise and casewise outliers to panel data and propose a robust estimator for the random effects model. This estimator modifies the feasible generalized least squares (FGLS) approach by incorporating robust procedures in the estimation steps. The performance of the proposed method is evaluated through simulations. Finally, an application to a real panel dataset illustrates the different estimate values obtained with each methodology.

Keywords: Casewise outliers · Cellwise outliers · FGLS · Panel data · Robust methods

Acknowledgements: This work is supported by CIDMA and CEAUL under the FCT (Portuguese Foundation for Science and Technology) projects UID/04106, UID/00006/2025 and UIDB/00006/2020, DOI: 10.54499/UIDB/00006/2020.

Aplicações em Ambiente, Clima, Geociências e Agricultura II

Clustering, time series and risk analysis for assessing water quality in a River Basin

A. Manuela Gonçalves [1,2][0000-0001-8491-6048], Irene Brito [1,2][0000-0002-7075-3265], and Ana Pedra [1,2]
mneves@math.uminho.pt, ireneb@math.uminho.pt, pg46704@alunos.uminho.pt

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

Abstract: Monitoring surface water quality in river basins is conditioned inherent risks due to uncertainties in hydrological and meteorological conditions, as well as anthropogenic, agricultural, and industrial pollution sources. Statistical methods are essential tools for analyzing and forecasting changes in water quality, as well as assessing the risk of water pollution. This study proposes a novel methodology that integrates clustering, risk theory, and time series analysis to assess and forecast surface water quality. The main objective is to develop an average risk index predictor for water pollution and evaluate whether the ranking derived from in-sample data can be used to forecast future pollution risk. The methodology is applied to monthly surface water quality data from monitoring stations in the Douro River basin, Portugal. Considering the influence of hydrological and meteorological conditions—particularly flow variability—the analysis distinguishes between the dry season (May–October) and the wet season (November–April). A cluster analysis groups monitoring stations with similar characteristics. Several risk measures (like value at risk, and probability of excess) are calculated for each cluster to quantify pollution risk. Time series models, such as SARIMA and exponential smoothing methods, are then applied to provide forecasts. These predictions are compared with the cluster-based rankings obtained from in-sample data to assess the performance of the risk index predictor. The results demonstrate the potential of the proposed methodology for anticipating water pollution risks. The approach is adaptable and can be applied to other river basins facing similar environmental challenges.

Keywords: Clustering · Risk analysis · Surface water quality · Time series

Acknowledgements: A. Manuela Gonçalves and Irene Brito were partially financed by Portuguese Funds through FCT within the Project UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM). Ana Pedra thanks CMAT for the research fellowship (BI) UMINHO/BIM/2022/100.

References

[1] Brito, I., Gonçalves, A.M., Pedra, A.: Risk assessment for the surface water quality evaluation of a hydrological basin. Stochastic Environmental Research & Risk Assessment, **38**(11), 4527–4553 (2024). DOI:10.1007/s00477-024-02817-w
[2] Gonçalves, A.M., Costa, M.: Clustering and forecasting of dissolved oxygen concentration on a river basin. Stochastic Environmental Research & Risk Assessment, **25**(2), 151–163 (2011). DOI:10.1007/s00477-010-0429-5

Boost Fisher Scoring: A Robust Approach to Parameter Estimation in State-Space Models

F. Catarina Pereira [1][0000-0002-6545-2900], Marco Costa [2][0000-0001-7686-2430], and A. Manuela Gonçalves [3][0000-0001-8491-6048]
id9976@uminho.pt, marco@ua.pt, mneves@math.uminho.pt

[1] *University of Minho, Centre of Mathematics, 4710-057 Braga, Portugal*
[2] *University of Aveiro, Águeda School of Technology and Management, Centre for Research and Development in Mathematics and Applications, 3810-193 Aveiro, Portugal*
[3] *University of Minho, Department of Mathematics and Centre of Mathematics, 4710-057 Braga, Portugal*

Abstract: Parameter estimation plays a crucial role in statistical modeling [1]. However, it can be challenging in the presence of outliers, which may compromise the quality of predictions and the convergence of numerical methods. This work presents a novel approach for parameter estimation in state-space models, based on a modified version of the Fisher scoring method, enhanced by bootstrap techniques. The proposed algorithm, named Boost Fisher Scoring (BF), combines the efficiency of the classical method with nonparametric bootstrap to approximate the Fisher information matrix. This strategy improves numerical stability and leads to more reliable estimates of standard errors. In addition, a robust extension of the method, called BFout, was specifically developed to handle time series with outliers. This version performs resampling on standardized residuals after removing outliers, ensuring that the bootstrap samples are less contaminated.The performance of both BF and BFout was assessed through simulations across various scenarios, including different sample sizes, levels of variance, and degrees of autocorrelation. The methods were also applied to real-world meteorological data, involving daily maximum air temperature forecasts, where their practical advantages became evident. Overall, the results suggest that these methods provide more reliable inference, particularly in small-sample or contaminated-data settings, being an efficient and robust alternative to traditional maximum likelihood methods.

Keywords: Bootstrap · Outliers · Robust estimation · State-space models

Acknowledgements: FCP was financed by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM. MC was partially financed by Portuguese Funds through FCT within CIDMA under the project UID/04106. AMG was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM).

References

[1] Shumway, R.H., Stoffer, D.S.: Time series analysis and its applications: with R examples. Springer (2025). DOI:10.1007/978-3-031-70584-7

Forecasting of Pollen Concentrations in Évora

Ana Sapata [1,2], Anabela Afonso [3][0000-0002-5517-4855],
Célia M. Antunes [4,5][0000-0001-8906-1085] and José Saias [6][0000-0003-3025-0687]
ana_sapata@sapo.pt, aafonso@uevora.pt, cmma@uevora.pt, jsaias@uevora.pt

[1] *University of Évora, School of Science and Technology, Portugal*
[2] *Dectechnologies - Sistemas de Informação, Lda*
[3] *University of Évora, Research Center in Mathematics and Applications (CIMA), School of Science and Technology, Portugal*
[4] *University of Évora, School of Health and Human Development, Portugal*
[5] *University of Évora, Center for Sci-Tech Research in Earth System and Energy, CREATE, Portugal*
[6] *University of Évora, ALGORITMI Research Center/LASI, VISTA Lab, Portugal*

Abstract: The increase in respiratory diseases related to allergic reactions led the World Health Organization to declare allergies as a public health problem. Changes in pollen seasons, such as their intensity or duration, have an impact on the symptoms in people with allergies. Therefore, it is important to monitor and forecast the pollen concentrations. The primary objective of this work is to forecast pollen concentration in Évora based on meteorological variables. The methodology utilizes a 22-year daily dataset (2002-2023) from Évora, encompassing the variable in study and meteorological variables, including temperature, precipitation, relative humidity, and wind. Following robust data treatment for handling missing and anomalous values, an in-depth descriptive statistical analysis was performed. This analysis revealed a pronounced seasonality in pollen concentrations, with notable peaks between March and May, and a strong temporal autocorrelation. The modeling approach involves applying and comparing various techniques, ranging from classical statistical time series models (e.g., ARIMA, SARIMA, and Dynamic Regression), to Machine Learning algorithms such as Artificial Neural Networks. Model performance was rigorously evaluated using metrics such as the Mean Absolute Error and Root Mean Square Error, with the use of time-series adapted cross-validation strategies to ensure their robustness and generalization power. This study aims not only to deepen the understanding of pollen dynamics in Évora but also to provide more accurate forecasting tools that can be directly applied to public health management and assist individuals with allergies.

Keywords: Forecasting · Machine learning · Pollen concentrations · Time series

Acknowledgements: This work is partially supported by national funds through FCT - Fundação para a Ciência e Tecnologia under the project https://doi.org/10.54499/UIDB/04674/2020. This work is partially supported by national funds through FCT - Fundação para a Ciência e Tecnologia in the framework of the CREATE project UID/06107/2023.

Estatística Computacional I

Modelo de Previsão: uma Aplicação a uma Indústria de Calçado

Nadine Laranjeira[1], **M. Rosário Ramos** [1,2[0000-0001-9114-0807]]
nadinelaranjeira@gmail.com, MariaR.Ramos@uab.pt

[1] *Universidade Aberta, Portugal*
[2] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: O crescimento económico do setor secundário contribuiu para que as empresas se tornassem mais competitivas. Para alcançar o sucesso há necessidade de recorrer a previsões que contribuem para uma boa gestão empresarial. Na indústria, as previsões desempenham um papel crucial, permitindo um melhor planeamento, a antecipação de produções, a otimização de recursos e a redução de custos, resultando numa maior eficiência e na melhoria da produtividade.
Com o aumento da faturação surge a necessidade de utilização de ferramentas de análise de dados para apoio à decisão e antever resultados financeiros. Este trabalho teve como objetivo aplicar modelos de previsão de séries temporais a uma indústria de calçado sediada na região Norte de Portugal. Procurou-se um modelo para as vendas, i.e., a produção que é adquirida pelos clientes que comercializam os artigos de calçado. A análise teve por base a série das vendas mensais e a série das previsões mensais das encomendas que o cliente partilha no início de cada ano.
Considerando que as séries podem evidenciar tendência e/ou efeito sazonal (não estacionárias), recorreu-se aos modelos ARIMA e à diferenciação sazonal. Foi utilizada a regressão dinâmica, nomeadamente ARIMAX, para incorporar as previsões dos clientes no modelo. Quando necessário, foi aplicada a transformação Box-Cox. Foram estudadas as vendas de cinco artigos/tipos de calçado produzidos pela empresa. O modelo ARIMAX revelou-se o mais adequado para previsão em quatro dos artigos da produção e o modelo ARIMA para um dos artigos. Por fim, destacam-se algumas limitações e desafios colocados, relacionados com o tamanho da série e o período coberto, que abrangeu a situação da pandemia COVID-19. Todas as análises foram conduzidas no software R.

Keywords: ARIMA · Previsão · Regressão dinâmica · Vendas

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025 and UIDB/00006/2020.
DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Group Lasso for Finite Mixtures of Linear Regression Models: a Simulation Study

Ana Moreira [1[0009-0008-1779-8086]] and Susana Faria [1[0000-0001-8014-9902]]
id10866@uminho.pt, sfaria@math.uminho.pt

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

Abstract:
Finite Mixture Regression models provide a flexible tool for analyzing data that arise from heterogeneous populations, where the relationship between the dependent variable and explanatory variables may differ across latent subpopulations. In practical applications, these models often involve a large number of explanatory variables, making variable selection a critical aspect of model building. To address this, several regularization methods have been proposed. In particular, group-based penalization methods are well-suited for categorical data, as they respect the group structure inherent in such predictors. This study focuses on the problem of variable selection within mixtures of linear regression models under both low- and high-dimensional settings. We compare the performance of three penalization methods: the Least Absolute Shrinkage and Selection Operator (LASSO), the Group LASSO, and the Adaptive Group LASSO (AGLASSO). Our analysis includes scenarios where the number of variable groups grows with the sample size, as well as cases in which the number of groups surpasses the sample size. Through a comprehensive simulation study, we assess how different data configurations impact the performance of these techniques in identifying informative predictors. The results demonstrate that the AGLASSO consistently achieves superior performance across multiple scenarios.

Keywords: Group lasso · Mixtures of linear regression models · Penalized maximum likelihood estimation · Simulation study

Acknowledgements: This research at CMAT was supported by FCT - Fundação para a Ciência e a Tecnologia, I.P. by project reference 2022.12256.BD and https://doi.org/10.54499/2022.12256.BD identifier.

References

[1] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), **58.**(1.), 267-288 (1996). DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Wang, M., Tian, G.L.: Adaptive group Lasso for high-dimensional generalized linear models. Statistical Papers, **60.**,1469-1486 (2019). DOI:https://doi.org/10.1007/s00362-017-0882-z

[3] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology, **68.**(1.), 49-67 (2006). DOI:https://doi.org/10.1111/j.1467-9868.2005.00532.x

Analytical Properties of Kalman Predictor Derivatives in State-Space Models

Marco Costa [1,2][0000-0001-7686-2430] and Magda Monteiro [1,2][0000-0001-8585-4440]
marco@ua.pt, msvm@ua.pt

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

Abstract: This communication explores the analytical structure and recursive computation of the partial derivatives of Kalman filter predictors with respect to model parameters in linear Gaussian state-space models. These derivatives are fundamental in score-driven parameter estimation via maximum likelihood and other estimation approaches, such as the generalized method of moments based on 1st order Taylor polynomials. We present a complete recursive formulation for the derivatives of the one-step ahead predictor, filtered prediction, smoothed prediction, Kalman gain, and their associated covariance matrices. Under stationarity assumptions, the initial conditions for these derivatives are derived using matrix differential calculus, exploiting the structure of Kronecker products and the inversion of linear operators. A key result concerns the proportionality of the derivatives of Kalman predictors with respect to the process and observation noise variances in the univariate case. We prove that these derivatives are linearly related across all time steps, with proportionality factors determined by the inverse of the ratio of variances. These relationships also extend to the Kalman and smoothing gains, highlighting deep structural regularities in the filter's sensitivity to variance components. All expressions are systematically organized in recursive form, supporting direct implementation in numerical algorithms. While numerical approximations such as finite differences or automatic differentiation are available, the analytical approach ensures greater precision and computational efficiency, particularly in real-time or high-dimensional settings.

Keywords: Kalman filter · Predictor derivatives · Recursive computation · State-space models

Acknowledgements: Marco Costa and Magda Monteiro were partially financed by Portuguese Funds through FCT within CIDMA under the project UID/04106.

Estatística Espacial I

Supervised Statistical Learning Methods in the Presence of Spatial Correlation

Beatriz Ferreira[1] and Raquel Menezes [1[0000-0001-5552-917X]]
pg52211@alunos.uminho.pt, rmenezes@math.uminho.ptt

[1] *CMAT – Centre of Mathematics (CMAT), University of Minho, Guimarães, Portugal*

Abstract: This project investigates the application and performance of supervised machine learning methods for data with spatial correlation, focusing on Random Forest Regression Kriging (RFRK) and Gaussian Process Boosting (GPB). Spatially correlated data often exhibit nonstationarity, heterogeneity, and spatial dependence that challenge classical model assumptions, making it essential to explore machine learning techniques capable of handling these complexities while ensuring predictive accuracy. Using simulated data, we assess each method's predictive performance and variable selection capabilities under both linear and nonlinear relationships. Under linear settings, Kriging performs better, with RFRK and GPB producing similar covariance parameter estimates (τ^2, σ^2, ϕ) for Gaussian and exponential spatial correlation structures. However, Kriging's performance declines under nonlinear conditions, showing reduced accuracy and higher prediction error, whereas GPB maintains strong predictive power and more robust behavior. Both RFRK and GPB effectively identify the most relevant predictors of the response variable. To validate and extend these findings, we are applying these methods to real-world data from scientific campaigns conducted by the Portuguese Institute for the Sea and Atmosphere (IPMA), aiming to characterize the spatial distribution of pelagic marine species based on environmental predictors and to further evaluate the practical applicability of these approaches in realistic prediction tasks.

Keywords: Boosting · Geostatistics · Random Forest

Acknowledgements: This study was supported by the Portuguese Foundation for Science and Technology (FCT), provided through the Centre of Mathematics via the project UID/00013 and the Individual Scholarship UMINHO/BIM/2024/130.

References

[1] Sigrist, F.: Gaussian Process Boosting. *Journal of Machine Learning Research* (2022). https://doi.org/10.48550/arXiv.2004.02653

[2] Dumelle, M., Higham, M., Ver Hoef, J.M.: spmodel: Spatial statistical modeling and prediction in R. *PLoS ONE* (2023). https://doi.org/10.1371/journal.pone.0281920

Spatial Analysis of Ascending Thoracic Aortic Aneurysms

Alda Carvalho [1,2][0000-0003-2642-4947], Rodrigo Valente [3][0000-0003-0871-0683], José Xavier [3][0000-0002-7836-4598], António Tomás [4][0000-0002-9144-2669] and Katalina Oviedo Rodríguez [5][0000-0001-8869-4067]
alda.carvalho@uab.pt, rb.valente@campus.fct.unl.pt, jmc.xavier@fct.unl.pt, acruztomas@gmail.com, katalina.oviedo.rodriguez@una.ac.cr

[1] *DCeT, Universidade Aberta, Lisbon, Portugal*
[2] *CEMAPRE/ISEG Research, Universidade de Lisboa, Lisbon, Portugal*
[3] *UNIDEMI, NOVA School of Science and Technology, Lisbon, Portugal*
[4] *Department of Cardiothoracic Surgery, Santa Marta Hospital, Lisbon, Portugal*
[5] *Escuela de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional Heredia, Costa Rica*

Abstract: Cardiovascular desease are the leading cause of death in adult over 65. Within Ascending Thoracic Aortic Aneurysms (ATAA) rupture is a serious asymptomatic phatology than cause death if not treated. Current clinical criteria based on maximum diameter fail to reliably predict rupture risk. ATAAs arise from progressive weakening and dilation of the thoracic aorta, influenced by age, genetics, and lifestyle factors. This talk presents a spatial statistical approach to analyze ATAAs and their biomechanical behavior, using computed tomography angiography (CTA) scans from 87 patients. Measurements extracted along the aortic centerline were used to compute experimental variograms for key variables such as maximum diameter (the clinical standard) and cross-sectional area. Focusing on systolic (35%) and diastolic (75%) phases of the cardiac cycle, we analyzed spatial structural differences between systole and diastole, revealing patterns linked to aortic wall geometry. Excluding a few atypical cases, two predominant variogram types emerged, strongly correlated with aneurysm location and extension. These findings suggest spatial statistical patterns provide valuable insights beyond diameter alone, which may improve risk assessment in ATAA patients.

Keywords: Ascending aortic aneurysm · CTA scans · Spatial statistics · Variogram

Acknowledgements: The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT - MCTES) for its financial support via the project AneurysmTool (DOI: 10.54499/PTDC/EMD-EMD/1230/2021).

Modelling Wildfires in the UK Using Spatio-Temporal Point Processes

Gordon Andrew Hannah[1][ORCID: 0009-0008-0401-8511]
gah22@st-andrews.ac.uk

[1] *University of St Andrews, United Kingdom*

Abstract: Wildfire events are increasing throughout the globe. Climate change, land use pressures, and human activity are contributing to these events occurring more often and with greater severity. These events pose a severe risk to life, cause ecological and environmental damage, and affect forestry-based economic activity.

It is therefore paramount to develop models that can help predict under what conditions such destructive events occur. This study explored the spatial distribution of wildfire ignitions in the United Kingdom using a point process modelling framework, with particular attention to spatial autocorrelation and latent spatial structure.

Log-Gaussian Cox processes (LGCPs) were applied to model the spatial distribution of wildfire ignitions using the inlabru package in R, which builds on the integrated nested Laplace approximation (INLA) framework. This approach allowed for the modelling of unobserved spatial variation through a latent Gaussian random field and enabled efficient inference even in complex spatial models.

The analysis focused on comparing spatial patterns of ignition with variation both spatially and temporally. The modelling framework supported the representation of spatial heterogeneity and offered insight into structural patterns in the ignition data. This approach provides a foundation for further refinement and interpretation as model development progresses.

By applying a consistent spatial framework to wildfire data from the UK, this study contributes to the wider effort to understand fire occurrence through statistically rigorous and spatially aware methods, capable of adapting to evolving fire-prone landscapes.

Keywords: Point processes · Spatio-temporal modelling · Wildfires

Acknowledgement: Thank you to Dr Regina Baltazar Bispo for their continued support and direction throughout this research.

Métodos Não Paramétricos

Rank and Related Tests: A Randomization Procedure for Grouping Factor Levels in Cocoa Breeding Experiments

Kwaku Opoku-Ameyaw [1,2][0000-0001-8606-6155], **Célia Nunes** [2,3][0000-0003-0167-4851], and Manuel L. Esquível [4][0000-0003-4991-7568]
kwaku.opokuameyaw@crig.org.gh, celian@ubi.pt, mle@fct.unl.pt

[1] *Cocoa Research Institute of Ghana, New Tafo-Akim, Ghana*
[2] *CMA - Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal*
[3] *Department of Mathematics, University of Beira Interior, Covilhã, Portugal*
[4] *Department of Mathematics, Nova School of Science and Technology and Nova Math, Universidade Nova de Lisboa, Caparica, Portugal*

Abstract: Rank tests play a crucial role in statistical analysis by providing a robust approach for analyzing data that do not meet the assumptions of conventional parametric statistics. Their versatility makes them applicable to a broad spectrum of contexts, leading to their widespread adoption across diverse research areas, including agriculture and finance, among others. In the field of statistics, grouping of levels within a factor is a fundamental task that underpins meaningful inferences. In 2023, [1] proposed a nonparametric test for grouping the levels of a factor, using the Edgeworth series expansion to obtain the distribution's quantiles. These quantiles were then used to build confidence intervals for testing the hypothesis, serving as an alternative to deriving the exact distribution of the test statistic, which is often challenging. Building upon this framework, the present study proposes an alternative randomization procedure. inspired by Fisher's randomization method, to derive the exact distribution of the test statistic, with a specific focus on both univariate and multivariate settings. The proposed approach is applied to a cocoa breeding experiment conducted in Ghana, to evaluate the performance of adaptability of twelve different cocoa varieties on four types of acidic soils.

Keywords: Cocoa breeding experiment · Grouping of factor levels · Nonparametric test · Randomization method · Univariate and multivariate cases ·

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the Center of Mathematics and Applications (UID/00212) and NOVA Math (UID/00297).

References

[1] Opoku-Ameyaw, K., Nunes, C., Esquivel, M. L., Mexia, J.T.: CMMSE: a nonparametric test for grouping factor levels: an application to cocoa breeding experiments in acidic soils. Journal of Mathematical Chemistry, **61**(3), pages 652-672 (2023). DOI:10.1007/s10910-022-01431-x

Métodos de agregação: contributo dos estimadores baseados em entropia

Ana Helena Tavares[1,2][0000-0003-4632-3561], Maria Costa[1,3][0000-0002-4776-6375] e Pedro Macedo[1,3][0000-0002-4371-8069]
ahtavares@ua.pt, lopescosta@ua.pt, pmacedo@ua.pt

[1] *Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA)*
[2] *Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro*
[3] *Departamento de Matemática, Universidade de Aveiro*

Abstract: Os métodos de agregação, concebidos para lidar com dados em larga escala, baseiam-se na subdivisão do conjunto de dados em grupos, na obtenção de estimativas em cada grupo e, posteriormente, na combinação desses resultados para produzir estimativas finais mais estáveis e precisas. No entanto, em contextos mal condicionados, a estabilidade destes métodos pode ser comprometida devido à presença de colinearidade.
Este trabalho compara o desempenho de três métodos de agregação, Bagging, Magging e Neagging [1], em combinação com três estimadores, no contexto de um modelo de regressão linear mal condicionado. Os estimadores considerados incluem um estimador clássico (OLS) e dois estimadores baseados em entropia (GME e W-GME). O estudo de simulação abrangeu 48 cenários distintos, resultantes da combinação de diferentes níveis de colinearidade, número de grupos, número de observações por grupo, suporte dos parâmetros desconhecidos e distribuição do erro aleatório.
Os resultados do estudo confirmam que o desempenho dos métodos de agregação depende fortemente do estimador utilizado nos subconjuntos amostrados. Destaca-se, em particular, o desempenho excecional do método de agregação Magging aliado ao estimador W-GME, mesmo usando suportes amplos (o que mimetiza a ausência de informação prévia sobre os parâmetros desconhecidos). Adicionalmente, evidencia-se a importância dos métodos baseados em entropia e os contextos em que estes superam as técnicas clássicas, fornecendo orientações para a sua aplicação prática.
As metodologias de agregação são também aplicadas a um conjunto de dados reais, por forma a ilustrar a sua utilidade prática e demonstrar a precisão dos estimadores baseados em entropia.

Keywords: Colinearidade · Entropia · Regressão Linear · Simulação

Acknowledgements: O trabalho inicial contou com a colaboração de Ana Silva, Tiago Freitas e Rui Costa. É parcialmente suportado pelo CIDMA ao abrigo do Programa de Financiamento Plurianual de Unidades de I&D da FCT (Fundação para a Ciência e a Tecnologia).

References

[1] Costa, M.C.; Macedo, P.; Cruz, J.P. Neagging: An Aggregation Procedure Based On Normalized Entropy. In: Proceedings of the International Conference on Numerical Analysis and Applied Mathematics, 2022.

Métodos de Reamostragem na Estimação do Índice Extremal

Dora Prata Gomes [1[0000-0002-5165-2346]] e M. Manuela Neves[2[0000-0003-2468-3857]]
dsrp@fct.unl.pt, manela@isa.ulisboa.pt

[1] *Centro de Matemática e Aplicações (NOVA Math) e Departamento de Matemática, NOVA FCT*
[2] *Centro de Estatística e Aplicações (CEAUL) e Instituto Superior de Agronomia, Universidade de Lisboa*

Abstract: A estimação de parâmetros associados a eventos raros e/ou extremos é fundamental em diversas áreas aplicadas, como as ciências ambientais, a engenharia e finanças. Em particular, o índice de valores extremos, ξ, e o índice extremal, θ, desempenham um papel central na caracterização da cauda das distribuições e da dependência temporal de eventos extremos. No entanto, a obtenção de estimativas precisas destes parâmetros continua a representar um desafio, devido à sensibilidade face à escolha do número k de observações mais extremas consideradas. Este trabalho propõe uma abordagem adaptativa para a estimação de θ, baseada na aplicação dos métodos de reamostragem *bootstrap* e *jackknife*. É desenvolvido um algoritmo para a escolha automática de k que equilibra o viés e variância das estimativas. A metodologia é avaliada através de um extenso estudo de simulação, incluindo casos com dependência temporal e aplicada a conjuntos de dados reais de séries ambientais.

Keywords: Escolha adaptativa · Índice extremal · Métodos de reamostragem

Acknowledgements: Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito dos projetos UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) e UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (Centro de Matemática e Aplicações) e do projeto UIDB/00006/2020 (CEAUL).

References

[1] Neves, M.M., Gomes, M.I., Figueiredo, F. and Prata Gomes, D.: Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation. Journal of Statistical Theory and Practice **1 9**(1), 184-199 (2015). DOI:10.1080/15598608.2014.890984

[2] Prata Gomes, D. and Neves, M. M: Revisiting Estimation Methods for Some Parameters of Rare Events. In: Proceedings New Frontiers in Statistics and Data Science, pp. 409-421 (2025). DOI:10.1007/978-3-031-68949-9_30

Simulation Study on Projection-Based Goodness-of-Fit Tests for Generalized Linear Models

Rui Costa-Miranda [1[0009-0002-0308-787X]], Rita Gaio [1[0000-0003-3906-0775]] and Wenceslao González-Manteiga [2[0000-0002-3555-4623]]
rui.miranda@med.up.pt, argaio@fc.up.pt, wenceslao.gonzalez@usc.es

[1] *Faculty of Sciences of the University of Porto and Centre of Mathematics of the University of Porto, Portugal*
[2] *Faculty of Mathematics of the University of Santiago de Compostela, Spain*

Abstract: Goodness-of-fit tests are essential tools for validating the adequacy of generalized linear models (GLMs). However, many well-established test statistics rely on data-dependent asymptotic distributions and are locally sensitive to model estimation errors. When contrasting composite null hypotheses, this typically requires bootstrap procedures involving repeated re-estimations of the model, which can be computationally intensive and result in a lack of power for high-dimensional settings. To overcome these limitations, we investigate a recently proposed fast-bootstrap goodness-of-fit test [2]. It is based on a Crámer–von Mises-type test statistic constructed from squared Neyman orthogonal kernels integrated with respect to a Gaussian process distribution. Focusing on count data models, we conduct a simulation study to compare the finite-sample performance of this test against the preceding projection-based method, which is not robust to estimation errors [1]. We examine empirical size and power under various null and alternative hypotheses involving different conditional moment restrictions. The results suggest that the orthogonal Gaussian-process-based approach offers improved robustness and computational efficiency, making it a promising alternative for model checking in GLMs.

Keywords: Conditional moment restrictions · Generalized linear models · Goodness-of-fit · Neyman orthogonal kernels

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

References

[1] Escanciano, J.C.: A consistent diagnostic test for regression models using projections. Econometric Theory, **22**(6), 1030-1051 (2006). DOI:10.1017/S0266466606060506

[2] Escanciano, J.C.: A Gaussian process approach to model checks. The Annals of Statistics, **52**(5), 2456-2481 (2024). DOI:10.1214/24-AOS2443

A Delta Sequence Class of Density Derivative Estimators for Circular Data

Carlos Tenreiro [1][0000-0002-5495-6644]
tenreiro@mat.uc.pt

[1] *CMUC, DMUC, University of Coimbra*

Abstract: Given an independent and identically distributed sample of angles $X_1, \ldots, X_n \in [0, 2\pi[$ from some absolutely continuous circular random variable X with unknown probability density function f, we are interested in this work in estimating the r-order derivative of f, with $r \in \mathbb{N}$, by using a general class of estimators, called delta sequence estimators, which are defined, for $\theta \in [0, 2\pi[$, by

$$\hat{f}_{r,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \delta_{r,n}(\theta - X_i),$$

where $\delta_{r,n} : \mathbb{R} \to \mathbf{R}$, for $n \in \mathbf{N}$, is a sequence of periodic functions with period 2π, called delta function sequence, which satisfies the conditions

(Δ.1) For all $n \in \mathbb{N}$ we have $\int_{-\pi}^{\pi} \delta_{r,n}(y)^2 dy < \infty$, and

$$\int_{-\pi}^{\pi} y^{\ell} \delta_{r,n}(y) dy = (-1)^r r! \delta_0(r - \ell) + o(1), \text{ for } \ell = 0, 1, \ldots, r;$$

(Δ.2) $\limsup_{n \to +\infty} \int_{-\pi}^{\pi} |y|^r |\delta_{r,n}(y)| dy < \infty$;

(Δ.3) $\sup_{\lambda < |y| \leq \pi} |\delta_{r,n}(y)| \to 0$ as $n \to +\infty$, for all $0 < \lambda < \pi$.

This class of estimators includes not only the estimator of $f^{(r)}$ based on the standard kernel density estimator initially studied by Beran (1979), Hall et al. (1987), and Bai et al. (1988), but also the estimator of $f^{(r)}$ based on the Parzen-Rosenblatt type density estimator proposed and studied in Tenreiro (2022, 2025).

Keywords: Circular data · Density derivatives · Kernel estimation

Acknowledgements: Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324.

References

[1] Bai, Z.D., Rao, C.R., and Zhao, L.C.: Kernel estimators of density function of directional data. Journal of Multivariate Analysis, **27**, 24–39 (1988).
[2] Beran, R. Exponential models for directional data. The Annals of Statistics, **7**, 1162–1178 (1979).
[3] Hall, P., Watson, G.S., Cabrera, J. Kernel density estimation with spherical data. Biometrika, **74**, 751–762 (1987).
[4] Tenreiro, C. Kernel density estimation for circular data: a Fourier series-based plug-in approach for bandwidth selection. Journal of Nonparametric Statistics, **34**, 377–406 (2022).
[5] Tenreiro, C. A note on a Parzen–Rosenblatt type density estimator for circular data. In: *New Frontiers in Statistics and Data Science (SPE2023, Guimarães, Portugal, October 11-14)* Henriques-Rodrigues, L. et al. (eds), Springer Proceedings in Mathematics & Statistics 469, pp. 1–11, 2025.

Análise de Sobrevivência

Avanços em Modelos de Sobrevivência Multi-estado Não-Markovianos: Revisão Sistemática e Perspetivas Futuras

Marta Azevedo [1][0009-0003-4412-6991], Luís Meira-Machado [1][0000-0002-8577-7665] e Carla Moreira [1][0000-0002-0570-0650]

marta.vasconcelos4@gmail.com, lmachado@math.uminho.pt, carlamgmm@gmail.com

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

Abstract: Nesta comunicação, apresentamos uma revisão sistemática de cerca de 50 artigos pioneiros em análise de sobrevivência multi-estado não-Markoviana — um campo em rápida expansão na bioestatística que ultrapassa as limitações dos modelos Markovianos clássicos. Após uma pesquisa exaustiva em bases de dados, identificámos abordagens inovadoras, como processos semi-Markov, modelos doença–morte, análises 'landmark' e métodos de ponderação por probabilidade inversa, capazes de lidar com covariáveis dependentes do tempo, censura por intervalos e dados truncados à esquerda.

Os resultados destacam três eixos metodológicos principais: (i) Estimadores flexíveis e Bayesianos para acomodar formas complexas de risco; (ii) Inovações computacionais que permitem escalabilidade; (iii) Aplicações de alto impacto em oncologia, doenças crónicas e registos eletrónicos de saúde.

Concluímos que, embora estas metodologias ofereçam alternativas poderosas aos modelos convencionais, persistem lacunas em protocolos de validação padronizados e em implementações de software acessíveis. Por fim, apontamos três desafios para investigação futura: diagnóstico robusto de não-Markovianidade, seleção de covariáveis em cenários de alta dimensionalidade e integração com abordagens de machine learning.

Este trabalho traça o estado-da-arte e abre caminho para o desenvolvimento de ferramentas mais rigorosas e práticas na análise multi-estado não-Markoviana em ambientes biomédicos.

Keywords: Análise de sobrevivência · Modelo multi-estado · Não-Markoviano · Revisão sistemática

Acknowledgements: Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., ao abrigo do Contrato-Programa UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM), e do projeto com a referência 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] Putter, H., Fiocco, M., Geskus, R. B.: Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**(11), 2389–2430 (2007).

[2] Aalen, O. O., Borgan, Ø., Gjessing, H. K.: *Survival and Event History Analysis: A Process Point of View*. Springer (2008).

[3] Datta, S., Satten, G. A.: Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, **58**(4), 792–802 (2002). https://doi.org/10.1111/j.0006-341X.2002.00792.x

Evaluating the Linearity of a Covariate in Shared-Parameter Joint Models

Xavier Piulachs [1[0000-0003-2150-6273]], Anouar El Ghouch [2[0000-0002-3805-726X]], and Ingrid Van Keilegom [2,3[0000-0001-8827-7642]]

xavier.piulachs@upc.edu,
anouar.elghouch@uclouvain.be,
ingrid.vankeilegom@kuleuven.be

[1] *Department of Statistics and Operations Research, Polytechnic University of Catalonia, Spain*
[2] *ISBA - Institute of Statistics, Biostatistics and Actuarial Sciences, Catholic University of Louvain, Belgium*
[3] *ORSTAT - Research Centre for Operations Research and Statistics, KU Leuven, Belgium*

Abstract: Shared-parameter joint modeling is a useful technique for associating longitudinal and time-to-event data [1]. When the focus is on the survival outcome, the conditional logarithm of the hazard function is typically assumed to be linearly related over time to a set of explanatory covariates, among other terms. However, this assumption is restrictive and may lead to misleading results. Our aim is to easily assess this modeling hypothesis for any continuous fixed covariate. To do so, we examine the appropriateness of a nonparametric testing procedure based on a penalty-modified Akaike information criterion [2, 3]. An extensive numerical study is conducted to check the validity of the test within the joint modeling framework, while determining the extent of deviation from linearity in the covariate effect. Moreover, once a deviation is detected, we examine the improvement in the model's predictive performance. The usefulness of the testing procedure is illustrated with a clinical trial of HIV-infected subjects, focusing on the effect of nadir CD4 cell count in a predictive joint model for time to immune recovery.

Keywords: Akaike information criterion · Joint model · Nonlinear covariate · Order selection test

Acknowledgements: This work was partially supported by Belgian scientific research funding through the FWO and F.R.S.- FNRS (via the Excellence of Science Program, project ASTeRISK, grant no. 40007517) and through the FWO (via the Senior Research Projects, grant no. G047524N).

References

[1] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time Data. *Biostatistics* 1, no. 4 (2000): 465–480.
[2] J.D. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag: New York, USA (1997).
[3] M. Aerts, G. Claeskens, and J.D. Hart. Testing the Fit of a Parametric Function. *Journal of the American Statistical Association* 94, no. 447 (1999): 869–879.

Using A Joint Model for Longitudinal and Time-to-Event Data to Estimate the Causal Effect of Liver Transplantation on Survival in Hepatocellular Carcinoma Patients

Pedro Miranda Afonso[1][0000-0001-6708-9597], Hao Liu[2], Michele Molinari[3], and Dimitris Rizopoulos[1][0000-0001-9397-0900]

p.mirandaafonso@erasmusmc.nl, proliuhao@gmail.com, molinarim@upmc.edu, d.rizopoulos@erasmusmc.nl

[1] *Department of Biostatistics, Erasmus University Medical Center*
[2] *Starzl Transplant Institute, University of Pittsburgh Medical Center*
[3] *J.C. Walter Jr Transplant Center, Houston Methodist Hospital*

Abstract: Liver transplantation (LT) is the only curative treatment for selected patients with unresectable hepatocellular carcinoma (HCC). Due to organ scarcity, patients must wait for a suitable graft, during which they may become ineligible due to tumour progression. A predictive model identifying patients at highest risk of waitlist dropout and those who would benefit most from LT could improve organ allocation. Transplant-related survival benefit, defined as the additional survival time gained from LT compared to waitlist survival, provides a comprehensive metric to guide allocation. Estimating this causal effect requires addressing observational data and time-varying confounders. We developed a joint model for longitudinal and time-to-event data that predicts individualised transplant-related survival benefit in HCC patients. Unlike the G-formula, structural marginal models or targeted maximum likelihood estimation, our model makes stronger assumptions about the biomarker measurement process but remains non-parametric for competing processes like censoring and visit times. We analysed data from 7,471 HCC patients listed in the US Scientific Registry for Transplant Recipients between 2012 and 2022, of whom 4,786 received a liver. Our model associates pre-transplant trajectories of three well established predictors—serum level of tumour α-fetoprotein level, tumour burden score, and model for end-stage liver disease score—with the risk of death before and after transplantation. Dynamic updates enable real-time refinement of predictions and identification of patients most likely to benefit from transplantation. This model represents an advancement in optimizing liver transplant decisions, improving overall survival for waitlisted HCC patients. The model is available in the R package `JMbayes2`.

Keywords: Conterfactual prediction · Joint model · Multivariate longitudinal data · Shared-parameter model

A Bayesian Approach for Modeling Time-to-Event Distal Outcomes

Leila Denise A F Amorim [1,2][0000-0002-1112-2332], Marcos Aurélio Eustorgio Filho [2,3][0000-0003-4596-5896], and Lilia Carolina C. Costa [1,2][0000-0001-5107-2723]

leiladen@ufba.br, marcoseust@gmail.com, liliacosta@ufba.br

[1] *Department of Statistics, Federal University of Bahia, Brazil*
[2] *Laboratory of Causal Inference and Applications (LInCa), Brazil*
[3] *Master's Program of Mathematics, Federal University of Bahia, Brazil*

Abstract: Models with distal outcomes are utilized to estimate the effects of unobservable (latent) characteristics on observed dependent variables, while also accounting for the influence of additional predictors. Although these methods introduce mathematical complexity, they allow for the capture of multifaceted effects—such as health habits and behaviors—which can be assessed through various indicators. Distal outcome models can be defined as extensions of latent class analysis (LCA), a finite mixture model that identifies and classifies unobservable subgroups (or classes) based on observed response patterns. This paper explores recent advancements in the modeling of categorical latent variables and censored time-to-event outcomes. We examine how a categorical latent variable affects time-to-event occurrences using the Cox proportional hazards model. Additionally, we propose a Bayesian inference approach to jointly estimate the parameters of both latent class analysis (LCA) and the Cox model, employing both one-step and three-step methods. Simulation studies were conducted to assess the properties of two estimators, called Simplified Bayesian Modal (BSM) and Simultaneous Bayesian (BS), both of which were proposed in this work for analyzing distal responses defined by censored failure times. The findings from the simulation studies indicate that the Simultaneous Bayesian Method (BS) significantly reduces bias in estimating the effect associated with latent classes on the distal outcome. Both methods enable the incorporation of additional observed predictors into the model.

Keywords: Bayesian methods · Distal outcomes · Latent class analysis · Survival analysis

Acknowledgements: This research was supported in part by the Research Foundation of the State of Bahia (FAPESB: Fundação de Amparo à Pesquisa do Estado da Bahia, grant term APP0021/2023) and by the National Council for Scientific and Technological Development (CNPq: Conselho Nacional de Desenvolvimento Científico e Tecnológico, grant term 403465/2023-0), both from Brazil. We also acknowledge the support received by M.A.E.F from CNPq.

Unveiling the Operational Reliability of Coastal Tide Gauges: A Comparative Survival Analysis Enhanced with Statistical and Machine Learning Approaches

Dora Carinhas [1,2[0000-0003-4922-9271]], Paulo Infante [2[0000-0002-1644-9502]], and António Martinho [3[0000-0002-1116-5607]]
dora.carinhas@hidrografico.pt; pinfante@uevora.pt
antonio.martinho@marinha.pt

[1] *Instituto Hidrográfico*
[2] *Universidade de Évora, Escola de Ciências e Tecnologia, Centro de Investigação em Matemática e Aplicações*
[3] *CINAV, Escola Naval*

Abstract: The reliability of sea level observation systems is critical for ensuring continuous and accurate data for coastal studies and extreme event monitoring. This work analyses the time between failures of two tide gauge stations (Leixões and Sines), using survival analysis to evaluate and compare their operational performance.
Non-parametric methods (Kaplan-Meier estimator [1]) and classical parametric models (Exponential, Weibull, Log-Normal [2]) were fitted via maximum likelihood estimation. Model performance was assessed using survival curves, diagnostic plots, and the Akaike Information Criterion (AIC). The Weibull distribution provided the best parametric fit in both cases, although Kaplan-Meier remains robust for small sample sizes and irregular patterns.
Reliability indicators such as the Mean Time Between Failures (MTBF) and the Crow-AMSAA model were used to quantify system performance and identify non-homogeneous failure trends. The results highlighted notable operational differences between the stations. To enhance the modelling framework, we explore the potential of semi-parametric models (e.g., Cox proportional hazards) and machine learning techniques, namely random survival forests [3], which accommodate non-linearities and interaction effects in time-to-failure data. These approaches are particularly valuable in environmental systems with complex operational dynamics.
Our findings support the adoption of advanced statistical and machine learning tools in the reliability assessment of environmental monitoring networks.

Keywords: Crow-AMSAA · Kaplan-Meier · Random survival forests · Reliability modelling · Weibull distribution

References

[1] Klein, J. P., Moeschberger, M. L.: Survival Analysis: Techniques for Censored and Truncated Data. Springer (2003).
[2] Meeker, W. Q., Escobar, L. A.: Statistical Methods for Reliability Data. Wiley (1998).
[3] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S.: Random survival forests. The Annals of Applied Statistics, **2**(3), 841–860 (2008). https://doi.org/10.1214/08-AOAS169

Extremos

Weibull Tail Coefficient Estimation via Linear Combinations

M Ivette Gomes [1[0000-0002-2903-6993]], Frederico Caeiro [2[0000-0001-8628-7281]], Fernanda Figueiredo [1,3[0000-0003-0255-4106]], and Lígia Henriques-Rodrigues [4[0000-0003-4881-4188]]
migomes@fc.ul.pt, fac@fct.unl.pt, otilia@fep.up.pt, ligiahr@uevora.pt

[1] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *NOVA School of Science and Technology (NOVA FCT) and CMA, NOVA University of Lisbon, Campus de Caparica, Caparica, Portugal*
[3] *University of Porto, School of Economics and Management, Porto, Portugal*
[4] *School of Science and Technology (ECT-UE) and CIMA, University of Évora, Évora, Portugal*

Abstract: The *Weibull tail-coefficient* (WTC) is the reciprocal of the index of regular variation in a regularly varying cumulative hazard function. The associated right-tails, like the normal *right tail function* (RFT), fall within the max-domain of attraction of an *extreme value* (EV) distribution with a null EV *index* (EVI), the so-called Gumbel RTF, but they exhibit a penultimate (or pre-asymptotic) behaviour closer to a Max-Weibull or to a Fréchet RTF, depending on whether the WTC is smaller or greater than one, respectively. Due to the specific nature of the WTC, and its deep link to a positive EVI, WTC-estimators are closely linked to the classical Hill estimator in [2]. Several generalised means have recently been used with success in the estimation of a positive EVI and of the WTC (see [1], among others). For the estimation of the WTC, we now advance with the use of *asymptotically best linear* (ABL), already considered in the EVI-estimation. The performance of the new ABL WTC-estimators for finite samples is illustrated through Monte-Carlo simulations and real data applications.

Keywords: Extreme value theory · Linear combinations · Weibull tail coefficient

Acknowledgements: Research partially funded by HiTEc Cost Action CA21163 and by national funds through FCT—Fundação para a Ciência e a Tecnologia under the projects: UID/00006/2025 and UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020) (CEAUL); UID/00297/2025 and UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) (CMA); UID/04674/2025 and UIDB/04674/ 2020 (https://doi.org/10.54499/UIDB/04674/2020) (CIMA).

References

[1] Caeiro, F., Henriques-Rodrigues, L., Gomes, M.I.: The use of Generalized Means in the estimation of the Weibull tail coefficient. Computational and Mathematical Methods, **2022**, Article ID 7290822 (2022). DOI:10.1155/2022/7290822
[2] Hill, B.: A simple general approach to inference about the tail of a distribution. Ann. Statist., **3**, 1163-1174 (1975). DOI:10.1214/aos/1176343247

Testing the Domain of Attraction for Maxima

Frederico Caeiro [1[0000-0001-8628-7281]], M Ivette Gomes [2[0000-0002-2903-6993]], Lígia Henriques-Rodrigues [3[0000-0003-4881-4188]], and Cláudia Neves [2,4[0000-0003-1201-5720]]
migomes@fc.ul.pt, fac@fct.unl.pt, ligiahr@uevora.pt, claudia.neves@kcl.ac.uk

[1] *NOVA School of Science and Technology (NOVA FCT) and CMA, NOVA University of Lisbon, Campus de Caparica, Caparica, Portugal*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *School of Science and Technology (ECT-UE) and CIMA, University of Évora, Évora, Portugal*
[4] *King's College London, London, United Kingdom*

Abstract: The extreme value theorem establishes that the limiting distribution for linearly normalised partial maxima of a sequence of i.i.d. random variables is the Generalised Extreme Value distribution. Through the von-Mises parametrisation, this distribution is known to unity the only possible three extreme value distributions arising as limit: Gumbel, Fréchet, and Weibull distributions, of which the Gumbel distribution is especially appealing for the straightforward statistical inference it provides. In this talk, we will address the problem of testing whether the true underpinning distribution F to the sample data belongs to the Gumbel domain of attraction, against an alternative extreme value distribution. A Monte Carlo simulation mechanism is employed to estimate the critical values of the test as well as to assess the power of the tests of hypotheses under study.

Keywords: Extreme value theory · Max-domain of attraction · Statistical hypothesis test

Acknowledgements: Research partially funded by national funds through FCT—Fundação para a Ciência e a Tecnologia under the projects: UID/00006/2025 and UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020) (CEAUL); UID/00297/2025 and UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020), UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (CMA); UID/04674/2025 and UIDB/04674/2020 (https://doi.org/10.54499/UIDB/04674/2020) (CIMA).

References

[1] Neves, C., Fraga Alves, M.I.: Testing Extreme Value Conditions – an Overview and Recent Approaches. REVSTAT-Statistical Journal, 6(1), 83–100 (2008). DOI: https://doi.org/10.57805/revstat.v6i1.59

[2] Tiago de Oliveira, J. and Gomes, M.I. (1984). Two statistics for choice of univariate extreme value models. In Statistical Extremes and Applications (J. Tiago de Oliveira, Ed.), D. Reidel, Dordrecht, pp. 651–668. DOI:https://doi.org/10.1007/978-94-017-3069-3_50

The Extremal Index P-Estimator: a Photovoltaic Energy Data Application

M. Cristina Miranda [1,2,3][0000-0002-4642-5683],
Manuela Souto de Miranda [2][0000-0002-5703-5082], Conceição Amado [4][0000-0001-6664-6486],
M. Ivette Gomes [3][0000-0002-2903-6993],
cristina.miranda@ua.pt, manuela.souto@ua.pt,
conceicao.amado@tecnico.ulisboa.pt, migomes@ciencias.ulisboa.pt

[1] *ISCA, University of Aveiro, Portugal*
[2] *CIDMA, University of Aveiro, Portugal*
[3] *CEAUL, University of Lisbon, Portugal*
[4] *CEMAT, IST, University of Lisbon, Portugal*

Abstract: The extremal index is a parameter that arises in the context of an extreme value distribution. It is associated with the degree degree of dependence of data under certain local dependence conditions. In the limit, extreme values tend to occur in clusters, separated from each other by a reasonable number of lower values. One possible interpretation of the extremal index is the limit of the reciprocal mean cluster size. Highly dependent data tend to exhibit dense clusters and lower values of the extremal index, which can take values between zero and one. Independent or asymptotically independent data will have an extremal index equal to or close to one. It can also be interpreted as the proportion of non-zero inter-exceedances times above some high threshold. This interpretation has led the authors to the recent proposal of a a proportion estimator (P-estimator). In this study, we apply the P-estimator to historical solar energy data and compare its performance with established methods.

Keywords: Clusters of exceedances · Extremal index · P-estimator

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

Deteção de Mudanças de Estruturas em Séries Temporais

Ludomilo Almeida [1[0009-0007-8433-337X]], Dulce Gomes [1,2[0000-0001-8085-1945]] e Lígia Henriques-Rodrigues [1,2[0000-0003-4881-4188]]
m57338@alunos.uevora.pt, dmog@uevora.pt, ligiahr@uevora.pt

[1] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal*
[2] *CIMA – Centro de Investigação em Matemática e Aplicações da Universidade de Évora, Portugal*

Abstract: Mudanças de estrutura em séries temporais referem-se a pontos a partir dos quais a série passa a apresentar diferentes propriedades estatísticas, tais como mudanças de nível, de tendência ou na forma da distribuição. A correta identificação desses pontos pode ser particularmente difícil em situações de mudanças subtis e ainda mais complexa na presença de *outliers*, que dão origem a falsas mudanças. Neste trabalho, propomos efetuar um estudo de simulação de Monte Carlo para avaliar a eficácia dos métodos apresentados na deteção dos pontos de quebra. As abordagem tradicionais em séries temporais consideram que os resíduos são normais ou aproximadamente normais. Com o objetivo de avaliar a robustez e eficácia dos métodos disponíveis na literatura utilizar-se-ão distribuições de cauda pesadas — limitação comum à maioria dos métodos usados na deteção de quebras de estrutura [1, 2] e em séries com forte dependência – em alternativa à normalidade e na presença ou ausência de *outliers*. Para além disso, investigamos ainda a deteção de mudanças na forma da distribuição generalizada de valores extremos [3]. A deteção eficiente de mudanças de estrutura é essencial em áreas como a climatologia, a economia e a saúde pública, permitindo obter melhores previsões e fornecendo ferramentas que auxiliem no processo de tomada de decisão.

Keywords: Distribuição generalizada de valores extremos · Mudança de estrutura · Séries temporais

Acknowledgements: LA é financiado com uma bolsa de investigação da Fundação Calouste Gulbenkian. DG e LHR são financiadas por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, no âmbito dos projeto UIDB/04674/2020, DOI:10.54499/UIDB/04674/2020.

References

[1] Killick, R. and Eckley, I. A. (2014). Changepoint: An R package for changepoint analysis. Journal of statistical software, 58:1–19. DOI:10.18637/jss.v058.i03
[2] Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002): An R package for testing for structural change in linear regression models. Journal of statistical software, 7:1–38. DOI:10.18637/jss.v007.i02
[3] Kojadinovic, I. and Naveau, P. (2017): Detecting distributional changes in samples of independent block maxima using probability weighted moments. Extremes, 20:417–450. DOI:10.1007/s10687-016-0273-1

A novel approach to model long time survival times in highly censored data

Eduardo Janotti Cavalcante [1[0000-0001-9101-4464]], Antonio Carlos Pedroso de Lima [1[0000-0003-0617-328X]], and **Lígia Henriques-Rodrigues** [2[0000-0003-4881-4188]]
eduardjanotti@gmail.com, acarlos@ime.usp.br, ligiahr@uevora.pt

[1] *Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil*
[2] *School of Science and Technology (ECT-UE) and CIMA, University of Évora, Évora, Portugal*

Abstract: Extreme Value Theory (EVT) plays a crucial role in modeling rare events, particularly through the Peaks-Over-Threshold approach and the Generalized Pareto Distribution (GPD). However, traditional implementations face significant challenges when applied to incomplete data (such as censored observations) and long-term survival scenarios, where standard assumptions often fail to hold. Despite recent advances, key issues remain—especially regarding the selection of an appropriate threshold for the validity of the GPD and the integration of modern statistical techniques.

In this talk, we introduce a novel method based on a controlled Gaussian process that mimics the tail behavior of the GPD for large values. Specifically, the proposed process is constructed to satisfy the second Extreme Value Theorem (also known as the Pickands–Balkema–de Haan theorem), which states that, above a high threshold, the distribution of exceedances converges to a member of the GPD family.

Our approach extends both extreme value and survival models, reducing reliance on restrictive assumptions and offering greater flexibility by eliminating the need for explicit threshold estimation. Instead, it models the entire data range within a unified framework.

We will show that the model can naturally estimate the tail index, even in cases where regularity conditions are violated (e.g., when the tail index $\xi \leq -1$). Simulations and applications to highly censored datasets will be presented to highlight the model's practical advantages.

Keywords: Gaussian process · Gerenalized Pareto distribution · Survival analysis

Acknowledgements: EJC is partially financed by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. LHR is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UIDB/04674/2020 and UIDB/04674/2025 (https://doi.org/10.54499/UIDB/04674/2020).

Ciência de Dados II

Synthetic Integer-Valued Times Series Generation: an Experimental Study

Isabel Silva [1[0000-0002-6307-3456]], Isabel Pereira [2[0000-0002-5152-546X]], and Maria Eduarda Silva [3[0000-0003-2972-2050]]
ims@fe.up.pt, isabel.pereira@ua.pt, mesilva@fep.up.pt

[1] *Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, and CIDMA, Portugal*
[2] *Departamento de Matemática, Universidade de Aveiro and CIDMA, Portugal*
[3] *LIADD-INESC TEC, Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-464 Porto, Portugal*

Abstract: The analysis of integer-valued time series and, in particular, count time series has received wide attention in the literature in the past decades. Several models and approaches have been proposed to cater for the characteristics displayed by the data, such as trend, seasonality, over-dispersion, outliers, negative serial correlation, heterocedasticity and censoring, to name just a few. However, the number of count time series available is limited, hindering the effective evaluation of models and approaches. In fact, often the goodness of fit and advantages of a new model are illustrated with one or two time series. In this work, we propose an approach to simulate sets of time series of counts with diverse and controllable characteristics, relying on Mixture INAR (MixINAR) models and a feature space. First, we validate a set of statistical features appropriate to characterize time series of counts. Then we use MixINAR models to simulate sets of time series and compare the synthetic data with benchmarking time series data sets.

Keywords: Count Time Series · Features · Mixture Models

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020, DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).
Furthermore, this work is partially supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT), reference UID/04106/2025.

Modelação de Eventos Raros em Sinistralidade Rodoviária: Comparação de Técnicas de Reamostragem e Algoritmos de Classificação

Lorena Santos [1], Gonçalo Jacinto [1,2[0000-0002-3292-2208]], Anabela Afonso [1,2[0000-0002-5517-4855]] e Paulo Infante [1,2[0000-0002-1644-9502]]
m53481@alunos.uevora.pt, gjcj@uevora.pt, aafonso@uevora.pt, pinfante@uevora.pt

[1] *Universidade de Évora, Escola de Ciências e Tecnologia, Portugal*
[2] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Portugal*

Abstract: Este trabalho analisa o impacto de diferentes técnicas de reamostragem na modelação preditiva de eventos raros em sinistralidade rodoviária grave, utilizando dados reais do distrito de Setúbal entre 2016 e 2023. Dado o forte desbalanceamento da variável dependente (apenas 2,3% dos casos envolvem mortos ou feridos graves), foram comparadas duas metodologias: ROSE, que gera amostras sintéticas com base em suavização de kernel, e SMOTENC, que adapta o algoritmo SMOTE para conjuntos com variáveis categóricas e numéricas. Estas técnicas foram testadas sob diversas estratégias de balanceamento (oversampling total, parcial e combinado) e volumes de dados (10.000, 35.000 e 85.000 observações), avaliando-se cinco algoritmos de classificação: Regressão Logística, Naive Bayes, C5.0, Random Forest e XGBoost, com validação cruzada estratificada e métricas como F1-score e AUC. Os resultados evidenciam que o SMOTENC é mais eficaz na Regressão Logística, enquanto o ROSE potencia o desempenho de modelos baseados em árvores, com destaque para o algoritmo C5.0, que obteve resultados praticamente perfeitos em todos os cenários testados. Verificou-se ainda que o aumento do volume de dados contribui sistematicamente para a melhoria da capacidade preditiva, independentemente da técnica ou modelo utilizado. Estes achados reforçam a importância de uma escolha criteriosa da abordagem de reamostragem em função do algoritmo de classificação, especialmente em contextos de elevada desproporção entre classes, e demonstram o potencial destas metodologias na identificação de padrões críticos em sistemas complexos de segurança rodoviária.

Keywords: Algoritmos de classificação · *Machine learning* · Reamostragem · Sinistralidade rodoviária

Acknowledgements: Este trabalho é parcialmente financiado pelo CIMA (Centro de Investigação em Matemática e Aplicações, Universidade de Évora), financiado pela FCT (Fundação para a Ciência e a Tecnologia), projeto UID/MAT/04674/2020.

Statistical Techniques for Real-Time Digital Twins and the Industrial Metaverse

Cecilia Castro [1][0000-0001-9897-8186]
cecilia@math.uminho.pt

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

Abstract: Digital twins and industrial metaverse workspaces integrate physics-based and data-driven models with heterogeneous, high-frequency telemetry to support low-latency decisions. This talk explains where statistics fits in that pipeline and how it operates in practice. First, we frame state–space modeling inside a digital twin: sequential estimators—Kalman Filter (KF), Extended Kalman Filter (EKF), and Particle Filters (PF)—produce latent state estimates *with uncertainty* from streaming measurements; we outline when each is appropriate and how uncertainty drives downstream actions. Second, we cover online change detection by monitoring innovations/residuals, using CUSUM with thresholds calibrated by Average Run Length (ARL) and, when closed-form calibration is unavailable, an online residual bootstrap to adapt thresholds while preserving low false-alarm rates. Third, we connect online Maximum Likelihood (ML) to the twin: Recursive Least Squares (RLS) updates linear parameters, and online Expectation–Maximization (EM) leverages filter-based expectations to update latent-parameter models—keeping the twin calibrated during operation and handing consistent estimates to decision layers (including Model Predictive Control, MPC). Brief case vignettes—such as BMW Group's virtual factory in NVIDIA Omniverse and power-generation digital twins—serve as engineering blueprints linking sensor streams and interaction logs to estimation, detection, and control. The goal is to show, concretely, how statistical methods enable trustworthy, real-time inference in digital twins and metaverse-facing applications.

Keywords: Digital Twins · Industrial Metaverse · Kalman Filtering · Real-Time Inference · Sequential Change Detection

Acknowledgements: This work was partially supported by the Centro de Matemática da Universidade do Minho (UID/00013).

References

[1] Doucet, A., de Freitas, N., Gordon, N. (eds.): *Sequential Monte Carlo in Practice*. Springer (2001). https://doi.org/10.1007/978-1-4757-3437-9

[2] Särkkä, S., Svensson, L.: *Bayesian Filtering and Smoothing* (2nd ed.). Cambridge University Press (2023). https://doi.org/10.1017/9781108917407

[3] Cheng, J., Chen, H., Xue, Z., Huang, Y., Zhang, Y.: An online exploratory maximum likelihood estimation approach to adaptive Kalman filtering. IEEE/CAA Journal of Automatica Sinica **12**(1), 228–254 (2025). 10.1109/JAS.2024.125001

[4] Castro, C., Leiva, V., Basso, F.: A Data-Driven Systematic Review of the Metaverse in Transportation: Current Research, Computational Modeling, and Future Trends. Computer Modeling in Engineering & Sciences **144**(2), 1481–1543 (2025). https://doi.org/10.32604/cmes.2025.067992

Estatística Espacial II

Imputation for Net Income on Rotating Panel Data: an Approach with Conditional Autoregressive (CAR) Models on Portuguese Labor Survey

Antonio Loría-García [1[0000-0002-5523-0941]], Lígia Henriques-Rodrigues [1[0000-0003-4881-4188]], and Pedro Campos [2[0000-0001-5495-9434]]

antonio.loria@uevora.pt, ligiahr@uevora.pt, pcampos@fep.up.pt

[1] *CIMA –Centro de Investigação em Matemática e Aplicações, Universidade de Évora, Portugal*
[2] *LIAAD-INESC TEC, School of Economics and Management, Universidade do Porto*

Abstract: Net income is a key variable of labor force surveys. When sampling includes a rotating panel, net income has considerable high rates of missing values and attrition, according to the number of rotations of the sample. This work presents an application of imputation through spatial statistics with conditional autoregressive models (CAR) for net income on rotating panel data. CAR models try to capture spatial dependence incorporating a neighborhood structure based on adjacency or distance-based weights. The results are presented for quarterly data from the Portuguese Labor Survey of 2023-2024, with modeling using covariates of age, sex, academic degree, months of employment and estimates of spatial dependence.

Keywords: Attrition · Conditional autoregressive models · Imputation · Net income · Rotating panel data

Acknowledgements: This work is partially financed by Centro de Investigação em Matemática e Aplicações (CIMA), Universidade de Evora, Portugal, under reference CIMA/BD1/2023, projeto UIDP/04674/2020, Financiamento Plurianual 2020-2023 do CIMA, Fundação para a Ciência e a Tecnologia (FCT/MCTES) and the mobility program for postgraduate studies of Universidad de Costa Rica (UCR).

References

[1] Besag, J.: Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological), **36. Vol.**(2.), 192–236 (1974).
[2] Bivand, R., Pebesma, E., Gómez-Rubio V.: Applied Spatial Data Analysis with R. Springer New York, NY (2013). DOI:https://doi.org/10.1007/978-1-4614-7618-4

Visual Spatial Learning: Single-Field Spatial Interpolation Using Convolutional Neural Networks

Daniel Tinoco [1,2], Raquel Menezes [1[0000-0001-5552-917X]], and Carlos Baquero [2[0000-0002-3933-6850]]
danieltinoco@fe.up.pt, rmenezes@math.uminho.pt, cbm@fe.up.pt

[1] *Centre of Mathematics (CMAT), University of Minho, Guimarães, Portugal*
[2] *DEI-FEUP & INESC TEC, University of Porto, Porto, Portugal*

Abstract: Predicting a complete spatially correlated field from sparse observations is a fundamental challenge in spatial statistics and environmental modeling. Classical interpolation methods such as Kriging rely on Gaussian process assumptions and variography, which can limit their effectiveness in non-stationary settings and require substantial domain expertise. In this work, we leverage an architecture based on convolutional neural networks (CNNs) for spatial interpolation that is trained and applied to a single partially observed field, without access to external data or prior fields. The model is supervised directly on the observed locations and learns to predict values at unobserved points on the user defined grid. Unlike Kriging, our method does not require explicit covariance modeling or variogram estimation, and it can flexibly capture local spatial patterns in a data-driven manner. We evaluated the approach on both synthetic and real-world spatial datasets, showing that it achieves competitive or improved performance relative to Kriging, particularly in cases with non-stationary or complex spatial structures. This work demonstrates the potential of CNNs for single instance spatial interpolation under sparse supervision, offering a practical alternative to classical geostatistical methods.

Keywords: Convolutional neural networks · Single-field learning · Sparse supervision · Spatial interpolation · Spatial statistics

Acknowledgements: The first and second authors received support from the Portuguese Foundation for Science and Technology (FCT), provided through UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM). The first author also received additional support from FCT through the Individual PhD Scholarship 2024.06508.BDANA. The third author received support from national funds through FCT, under the reference UIDB/50014/2020.

A Zero-inflated Spatio-temporal Approach for Joint Modeling of Fishery-depedent and Fishery-independent Data to Understand Fish Distribution

Daniela Silva [1,2][0000-0002-5597-2822], Raquel Menezes [2], Gonçalo Araújo [3], Ana Teles-Machado [4], Renato Rosa [5], Ana Moreno [1], Alexandra Silva [1], and Susana Garrido [1]

daniela.dasilva@ipmat.pt, rmenezes@math.uminho.pt, goncalo.araujo@novasbe.pt, ammachado@fc.ul.pt, renato.rosa@fe.uc.pt, amoreno@ipma.pt, asilva@ipma.pt, susana.garrido@ipma.pt

[1] *Portuguese Institute for the Sea and Atmosphere (IPMA)*
[2] *Centre of Mathematics, University of Minho*
[3] *Nova School of Business and Economics, Nova University Lisbon*
[4] *Instituto Dom Luiz (IDL), University of Lisbon*
[5] *Centre for Business and Economics Research, University of Coimbra*

Abstract: Effective management of marine ecosystems relies on statistical tools to accurately characterize species distributions, essential for sustainable fisheries. We present a novel six-layer joint spatio-temporal model that integrates fishery-independent (FID) and fishery-dependent data (FDD) while addressing the specific features of each source, including preferential sampling (PS) in FDD. The model accommodates zero-inflated (ZI) data, decouples presence-absence and biomass processes, and incorporates spatial and temporal dependencies. It also adjusts for varying measurement processes across data types through a catchability component, capturing vessel-specific efficiency.

Model inference is based on stochastic partial differential equations (SPDE) and Laplace approximation, enabling computationally efficient estimation in complex settings. Simulation studies demonstrate the model's robustness across a range of PS scenarios and sample configurations, confirming its capacity to recover true parameters and detect PS patterns.

We apply the model to European sardine (*Sardina pilchardus*) off the Portuguese coast (2013–2018), combining FID and FDD with environmental covariates. Results reveal biomass hotspots not identified by single-source models and elucidate ecological relationships, such as the effects of sea surface temperature and chlorophyll-a.

This work advances spatial statistics by addressing challenges inherent to integrating heterogeneous ecological data. The proposed framework is scalable and broadly applicable to other domains (e.g., epidemiology and public health) where PS and ZI data are prevalent. Our approach underscores the value of methodological innovation in tackling real-world problems through interdisciplinary modeling.

Keywords: Fish data · Integrating data sources · Preferential sampling · Spatio-temporal modeling · Species distribution model

Acknowledgements: This study received support from the SARDINHA2030 project (MAR-111.4.1-FEAMPA-00001), and the project UID/00013 through the Centre for Mathematics of University of Minho (CMAT/UM).

Bioestatística e Epidemiologia I

Integrating Supervised and Unsupervised Learning for Variant Post-Filtering in Whole-Genome Sequencing

Vera Pinto [1,2][0000-0002-0724-5612], Lisete Sousa [1,2][0000-0002-2114-720X], and Carina Silva [1,3][0000-0003-1021-7935]
vgpinto@ciencias.ulisboa.pt, lmsousa@ciencias.ulisboa.pt, carina.silva@estesl.ipl.pt

[1] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa*
[2] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa.*
[3] *ESTeSL—Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa 1990-096 Lisboa*

Abstract: Accurate distinction between true and false variant calls is a critical step in whole-genome sequencing (WGS) analysis pipelines. Traditional approaches, such as GATK's Variant Quality Score Recalibration (VQSR), rely on Gaussian Mixture Models (GMM) and predefined thresholds. These methods often assume that variant quality features follow specific distributions and may not generalize well to datasets with different sequencing properties or coverage profiles, limiting their adaptability to different scenarios. In this work, we present a unified and automated post-filtering framework that combines six models, both supervised and unsupervised, to classify variants based on features extracted from variant call format files (VCF) as annotations. These include GMM, Bayesian Gaussian Mixtures (BGM), logistic regression (LR), random forests (RF), and LightGBM (LGBM), the latter also tuned via Bayesian optimization. All models were trained and tested using the NA12878 truth set, using chromosome 20 and WGS. The results show that tree-based methods outperform both probabilistic and linear models in classifying true versus false variant calls. Unsupervised models, GM and BGM, still offered competitive performance. This study introduces an automated pipeline for variant post-filtering that allows flexible trade-offs between precision and sensitivity.

Keywords: Bayesian optimization · Gaussian Mixture Models · Supervised and unsupervised learning · Variant filtering · Whole-genome sequencing

Acknowledgements: This research was funded by FCT – Fundação para a Ciência e a Tecnologia through the Ph.D. grant UI/BD/153743/2022 DOI:10.54499/UI/BD/153743/2022., and was also partially financed by CEAUL through FCT under the project UID/00006/2025. DOI:10.54499/UIDB/00006/2020.

Cutoff-Free Sero-Epidemiological Analysis of Infectious Diseases

Nuno Sepúlveda [1,2][0000-0002-8542-1706]
nuno.sepulveda@pw.edu.pl

[1] *Faculty of Mathematics & Information Science, Warsaw University of Technology, Poland*
[2] *CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal*

Abstract: Sero-epidemiological studies of infectious diseases often aim to estimate seroprevalence and fit reversible catalytic models to reconstruct the transmission history of a pathogen within a population. Traditionally, these analyses rely on binary serological data, classifying individuals as seropositive or seronegative based on an estimated or predefined cutoff in the antibody level distribution. This approach assumes perfect classification accuracy—an assumption that rarely holds true. In reality, the use of a cutoff leads to misclassification, akin to diagnosing disease with an imperfect diagnostic test, introducing bias into both seroprevalence estimates and model-based inferences. In this talk, I argue that the cutoff-based approach is unnecessary and potentially misleading. I propose an alternative method based on fitting a two-component mixture model to the continuous antibody measurements, with the mixing weight corresponding to the seropositive group interpreted as the seroprevalence. This model-based framework allows for the direct estimation of seroprevalence without dichotomizing the data and enables downstream modeling, such as fitting reversible catalytic models, in a more statistically principled way. I will illustrate the proposed methodology using data from a sero-epidemiological study of malaria.

Keywords: Cutoff-free approach · Epidemiology · Finite mixture models · Reversible catalytic model · Seroprevalence

Acknowledgements: The author is funded by the European Union (Horizon Europe, project reference 101057665). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or HADEA. Neither the European Union nor the granting authority can be held responsible for them.

Comparative Study of the Most Used Growth Models Applied to Weight in Infants Aged 0 to 2 Years

Marta Alves [1,2,3][0000-0002-7421-8550], Marisol Garzón [4][0000-0001-7793-6948], Bruno Heleno [3][0000-0002-3943-1858], Ana Luísa Papoila [1,2,3][0000-0002-2918-8364] e Carlos Brás-Geraldes [1,5][0000-0002-1551-6531]

marta.alves@ulssjose.min-saude.pt, garzon.marisol1@gmail.com, bruno.heleno@nms.unl.pt, ana.papoila@nms.unl.pt, carlos.geraldes@isel.pt

[1] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *Centro de Investigação - Gabinete de Análise Epidemiológica e Estatística, Unidade Local de Saúde São José; Centro Clínico Académico de Lisboa, Portugal*
[3] *NOVA Medical School|Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Portugal*
[4] *Pediatra no Hospital Fernando Fonseca*
[5] *Instituto Superior de Engenharia de Lisboa do Instituto Politécnico de Lisboa, Portugal*

Abstract: Monitoring a child's growth through routine assessments over time provides valuable insights into their overall health. In child development, growth is widely recognized as a key indicator for the early diagnosis of potential diseases and in the nutritional status evaluation. Over the years, several models have been proposed to analyze growth patterns. A scoping review conducted in this study—focused on infants aged 0 to 2 years—identified the most commonly used models: Jenss-Bayley, Count, Reed 1st order, fractional polynomials, and SuperImposition by Translation and Rotation (SITAR). These models were compared with the Generalized Additive Models for Location, Scale, and Shape (GAMLSS), which are typically used by the World Health Organization for modeling growth data. For this purpose, data of 414 infants aged 0 to 2 years, collected as part of a birth cohort study conducted in São Tomé Island (March 2013 to July 2015) were analyzed. Model fit and complexity were assessed using Akaike Information Criterion and Bayesian Information Criterion. Predictive accuracy was evaluated using the mean absolute error, the mean squared error, and the normalized mean squared error. Finally, the Wasserstein distance was applied to compare the distribution of estimated weights from each model to the observed weight distribution. Results showed that SITAR provided the best goodness-of-fit, while GAMLSS obtained the best predictive accuracy for both sexes. Regarding the Wasserstein distance, SITAR again performed best in approximating the distribution of observed weights, for both sexes.

Keywords: Body weight · GAMLSS · Growth models · SITAR

Stratification in ME/CFS: Association Between Domain-Specific Severity Profiles and Herpesvirus Antibody Responses

João Malato [1,2][0000-0003-4389-1483], Luís Graça [1][0000-0001-6935-8500], Ji-Sook Lee [3][0000-0003-1747-9700], Jacqueline M. Cliff [3][0000-0002-5653-1818], Luis Nacul [4,5][0000-0003-1411-8088], Eliana M. Lacerda [4][0000-0002-5077-7868], and Nuno Sepúlveda [2,6][0000-0002-8542-1706]

joao.malato@gimm.pt, luis.graca@gimm.pt, sook.lee@brunel.ac.uk, jacqueline.cliff@brunel.ac.uk, luis.nacul@lshtm.ac.uk, eliana.lacerda@lshtm.ac.uk, nuno.sepulveda@pw.edu.pl

[1] *GIMM, Faculdade de Medicina, Universidade de Lisboa, Portugal*
[2] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *CIRTM, Brunel University of London, United Kingdom*
[4] *LSHTM, University of London, United Kingdom*
[5] *CCDP, BC Women's Health Research Institute, British Columbia, Canada*
[6] *Politechnika Warszawska, Poland*

Abstract: Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a chronic, debilitating condition with unknown aetiology and pathophysiology. Diagnosis relies primarily on symptom assessment and exclusion of other fatigue-inducing illnesses. The absence of standardised criteria and natural symptom variability reported by patients result in a heterogeneous diagnosed population. Stratifying suspected cases into subgroups could help identify more homogeneous profiles for research and treatment.

We analysed data from the UK ME/CFS Biobank to study the association between symptom clusters and herpesvirus antibody responses, enhancing our understanding of how related symptoms could explain the viral origin of the disease. 47 symptoms in 241 ME/CFS patients and 106 healthy controls were used to measure intra- and inter-rater agreement, estimating individual entropy and pairwise Cohen's κ coefficients, respectively. Latent class analysis identified severity-based subgroups across seven domains (immunological, neuroendocrine, PEM, autonomic, neurocognitive, neurophysiological, and pain), with optimal class number selected using the Akaike information criterion. Lastly, plasma IgG concentration and seropositivity for six herpesviruses were compared among clusters using non-parametric tests, with multiple-testing correction.

We found significant associations between severity-based subgroups from clinical domains and herpesvirus IgG levels. Notably, higher HSV-1 antibody titres were observed in more severe autonomic and neurocognitive subgroups compared to milder groups and controls. Similar trends were found for HSV-2 and EBV for other domains. These findings support domain-specific stratification as a means to reduce phenotypic noise in ME/CFS research and suggest possible links between symptom severity and latent herpesvirus activity, knowledge that may be used for targeted treatments and biomarker discovery.

Keywords: Herpesvirus serology · Latent class analysis · ME/CFS · Patient stratification · Symptom domains

Estatística Espacial III

Joint Modeling of Spatial Intensity, Detectability and Marks in Caribou Surveys using `inlabru`

Iúri J. F. Correia [1,2][0000-0002-9718-0014], Soraia A. Pereira [1,3][0000-0002-7336-1320], Tiago A. Marques [1,4,5][0000-0002-2581-1972], Christine Cuyler [6][0000-0002-2820-8749], and Marta M. Rufino [1,7][0000-0002-0734-7491]
ijcorreia@fc.ul.pt

[1] *Centro de Estatística e Aplicações – Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa (FCUL), Lisboa, Portugal*
[2] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal*
[3] *Centro de Matemática, Universidade do Minho, Guimarães, Portugal*
[4] *Centre for Research into Ecological and Environmental Modelling, University of St Andrews, Scotland*
[5] *Departamento de Biologia Animal, FCUL, Portugal*
[6] *Greenland Institute of Natural Resources, P.O. Box 570, Nuuk, Greenland*
[7] *Instituto Português do Mar e da Atmosfera, 1495-165 Lisboa, Portugal*

Abstract: Spatial point process models for ecological survey data must jointly address imperfect detection, marked observations, and underlying spatial structure. In this talk, we present a Bayesian marked point-process framework, implemented in `inlabru`, that simultaneously models spatial intensity, detectability, and marks. We represent observed caribou group locations as a log-Gaussian Cox process (LGCP) whose intensity is thinned by a detection-probability function. Group size is treated as a stochastic mark associated with each detected group. Spatial heterogeneity in abundance and detectability is captured by a latent Gaussian random field defined over a triangular mesh and approximated via the Stochastic Partial Differential Equation approach (SPDE). Model fitting proceeds via Integrated Nested Laplace Approximations (INLA) within the `inlabru` interface, yielding fast and accurate posterior inference for both fixed-effect covariates and hyperparameters governing spatial smoothness. We illustrate this methodology with aerial wildlife survey data, demonstrating improved estimation of spatial abundance and detection surfaces relative to traditional Distance Sampling (DS) and Density Surface Modelling (DSM). Estimates from the joint model ($\widehat{CV} = 0.035$), compared with those from DS ($\widehat{CV} = 0.086$) and DSM ($\widehat{CV} = 0.037$), achieved lower uncertainty overall.

Keywords: Detection function · Inlabru · Log-Gaussian Cox process · Marked spatial point process · SPDE

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UI/BD/152236/2021 (DOI: 10.54499/UI/BD/152236/2021) and UID/00006/2025 (DOI: 10.54499/UID/00006).

Modelling Street Crime in Almada, Portugal, Using Point Processes

Inês Oliveira [1[0009-0000-0035-0738]], Paula Simões [2,3[0000-0002-1074-1297]], and Isabel Natário [1,3[0000-0001-6020-9373]]

ild.oliveira@campus.fct.unl.pt, pc.simoes@fct.unl.pt, icn@fct.unl.pt

[1] *Nova School of Science and Technology, NOVA University of Lisbon, Caparica, Portugal*
[2] *Military Academy Research Center, Military University Institute (CINAMIL), Lisbon, Portugal*
[3] *Center for Mathematics and Applications (NOVA MATH), NOVA University of Lisbon, Caparica, Portugal*

Abstract: Crime exhibits a pronounced spatial concentration, with a small proportion of locations, designated by "hotspots", accounting for a significant share of criminal events, as established by Weisburd's law of crime concentration. This universal phenomenon underscores the potential of spatial and spatio-temporal analysis in informing effective crime prevention strategies. In Portugal, the Guarda Nacional Republicana (GNR) recognises the relevance of intelligence-led policing to optimize patrol allocation and enhance crime prevention, and is giving the first steps in the application of advanced statistical methods to Portuguese crime data.
This study focuses on crime records from Almada, a municipality in the Lisboa e Vale do Tejo region, which has a crime rate above the national average. Using data from 2022–2023, provided by the GNR's information division, we aim to investigate the spatial distribution of crime and identify patterns that can inform policing strategies. The research adopts a spatial point pattern analysis framework, employing a spatio-temporal Log-Gaussian Cox Process (LGCP) model. This approach is particularly suitable for clustered point patterns, as it may account for stochastic dependencies, temporal trends, and spatially varying socio-economic and environment factors.
The methodology involves Bayesian modelling with inference conducted using the Integrated nested Laplace approximation (INLA) method with the Stochastic Partial Differential Equation (SPDE) approach. Key objectives include estimating spatial dependencies between crime events, assessing the influence of socio-economic covariates, and identifying temporal dynamics within crime hotspots.
Results will be interpreted through the lens of criminological theory to provide actionable insights for evidence-based policing, contributing to organisational performance by increasing the efficiency of the GNR's patrolling activities on national territory.

Keywords: Crime data · INLA · Log-Gaussian Cox processes · Spatio-temporal statistics

Acknowledgements: This work was supported by national found through FCT - Fundação para a Ciência e a Tecnologia, I.P., under projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications). We thank the GNR data

Sobre a Validação Cruzada em Dados Dependentes

Isabel Natário [1,2][0000-0001-6020-9373], Ricardo Coelho [1][0000-0002-8791-2840]
icn@fct.unl.pt, rpe.coelho@campus.fct.unl.pt

[1] *NOVA Math - Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal*
[2] *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal*

Abstract: Na modelação estatística, a capacidade preditiva do modelo é frequentemente avaliada pelos chamados métodos de validação cruzada, que usam (parte) dos próprios dados para confirmarem a referida qualidade, na inviabilidade de se obterem mais dados para tal. Neste processo, o modelo é estimado com parte da amostra (grupo de treino) e a sua capacidade preditiva avaliada nos restantes elementos da amostra (grupo de teste ou de validação). Destacam-se nestes métodos o chamado procedimento de validação cruzada que deixa um elemento de fora de cada vez ou o procedimento de validação cruzada que deixa de fora um de k grupos de elementos de cada vez. Estes procedimentos assentam na hipótese de que há independência entre os elementos no grupo de treino e os do grupo de teste, o que não acontece em dados de natureza espacial ou temporal, em que os dados tendem a agrupar-se, sobre-representando alguns locais/períodos de tempo e sub-representando outros. Assim, pode acontecer que o grupo de validação não seja representativo do grupo de teste, enviesando o resultado. Ou ainda, caso os grupos de validação e teste resultantes da partição aleatória sejam muito próximos, tal pode resultar em modelos sobre-ajustados, com resultados excessivamente optimísticos. Neste trabalho apresenta-se um estudo de simulação para ilustrar a magnitude do problema e elencam-se algumas estratégias possíveis para ultrapassar a questão, possibilitando a adequada avaliação da capacidade preditiva de modelos para dados de natureza dependente.

Keywords: Dados dependentes · Espaço · Tempo · Validação cruzada

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projeto UIDB/00297/2020, DOI: 10.54499/UIDB/00297/2020, e UIDP/00297/2020, DOI: 10.54499/UIDP/00297/2020 (Centro de Matemática e Aplicações)

References

[1] Adin, A., Krainski, E.T., Lenzi, A., Liu, Z., Martínez-Minaya, J., Rue, H.: Automatic cross-validation in structured models: Is it time to leave out leave-one-out?. Spatial Statistics **62**, 100843 (2024). DOI:https://doi.org/10.1016/j.spasta.2024.100843

[2] Wang, Y., Khodadadzadeh, M., Zurita-Milla, R.: Spatial+: A new cross-validation method to evaluate geospatial machine learning models. International Journal of Applied Earth Observation and Geoinformation **121**, 103364 (2023). DOI:https://doi.org/10.1016/j.jag.2023.103364

Modeling the Spatial Distribution of Dinosaur Fossil Records

Carolina S. Marques [1[0000-0002-5936-9342]], Soraia Pereira [1,2[0000-0002-7336-1320]], Emmanuel Dufourq [3[0000-0002-6986-3408]], Elisabete Malafaia [4,5,6], Pedro Mocho [4,5,6,7[0000-0002-3348-5572]], Joana Órfão [4,5,6[0000-0002-9732-4911]], and Vanda F. Santos [4,6,8[0000-0001-5842-7857]]

csmarques@fc.ul.pt,soraia.pereira@math.uminho.pt,emmanuel@cs.uni.edu, efmalafaia@fc.ul.pt,pdmocho@fc.ul.pt,orfao.jo@gmail.com,vafsantos@fc.ul.pt

[1] *CEAUL, FCUL, PT*
[2] *CMAT, PT; Departamento de Matemática, Universidade Minho, PT*
[3] *Computer Science, University of Northern Iowa, Iowa, USA*
[4] *IDL, FCUL, PT*
[5] *Grupo de Biología Evolutiva, UNED, ES*
[6] *Departamento de Geologia, FCUL, PT*
[7] *Dinosaur Institute, Natural History Museum, US*
[8] *PaleoIbérica Research Group, University of Alcalá, ES*

Abstract: Predicting the spatial distribution of dinosaur fossil records presents significant challenges due to different aspects, such as geological heterogeneity, variable sampling effort, and spatial autocorrelation. In this study, we integrate machine learning (ML) classifiers and a Bayesian spatial point process model to address these issues. We obtained records of dinosaur fossils from the Paleobiology Database and processed them to classify observations into three categories: eggs, bones, and footprints. Pseudo-absence points were generated to train ML models. We paired the resulting data with different covariates such as lithology, slope, and impermeability. We first applied ML algorithms (e.g., Random Forest (RF), Gradient Boosting (GB), Neural Network). RF and GB consistently achieved superior classification accuracy, capturing complex nonlinear covariate effects. In parallel, we fitted a Bayesian Log-Gaussian Cox Process (LGCP) to explicitly model observation uncertainty and spatial dependence, presenting full posterior intensity surfaces with credible intervals. Comparative analysis shows that ML classifiers provide higher raw predictive performance but lack direct quantification of spatial uncertainty, whereas the LGCP offers principled inference on spatial structure and credible uncertainty bounds for predicted intensities. We therefore propose a hybrid workflow: (1) use RF for initial variable selection and preliminary suitability mapping, and (2) apply the Bayesian LGCP to produce final intensity surfaces with uncertainty quantification. This integrated strategy helps create a suitability map for each type of fossil record, aiding paleontologists in discovering new fossil sites and thereby contributing to advances in the field.

Keywords: Dinosaur fossil record · Machine Learning · Presence-only data · Spatial distribution · Spatial point process model

Acknowledgements: The authors thank Afonso Barrocal, Afonso Ferreira and Alexandre Fonseca for the initial discussion about the study. This work was supported by Portuguese government funds through FCT (UI/BD/154258/2022; UI/BD /151441/2021; CEECIND/01770/2018; UID/00006/2025; UIDB/00006/2020; UID/ 50019/2025; UIDB/50019/2020; LA/P/0068/2020, CEECIND/00726/2017). ED was funded by the Carnegie Corporation of New York.

Aplicações em Econometria, Finanças e Gestão

Stochastic Differential Equation Harvesting Models and Effects of Parameter Estimation Errors

Carlos A. Braumann [1,2][0000-0003-2721-9750] and Nuno M. Brites [3][0000-0002-5719-6310]
braumann@uevora.pt, nbrites@iseg.ulisboa.pt

[1] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Portugal*
[2] *Universidade de Évora, Escola de Ciências e Tecnologia, Portugal*
[3] *ISEG/UL – Universidade de Lisboa, Department of Mathematics; REM – Research in Economics and Mathematics, CEMAPRE, Portugal*

Abstract: A harvested population in a random environment is modeled using a stochastic differential equation (SDE). We consider the optimal variable effort harvesting policy (based on stochastic optimal control), which has serious implementation problems, and the sub-optimal constant effort policy. For such policies, we study the consequences of parameter estimation errors on the harvesting efforts and on the prediction accuracy and profit losses of the corresponding harvesting profits. Using population and economic data from a real fishery for illustration, the sensitivity to parameter estimation errors of the estimated efforts and of the predicted and the real harvesting profits will highlight which parameters will most benefit from investments on estimation accuracy.

Keywords: Effects on efforts and profits · Harvesting model · Parameter estimation errors · Stochastic differential equations · Variable and constant effort policies

Acknowledgements: C.A. Braumann is a member of the Centro de Investigação em Matemática e Aplicações, supported by Fundação para a Ciência e a Tecnologia - FCT (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.

Predicting Daily Euro-Dollar Exchange Rate with SARIMA, LSTM and Decomposition-based models

Vasco Carneiro [1][0009-0005-9341-6084]
vascocarneiro@tecnico.ulisboa.pt

[1] *Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Abstract: The study investigates the task of forecasting the daily Euro-Dollar exchange rate using both traditional statistical models and deep learning techniques. Specifically, it compares the performance of Seasonal AutoRegressive Integrated Moving Average (SARIMA) models and Long Short-Term Memory (LSTM) networks. In addition to direct forecasting approaches, the study explores decomposition-based methods, such as Singular Spectrum Analysis (SSA) and Seasonal- Trend Decomposition using Loess (STL), where the time series is separated into components—trend, seasonality, and residuals—each modeled individually before recombination. Three incrementally constructed datasets are employed: the first contains only the exchange rate; the second incorporates macroeconomic indicators, such as GDP and CPI; and the third uses Principal Component Analysis (PCA) for dimensionality reduction, following feature engineering. The study prioritizes directional prediction over point forecasting, focusing on whether the exchange rate will increase or decrease on the following day. The analyzed period spans from April 12, 1999, to August 1, 2022. Findings show that LSTM models outperform SARIMA across all datasets. The dataset featuring macroeconomic indicators yielded the best performance in the direct modeling approach. Consequently, it was selected for the STL decomposition experiments, allowing LSTM models to capture underlying patterns more effectively and further improve forecasting results. The best-performing models were then applied to more recent data, the 2022–2024 period, achieving a directional accuracy of 57.01% in predicting whether the exchange rate would increase or decrease on the following day, demonstrating the value of decomposition-based deep learning approaches in financial time series forecasting.

Keywords: Euro-Dollar Exchange Rate Forecast · LSTM · STL Decomposition

References

[1] Ghalayini, Latife: Modeling and forecasting the US dollar/euro exchange rate. International Journal of Economics and Finance, **Vol. 6**(num. 1), pages (194–207) (2014). https://epe.bac-lac.gc.ca/100/201/300/intl_journal_economics_finance/2014/IJEF-V6N1-ALL.pdf#page=197

[2] Yıldırım, Deniz Can and Toroslu, Ismail Hakkı and Fiore, Ugo: Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. Financial innovation, **Vol. 7**, pages (1–36) (2021). https://link.springer.com/content/pdf/10.1186/s40854-020-00220-2.pdf

Distinguishing Repeat and First-Time Hotel Guests: A Comparative Analysis of Classification Models

Sílvia Maria Dias Pedro Rebouças [1,2,3][0000-0002-8475-9748], Inês Eusébio Gonçalves [1], Conceição Ribeiro [4,5][0000-0003-0185-3200], Tiago Miguel Pereira Candeias [1,2][0000-0002-5254-6262] and Miguel Nuno Portugal [1,6][0000-0001-9222-5490]
p5974@ismat.pt, inesgoncalves_ig@hotmail.com, cribeiro@ualg.pt, p4200@ismat.pt, p58169@ismat.pt

[1] *ISMAT – Instituto Superior Manuel Teixeira Gomes,*
[2] *COPELABS – Cognitive and People-centric Computing R&D Unit*
[3] *ESGHT/UAlg – Escola Superior de Gestão, Hotelaria e Turismo/Universidade do Algarve*
[4] *ISE/UAlg - Instituto Superior de Engenharia/Universidade do Algarve*
[5] *CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa*
[6] *CiTUR – Centre for Tourism Research, Development and Innovation*

Abstract: Understanding guest recurrence is essential for strategic decision-making in the hospitality industry, particularly in high-end establishments. This study aims to distinguish repeat hotel guests from first-time visitors through descriptive statistical analysis and the application of supervised machine learning techniques. Data were collected from a luxury hotel with Michelin-starred restaurant services, covering all 4,370 reservations recorded between 2018 and 2023. Descriptive statistics were used to characterise booking patterns, demographic profiles, and spending behaviours, allowing for an initial comparison between guest types. Subsequently, several classification models were developed and evaluated to predict guest recurrence based on variables available at the time of booking. The models tested included Logistic Regression, Decision Trees, Random Forest, k-Nearest Neighbours, Support Vector Machines, Neural Networks, XGBoost, and LightGBM. Performance metrics such as accuracy, sensitivity, specificity and AUC were calculated using cross-validation and separate test sets. Random Forest and LightGBM achieved the best performance, with AUC values above 0.84. Variables with the highest predictive relevance included guest nationality and room category, followed by expenditures on food and lodging. The results offer a valuable statistical foundation for implementing targeted marketing, improving guest retention strategies, and supporting revenue management decisions in the luxury hospitality sector.

Keywords: Hospitality Management · Hotel Guest Classification · Machine Learning · Predictive Modeling · Repeat Customers

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025 and UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020.

Aplicação de Redes Neuronais Artificiais à Previsão de Preços de Criptomoedas

José Cruz [1][0000-0002-5894-2733] e Tiago A. Marques [2,3,4][0000-0002-2581-1972]
jose.cruz@universidadeeuropeia.pt, tiago.marques@st-andrews.ac.uk

[1] *Universidade Europeia*
[2] *Centro de Estatística e Aplicações da Universidade de Lisboa*
[3] *Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland*
[4] *Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Portugal*

Abstract: As Redes Neuronais Artificiais (RNA) têm emergido como ferramentas promissoras na modelação de séries temporais financeiras, destacando-se pela sua capacidade de capturar relações não lineares complexas e de processar grandes volumes de dados sem pressupostos estritos de linearidade. Este estudo propõe a aplicação de RNA à previsão de preços de criptomoedas, com ênfase no Bitcoin. A análise será baseada em dados diários históricos, complementados por variáveis macroeconómicas e indicadores financeiros, permitindo testar diferentes arquiteturas de redes neuronais. Os resultados obtidos serão comparados com os de modelos econométricos tradicionais, como ARIMA e GARCH. A avaliação do desempenho preditivo será efetuada com recurso a métricas como o Root Mean Squared Error (RMSE) e Mean Absolute Error (MAE). Espera-se que as RNA apresentem vantagens face aos modelos convencionais, especialmente em contextos de elevada volatilidade e não linearidade. Este trabalho visa contribuir para a compreensão do potencial das técnicas de aprendizagem automática na previsão de ativos digitais, oferecendo uma perspetiva comparativa entre abordagens clássicas e modelos de machine learning aplicados ao domínio financeiro.

Keywords: Redes neuronais artificiais · Mercados financeiros· Previsão

Séries Temporais III

Forecasting Hotel Demand

Clara Cordeiro [1[0000-0002-1026-6078]], Nuno António [2[0000-0002-4801-2487]], and Sara Galguinho [2[0009-0008-6013-1113]]
ccordei@ualg.pt, nantonio@novaims.unl.pt,20220682@novaims.unl.pt

[1] *FCT-DM, Universidade do Algarve; CIDMA-Centro de Investigação e Desenvolvimento em Matemática e Aplicações da Universidade de Aveiro; CEAUL-Centro de Estatística e Aplicações, Universidade de Lisboa, Portugal*
[2] *NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal*

Abstract: Accurately forecasting demand is crucial for the hospitality sector. Common data sources used in hotel demand forecasting are historical and advanced. Historical data reflects the occupancy rate, which can be measured on a daily basis, while advanced booking data captures booking behaviour for each day leading up to arrival. This study adopts a different approach by combining both types of data, using an ensemble of time series models and machine learning techniques. A case study based on hotel data demonstrates that the proposed combined method produces more accurate forecasts than conventional forecasting techniques. These findings highlight the benefits of integrating multiple approaches and data sources in revenue management, offering valuable insights for the hospitality industry.

Keywords: Hotel demand · Machine Learning · Revenue management · Time series models

Acknowledgements: This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS – Nuno António.
This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., within the scope of project UID/04106 - Clara Cordeiro.

References

[1] Ampountolas, A.: Addressing complex seasonal patterns in hotel forecasting: a comparative study. Journal of Revenue and Pricing Management, **24**(2), 143–152 (2024). DOI:10.1057/s41272-024-00494-6
[2] António, N., Nunes, L.: Predicting hotel bookings cancellation with a machine learning classification model. In: Proceedings of the 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 1049–1054, 2017. DOI:10.1177/1938965519851466
[3] Wang, X., Hyndman, R. J., Li, F., Kang, Y.: Forecast combinations: An over 50-year review. International Journal of Forecasting, **39**(4), 1518-1547 (2023). DOI:10.1016/j.ijforecast.2022.11.005

INAR Models with Structural Breaks: a CUSUM-Guided Maximum Likelihood Approach

Magda Monteiro [1,2][000-0001-8585-4440] and Isabel Pereira [2,3][0000-0002-5152-546X]
msvm@ua.pt, isabel.pereira@ua.pt

[1] *Águeda School of Technology and Management, University of Aveiro*
[2] *Center for Research and Development in Mathematics and Applications, University of Aveiro*
[3] *Department of Mathematics, University of Aveiro*

Abstract: Count time series are essential in domains, such as healthcare, finance, and transportation for modeling discrete events over time. A popular framework for stationary count data modeling is the INAR (Integer-valued Autoregressive) model, which uses a thinning operator and innovation distributions, like Poisson, negative binomial, or geometric, the latter two being particularly effective for handling overdispersed data. However, these models often assume time-invariant parameters, an assumption frequently violated in real-world applications, such as during disease outbreaks where case counts follow dynamic patterns. This work focuses on an INAR model framework to handle structural breaks, using a frequentist approach centered on maximum likelihood estimation. Structural changes in model parameters are identified through a CUSUM-based procedure, followed by a targeted grid search window centered at the candidate breakpoint detected by the test. This methodology enables localized refinement of change point positions, while preserving computational tractability. The proposed approach is applied to real-world health indicator data, showcasing its effectiveness in capturing temporal shifts. A simulation study was carried out to compare this breakpoint detection strategy with an approach based on visual detection with a grid search.

Keywords: CUSUM test · INAR models · Maximum likelihood estimation · Overdispersion · Structural breaks

Acknowledgements: This work is partially supported by CIDMA under the Portuguese Foundation for Science and Technology, reference UID/04106 (FCT, https://ror.org/00snfqn58).

Structural Breaks in Overdispersed INAR Models: A Bayesian Approach

Isabel Pereira [1,2][0000-0002-5152-546X], Magda Monteiro [1,3][000-0001-8585-4440], and Maniha Zafar [1]
isabel.pereira@ua.pt, msvm@ua.pt, maniha.zafar@ua.pt

[1] *Department of Mathematics, University of Aveiro*
[2] *Center for Research and Development in Mathematics and Applications, University of Aveiro*
[3] *Águeda School of Technology and Management, University of Aveiro*

Abstract: Integer-valued autoregressive (INAR) models are widely used for modeling time series of counts in diverse fields such as epidemiology, finance, and public safety. These models are particularly useful due to their ability to accommodate the discrete nature of count data and to incorporate important features such as equidispersion and overdispersion. However, traditional INAR models often assume time-invariant parameters, which may not reflect the dynamic behavior of real-world processes, such as shifts due to policy changes or phases of an epidemic.
This work investigates INAR models with structural breaks, focusing on scenarios where model parameters change across different regimes. We consider overdispersed innovations and adopt a fully Bayesian approach for parameter estimation and change-point detection. The methodology employs updated Markov Chain Monte Carlo (MCMC) techniques, incorporating hidden Markov chains to identify latent regimes.
A comprehensive simulation study is conducted under varying conditions, including different proportions (length) for the duration of distinct regimes in the time series and diverse distributional characteristics of the data. Finally, the proposed approach is applied to a real-world dataset involving health indicators, illustrating the practical value of the methodology in capturing complex dynamics and structural shifts in count time series.

Keywords: INAR models · MCMC · Overdispersion · Structural breaks

Acknowledgements: This work is partially supported by CIDMA under the Portuguese Foundation for Science and Technology, reference UID/04106/2025 (FCT, https://ror.org/00snfqn58).

Data-Driven Fragmented Autocorrelation for Improved Time Series Clustering

Jorge Caiado [1[0000-0002-0405-1695]] and Nuno Crato [1[0000-0002-3572-0396]]
jcaiado@iseg.ulisboa.pt, ncrato@iseg.ulisboa.pt

[1] *ISEG-Lisbon School of Management and Economics and REM/CEMAPRE, Universidade de Lisboa, Portugal*

Abstract: Time series clustering requires distance metrics capable of capturing meaningful patterns. Traditional methods using full autocorrelations (ACF) or all periodogram ordinates can be inefficient due to irrelevant lags or frequencies. While recent fragmented approaches improve focus, they assume prior structural knowledge. In this paper, we propose a data-driven fragmentation method that automatically selects statistically significant ACF and PACF lags, filtering noise and enhancing clustering accuracy. Tests on simulated and real-world economic data (GDP, inflation, military spending, fertility rates) show superior performance over conventional metrics, especially when underlying dynamics are unknown.

Keywords: ACF · Distance metric · Economic indicators· Fragmented autocorrelation · Time series clustering

Ciência de Dados III

The Work Climate Questionnaire (WCQ) for Volunteer Settings: Psychometric Properties in a Portuguese Sample

Ricardo Baptista [1[0000-0002-3365-2094]], **Conceição Ribeiro** [2,3[0000-0003-0185-3200]], Rita dos Santos [1,4[0000-0002-3278-8424]], Marta Brás [1,4[0000-0001-7430-1939]], M. Dulce Estêvão [5[0000-0002-7151-8363]], Claúdia Carmo [1,4[0000-0002-7301-349XD]], Saul Neves de Jesus [1,4[0000-0003-2019-1011]], José Tomás da Silva [6[0000-0002-2782-6780]],and Cátia Martins [1,4[0000-0002-1819-8516]]

a64413@ualg.pt, cribeiro@ualg.pt, rasantos@ualg.pt, mbras@ualg.pt, mestevao@ualg.pt, cgcarmo@ualg.pt; snjesus@ualg.pt, jtsilva@fpce.uc.pt, csmartins@ualg.pt

[1] *Faculdade de Ciências Humanas e Sociais, Universidade do Algarve*
[2] *Instituto Superior de Engenharia, Universidade do Algarve*
[3] *CEAUL, Faculdade de Ciências, Universidade de Lisboa*
[4] *Centro Universitário de Investigação em Psicologia, Universidade do Algarve*
[5] *Escola Superior de Saúde, Universidade do Algarve*
[6] *Centro de Estudos Sociais, Faculdade de Psicologia e de Ciências da Educacação, Universidade de Coimbra*

Abstract: In the volunteering setting, the recruitment and maintenance of volunteers are daily challenges. Within the framework of Self-determination Theory, the degree of autonomy satisfaction and support influences individuals' motivation and engagement. The aim of this study was to analyse the psychometric properties of Work Climate Questionnaire (WCQ) in the volunteering setting. A total of 237 volunteers, mainly female (72.2%), aged from 13 to 81, essentially working in the social area (66%) participated. The factorial structure of the Portuguese version of the WCQ was examined through Confirmatory Factor Analysis (CFA). Given the potential impact of Maximum Likelihood (ML) or Generalized Minimum Squares estimation methods on fit indices and parameter estimates, the Diagonally Weighted Least Squares (DWLS) estimation method was employed via the *lavaan* package to accommodate the non-normal distribution of item responses. Overall, the WCQ, with its 14-item and 6-item variants, exhibits robust psychometric properties, making it a valuable tool for researchers and practitioners in the field of volunteering.

Keywords: Autonomy support · Diagonally Weighted Least Squares · Psychometric · Self-Determination Theory · Volunteering

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project CIP - Refª UID/PSI/04345/ 2020 and under the projects UIDB/00006/2025, UIDB/00006/2020 (DOI: 10.54499/. UIDB/ 00006/2020).

Neural Network Binary Predictions Explanation through Propensity Score Methodology

Luís Garcez [1[0000-0002-8637-7946]], João Telhada [1[0000-0001-6353-6329]], and Eduardo Severino [1[0000-0001-7743-0111]]

luisgarcez1@gmail.com, jmtelhada@ciencias.ulisboa.pt, jeseverino@ciencias.ulisboa.pt

[1] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: Neural Networks (NNs) have become indispensable tools in predictive modeling due to their remarkable accuracy and flexibility across diverse domains, including medicine, finance, and engineering. However, despite their successes, neural networks are notoriously opaque, often labeled as "black box" models due to their complex internal structures that lack intuitive interpretability. This opacity poses significant challenges for applications that require a clear understanding and justification of model decisions, such as healthcare diagnostics and regulatory compliance. The aim of this work is to introduce a novel approach to provide a more interpretable explanation for NNs predictions, based on Propensity Score (PS) Methodology Inverse Probability Treatment Weighting (IPTW). For each predictor, this methodology compares the predictions made between predictor variable levels on weigthed populations with similar characteristics regarding all the other predictor variables. This methodology computes an Average Prediction Effect (APE) for each predictor variable, quantifying the average effect on model predictions for the predicted population.

This novel integration of propensity score methods into neural network interpretability delivers absolute and interpretable explanations on how specific predictors influence, on average, predictions.

Keywords: Explainable AI · Neural networks · Propensity score methodology

References

[1] Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55 DOI: 10.2307/2335942

[2] Jian-Xun Mi, Xilai Jiang, Lin Luo, Yun Gao (2024). Toward explainable artificial intelligence: A survey and overview on their intrinsic properties. Neurocomputing. DOI:10.1016/j.neucom.2023.126919

Modelos Preditivos para Dados Longitudinais: Um Estudo de Simulação

Elsa Soares [1[0000-0003-0612-5202]], Inês Sousa [1[0000-0002-2712-1713]]
id10725@alunos.uminho.pt, isousa@math.uminho.pt

[1] *Centro de Matemática, Escola de Ciências, Universidade do Minho*

Abstract:
Dados longitudinais, caracterizados por medições repetidas ao longo do tempo em cada unidade amostral, são cada vez mais comuns em áreas como ciências sociais, económicas, comportamentais, biológicas e outras. Estes dados distinguem-se dos estudos transversais por permitirem acompanhar a evolução individual ao longo do tempo, possibilitando uma análise mais detalhada da evolução de fenómenos dinâmicos. Os modelos estatísticos clássicos, como os modelos mistos generalizados, mantêm-se amplamente utilizados, oferecendo resultados robustos quando as relações entre variáveis são relativamente simples ou conhecidas. No entanto, quando é necessário modelar múltiplas variáveis e captar padrões temporais complexos ou não lineares, surgem limitações nestas abordagens paramétricas [1].
Neste contexto, técnicas de *machine learning* revelam-se promissoras para tarefas preditivas, ao lidarem com elevada dimensionalidade, relações não lineares e padrões desconhecidos a priori em dados longitudinais. Contudo, a sua aplicação a este tipo de dados enfrenta desafios como a correlação entre medições do mesmo indivíduo, a heterogeneidade das trajetórias e a existência de diferentes fontes de variabilidade, exigindo abordagens metodológicas específicas [2].
Este trabalho apresenta um estudo de simulação que compara o desempenho preditivo de modelos estatísticos e de *machine learning* em dados longitudinais com múltiplas fontes de variabilidade — incluindo efeitos aleatórios individuais, processos gaussianos no tempo e erro branco independente. Simulam-se trajetórias de 100 indivíduos ao longo de 20 momentos temporais. Os resultados permitem avaliar a robustez e precisão dos modelos, oferecendo orientações práticas para a sua aplicação em contextos longitudinais complexos.

Keywords: Dados longitudinais · *Machine learning* · Modelos mistos · Simulação · Variabilidade

Acknowledgements: Agradecemos à Fundação para a Ciência e a Tecnologia (FCT) pelo apoio financeiro concedido através da bolsa de doutoramento com referência UI/BD/154394/2023.

References

[1] Sheetal, A., Jiang, Z., Di Milia, L.: Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. Applied Psychology **72**(3), 1339–1364 (2023).
[2] Cascarano, A., Mur-Petit, J., Hernández-González, J. *et al.*: Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review **56**(Suppl 2), 1711–1771 (2023). https://doi.org/10.1007/s10462-023-10561-w

A Principal Component Analysis for Ordinal Data

Hugo Alonso [1,2[0000-0002-1599-5392]] and Adelaide Freitas [2,3[0000-0002-4685-1615]]
hugo.alonso@upt.pt, adelaide@ua.pt

[1] *Portucalense University/Research on Economics, Management and Information Technologies (REMIT), Rua Dr. António Bernardino de Almeida, 541, 4200-072 Porto, Portugal*
[2] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*
[3] *Department of Mathematics, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal*

Abstract: Large datasets are common and can be challenging to interpret. Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of such datasets, making them easier to understand while preserving as much of the original information as possible [1]. In some cases, datasets consist solely of ordinal variables measured on the same scale. A common example is job interviews, where candidates are evaluated across multiple criteria using the same Likert scale. In this work, we propose a PCA approach specifically designed for ordinal data measured on the same scale. The idea is to replace Pearson's correlation matrix, used in standard PCA for quantitative variables, with a matrix based on r_{int}, a measure of association between ordinal variables introduced in [2, 3]. Although r_{int} can be computed regardless of whether the ordinal variables share the same scale or not, the calculation formula is simpler when the scales are equal, making it more practical to use in such cases. We present results from an empirical study in which we compare the performance of our r_{int}-based PCA with PCA based on Spearman's and Kendall's correlation coefficients, using several datasets consisting exclusively of ordinal variables measured on the same scale.

Keywords: Ordinal data · Principal Component Analysis · Rank correlation

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

References

[1] Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, **374**, pages 1-16 (2016). DOI:10.1098/rsta.2015.0202
[2] Pinto da Costa, J.F., Alonso, H., Cardoso, J.S.: The unimodal model for the classification of ordinal data. Neural Networks, **21**, pages 78-91 (2008). DOI:110.1016/j.neunet.2007.10.003
[3] Pinto da Costa, J.F., Alonso, H., Cardoso, J.S.: Corrigendum to "The unimodal model for the classification of ordinal data" [Neural Netw. 21 (2008) 78-79]. Neural Networks, **59**, pages 73-75 (2014). DOI:10.1016/j.neunet.2014.06.003

Probabilidade e Processos Estocásticos

Unravelling Causal Dependencies in Climate Indices Using Mutual Information Rate Decomposition

Helder Pinto [1[0000-0002-0455-6466]], Susana Barbosa [2[0000-0003-2198-3715]], Maria Eduarda Silva [3[0000-0003-2972-2050]] and Ana Paula Rocha [1[0000-0003-3218-7001]]
helder.pinto@fc.up.pt, susana.a.barbosa@inesctec.pt, mesilva@fep.up.pt, aprocha@fc.up.pt
[1] *CMUP LASI, Dep. Matemática, Fac. Ciências, Univ. do Porto, Portugal*
[2] *INESC TEC, Porto, Portugal*
[3] *LIADD-INESC TEC, Faculdade de Economia, Univ. do Porto, Portugal*

Abstract: The North Atlantic Oscillation (NAO) is a key pattern of climate variability, extensively studied for its substantial influence on the North Atlantic and surrounding regions. To better understand interactions between climate systems, causality analysis has gained increasing attention — particularly in assessing the impact of sea surface temperature on the NAO [1]. This study examines the NAO's relationships with several major climate indices, including the Pacific–North American (PNA) index, Arctic Oscillation (AO), Atlantic Multidecadal Oscillation (AMO), Pacific Decadal Oscillation (PDO), Tropical North Atlantic (TNA), Niño3.4 index, and the Quasi-Biennial Oscillation (QBO). These interactions are analyzed using the Mutual Information Rate (MIR), an information-theoretic metric capable of capturing both linear and nonlinear dependencies over time [2]. Furthermore, MIR can be decomposed into components related to Transfer Entropy, allowing for the quantification of information flow between indices and the identification of the dominant direction of interaction. Our results offer insights into climate system interactions and suggest that information-theoretic approaches are useful for exploring potential drivers of the NAO.

Keywords: Climate Analysis · Mutual Information Rate · North Atlantic Oscillation

Acknowledgements: H.P and A.P.R supported by CMUP, financed by FCT – Fundação para a Ciência e Tecnologia, I.P., projects UID/00144-Centro de Matemá-tica da Universidade do Porto. H.P. thanks FCT, for the Ph.D. Grant 2022.11423.BD.

References

[1] Vannitsem, Stéphane, X. San Liang, and Carlos A. Pires. "Nonlinear causal dependencies as a signature of the complexity of the climate dynamics." EGUsphere 2024 (2024): 1-22. DOI: 10.5194/egusphere-2024-3308

[2] Pinto, Hélder, Y. Antonacci, C. Barà, R. Pernice, I. Lazic, L. Faes, and A. P. Rocha. "Estimating the Mutual Information Rate of Short Time Series from Coupled Dynamic Systems." Under review in Communications in Nonlinear Science and Numerical Simulation (2024). DOI: 10.2139/ssrn.5118280

A New Approach to the Delta Approximation Method for Mixed Stochastic Differential Equation Models

Patrícia A. Filipe [1,2,3][0000-0003-3664-7239], Gonçalo Jacinto [1,4][0000-0002-3292-2208], Carlos A. Braumann [1,4][0000-0003-2721-9750]
patricia.filipe@iscte-iul.pt, gjcj@uevora.pt, braumann@uevora.pt

[1] *CIMA, Instituto de Investigação e Formação Avançada, Universidade de Évora, Portugal*
[2] *Iscte Business School, Iscte-Instituto Universitário de Lisboa, Portugal*
[3] *BRU-Iscte, Iscte-Instituto Universitário de Lisboa, Portugal*
[4] *Escola de Ciências e Tecnologia, Universidade de Évora, Portugal*

Abstract: Stochastic differential equation (SDE) models are well-suited for describing individual growth in randomly fluctuating environments and have been successfully applied to model cattle weight. To capture animal-specific variability, we extend these models to mixed effects SDEs by allowing key parameters, such as growth rate and asymptotic size at maturity, to vary across individuals. We consider a transformation of weight that results in a Ornstein-Uhlenbeck model.
Because the likelihood function for mixed SDE models often lacks a closed-form expression, we recently proposed a new approximation technique, denominated the Delta method. This approach is based on maximizing an approximate likelihood obtained by replacing the exponential function within the integrand with a second-order Taylor expansion. Notably, the method accommodates irregular observation times, a common limitation in existing software.
We have recently developed a refinement of the Delta method that further improves its accuracy by applying the second-order Taylor expansion earlier, more specifically, to the argument of the exponential function rather than the exponential itself. This adjustment yields better estimation results, particularly when both parameters are treated as random effects. We demonstrate the performance of the improved method through applications to both simulated data and real weight records from a large heterogeneous sample of Mertolengo cattle.

Keywords: Delta method · Maximum likelihood estimation · Mixed models · Stochastic differential equations

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/04674/2020. DOI: 10.54499/UIDB/04674/2020.

References

[1] Jamba, N.T., Jacinto, G., Filipe, P.A., Braumann, C.A.: Likelihood Function through the Delta Approximation in Mixed SDE Models. Mathematics , **10**, 385 (2022). DOI:https://doi.org/10.3390/math10030385
[2] Jamba, N.T., Jacinto, G., Filipe, P.A., Braumann, C.A.: Estimation for stochastic differential equation mixed models using approximation methods. AIMS Mathematics, **9**4, 7866 – 7894 (2024). DOI:https://doi.org/10.3934/math.2024383

Longitudinal Count Data: A Simulation Study using R

M. Helena Gonçalves [1,3] [0000-0002-6990-7239] and
M. Salomé Cabral [2,3] [0000-0003-4462-4811]
mhgoncal@ualg.pt, mscabral@fc.ul.pt

[1] *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro-Portugal*
[2] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa-Portugal*
[3] *Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa-Portugal*

Abstract: In longitudinal count data studies, measurements are collect at several time points on each individual in one or more treatment groups. In such cases, repeated measurements are made on the same individual over time and correlation is usually present among response variables for a given individual. The generalized linear mixed models (GLMMs) account for that correlation by the inclusion of random effects in the linear predictor, although GLMMs assumes that the measurements of the same individual are independent conditional to the random effects and covariates, which may be not true. The methodology implemented in the R package cold is based on the likelihood approach and a serial dependence AR1 model is incorporated allowing that the dependence between repeated measures is considered in terms of numerical analysis. The dependence between repeated measures is ignored in the traditional approach implemented in the R package lme4 that only allow an independent structure. A simulation study is used to compare the aforementioned R packages.

Keywords: Repeated measurements · Random effects · Generalized linear mixed models · Serial dependence

Métodos Bayesianos

Modelling Mobility Data during COVID-19 in Portugal with `R-INLA`

André Brito [1,2,4][0009-0006-1345-816X], Ausenda Machado [4,5][0000-0002-1849-1499], Ana Paula Rodrigues [4][0000-0003-2264-4723], Paula Patrício [1,2][0000-0001-6766-4336], Regina Bispo [2,3][0000-0002-6723-2557]

anm.brito@campus.fct.unl.pt, ausenda.machado@insa.min-saude.pt, ana.rodrigues@insa.min-saude.pt, pcpr@fct.unl.pt, r.bispo@fct.unl.pt

[1] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA FCT), Portugal*
[2] *Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), Portugal*
[3] *School of Mathematics and Statistics and Centre for Research into Ecological and Environmental Modelling University of St Andrews, UK*
[4] *Department of Epidemiology, National Institute of Health Doctor Ricardo Jorge, Portugal*
[5] *Comprehensive Health Research Center (CHRC), NOVA University Lisbon, Portugal*

Abstract: This study models human mobility patterns in mainland Portugal during the COVID-19 pandemic using Google Community Mobility Reports. A hierarchical spatio-temporal modelling framework was developed, incorporating covariates such as day of the week, holidays, lockdown periods, stringency index, and temperature to explain variability in mobility across three categories: Workplaces, Retail and Recreation, and Transit Stations. The model was estimated using a Bayesian approach through Integrated Nested Laplace Approximation (INLA). Results demonstrate that linear trends and seasonality significantly shape mobility patterns, while holiday effects, lockdowns, and stringency measures play substantial roles in explaining abrupt changes. Temperature was found to have a modest positive effect on mobility. Although the model performed well on unseen data, some limitations emerged, including its reduced ability to forecast the magnitude of holiday-related spikes and the changing dynamics of post-pandemic mobility. The hierarchical model structure proved well-suited to capture spatial and temporal dependencies, and INLA's efficiency enabled the estimation of complex models on a large dataset.

Keywords: COVID-19 · Google mobility reports · INLA · Mobility · Spatio-temporal models

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). André Brito is financed through a FCT PhD Scholarship 2024.00664.BDANA.

The Logistic-Normal distribution: a powerful prior on the simplex

Rui Martins [1][0000-0003-1862-7066]
rmmartins@fc.ul.pt

[1] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract: This work was originated by the question: "Are women's and men's football leagues equally predictable?". The outcome of a football match – Home-Win, Draw, or Away-Win – can be modeled as a realization of a multinomial random variable with three mutually exclusive events. Traditional approaches often use the Dirichlet distribution as a prior for the multinomial proportions, primarily due to its conjugacy and simplicity. In this work, we propose an alternative: the Logistic-Normal distribution, a multivariate prior for proportions that has received limited attention in this context, whose appeal lies in its connection to the multivariate normal distribution, as it is derived by applying a multivariate logistic transformation to variables assumed to follow a multivariate normal distribution.
We develop models to analyze data from the main Portuguese women's and men's football leagues, spanning seven complete seasons (2016–2017 to 2022–2023). The structure involves probability vectors that sum to 1 framed within the Aitchison Geometry on the Simplex. The estimating framework estimates latent team-specific strengths, accounts for variability across seasons and rounds, and investigates the influence of home advantage. Furthermore, we introduce two novel metrics (League Outcome Predictability and League Probabilistic Dominance) aimed at assessing the competitiveness of football leagues and the overall unpredictability of sports leagues.

Keywords: Compositional data · Dynamic models · Football prediction · Logistic-Normal prior · Sports comparison

Acknowledgements: This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through the CEAUL projects UID/00006/2025, DOI: 10.54499/UIDB/00006/2020 and DOI: 10.54499/UIDP/00006/2020.

References

[1] Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), **44**(2), 139–160 (1982). doi.org/10.1111/j.2517-6161.1982.tb01195.x

[2] Martins, R.: An underrated prior distribution for proportions. The logistic–normal for dynamical football predictions. In: In International Workshop on Statistical Modelling, pp. 121–127, 2024. doi.org/10.1007/978-3-031-65723-8_19

Mapping Urban Fire Intensity in Portugal: A Bayesian Approach with INLA and SPDE

Nádia Bachir [1][0000-0002-1355-0090], Regina Bispo [1,2][0000-0002-6723-2557], and Lígia Henriques-Rodrigues [3,4] [0000-0003-4881-4188]
n.bachir@campus.fct.unl.pt, r.bispo@fct.unl.pt, ligiahr@uevora.pt

[1] *Center for Mathematics and Applications (NOVA Math), NOVA School of Sciences and Technology, NOVA University of Lisbon, Portugal*
[2] *School of Mathematics and Statistics and Center for Research into Ecological and Environmental Modeling, University of St Andrews, Scotland*
[3] *CIMA - Research Center in Mathematics and Applications, University of Évora, Portugal*
[4] *Department of Mathematics, University of Évora, Portugal*

Abstract: Understanding the spatial distribution of urban fire events is essential for assessing fire risk and informing prevention strategies. This study explores the spatial patterns of urban fires in Portugal using Bayesian hierarchical modelling, combining the Integrated Nested Laplace Approximation (INLA) with the Stochastic Partial Differential Equation (SPDE) approach. Fire intensity is modelled through a Log-Gaussian Cox Process (LGCP) with a Poisson likelihood. The spatial component is represented by a latent Gaussian field defined over a continuous domain, allowing us to capture spatial heterogeneity in fire intensity. This framework allows for flexible specification, including the potential integration of relevant covariates to better understand factors influencing fire distribution. The resulting model produces spatial predictions of fire intensity across the study area. A posterior mean intensity map is generated over a regular grid, offering a detailed visualisation of high-risk zones. This approach provides a robust and computationally efficient method for modelling spatial point processes in a Bayesian framework. By leveraging INLA and SPDE, the model delivers interpretable results with high spatial resolution, contributing to a better understanding of urban fire dynamics in Portugal.

Keywords: Bayesian Spatial Modelling · INLA-SPDE · Urban Fires

Acknowledgements: NB is funded by the individual research grant 2023.04213.BD (https://doi.org/10.54499/2023.04213.BD).
RB is funded by 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).
LHR is funded by project UIDB/04674/2020 (https://doi.org/10.54499/UIDB/04674/2020) (Research Center in Mathematics and Applications).

Estatística Computacional II

Predicting Model Degradation under Noisy Conditions: A Robustness Study for Regression Problems

Catarina Fernandes [1[0009-0007-7624-7949]], Fátima Rbaibi [1[0009-0002-6643-7195]], Isabela Alves [1[0009-0001-6178-7033]], Nelson Vieira [2[0000-0001-8756-4893]], and **Luís Silva** [2[0000-0001-9677-4315]]
{clfernandes, fatimarbaibi, isabelaalves, nvieira, lmas}@ua.pt

[1] *Department of Social, Political and Territorial Sciences, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal*
[2] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*

Abstract: The robustness of machine learning models to data disturbances is essential for their application in real-world contexts such as healthcare, finance, and industry. This work proposes a systematic approach to assess the robustness of supervised regression models to the introduction of noise in predictor variables. Ten public datasets from different domains (e.g., insurance, prices, wages) and four regression models – Ridge Regression, K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting (GBM) – were used.
Gaussian noise (numerical variables) and permutations (categorical variables) were applied with different intensities and proportions of altered instances. Performance loss was measured by the percentage change in the Root Mean Square Error (RMSE) between tests with and without noise. Furthermore, the maximum tolerance to noise before occurring a substantial performance decay was analyzed. The results reveal clear differences between the models: Ridge and KNN demonstrated greater robustness, whereas RF and GBM exhibited higher sensitivity to perturbations.
Additionally, a clustering strategy based on four structural descriptors of the data was proposed to predict average robustness loss curves. Validation with two test datasets showed that, while effective for stable models (Ridge, KNN), the prediction was less accurate for more complex models (RF, GBM), especially in atypical scenarios. It is concluded that the methodology can support preliminary decisions in contexts with data uncertainty, but its application requires caution when involving models sensitive to the data structure.

Keywords: Clustering · Noise robustness · Regression models

Acknowledgements: This work is supported by CIDMA under the Portuguese Foundation for Science and Technology (FCT, https://ror.org/00snfqn58) Multi-Annual Financing Program for R&D Units. Project ref. UID/04106.

References

[1] Padró-Ferragut, C., Ramírez-Quintana, M.J., Martínez-Plumed, F.: Noise Tolerance and Robustness Ranking in Machine Learning Models. In: Julian, V., et al. Intelligent Data Engineering and Automated Learning 2024. Lecture Notes in Computer Science, vol 15347, pp. 96–108, 2025. Springer. DOI: 10.1007/978-3-031-77738-7_9

Modelling Distributional Data as Matrix-valued Data

Marcus Mayrhofer [1[0000-0002-3430-8308]], **Paula Brito** [2[0000-0002-2593-8818]], A. Pedro Duarte Silva [3[0000-0003-1378-2403]], Peter Filzmoser [1[0000-0002-8014-4682]]
marcus.mayrhofer@tuwien.ac.at, mpbrito@fep.up.pt, psilva@ucp.pt, peter.filzmoser@tuwien.ac.at

[1] *Institute of Statistics and Mathematical Methods in Economics, TU Wien*
[2] *Faculdade de Economia, Universidade do Porto & LIAAD INESC TEC*
[3] *Universidade Católica Portuguesa, Católica Porto Business School & CEGE*

Abstract: We consider symbolic data, where units are described by histogram or interval-valued variables $Y_j, j = 1, \ldots, p$. In our model, each observed distribution is represented by a central statistic C, and the logarithm transformation of inter-quantile ranges, denoted $R_h^*, h = 1, \ldots, m$, for a chosen set of quantiles, where m is the number of considered intervals. Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles ; interval-valued data are represented by midpoints and log-ranges ($m = 1$). Multivariate Normal distributions are then assumed for the whole set of indicators. Furthermore, we consider alternative structures of the variance-covariance matrix. In this work we model these data as matrix-valued, represented as a tensor of dimension $n \times p \times (m + 1)$, $X \sim MN(M, \Sigma_{var}, \Sigma_{ind})$, where

- Σ_{var} is $p \times p$ and gathers variances and covariances between the variables Y_j
- Σ_{ind} is $(m + 1) \times (m + 1)$ and gathers variances and covariances between the considered indicators $C, R_1^*, \ldots, R_m^*$

In this model, the global covariance matrix Σ is written as $\Sigma = \Sigma_{ind} \otimes \Sigma_{var}$. This implies that we assume that covariances between the different indicators are constant across variables, and covariances between the different variables are constant across indicators, thereby obtaining a more parsimonious model. The different covariance configurations correspond to setting Σ_{ind} and/or Σ_{var} as block-diagonal matrices. The Matrix Minimum Covariance Determinant (MMCD) method accounts for the matrix-variate data structure and robustly estimates the mean matrix M, as well as the row-wise Σ_{var} and column-wise Σ_{ind} covariance matrices.
Robust Mahalanobis distances based on MMCD estimators then allow for outlier detection. Using the concept of Shapley values for outlier explanation enables the decomposition of the squared Mahalanobis distances into contributions of the variables, indicators, and individual cells of the matrix-valued observations.
Applications to real data put in evidence the interest of the proposed approach for robust multivariate symbolic data analysis.

Keywords: Explainable AI · Histogram data · Matrix-valued data · Shapley values

Acknowledgements: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within projects UID/50014/2023 (https://doi.org/10.54499/UID/50014/2023), and UID/GES/00731/2019.

Maximum Likelihood Estimation of the Parameters of the Power-Normal Distribution

Rui Gonçalves [1[0000-0002-7092-5442]]
rjasg@fe.up.pt

[1] *Depto de Eng. Civil e Georrecursos, Faculdade de Engenharia, Universidade do Porto*

Abstract: The power-normal distribution (PN) is a family of distributions that includes the normal and lognormal distributions. The PN distribution is linked to the Box-Cox (BC) transformation in the sense that in an ideal situation, when BC transformation is successful then the original scale data is PN distributed. The BC transformation is only applicable if data is positive which means that transformed data is left truncated normal instead of normal. However, if the center of the transformed data distribution is far away from the truncation point then data can be considered normal. Box and Cox suggested that the estimation of the BC (and PN) λ parameter could be made ignoring the truncation caused by the positivity condition. That may be acceptable to get approximate normality but as we show is devastating when it comes to parameter estimation. Although the Maximum Likelihood (ML) estimator of λ has a normal distribution it doesn´t have a closed form expression. Estimates are found using numerical methods which depend on several convergence criteria.
In this work, we use simulated data sets to perform exact ML estimation of the parameters in the PN distribution. We present two algorithms: one that jointly estimates all three parameters — μ, σ, and λ — and another that focuses exclusively on estimating λ. We compare and discuss the results of these estimation procedures. Considering the practical relevance of the BC transformation, we also apply the methods to bootstrapped data sets and compare the outcomes with those obtained from the original simulated data.

Keywords: Box-Cox transformation · Maximum Likelihood Estimation· Power-Normal

References

[1] Gonçalves, R., The power-normal distribution. In: AIP Conf. Proc. 2116 (1): 110009 (2019) DOI:org/10.1063/1.5114102
[2] Gonçalves, R. The power-normal distribution. Journal of Physics, 1334, pp:012014, Ano 2019. DOI:10.1088/1742-6596/1334/1/012014
[3] Gonçalves, R. Exact vs approximated ML estimation for the box-cox transformation. AIP Conf. Proc. 3094, 500034 (2024). DOI:10.1063/5.0211637
[4] Goto, M., Inoue, T. and Tsuchya. On the estimation of parameters in power-normal distribution. Bull. of Inf. and Cybernetics, 21(1-2).pp 41-53, (1984).

Bioestatística e Epidemiologia II

Unveiling the Power of the Aranda-Ordaz Link in GAMLSS: A Comparative Study for Binary Data

Neto Pascoal [1,2][0000-0001-8616-0520], Eunice Carrasquinha [1,2][0000-0003-3465-4347], and Carlos Geraldes [1,3][0000-0002-1551-6531]
npascoal@fc.ul.pt, eitrigueirao@ciencias.ulisboa.pt, carlos.geraldes@isel.pt

[1] *CEAUL — Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *DEIO — Departamento de Estatística e Investigaçõ Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *ISEL — Instituto Superior de Engenharia de Lisboa, Portugal*

Abstract: This study explores the potential of the Aranda-Ordaz (AO) transformation within the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework to recover structures of models with unknown link functions. Simulated datasets were generated under scenarios where the true data-generating process followed classical link functions. A GAMLSS model with the flexible AO link was then fitted to assess its capacity to approximate both partial effects and the underlying link function of the unknown model. The parameter ψ was estimated using penalized likelihood, median log-likelihood, and bootstrap resampling methods, and model performance was assessed. The results demonstrate that the AO link achieves a close approximation to the true model, yielding superior fit and predictive accuracy. These findings reinforce the importance of flexible link functions in GAMLSS and highlight the AO transformation as a powerful tool for uncovering complex model structures in scenarios where the true link is unknown.

Keywords: Aranda-Ordaz · Binary data · GAMLSS Models · Link function · Predictive modeling

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UID/00006/2025, UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020) and UI/BD/154312/2022. DOI: 10.54499/UI/BD/154312/2022 (https://doi.org/10.54499/UI/BD/154312/2022).

References

[1] Aranda-Ordaz, F. J.: An extension of the proportional-hazards model for grouped data. In: Biometrics, pp. 109–117, 1983. https://doi.org/10.2307/2530811
[2] Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., De Bastiani, F.: Flexible regression and smoothing: Using GAMLSS in R. CRC Press, 2017. https://doi.org/10.1201/b21973
[3] Hastie, T. J., Tibshirani, R. J.: Generalized Additive Models. Chapman and Hall/CRC Press, London, UK (1990). https://doi.org/10.1201/9780203753781

Analyzing Vaccination Risks Compared to Infection Risks: a Game Theory Perspective Considering Reinfection

José Martins [1,2][0000-0002-0556-7861], and Alberto Pinto [1,3][0000-0003-2953-6688]
jmmartins@ipleiria.pt, aapinto@fc.up.pt

[1] *LIAAD-INESC TEC, Portugal*
[2] *School of Technology and Management, Polytechnic of Leiria, Portugal*
[3] *Faculty of Sciences, University of Porto, Portugal*

Abstract: For diseases where vaccination is not mandatory, individuals take into account multiple factors when deciding whether or not to get vaccinated. Their decisions constitute vaccination strategies that are influenced by the morbidity risks associated with both the vaccine and the infection, as well as by the the probabilities of becoming infected, which vary over time with the course of the disease and the decisions of all other individuals.
In 2017, Martins and Pinto introduced the evolutionary vaccination dynamics for a homogeneous population vaccination strategy. In this work, we introduce the dynamics of the morbidity relative risk, defined as the ratio between the vaccine morbidity risk and the infection morbidity risk. Using the basic reinfection epidemiological SIRI model, we study the evolution of the homogeneous population vaccination strategy, which defines the vaccination coverage level attained, when the morbidity relative risks evolve over time. Depending on the parameters of the epidemic model, we observe the emergence of a Hopf bifurcation in the ODE system describing the vaccination and the morbidity relative risk dynamics. This implies the existence of limit cycles in vaccination coverage instead of a single and stable vaccination level.

Keywords: Game theory · Reinfection · Risks · SIRI model · Vaccination

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2023 (https://doi.org/10.54499/UID/50014/2023).

Metascience in Ecology: The Role of Hypothesis Testing in Ecology

Gabriela Xavier-Quintais [1,3][0000-0003-4896-1225],
Tiago A. Marques [1,2][0000-0002-2581-1972], Daniel Lakens [3][0000-0002-0247-239X]
fc56156@alunos.fc.ul.pt, tamarques@ciencias.ulisboa.pt, D.Lakens@tue.nl

[1] *Faculty of Science of the University of Lisbon (FCUL)*
[2] *University of St Andrews (UstA)*
[3] *Eindhoven University of Technology (TU/e)*

Abstract: The use of statistics in scientific research has become predominant in ecology. In order not to overvalue statistics at the detriment of the field of knowledge for which it is being used, it is essential to understand both its potential and its limitations in ecology, so as to use statistics for ecology and not ecology for statistics. The *p-value* is a probability widely used in ecological research as a criterion for determining "statistical significance", and in association with hypothesis testing, it is sometimes referred to as one of the most influential and transformative concepts in modern science. However, the interpretation and uses of hypothesis testing and the *p-value* have raised many questions, with the incompatibility between "statistical significance" and "biological significance" being one of the biggest problems. On the other hand, there are some constraints specific to the field of ecology: observational studies (as opposed to randomized studies), limitations on sample sizes, high variability in the measurements of variables used to think about ecological phenomena, etc., which may limit how hypothesis testing is used. It is within this context that in this conference, we will present the results of a meta-scientific study (Manual Analysis: review of 110 papers; Automatic Analysis: review of 1697 papers in 11 different scientific journals) that aims to encourage reflection on how hypothesis testing is being used in the field of ecology, and to stimulate joint thinking about what needs to be changed.

Keywords: Misinterpretations · Null hypothesis · *P-value* ·

Posters

—**Pósteres**—

Sessão de Posters I

Adaptação e validação da escala *Motives for Online Gaming Questionnaire* (MOGQ) numa amostra de estudantes universitários

Elisete Correia [1[0000-0002-1121-2792]], Ana Luisa Lopes [2[0000 0002 5600 6116]] e Ana Paula Monteiro [3[0000-0002-4082-1474]]
ecorreia@utad.pt, analdslopes@outlook.com, apmonteiro@utad.pt

[1] *CEMAT, Departamento de Matemática, UTAD, Vila Real, Portugal*
[2] *Departamento de Educação e Psicologia, UTAD, Vila Real, Portugal*
[3] *CIIE, Departamento de Educação e Psicologia, UTAD, Vila Real, Portugal*

Abstract: Os jogos online tornaram-se uma atividade de lazer popular que abrange diversas idades e ambos os sexos. Embora essa atividade traga algumas vantagens, a comunidade científica identificou casos de dependência relacionados a esses jogos. O estudo das motivações revela-se essencial para compreender a origem e o desenvolvimento deste comportamento potencialmente aditivo. Assim, o objetivo do presente estudo foi adaptar e validar a escala *Motives for Online Gaming Questionnaire* (MOGQ) numa amostra de estudantes portugueses, que contempla sete motivações que estão, na base dos jogos *online*: social, escape, competição, coping, desenvolvimento de capacidades, fantasia e passatempo. A amostra foi constituída por 501 estudantes universitários, sendo 305 do sexo feminino e 196 do sexo masculino, com idades compreendidas entre os 17 e os 50 anos. A validação da escala apresentou bons índices de consistência interna e a análise fatorial confirmatória revelou uma qualidade de ajustamento aceitável. Os resultados obtidos indicam que a versão portuguesa da escala MOGQ é fiável e consiste num bom instrumento para a investigar as motivações que levam as pessoas a jogar jogos online.

Keywords: Análise fatorial confirmatória · Escala · Jogos online · Motivação

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto CEMAT/IST-ID [grants no. UIDB/04621/2020 e UIDP/04621/2020] e CIIE [grants no. UIDB/00167/2020 e UIDP/00167/2020]

A Python-Based Guide to Modeling Interval-Censored Time-to-Event Data

Rui Alves[1[0009-0004-7004-4962]], Luís Machado [1[0000-0002-8577-7665]], and Carla Moreira [1[0000-0002-0570-0650]]

ruimiguelalves03@gmail.com, lmachado@math.uminho.pt, carlamgmm@gmail.com

[1] *CMAT – Centro de Matemática, Escola de Ciências, Universidade do Minho, Portugal*

Abstract: In clinical and epidemiological research, events are often observed only within time intervals, giving rise to interval-censored data. This form of censoring poses specific methodological challenges that standard right-censoring techniques are not equipped to handle. In this work, we present a practical tutorial on analyzing interval-censored survival data using the Python programming language. We introduce essential tools for nonparametric estimation, discuss approaches for comparing survival functions across groups, and demonstrate the use of regression models tailored for interval-censored outcomes. The tutorial is illustrated with synthetic data reflecting real-world clinical scenarios and makes use of Python libraries such as 'lifelines' and 'scikit-survival'. Our aim is to provide applied researchers with a robust foundation for implementing interval-censored survival analyses in Python, bridging the gap between statistical theory and modern computational practice.

Keywords: Interval censoring · Python programming · Survival analysis · Turnbull estimator

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] Davidson-Pilon, C.: lifelines: survival analysis in Python. Journal of Open Source Software, **4**(40), 1317 (2019). https://doi.org/10.21105/joss.01317

[2] Pölsterl, S.: scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research, **21**(212), 1–6 (2020). http://jmlr.org/papers/v21/20-729.html

Analyzing Interval-Censored Survival Data: A Practical Guide Using R

Luís Machado [1][0000-0002-8577-7665], Carla Moreira [1][0000-0002-0570-0650], Rui Alves[1] and Aurélio Sidumo[1]
lmachado@math.uminho.pt, carlamgmm@gmail.com

[1] *CMAT – Centro de Matemática, Escola de Ciências, Universidade do Minho, Portugal*

Abstract: Interval censoring frequently arises in biomedical and clinical research when the exact timing of an event is unknown but is known to have occurred within a specific time interval. This type of censoring presents distinct challenges in survival analysis and often requires more advanced methods than those typically applied to right-censored data. In this work, we provide a comprehensive and practical tutorial on the analysis of interval-censored survival data using the R programming language. We begin with nonparametric estimation of survival functions and proceed to the comparison of survival curves across groups. We also explore the use of parametric and semiparametric regression models for analyzing interval-censored outcomes through applied examples. Throughout the tutorial, each method is illustrated using a synthetic dataset inspired by real clinical scenarios, focusing on methodological understanding and practical implementation. This work aims to serve as a resource for researchers and practitioners engaged in the robust modeling of interval-censored time-to-event data within clinical and epidemiological research.

Keywords: Interval censoring · R software · Survival analysis · Turnbull estimator

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

[1] Turnbull, B.W.: The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological), **38**(3), 290–295 (1976).
[2] Gómez, G., Calle, M.L., Oller, M., Langohr, R.: Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling, **9**(4), 259–297 (2009). https://doi.org/10.1177/1471082X0900900402
[3] Fay, M.P., Shaw, P.A.: `interval`: Analysis of interval-censored data. R package version 2.1.1. https://CRAN.R-project.org/package=interval (2023).
[4] Fay, M.P., Shaw, P.A.: Exact and asymptotic weighted logrank tests for interval-censored data: The `interval` R package. Journal of Statistical Software, **36**(2), 1–34 (2010). https://doi.org/10.18637/jss.v036.i02

Survival Analysis of COVID-19 Symptom Resolution in a Portuguese Cohort

Joana Pinto Costa[1[0000-0002-4461-4878]], Leandro Duarte[2], Luís Machado [2[0000-0002-8577-7665]], Ana Paula Amorim[2[0000-0003-3957-1129]], Margarida Tavares[1[0000-0003-4518-2197]], Paula Meireles[1[0000-0001-9055-7491]] and **Carla Moreira** [2[0000-0002-0570-0650]]

joanapintodacosta@ispup.up.pt, pg45191@alunos.uminho.pt, lmachado@math.uminho.pt, apamorim@math.uminho.pt, mftavares@ispup.up.pt, paula.meireles@ispup.up.pt, d8434@math.uminho.pt

[1] *EPIUnit ITR, Instituto de Saúde Pública da Universidade do Porto, Universidade do Porto, Portugal*
[2] *CMAT – Centro de Matemática, Escola de Ciências, Universidade do Minho, Portugal*

Abstract: COVID-19 affects individuals differently, with some experiencing persistent symptoms long after the initial infection. This study followed 2,777 adults with confirmed SARS-CoV-2 infection to estimate the time to symptom resolution and identify influencing factors. Data were collected on sociodemographic, clinical, and infection-related characteristics. Survival analysis and Cox regression models were used to assess symptom resolution over time. At follow-up, 36.5% of participants still reported unresolved symptoms. Faster recovery was associated with being male, having higher education levels, and a more positive perception of income. In contrast, slower recovery was linked to middle-aged individuals, comorbidities, and hospitalization during the acute phase. These findings highlight the prolonged impact of COVID-19 for a substantial portion of patients and the importance of addressing risk factors that contribute to delayed recovery.

Keywords: COVID-19 · Post-COVID-19 condition · Survival analysis

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).
We want to thank the Centro Hospitalar Universitário de São João (CHUSJ) for providing the information on which this study is based.

A gap time model based on a defective distribution with a time-varying recurrence-free proportion

Ivo Sousa-Ferreira[1,2][0000-0001-5526-3594], Ana Maria Abreu[1,3][0000-0002-6155-8492], Cristina Rocha[2][0000-0001-7162-4820]
ivo.ferreira@staff.uma.pt, abreu@staff.uma.pt, cmrocha@ciencias.ulisboa.pt

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

Abstract: Advances in biomedical research focused on recurrent events, such as disease relapses, have led to improved survival. Consequently, some patients no longer experience recurrences of the event of interest, even when followed for a sufficiently long period. In this setting, estimating the cure rate is clinically important, particularly when aiming to improve it. Most existing long-term survival models assume a constant cure rate. However, it may be more realistic to consider that a patient's probability of becoming recurrence-free varies over time.
To address this issue, we propose a parametric rate model for gap times between recurrent events based on a defective distribution. The model is formulated in terms of the conditional distribution of a gap time given the previous recurrence time. Within this framework, specifying an improper distribution naturally yields a time-varying recurrence-free proportion. The proposed model includes a special sub-model that is useful for testing the presence of recurrence-free individuals in the population. Parameter estimation is carried out using the maximum likelihood method under a right-censoring mechanism. The properties of the resulting estimators, along with the effectiveness of the likelihood ratio test, are assessed through simulation studies. An application to a real data set shows the practical relevance of the new model.

Keywords: Defective distribution · Gap times · Long-term survival model · Recurrent events · Time-varying recurrence-free proportion

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UID/00006/2025 and UIDB/00006/2020 (DOI:10.54499/UIDB/00006/2020) (CEAUL – Centro de Estatística e Aplicações); and the projects UID/04674/2025 and UIDB/04674/2020 (DOI:10.54499/UIDB/04674/2020) (Center for Research in Mathematics and Applications (CIMA) related to the Statistics, Stochastic Processes and Applications (SSPA) group).

Forecasting Seasonal Influenza in Portugal Using Pharmacy Sales: A Logistic Growth Model Approach.

João Brandão [1[0009-0000-9606-3273]], Rúben Pereira [1[0000-0003-4216-3150]], Zilda Mendes [1[0009-0009-3449-6503]], and António Teixeira Rodrigues [1,2,3[0000-0002-8161-9264]]

joao.brandao@anf.pt, ruben.pereira@anf.pt, zilda.mendes@anf.pt, antoniot.rodrigues@anf.pt

[1] *Center for Health Evaluation & Research/Infosaúde, National Association of Pharmacies (CEFAR/IF, ANF), Lisbon, Portugal*
[2] *Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal*
[3] *ICVS/3B's-PT Government Associate Laboratory, Braga/Guimarães*

Abstract: The monitoring of widespread community transmission, such as seasonal influenza, is essential for public health decision-making. Early outbreak detection and accurate severity prediction are used to manage disease control measures and to anticipate the epidemic's impact on the healthcare system. Traditional surveillance of influenza cases (e.g., physician diagnoses) naturally lag behind real transmission trends. As such, new surveillance systems have focused on sales data from over-the-counter products from community pharmacies. Most notably, the HiCorr project in Portugal can anticipate traditional data sources by approximately two weeks [1].
This project aims to enhance HiCorr by generating sales-based forecasts as a proxy for real influenza cases in Portugal. Specifically, we explore the use of logistic growth models, a method successfully applied during the COVID-19 pandemic to model epidemic dynamics. Our analysis uses time series sales data from a subset of over-the-counter products previously shown to be strongly correlated with primary care visits during past flu seasons. A multi-year analysis will be conducted to evaluate the forecasting performance across different seasonal influenza contexts.
Preliminary results show that the logistic growth model effectively captures the sigmoid pattern observed in cumulative sales data across all analyzed flu seasons. From a forecasting perspective, the application of parameter value constraints led to results that closely match observed trends for most seasons. On average, accurate forecasts regarding peak and intensity were achieved using only the first eight weeks of each season, approximately five weeks prior to the sales peak.

Keywords: Forecast · Logistic Growth Modeling · Pharmacoepidemiology

Acknowledgements: This project was developed in partnership with ACES Oeste Sul, extending to the Associação Nacional de Médicos de Saúde Pública (ANMSP).

References

[1] Pereira, R. et al.: Flu surveillance in Portugal using over-the-counter sales from community pharmacies. In: Proceedings of the 6th Statistics on Health Decision Making, p. 49, 2024. https://doi.org/10.34624/jshd.v6i1.37036

Estimation of the bivariate distribution function for interval censored data

Gustavo Soutinho [1][0000-1234-5678-9012], Luís Meira-Machado [2][0000-0002-8577-7665], and Marta Azevedo [2][0009-0003-4412-6991]
gustavo.soutinho@upt.pt, lmachado@math.uminho.pt,
marta.vasconcelos4@gmail.com

[1] *REMIT - Research on Economics, Management and Information Technologies, Universidade Portucalense, Portugal*
[2] *CMAT - Centro de Matemática, Universidade do Minho, Portugal*

Abstract: Analyzing time-to-event data is crucial across fields like medicine, engineering, and social sciences to understand underlying processes and support decision-making. A common issue is *interval censoring*, where events are known to occur within specific intervals, but their exact timing is unknown.
In some studies, individuals may experience multiple events, and the time between them—known as *gap times*—is of particular interest. While much research focuses on right-censored event times, few studies address scenarios where one or both events are interval censored.
This paper introduces new estimation methods that are based on the Turnbull estimator of survival, aimed at addressing the gap in literature regarding interval-censored events. Specifically, we explore the possibility of comparing the new estimators with methods based on the imputation of the event time. These imputation methods include estimating the event time as the midpoint of the interval, the point to the right of the interval, and the point to the left of the interval. Through empirical evaluation and simulation studies, we aim to provide insights into the relative performance and suitability of these estimation approaches for analyzing gap times with interval-censored data.

Keywords: Gap times · Interval censoring · Nonparametric estimation

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/05105: REMIT – Investigação em Economia, Gestão e Tecnologias da Informação.

References

[1] Cook, R.J., Lawless, J.F.: The analysis of recurrent event data. Springer, New York (2007). DOI:10.1007/978-0-387-69810-6
[2] de Uña-Álvarez, J., Meira-Machado, L.: A simple estimator of the bivariate distribution function for censored gap times. Stat Probab Lett, **78**, 2440–2445 (2008). DOI:10.1016/j.spl.2008.02.031
[3] Moreira, A., Meira-Machado, L.: Estimation of the bivariate distribution function for censored gap times. Commun Stat Simul Comput, **46**(1), 275–300 (2014). DOI:10.1080/03610918.2014.963609

Early Identification of Eating Disorders: The Contribution of Statistics to Clinical Decision Support

Vânia Almeida [1], Luís Machado [1,2[0000-0002-8577-7665]], and Andreia Gonçalves [3]
rafaelaalmeida.1611@gmail.com, lmachado@math.uminho.pt

[1] *Escola de Ciências, Universidade do Minho, Portugal*
[2] *CMAT – Centro de Matemática, Escola de Ciências, Universidade do Minho, Portugal*
[3] *Hospital da Senhora da Oliveira Guimarães*

Abstract: Eating disorders (ED) represent a growing public health issue, characterized by severe disturbances in eating behavior and a dysfunctional relationship with body image. Despite their clinical impact and increasing incidence, early detection still relies heavily on subjective methods, such as clinical interviews or self-report questionnaires, often leading to delayed diagnosis.
This study aims to integrate statistical modeling into clinical decision-making by developing predictive models for early detection of ED. Using a real dataset of 398 physically active adults, the study applies four classification algorithms: Logistic Regression, Decision Trees, Random Forest, and XGBoost. Results indicate that Body Image Distortion (BID) is the most influential predictor across all models. Among the algorithms, XGBoost demonstrated the best predictive performance, supporting its suitability for clinical risk assessment.

Keywords: Body image distortion · Clinical decision support · Eating disorders · Machine learning · Predictive modeling

References

[1] Breiman, L.: Random Forests. *Machine Learning*, **45**(1), 5–32 (2001).
[2] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
[3] Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, **29**(5), 1189–1232 (2001).
[4] Alcaraz-Ibáñez, M., Paterna, A., Sicilia, Á., Griffiths, M.D.: A Systematic Review and Meta-Analysis on the Relationship between Body Dissatisfaction and Morbid Exercise Behaviour. *International Journal of Environmental Research and Public Health*, **18**(2), 585 (2021).
[5] Corazza, O. et al.: The Emergence of Exercise Addiction, Body Dysmorphic Disorder, and Other Image-Related Psychopathological Correlates in Fitness Settings. *PLOS ONE*, **14**(4), e0213060 (2019).

A Comprehensive Exploratory Analysis on Pancreatic Adenocarcinoma

Isabel Fonseca [1[0009-0001-4022-3669]], Tiago Stoffel [1[0009-0009-9578-8040]], Raquel Mugeiro Silva [1[0009-0009-8019-0994]] and Eunice Carrasquinha [1,2[0000-0003-3465-4347]]
fc64397@alunos.fc.ul.pt, fc52949@alunos.fc.ul.pt, fc57945@alunos.fc.ul.pt, eitrigueirao@ciencias.ulisboa.pt

[1] *DEIO, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa 1749-016,Portugal*
[2] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa 1749-016, Portugal*

Abstract: Pancreatic cancer is a highly lethal gastrointestinal cancer characterized by a low 5-year survival rate associated with late stage diagnosis of the disease, usually detected by symptomatology [1]. The incidence and mortality due to this type of cancer are increasing on a worldwide scale [2]. In this work, we explore clinical data on pancreatic adenocarcinoma (PAD) from The Cancer Genome Atlas (TCGA), a public repository of genomic, transcriptomic, and clinical information across various cancer types. The main objective is to conduct an in-depth exploratory data analysis to identify trends and patterns in PAD, investigate potential associations between clinical variables, and detect subgroups of patients with similar profiles. For this purpose, dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (T-SNE), and Uniform Manifold Approximation and Projection (UMAP) were applied, followed by clustering methodologies. Based on the results obtained, classification models will be used to identify which clinical variables may be most relevant for predicting patient outcomes or subtypes.

Keywords: Classification models · Pancreatic cancer · Principal component analysis · T-SNE

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025 and UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020)

References

[1] Andersson, R., Haglund, C., Seppänen, H., Ansari, D.: Pancreatic cancer – the past, the present, and the future. Scandinavian Journal of Gastroenterology, **57**(10), 1169–1177 (2022). DOI: 10.1080/00365521.2022.2067786

[2] Hu, J.-X., Zhao, C.-F., Chen, W.-B., Liu, Q.-C., Li, Q.-W., Lin, Y.-Y., Gao, F.:Pancreatic cancer: A review of epidemiology, trend, and risk factors. World Journal of Gastroenterology, **27**(27), 4298–4321 (2021). DOI: 10.3748/wjg.v27.i27.4298

Os Padrões Espaço-temporais da Criminalidade de Rua e o Patrulhamento da Guarda Nacional Republicana

Duarte Branco [1][0009-0006-6186-0137], Ana Romão [1,2,4][0000-0002-9791-5172] e **Paula Simões** [2,3][0000-0002-1074-1297]

branco.dcs@gnr.pt, anaromao74@gmail.com, paula.simoes@academiamilitar.pt

[1] *Academia Militar, Instituto Universitário Militar, Portugal*
[2] *CINAMIL- Exército, Portugal*
[3] *CMA - Universidade NOVA de Lisboa, Portugal*
[4] *CICS.NOVA - Universidade Nova de Lisboa, Portugal*

Abstract: O mapeamento do crime tende a indiciar uma concentração espacial, em que uma reduzida proporção de locais, designados por "hotspots", registam uma parte significativa dos eventos criminais [1]. Em Portugal, a Guarda Nacional Republicana (GNR) reconhece a importância do policiamento baseado em informações para otimizar o patrulhamento e melhorar a prevenção da criminalidade. Este estudo incide sobre os registos de criminalidade de 2019 a 2024, no Destacamento Territorial de Almada, integrado na região de Lisboa e Vale do Tejo, que apresenta uma taxa de criminalidade acima da média nacional. Pretendemos investigar a distribuição espacial do crime de rua, testando o Princípio da Concentração do Crime bem como identificar padrões espaço-temporais que possam informar as estratégias de policiamento. A metodologia adotada, numa primeira fase, aplica técnicas de estatística descritiva, com recurso à Curva de Lorenz e ao Coeficiente de Gini. Posteriormente, recorre a técnicas de estatística inferencial, com implementação do Spatial Point Pattern Test [2, 3]. Os resultados evidenciam que 50% do crime se distribui num intervalo entre 0,9% e 7% do território, comprovando o Princípio da Concentração do Crime [1]. Concomitantemente, quatro dos cinco locais estudados, dentro do Destacamento, apresentaram, em pelo menos um dos anos, um Coeficiente de Gini próximo de 0.70, o que remete para a concentração do crime em microespaços [3]. Verificou-se, ainda, que a criminalidade apresenta similaridades na forma como se distribui ano após ano, tendo em conta a análise de células espaciais com e sem eventos. Tais resultados evidenciam contributos para as atividades de patrulhamento.

Keywords: Criminalidade · Padrões espaço-temporais · Patrulhamento

References

[1] Weisburd, D.: The law of crime concentration and the criminology of place. Criminology **53**(2), 133–157 (2015). https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9125.12070

[2] Ha, O., Andresen, M.: Spatial patterns of immigration and property crime in Vancouver: A spatial point pattern test. Canadian Journal of Criminology and Criminal Justice **62**(4), 30–51 (2020).

[3] Bernasco, W. and Steenbeek, W.: More places than crimes: Implications for evaluating the law of crime concentration at place. Journal of Quantitative Criminology **33**,451–467 (2017). https://doi.org/10.1007/s10940-016-9324-7

Escore de Propensão: Uma Aplicação do Programa Bolsa Presença

Rosemeire L. Fiaccone [1[0000-0001-5439-151]],
Italo Estrela de Souza Sá [1,2[0009-0000-6269-0279]], and Marcelo M. Taddeo [1]
fiaccone@ufba.br, itallo_estrella@hotmail.com, marcelo.taddeo@gmail.com

[1] *Instituto de Matemática e Estatística - UFBA / Brasil*
[2] *Superintendência de Estudos Econômicos e Sociais da Bahia (SEI) / Brasil*

Abstract: A mensuração do impacto de políticas públicas implica a análise de diferenças que resultem de uma determinada intervenção em termos dos resultados observados. O programa Bolsa Presença é uma iniciativa do Governo da Bahia (Brasil), via Secretaria de Educação (SEC), cujo objetivo é reduzir a evasão de estudantes da rede pública estadual da Bahia através da concessão de um benefício financeiro às famílias cadastradas no CadÚnico e em condições de vulnerabilidade socioeconômica. Este trabalho tem como objetivo, portanto, avaliar o impacto do Programa Bolsa Presença na redução do abandono escolar e à melhoria do desempenho acadêmico de estudantes da rede estadual. Para isso, são utilizadas metodologias estatísticas de inferência causal, capazes de lidar com a complexidade de intervenções tempo-dependentes. Em particular, aplicam-se métodos de ponderação baseados em escores de propensão, como o IPW (ponderação pelo inverso da probabilidade de tratamento) e o OW (ponderação pela área de suporte comum), com o objetivo de ajustar as diferenças entre os grupos de alunos que receberam a intervenção e que não receberam ao longo do tempo. Considerando que os efeitos podem variar entre diferentes perfis da população, o estudo incorpora uma análise causal de subgrupos, com o intuito de identificar heterogeneidades nos impactos do programa considerando a intervenção no *baseline* e também a intervenção tempo-dependente através da utilização dos modelos estruturais marginais para captar não somente esses efeitos que variam em diferentes estágios da intervenção, como também a variação dos efeitos entre os diferentes perfis populacionais incorporando análise causal de subgrupo. Este tipo de análise é poderosa, pois nos permite lidar simultaneamente com características particulares do subgrupo e com a variação temporal da intervenção. Neste trabalho, além da análise empírica do problema descrito acima, apresentaremos a fundamentação teórica da metodologia utilizada e a ilustraremos via estudos de simulação.

Keywords: Análise subgrupo causal · Escore de propensão · Inferência causal

Linear Regression Analysis of Harmonized IgG Antibody Levels Against the SARS-CoV-2 Spike Protein: A Cohort Study in Healthcare Workers

Ana L. Saraiva [1,2][0009-0000-7322-0838], Vera Afreixo [1,3][0000-0003-1051-8084], Ausenda Machado [2,4][0000-0002-1849-1499] and Vânia Gaio [2,4][0000-0001-7626-4991]
ana.saraiva@insa.min-saude.pt, vera@ua.pt, ausenda.machado@insa.min-saude.pt, vania.gaio@insa.min-saude.pt

[1] *Department of Mathematics (DMAT), University of Aveiro, Aveiro, Portugal*
[2] *Department of Epidemiology. Instituto Nacional de Saúde Doutor Ricardo Jorge. Lisbon.*
[3] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal*
[4] *Public Health Research Center. Escola Nacional de Saúde Pública. Universidade NOVA de Lisboa. Lisbon. & Comprehensive Health Research Center. Universidade NOVA de Lisboa. Lisbon.*

Abstract: The emergence of COVID-19 in 2019 led to the rapid development of vaccines and diagnostic tests. To assess antibody responses in healthcare workers (HCWs), a cohort study was conducted between 2021 and 2022 across three Portuguese hospitals. Antibody levels were measured at six time points: pre-vaccination, post-vaccination, and at 3, 6, and 12 months after the vaccination, as well as after the booster dose. Each hospital used a different assay: Abbott's CMIA, Roche's Elecsys® ECLIA, and Siemens' ADVIA Centaur®, posing challenges for data comparability. The study aimed to harmonize serological data across these hospitals and to model antibody increases and decreases over time using linear regression. To ensure adequate conversion of antibody titers from different laboratory methods, quantile harmonization, and Deming regression were applied. After harmonization, three linear regressions were fitted: one for the increase between pre-vaccination and post-vaccination, another for the decrease between post-vaccination and 12 months after vaccination, and finally, one for the increase between 12 months after vaccination and after the booster dose. Models included variables such as prior infection, age, hospital, smoking status, contact with COVID-19 patients, and chronic conditions. In the phase-specific analysis, in addition to variations between hospitals in the regression of the last increase after the booster dose, it was observed that individuals over 50 years of age exhibited a superior immune response (811 550; IC 95%: 598 774, 1 024 327; $p < 0.001$). This higher percentage increase may be explained by initially lower levels, unlike younger individuals who had higher titers.

Keywords: Harmonization · Healthcare workers · IgG antibody · SARS-CoV-2

Dor Musculosquelética e Sono em Profissionais de Saúde de Reabilitação: Aplicação do Modelo Beta-Binomial Inflacionado em Zero

Rute Teixeira [1[0009-0009-4479-9495]], **João P. Martins** [1,2,3[0000-0002-0474-1397]], Simão Ferreira [4[0000-0001-8233-2217]], Lucimere Bohn [5,6,7[0000-0001-7988-968X]] e Leonor G. Miranda [1[0000-0003-2882-0919]]

10210515@ess.ipp.pt, jom@ess.ipp.pt, sprf@ess.ipp.pt, lucimerebohn@fade.up.pt, lmiranda@ess.ipp.pt

[1] *Escola Superior de Saúde, Instituto Politécnico do Porto, rua Dr. António Bernardino de Almeida, 400, 4200-072 Porto, Portugal*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *Center for Health Studies and Research, University of Coimbra, Portugal*
[4] *RISE-Health, Center for Translational Health and Medical Biotechnology Research (TBIO), ESS, Polytechnic of Porto, R. Dr. António Bernardino de Almeida, 400, 4200-072 Porto, Portugal*
[5] *Lusofóna University, Faculty of Psychology, Education and Sport, Porto, Portugal*
[6] *Research Center in Physical Activity, Health and Leisure (CIAFEL) and Laboratory for Integrative and Translational Research in Population Health (ITR), Porto, Portugal*
[7] *Faculty of Sports; University of Porto, Porto, Portugal*

Abstract: Os profissionais de saúde em reabilitação apresentam, frequentemente, dor musculoesquelética e distúrbios do sono que podem comprometer o bem-estar e o desempenho profissional. Este estudo comparou a qualidade do sono e a intensidade da dor entre terapeutas ocupacionais e fisioterapeutas, avaliando também fatores preditores da qualidade do sono. Para tal, foi aplicado um questionário online que incluiu instrumentos Nórdico-Musculoesquelético e PSQI. A análise utilizou o modelo beta-binomial inflacionado em zero. Os terapeutas ocupacionais reportaram maior dor nos punhos. A dor nos punhos e joelhos, bem como o sexo feminino, associaram-se a pior qualidade do sono. Conclui-se que monitorização e intervenção precoce são essenciais para estes profissionais.

Keywords: Dor musculosquelética · Escala discreta · Inflacionamento em zero · Modelo beta-binomial · Sono

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UID/00006/2025 e UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Avaliação da Fiabilidade e Eficiência de Métodos de Classificação Hierárquicos versus Não Hierárquicos

Rui Santos [1,2][0000-0002-7371-363X], João Paulo Martins [2,3,4][0000-0002-0474-1397], Miguel Felgueiras [1,2,5][0000-0001-5450-7374] e Susana Ferreira [1][0000-0002-1077-5980]
rui.santos@ipleiria.pt, jom@ess.ipp.pt, mfelg@ipleiria.pt, susfer@ipleiria.pt

[1] *Escola Superior de Tecnologia e Gestão, Politécnico de Leiria*
[2] *CEAUL – Centro de Estatística e Aplicações, Universidade de Lisboa*
[3] *Escola Superior de Saúde, Instituto Politécnico do Porto*
[4] *Centro de Estudos e Investigação em Saúde, Universidade de Coimbra*
[5] *Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Universidade de Aveiro*

Abstract: Os métodos de classificação baseados em testes compostos podem proporcionar poupanças significativas de recursos, mas a sua fiabilidade tende a diminuir à medida que o tamanho do grupo aumenta. Este estudo recorre a simulações em R para explorar os compromissos entre eficiência (medida pelo número de testes realizados) e fiabilidade em várias estratégias de classificação.
Avaliamos métodos de classificação hierárquicos e não hierárquicos, considerando diferentes configurações, como o número e dimensão dos subgrupos, bem como a presença ou ausência de *master pool*. As simulações incluem testes qualitativos (assumindo que a sensibilidade composta é igual à do teste individual) e testes quantitativos (considerando os efeitos de diluição), abrangendo diversas taxas de prevalência e dimensões de grupo. Analisamos também o impacto de diferentes distribuições da substância discriminante e vários níveis de qualidade dos testes.
Os resultados sublinham a importância de equilibrar eficiência e fiabilidade: as estratégias mais eficazes são geralmente aquelas que conseguem uma poupança substancial de testes sem comprometer a fiabilidade.

Keywords: Classificação · Efeito de diluição · Simulação · Testes compostos

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/00006/2025 e UIDB/00006/2020. https://doi.org/10.54499/UIDB/00006/2020.

References

[1] Martins, J.P., Santos, R., Sousa, R.: Testing the maximum by the mean in quantitative group tests. In: New Advances in Statistical Modeling and Applications, Springer-Verlag, pp. 55–63. (2014)
https://doi.org/10.1007/978-3-319-05323-3_5
[2] Santos, R., Pestana, D., Martins, J.P.: Extensions of Dorfman's Theory. In: Recent Developments in Modeling and Applications in Statistics, Springer-Verlag, Springer-Verlag, pp. 179–189. (2013)
https://doi.org/10.1007/978-3-642-32419-2_19
[3] Santos, R., Martins, J., Felgueiras, M., Ferreira, L.: Accuracy Measures for Binary Classification Based on a Quantitative Variable, REVSTAT-STAT J **17**(2), pp. 223–244. (2019) https://doi.org/10.57805/revstat.v17i2.266

Prevalence and Risk Factors of Subretinal Drusenoid Deposits in Age-related Macular Degeneration: The Coimbra Eye Study

Rita Coimbra [1[0000-0002-3977-4164]], Cláudia Farinha [1,2,3,4[0000-0003-4596-0913]], Alina Humenyuk [1[0000-0002-9105-3751]], Patrícia Barreto [1,4[0000-0002-3375-551X]], Rufino Silva [1,2,3,4[0000-0001-8676-0833]]

racoimbra@aibili.pt, cvfarinha@aibili.pt, ahumenyuk@aibili.pt, pbarreto@aibili.pt, rufino.silva@oftalmologia.co.pt

[1] *AIBILI - Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal*
[2] *Ophthalmology Department, Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra, Portugal*
[3] *Clinical Academic Center of Coimbra (CACC), Coimbra, Portugal*
[4] *University of Coimbra, Coimbra Institute for Clinical and Biomedical Research.*

Abstract: Age-related macular degeneration (AMD) is a degenerative disease and the leading cause of irreversible vision loss in people aged over 55, in western countries. AMD is a multifactorial disease, for which genetic, clinical and lifestyle factors contribute.

To explore the prevalence and associated risk factors for subretinal drusenoid deposits (SDD) in the Lousã cohort of the Coimbra Eye Study (AMD_LifeGene, NCT05735730).

A total of 389 eyes from 218 participants with AMD diagnosis were analyzed using the Rotterdam Classification. Multivariable logistic regression with generalized estimating equations (GEE) was used to assess associations between the presence of SDD and potential risk factors, including age, sex, BMI, smoking status, hypertension, diabetes, adherence to the Mediterranean diet and physical activity, and adjusted for inter-eye correlation.

SDD was present in 113 eyes (29.0%) of all AMD cases. Older age was significantly associated with increased odds of SDD (OR = 1.11; 95% CI: 1.05–1.17). High adherence to the Mediterranean diet was associated with a 71% reduction in the odds of SDD compared to low adherence (OR = 0.29; 95% CI: 0.12–0.73). The model showed a good discriminative ability (AUC = 0.776; 95% CI 0.723 to 0.830).

SDD are recognized as biomarkers of faster AMD progression and higher risk of developing late-stage disease. These results highlight the protective effect of high adherence to the Mediterranean diet in reducing the risk of SDD in AMD patients. Longitudinal studies are needed to further explore causal relationships between lifestyle factors, aging, and SDD development.

Keywords: Age-related macular degeneration · Generalized estimating equations · Risk factors

Dietary Patterns in a Population-Based Cohort: The Coimbra Eye Study

Alina Humenyuk [1[0000-0002-9105-3751]], Patrícia Barreto [1,2[0000-0002-3375-551X]], Cláudia Farinha [1,2,3,4[0000-0003-4596-0913]], Rita Coimbra [1[0000-0002-3977-4164]] and Rufino Silva [1,2,3,4[0000-0001-8676-0833]]

ahumenyuk@aibili.pt, pbarreto@aibili.pt, cvfarinha@aibili.pt, racoimbra@aibili.pt, rufino.silva@oftalmologia.co.pt

[1] *AIBILI - Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal*
[2] *University of Coimbra, Coimbra Institute for Clinical and Biomedical Research.*
[3] *Ophthalmology Department, Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra, Portugal*
[4] *Clinical Academic Center of Coimbra (CACC), Coimbra, Portugal*

Abstract: Age-related macular degeneration (AMD) stands as the leading cause of blindness in individuals over 55 years across Western countries. Emerging evidence suggests that diet plays a crucial role in influencing both the risk and progression of AMD. This study aimed to identify dietary patterns within the inland Lousã cohort of the Coimbra Eye Study (AMD_LifeGene, NCT05735730). Dietary intake data were collected from 1053 participants using a food frequency questionnaire, providing average daily intake estimates (g/day) for each food item. Twenty-six food groups were created based on nutritional similarity and analyzed using principal component analysis (PCA). The number of retained dietary patterns was determined by eigenvalues greater than 1.2, scree plot inflection points, and pattern interpretability. Varimax rotation was applied. Three distinct dietary patterns emerged. The first, a less healthy pattern, was marked by low intake of olive oil, vegetables, salad, and fruit, although it also featured low snack consumption. The second, a healthier pattern, was characterized by higher intake of yogurt and cereals, along with reduced intake of sugar, coffee, red meat, breads, starchy sides, and alcoholic beverages. The third, a fish-based pattern, featured higher intake of fish and lower consumption of sweets and soft drinks. These findings provide a current snapshot of dietary habits in the Lousã cohort and offer a valuable basis for assessing changes in dietary patterns compared to data collected a decade ago. Future analyses will also investigate whether adherence to these patterns is associated with the presence or absence of AMD.

Keywords: Age-related macular degeneration · Dietary patterns · Principal component analysis

Patient Engagement with a Digital Decision-Support Tool in Breast Cancer Surgery

Miguel Broes [1[0009-0007-7867-5242]],
Giovani Silva [1,2[0000-0002-7434-2383]], and Marília Antunes [1,3[0000-0002-1257-2829]]
fc54776@alunos.fc.ul.pt, giovani.silva@tecnico.ulisboa.pt,
marilia.antunes@ciencias.ulisboa.pt

[1] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[2] *Dep. Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[3] *Dep. Estatística e Investigação Operacional, FCUL, Universidade de Lisboa*

Abstract: The CINDERELLA Project (https://cinderellaproject.eu) is an innovative EU-funded initiative aimed at improving aesthetic outcomes and patient satisfaction in breast cancer surgery, carried out across five countries — Portugal, Italy, Germany, Poland, and Israel — with the goal of advancing shared decision-making in breast cancer treatment, reducing the need for corrective surgeries, and setting a new standard for aesthetic evaluation in healthcare. The project presents an innovative AI-powered tool—the CINDERELLA APP-developed to support shared decision-making in breast cancer care. The CINDERELLA trial evaluates the impact of a Digital Health intervention on patients' satisfaction with the aesthetic outcomes of locoregional treatment, focusing on how well patients' expectations align with actual results. It also explores the influence of the app on quality of life and psychological well-being. The APP includes Educational Modules with information about breast cancer, treatment options (especially surgery and reconstruction), and potential aesthetic outcomes. The modules were prepared in a patient-friendly language with multimedia, understandable by most patients, but with different levels of complexity and extension. In this way, it is important to assess the acceptability, usability, clinical relevance, and patient satisfaction of the app. In this work, we focus on patient engagement and interaction with the CINDERELLA APP during the first twenty-two months of recruitment in the intervention group. In particular, we explore usage patterns that depend on sociodemographic and clinical characteristics of the patients such as age, education, marital status, and type of surgery planned, among others.

Keywords: CINDERELLA project · Digital health · Usability patterns
Acknowledgements: This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement number 101057388. The work by GS and MA is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025, UIDB/00006/2025, current DOI: 10.54499/UIDB/00006/2020 - https://doi.org/10.54499/UIDB/00006/2020.

References

[1] Ge, X., Peng, Y., tu,D.: A threshold mixed-effects Tobit model for treatment-sensitive subgroup identification based on longitudinal measures with floor and ceiling effects and a continuous covariate. Journal of Statistical Computation and Simulation, **Vol. 94**(Issue 11), 2544–2563 (2024). https://doi.org/10.1080/00949655.2024.2344126

Classifying App Usage Data with Finite Mixture Models

Ana Sofia Barata [1[0009-0005-8960-2968]], Giovani Silva [1,2[0000-0002-7434-2383]], and Marília Antunes [1,3[0000-0002-1257-2829]]
fc62520@alunos.ciencias.ulisboa.pt, giovani.silva@tecnico.ulisboa.pt, marilia.antunes@ciencias.ulisboa.pt

[1] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[2] *Dep. Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[3] *Dep. Estatística e Investigação Operacional, FCUL, Universidade de Lisboa*

Abstract:
The CINDERELLA project (https://cinderellaproject.eu) is an innovative EU-funded initiative designed to enhance aesthetic outcomes and patient satisfaction in breast cancer surgery. The project integrates an AI-driven platform designed to help patients make informed choices about their surgical options by predicting aesthetic outcomes based on personal data and a repository of pre- and post-surgical images. The study is conducted in five countries, Portugal, Italy, Germany, Poland, and Israel, with the objective of advancing shared decision making in breast cancer treatment, reducing the need for corrective surgeries, and setting a new standard for aesthetic evaluation in healthcare. An app was developed within the CINDERELLA project, with the aim of, among other ones, providing women information about the disease, treatment, and answering several questions that often arise. Hence, analysis of app usage patterns is relevant for understanding user behavior. However, data on app usage time may not always reflect actual engagement. Users may leave the app open while engaged in other activities, leading to inflated usage times (outliers). Therefore, app usage data can be thought of as a mixture of two types of records: those that represent exact usage and those that are inflated. Finite mixture models offer a robust framework for addressing classification challenges when the underlying process is not fully observable. This study will apply and extend finite mixture modeling techniques to classify app usage time into exact and inflated categories.

Keywords: CINDERELLA project · Inflated usage times · Mixture models · Outliers

Acknowledgements: This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement number 101057388. The work by GS ans MA is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025, UIDB/00006/2025, current DOI: 10.54499/UIDB/00006/2020 - https://doi.org/10.54499/UIDB/00006/2020.

The Impact of AI-Assisted Decision-Making on Patient Satisfaction: A BREAST-Q Study

Carolina Horta [1[0009-0002-4851-883X]], Giovani Silva [1,2[0000-0002-7434-2383]], and Marília Antunes [1,3[0000-0002-1257-2829]], on behalf of the CINDERELLA Consortium
fc56707@alunos.fc.ul.pt, giovani.silva@tecnico.ulisboa.pt, marilia.antunes@ciencias.ulisboa.pt

[1] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[2] *Dep. Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[3] *Dep. Estatística e Investigação Operacional, FCUL, Universidade de Lisboa*

Abstract: The CINDERELLA project (https://cinderellaproject.eu) is an innovative EU-funded initiative aimed at improving aesthetic outcomes and patient satisfaction in breast cancer surgery. The project integrates an AI-driven platform designed to help patients make informed choices about their surgical options by predicting aesthetic outcomes based on personal data and a repository of pre- and post-surgical images. The study is carried out across five countries—Portugal, Italy, Germany, Poland, and Israel—with the goal of advancing shared decision-making in breast cancer treatment, reducing the need for corrective surgeries, and setting a new standard for aesthetic evaluation in healthcare. A key feature of the project is a clinical trial comparing standard pre-surgical counseling with the use of the CINDERELLA platform. The intervention arm uses the AI tool to educate patients, providing them with visual and objective assessments of likely surgical outcomes, while the control group receives conventional information. The BREAST-Q Version 2.0, a validated and widely used patient-reported outcome measure (PROM) specifically designed for patients undergoing breast surgery, was applied to assess the effect of the intervention on patient satisfaction and well-being. Patients completed the BREAST-Q questionnaire at four time points: pre-operatively and post-operatively at wound healing, 6 months, and 12 months. This work focuses on the analysis of longitudinal data using Tobit-like models to compare outcomes between the two arms of the trial.

Keywords: BREAST-Q · CINDERELLA Project · Decision making · Tobit models

Acknowledgements: This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement number 101057388. The work by GS and MA is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UID/00006/2025, UIDB/00006/2025, current DOI: 10.54499/UIDB/00006/2020 - https://doi.org/10.54499/UIDB/00006/2020.

References

[1] Ge, X., Peng, Y., Tu, D.: PA threshold mixed-effects Tobit model for treatment-sensitive subgroup identification based on longitudinal measures with floor and ceiling effects and a continuous covariate. Journal of Statistical Computation and Simulation, **94**(11), 2544-2563 (2024). https://doi.org/10.1080/00949655.2024.2344126

Intervalos de Referência por Métodos Indiretos

Lara Pereira [1], Beatriz Saraiva [1],Henrique Reguengo [3], Ricardo Ribeiro [3], Margarida Brito [1,2[0000-0001-7453-3289]] e Rita Gaio [1,2[0000-0003-3906-0775]]
up202408394@edu.fc.up.pt, up202408844@edu.fc.up.pt,
henrique.reguengo.sqc@chporto.min-saude.pt,
ricardoribeiro.sqc@chporto.min-saude.pt, mabrito@fc.up.pt, argaio@fc.up.pt

[1] *Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal*
[2] *Centro de Matemática da Universidade do Porto*
[3] *Centro Hospitalar Universitário de Santo António*

Abstract: Os intervalos de referência (IRs) são ferramentas fundamentais na interpretação de resultados laboratoriais, ao definirem limites para os valores de indivíduos saudáveis e assim permitirem detetar anomalias. Podem ser determinados por métodos diretos através de coortes de indivíduos saudáveis. Não sendo possível, por limitações logísticas, éticas e/ou financeiras, usam-se métodos indiretos (introduzidos por R. G. Hoffmann, 1963), partindo de dados laboratoriais.
Este estudo tem como objetivo explorar e comparar diferentes metodologias indiretas de estimação de IRs, recorrendo a dados simulados. A análise considera as vantagens e desvantagens de cada abordagem, bem como a sua adequação às características do conjunto de dados. Adicionalmente, é avaliado o desempenho de cada método.
O trabalho procura contribuir para uma aplicação mais informada e eficaz destas abordagens, com impacto direto na prática laboratorial e no auxílio da tomada de decisão clínica.

Keywords: Intervalos de referência · Métodos indiretos · Prática laboratorial

Acknowledgements: Rita Gaio e Margarida Brito foram parcialmente apoiadas pelo CMUP, membro do LASI, financiado por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto com referência UID/00144.

References

[1] Hoffmann, R.G.: Statistics in the practice of medicine. Jama **185**(11), 864–873 (1963). DOI:10.1001/jama.1963.03060110068020
[2] Ma, S., Yu, J., Qin, X., Liu, J.: Current status and challenges in establishing reference intervals based on real-world data. *Critical Reviews in Clinical Laboratory Sciences* **60**(6), 427–441 (2023). DOI:10.1080/10408363.2023.2195496

Functional Dependence at Admission as a Prognostic Factor in Palliative Care: A Survival Analysis

Nuno Domingues [1[0009-0006-9719-114X]], Joana Bragança [2,3,4[0000-0002-7795-4131]], Ivo Sousa-Ferreira [5,6[0000-0001-5526-3594]], and Tiago Dias Domingues [1,6[0000-0002-4034-4276]]
fc52631@alunos.ciencias.ulisboa.pt, joanafigbrag@gmail.com, ivo.ferreira@staff.uma.pt, tmdomingues@ciencias.ulisboa.pt

[1] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *Health Sciences Institute – Universidade Católica Portuguesa, Lisboa, Portugal*
[3] *CIIS – Centre for Interdisciplinary Research in Health, Lisboa, Portugal*
[4] *Hospital da Luz, Lisboa, Portugal*
[5] *Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal*
[6] *CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal*

Abstract: The present study aimed to assess the potential of functional dependence at admission, measured by the Barthel Index, as a predictor of mortality in oncologic (O) and non-oncologic (NO) palliative care patients. This retrospective study analysed data from 886 patients admitted to a palliative care unit at a private hospital in Lisbon, Portugal. Exact survival times, defined as the time from admission to death, were recorded for all patients. To investigate the effect of functional dependence on patients' survival, Cox proportional hazards (PH) models were fitted, adjusting for gender and age. The PH assumption was evaluated using the Schoenfeld residuals, applying both graphical methods and the Grambsch-Therneau test. When this assumption did not hold, the time axis was divided into intervals, and a piecewise Cox PH model was fitted. Survival analysis revealed that, among O patients, the effect of functional dependence on survival was non-proportional over time ($p < 0.001$). A piecewise Cox PH model indicated that total dependence at admission was associated with a significantly higher risk of death ($p < 0.001$) during the first 7 days of hospitalisation. In contrast, for NO patients, the PH assumption held, and the traditional Cox model showed that total dependence was associated with lower overall survival ($p = 0.009$). The findings underscore the importance of assessing functional dependence at admission as a prognostic factor in palliative care and motivate further research into the time-varying effects of functional dependence on survival outcomes.

Keywords: Non-proportional hazards · Oncology nursing · Palliative care · Proportional hazards model · Survival analysis

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UID/00006/2025 and UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Sessão de Posters II

Robustness Evaluation of Machine Learning Models in Genomic Prediction

Vanda M. Lourenço [1[0000-0001-8338-7279]], Joseph O. Ogutu [2[0000-0002-7379-0387]], and Hans-Peter Piepho [2[0000-0001-7813-2992]]
vmml@fct.unl.pt, jogutu2007@gmail.com, hans-peter.piepho@uni-hohenheim.de

[1] *Department of Mathematics and NOVA Math, NOVA FCT, NOVA University of Lisbon, Portugal*
[2] *Biostatistics Unit, University of Hohenheim, Germany*

Abstract: Accurate genomic prediction (GP) of breeding values is essential in modern plant and animal breeding programs. GP relies on thousands of molecular markers (e.g., Single Nucleotide Polymorphisms) distributed across the genome, requiring computational methods capable of handling high-dimensional data. In this context, machine learning (ML) has emerged as a powerful framework due to its flexibility and ability to model complex genetic architectures. While many studies have compared the predictive performance of individual ML algorithms, few have offered broader evaluations across diverse methodological approaches—particularly with respect to robustness under data contamination. Yet, in practical breeding applications, data quality is often imperfect, and accuracy depends not only on model fit but also on a method's ability to perform reliably in the presence of noise.
This study addresses these gaps by evaluating the predictive accuracy and robustness of a range of supervised ML methods. Using simulated data from an animal breeding population, we assess performance across varying levels of data contamination, focusing on both prediction accuracy and error metrics. The results offer new insights into the relative strengths and limitations of different ML approaches under realistic conditions. These findings provide practical guidance for selecting robust and effective methods for genomic prediction in breeding applications.

Keywords: Breeding studies · Genomic prediction · Machine learning · Robustness · Single Nucleotide Polymorphisms

Acknowledgements: This work was partially funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math - Center for Mathematics and Applications) and project REACTION - 2023.14934.PEX (https://doi.org/10.54499/2023.14934.PEX).

References

[1] Lourenço, V.M., *et al*: Genomic prediction using machine learning: A comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. BMC Genomics, 25(1):152, 2024. DOI:10.1186/s12864-023-09933-x
[2] Ogutu, J.O., Piepho, H-P.: Regularized group regression methods for genomic prediction: Bridge, mcp, scad, group bridge, group lasso, sparse group lasso, group mcp and group scad. BMC Proceedings, 8(5):1–9, 2014. DOI:doi.org/10.1186/1753-6561-8-S5-S7

Optimizing Herbicide Use in Precision Agriculture through Classification Models

Alexandre Aparecido da Silva [1][0009-0001-5892-4033] and Luiz Fernando Carvalho [1][0000-0002-3982-0917]

alexandreaparecido@alunos.utfpr.edu.br, luizfcarvalho@utfpr.edu.br

[1] *UTFPR – Universidade Tecnológica Federal Paraná*

Abstract: The presence of weeds poses one of the main challenges in sugarcane production, potentially causing yield losses exceeding 40% [1]. With the advancement of remote sensing technologies and applied statistics, it has become feasible to employ classification methods for the accurate mapping of infested areas. In this study, we propose the use of a statistical model based on classification algorithms, such as Random Forest, to differentiate various types of vegetation in cultivated fields. As inputs, four spectral bands were obtained using a multispectral camera mounted on an unmanned aerial vehicle (UAV). The modeling approach enabled the accurate identification of regions infested with invasive species such as Brachiaria decumbens and Panicum maximum, thus allowing for the site-specific application of herbicides. Field experiments demonstrated that the use of the model resulted in up to a 57% reduction in the application of chemical agents. The results underscore the potential of advanced statistical modeling as a decision-support tool for precision agriculture and sustainable field management.

Keywords: Modeling · Precision agriculture · Remote sensing

Acknowledgements: This work is partially financed by Federal University of Technology – Paraná (UTFPR), with additional assistance from the LAMAP and LTEX laboratories.

References

[1] Kuva, M.A., Gravena, R., Pitelli, R.A., Christoffoleti, P.J., Alves, P.L.C.A.: Interference periods of weeds in the sugarcane crop: III – Brachiaria decumbens and Panicum maximum. *Planta Daninha*, **21**(1), 37–44 (2003). https://doi.org/10.1590/S0100-83582003000100005

From Behaviour to Personas: A Machine Learning Approach to Understand Adaptive Thermal Comfort Strategies in Workspaces

Celina P. Leão [1[0000-0003-3725-5771]], Lumy Noda [2[0000-0002-0395-6206]], Amanda V. P. Lima [2[0000-0001-8406-210X]], and Solange Leder [2[0000-0003-3784-4461]]
cpl@dps.uminho.pt, lumynoda@gmail.com,
amandavieiraarquitetura@gmail.com, solange.leder@academico.ufpb.br

[1] *Centro ALGORITMI/LASI, University of Minho, Guimarães, Portugal*
[2] *Federal University of Paraíba, João Pessoa, Brazil*

Abstract: Building on previous statistical research into thermal comfort and adaptive behaviour during remote work in tropical climates, this study proposes a machine learning–driven framework for constructing personas—conceptual user profiles that represent patterns of behavioural adaptation. The analysis draws on a dataset of 174 participants monitored during remote work, encompassing office settings in residences located in João Pessoa, Brazil, a city with a hot and humid climate. Unsupervised learning algorithms, namely K-Means and Hierarchical Clustering implemented using Orange data mining software, were applied to behavioural, demographic, and subjective comfort data. The objective was to uncover latent subgroups of occupants who adopt similar strategies to regulate indoor thermal comfort, such as adjusting clothing, using ventilation, or other. These personas are not raw data but conceptual models that synthesise statistically distinct behavioural profiles derived from real-world measurements, including questionnaire responses and environmental variables (e.g., temperature, humidity). Drawing inspiration from prior work on occupant behaviour modelling and from applications of personas in fields such as tourism to interpret stakeholders' motivations [1], the resulting clusters were incorporated into supervised learning models to predict the number and type of actions taken to achieve thermal comfort. The results revealed that different personas, such as "Autonomous Adapters," "Passive Dependents," and "Multi-Strategists", demonstrate significantly different behavioural patterns. This differentiation enhanced the interpretability and predictive capacity of the models. This integrative approach demonstrates how statistical clustering, behavioural theory, and explainable AI can converge to generate actionable insights. Personas serve as a meaningful intermediate layer between raw data and design strategies, offering a novel pathway to more occupant-centred, energy-efficient building operations.

Keywords: Adaptive behaviour · Machine learning · Personas · Thermal comfort

Acknowledgements: This work has been supported by FCT within the R&D Unit Project Scope UID/00319/Centro ALGORITMI (ALGORITMI/UM).

References

[1] Karolita, D., McIntosh, J., Kanij, T., Grundi, J., Obie, H.O.: Use of Personas in Requirements Engineering: A Systematic Mapping Study. Information and Software Technology, **162**, 107264 (2023). doi:10.1016/j.infsof.2023.107264

A Comparison Between Location Models and Regression Structures

Gabriel Reis Macedo [1[0009-0001-9739-133X]] and Luiz F. Carvalho [1,2[0000-0002-3982-0917]]
gabrielreismacedo@alunos.utfpr.edu.br, luizfcarvalho@utfpr.edu.br

[1] *UTFPR – Universidade Tecnológica Federal do Paraná, Apucarana*

Abstract: The present work provides a comprehensive analysis of regression models, with a particular emphasis on those belonging to the location family. The central motivation arises from the observation that relatively simple probability distributions—often characterized by greater interpretability—can, in certain cases, outperform more complex distributional forms. For instance, the Reverse Gumbel (RG) distribution can be effectively incorporated into the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework, which enables the modeling of multiple distributional parameters, including mean, mode, variance, and other higher-order moments. To investigate this hypothesis, three empirical datasets are analyzed, comparing a range of location-based models with the RG distribution within the GAMLSS architecture. The primary objective is to illustrate that employing a simpler distributional form—such as the RG—within a flexible and robust regression framework, like GAMLSS [2], may yield superior performance and interpretability relative to more complex location models, including those from the xgamma family [1].

Keywords: Empirical datasets · Model comparison · Simple distributions · Statiscal analysis

Acknowledgements: This work is partially financed by Federal University of Technology – Paraná (UTFPR), with additional assistance from the LAMAP and LTEX laboratories.

References

[1] Cordeiro, G.M.; Altun, E.; Korkmaz, M.C.; Pescim, R.R.; Afify, A.Z.; Yousof, H.M. The xgamma Family: Censored Regression Modelling and Applications. *Revstat. Stat. J.* **2020**, *18*, 593–612. https://doi.org/10.57805/revstat.v18i4.63

[2] Ramires, T.G.; Nakamura, L.R.; Righetto, A.J.; Pescim, R.R.; Mazucheli, J.; Rigby, R.A.; Stasinopoulos, D.M. Validation of stepwise-based procedure in GAMLSS. *J. Data Sci.* **2021**, *19*, 96–110. https://doi.org/10.6339/JDS.202101_19(1).0006

Applying Clustering Approaches to GAMLSS

Iago Macarini [1][0009-0006-1597-3130] and
Luiz Fernando Carvalho [1][0000-0002-3982-0917]
iagomacarini@alunos.utfpr.edu.br, luizfcarvalho@utfpr.edu.br

[1] *UTFPR - Universidade Tecnológica Federal do Paraná - Apucarana*

Abstract: This research introduces c-GAMLSS, an extension of Generalized Additive Models for Location, Scale, and Shape (GAMLSS) [1], specifically designed for the statistical analysis of multimodal and highly distorted data [2]. The core innovation lies in incorporating a latent 'cluster' variable to explain the response variable, allowing all distributional parameters of the response to be modeled as functions of this new covariate, alongside other available features. The method employs a step-based approach for resource selection, and a comprehensive simulation study demonstrates c-GAMLSS's superior performance compared to traditional Gaussian mixture models. Furthermore, through four diverse data applications, c-GAMLSS consistently outperforms results obtained with mixture models, recently developed complex distributions, cluster-weighted models [3], and mixture of experts models, both when authentic explanatory variables are present and when they are not, showcasing its robustness and enhanced quality, even when utilizing simple distributions that can be readily extended to more sophisticated ones.

Keywords: GAMLSS · Latent variable · Mixture models · Multimodal data · Statistical modeling

Acknowledgements: This work is patially financed by Federal University of Technology - Paraná (UTFPR), wich additional assistance from the LAMAP and LTEX laboratories.

References

[1] Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C (Appl. Stat.), **54**(3), 507–554 (2005). DOI:10.1111/j.1467-9876.2005.00510.x.

[2] Ramires, T.G., Ortega, E.M.M., Cordeiro, G.M., Hens, N.: A bimodal flexible distribution for lifetime data. J. Stat. Comput. Simul., **86**(12), 2450–2470 (2016). DOI:10.1080/00949655.2015.1115047.

[3] Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P.D.: Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Stat. Methods Appl., **24**(4), 623–649 (2015). DOI:10.1007/s10260-015-0311-1.

Multiplicative Algebra of Random Variables, Contraction and Expansion

Dinis Pestana [1,2,3][0000-0001-8999-1354], Sandra Mendonça [2,4][0000-0003-3364-0357], and Neto Pascoal [2,5][0000-0001-8616-0520]
ddpestana@ciencias.ulisboa.pt, sandram@staff.uma.pt, polenepascoal@gmail.com

[1] *DEIO, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2] *CEAUL — Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *Instituto de Investigação Científica Bento da Rocha Cabral, Lisboa, Portugal* [4] *Departamento de Matemática — FCEE, Universidade da Madeira, Campus Universitário da Penteada, 9020-105, Funchal, Portugal*
[5] *Faculdade de Ciências e Tecnologia, Universidade Zambeze, Moçambique*

Abstract: Let X and Y with support [0,1] be independent random variables. Then $XY \prec X \prec \frac{X}{Y}$. If $X \sim Gaussian(0,1)$ and $Y \sim Uniform(0,1)$ the slash random variable X/Y expansion of X has heavier tails and is useful in robustness studies. If both X and Y have support [0,1], denoting $\widetilde{X} = 1 - X$ and $\widetilde{Y} = 1 - Y$, the minimum of X/Y and $\widetilde{X}/\widetilde{Y}$ and $X + Y \bmod 1$ are useful in data-augmentation and in the improvement of pseudo-random numbers generation when X and/or Y are standard Uniform. Some general results when X and/or Y are order statistics of the standard Uniform, or more generally Beta, BetaBoop, Kumaraswamy or Mendel random variables are investigated, both for the purpose of improving robustness studies and for modelling in the meta-analysis of genuine and fake p-values [1, 2].

Keywords: BetaBoop and Mendel random variables · Contraction and expansion of random variables · Data-augmentation · Fake p-values · Uniform order statistics

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

References

[1] Brilhante, M.F.; Mendonça, S.; Pestana, D.; Rocha, M. L.; Sequeira, F.; Velosa, S.: Fake p-Values and Mendel Variables: Testing Uniformity and Independence. In: International Conference on Mathematical Applications, Funchal, Portugal, 23–28 (2018). https://iknowd.org/wp-content/uploads/submissions/icma18/icma18_1_IKnowD_20181130174603.pdf

[2] Brilhante, M.F., Gomes, M.I., Mendonça, S.; Pestana, D.; Santos, R.: Meta-analysis of Genuine and Fake p-Values. J Stat Theory Pract **19**, (29) (2025). https://doi.org/10.1007/s42519-025-00445-3

Transformação de Dados na Análise Estatística: Desenvolvimento de uma Framework de Apoio à Decisão

João Correia [1[0000-0002-2769-0887]], M. Rosário Ramos [1,2[0000-0001-9114-0807]], P. Engrácia [3,4[0000-0002-2673-3216]]

joaopmcorreia@proton.me, MariaR.Ramos@uab.pt, Patricia.Martins.Engracia@iscte-iul.pt

[1] *Universidade Aberta, Portugal*

[2] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

[3] *ISCTE - Instituto Universitário de Lisboa, Portugal*

[4] *ISTAR-ISCTE - Centro de Investigação em Ciências da Informação, Tecnologias e Arquitetura, Portugal*

Abstract: As transformações de dados são funções aplicadas para modificar a distribuição e a escala de dados, de modo a satisfazer requisitos específicos e permitir análises estatísticas mais significativas.

Este trabalho apresenta uma análise abrangente dos diversos métodos de transformação de dados, com foco na sua aplicabilidade em contextos estatísticos. Reconhecendo a importância da preparação adequada dos dados, são exploradas as implicações que a escolha de transformações pode ter na qualidade e validade das análises.

Como principal contributo, foi desenvolvida uma "framework" de apoio à seleção de transformações de dados, integrando características fundamentais dos dados — como distribuição, tipo, formato e escala — e incluindo os requisitos específicos das análises a realizar. A "framework", direcionada a uma comunidade variada de utilizadores, foi complementada por uma ferramenta "web" interativa, desenvolvida em linguagem Python, concebida para agilizar e sistematizar o processo de escolha da transformação mais adequada.

A abordagem proposta é ilustrada com a aplicação a casos de estudo, permitindo avaliar o desempenho dos métodos em contextos variados. Os resultados evidenciam que a eficácia das transformações depende fortemente das características dos dados e dos objetivos analíticos. Demonstra-se assim a importância de adotar uma estratégia com critério e bem fundamentada na escolha dessas transformações, inclusive reconhecendo que, em certos casos, a melhor decisão pode ser não aplicar qualquer transformação.

Keywords: Análise estatística · Pré-processamento · Transformação de dados

Welch t and Power Means

Sílvio Velosa [1][0000-0002-7853-3822], Sandra Mendonça [1,2][0000-0003-3364-0357]
silviov@staff.uma.pt, sandram@staff.uma.pt

[1] *DM-FCEE, Universidade da Madeira*
[2] *Centro de Estatística e Aplicações (CEA/UL)*

Abstract: The familiar Welch t statistic is associated with an unbiased estimator of the variance of the difference between the sample means of two independent normal random samples. Ames and Oliveira [1] have expressed this as the arithmetic average of another set of unbiased estimators of the same variance, which arise in an alternative solution to the Behrens-Fisher problem due to Scheffé [2, 3]. Oliveira et al.[4] suggested replacing the arithmetic average with more general power means. We study the mean squared error of this family of estimators.

Keywords: Behrens-Fisher problem · Bias · Generalized means · Scheffé · Welch t

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

References

[1] Ames, M. H. More on the Means Comparison with Unequal Variances Problem, *Journal of Biopharmaceutical Statistics*, 6, 177–183, 1996. DOI:10.1080/10543409608835131

[2] Scheffé, H. On Solutions of the Behrens-Fisher Problem, Based on the t-Distribution. *The Annals of Mathematical Statistics*, 14, 35–44, 1943. DOI:10.1214/aoms/1177731490

[3] Scheffé, H. A Note on the Behrens-Fisher Problem. *The Annals of Mathematical Statistics*, 15, 430–432, 1944. DOI:10.1214/aoms/1177731214

[4] Oliveira, J. L. N., Velosa, S. F. An Identity between Welch's Approximate and Scheffé's Exact Solutions to the Behrens-Fisher Problem. *Estatística: Desafios Transversais às Ciências com Dados — Atas do XXIV Congresso da Sociedade Portuguesa de Estatística*, 183–195, 2021.

Estandardização em Modelos Lineares: Quando é Útil, Quando é Neutra e Quando Atrapalha?

Dulce Pereira [1][0000-0001-7281-4992] e Anabela Afonso [1][0000-0002-5517-4855]
dgsp@uevora.pt, aafonso@uevora.pt

[1] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Escola de Ciências e Tecnologia, Portugal*

Abstract: Em modelos lineares, quando os preditores apresentam escalas distintas, ou se pretende comparar a magnitude relativa dos coeficientes, é muito usual estandardizarem-se as variáveis independentes — através da sua transformação para média zero e desvio padrão unitário [1, 2]. No entanto, essa transformação é muitas vezes aplicada de forma automática, sem uma reflexão crítica sobre os seus efeitos na significância estatística, no desempenho do modelo e na interpretação substantiva dos resultados.

Neste estudo exploratório procura-se investigar em que situações a estandardização influencia significativamente os coeficientes estimados, a sua significância e a robustez dos modelos lineares. Para tal, analisam-se quatro cenários distintos: (i) dados simulados com escalas muito heterogéneas; (ii) presença controlada de multicolinearidade; (iii) variáveis naturalmente centradas; (iv) dados reais. A análise recorre a métricas como o R^2, erros padrão, variações nos coeficientes, estabilidade sob reamostragem (bootstrapping) e sensibilidade interpretativa.

Os resultados irão permitir identificar contextos em que a estandardização se revela benéfica, neutra ou contraproducente — sobretudo em modelos com elevada colinearidade ou quando se privilegia a interpretação direta dos coeficientes. No final, serão apresentadas diretrizes práticas para orientar a decisão de estandardizar (ou não), com base na estrutura dos dados, nos objetivos da análise e na função interpretativa do modelo.

Keywords: Coeficientes de regressão · Interpretação estatística · Modelos lineares · Multicolinearidade · Robustez do modelo

Acknowledgements: Este trabalho é parcialmente financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto « UIDB/04674/2020. »

References

[1] Gelman, A.: Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine **27**(15), 2865–2873 (2008). https://doi.org/10.1002/sim.3107

[2] Bring, J.: How to standardize regression coefficients. The American Statistician **48**(3), 209–213 (1994). https://doi.org/10.1080/00031305.1994.10476059

Regressão Quantílica com Efeitos Fixos e Mistos: Comparação de Funções Disponíveis no R

Anabela Afonso [1[0000-0002-5517-4855]] e Dulce G. Pereira [1[0000-0001-7281-4992]]
aafonso@uevora.pt, dgsp@uevora.pt

[1] *Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Escola de Ciências e Tecnologia, Portugal*

Abstract: Os modelos de regressão quantílica (RQ) são uma alternativa robusta à regressão linear tradicional, em especial na presença de observações atípicas ou quando os pressupostos de homocedasticidade e normalidade dos resíduos não são válidos [1]. Além disso, ao permitirem a estimação de diferentes quantis, são uma alternativa distribucionalmente mais informativa dos dados. Os modelos de efeitos mistos permitem incluir estruturas hierárquicas ou correlacionadas, sendo muito usados em dados longitudinais, em painel ou encaixados. Ao longo das últimas décadas tem-se observado um desenvolvimento metodológico que permite combinar estas duas abordagens, i.e., modelos RQ com efeitos mistos (e.g., [2, 3]).
No programa `R` [4], existem várias funções, em diferentes *package*, que permitem ajustar modelos RQ, com efeitos fixos e mistos, que usam diferentes abordagens para estimar os parâmetros de interesse. Contudo, a aplicação destes modelos em conjuntos de dados de grande dimensão e estruturas complexas continua limitada devido a restrições computacionais e metodológicas [3, 5].
Neste trabalho, pretendemos comparar o desempenho, o tipo de estruturas mistas admitidas, a estabilidade numérica, o tempo computacional das diferentes funções existentes no programa `R` para a estimação dos parâmetros dos modelos RQ, com efeitos fixos e mistos. Este estudo pretende fornecer orientações práticas para investigadores que usam este tipo de modelos.

Keywords: Efeitos fixos · Efeitos mistos · Regressão quantílica

Acknowledgements: Este trabalho é parcialmente financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto ■UIDB/04674/2020■.

References

[1] Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica **46**(1), 33–50.
[2] Geraci, M., Bottai, M.: Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics **8**(1), 140–154 (2007).
[3] Geraci, M., Bottai, M.: Linear quantile mixed models. Statistics and computing **24**, 461–479 (2014).
[4] R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[5] Wang, S., Cao, W., Hu, X., Zhong, H., Sun, W.: A selective overview of quantile regression for large-scale data. Mathematic **13**, 837 (2025).

Pacing Strategies in 800m and 1500m Freestyle: A Data-Driven Analysis from the 2024 Olympics

Joana Pinto [1] , Carlos J. Costa [1[0000-0002-1037-0561]]
l54843@aln.iseg.ulisboa.pt, cjcosta@iseg.ulisboa.pt

[1] *ISEG – Lisbon School of Economics and Management*

Abstract: This study applies statistical and machine learning techniques to explore pacing strategies in Olympic-level 800m and 1500m freestyle swimming. Using publicly available data from the 2024 Olympic Games, swimmer performances were reconstructed using lap-by-lap split times. From this, a set of pacing-related variables was derived, including start speed, end speed, pacing variability (coefficient of variation), and final sprint segment.
The study first employs hierarchical clustering to explore patterns in swimmer behaviour and uncover naturally occurring pacing strategies. This technique is used to visually and statistically interpret common profiles such as parabolic, positive, and negative pacing patterns.
In a second phase, the goal is to classify swimmers into pacing strategy types based on their pacing profiles and demographic variables (e.g., gender and age), employing various methods, like SVM, Random Forest and Neural Networks, and evaluating them to identify the best option.
This research may provide coaches and sports scientists with tools for deeper insights into race dynamics and athlete decision-making.

Keywords: Hierarchical clustering · Machine learning · Olympic games 2025 · Pacing strategies · Swimming performance

Relação entre o Microbioma Uterino e a Adesão à Dieta Mediterrânica: Metodologias Estatísticas

Laura Vieira [1], Analuce Canha Gouveia [2,3,4[0000-0001-6411-4195]] e Délia Gouveia Reis [1,5[0000-0002-5087-3120]]
2169724@student.uma.pt, analuce.canha@um.es, delia.reis@staff.uma.pt

[1] *Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal*
[2] *Departamento de Fisiologia, Faculdade de Veterinária, Universidad de Murcia, Espanha*
[3] *Departamento de Bioquímica e Biologia Molecular, Universidad de Granada, Espanha*
[4] *Instituto de Investigación Biosanitaria ibs.GRANADA, Espanha*
[5] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*

Abstract: O útero foi, durante muito tempo, considerado um órgão estéril. No entanto, com os avanços na sequenciação do RNA ribossomal 16S, tornou-se evidente a presença de um microbioma uterino, potencialmente relevante na saúde reprodutiva e ginecológica. Sabe-se que a dieta mediterrânica pode modular a microbiota intestinal, mas o seu impacto no microbioma do útero continua pouco explorado. Neste trabalho, fazemos uma revisão crítica das metodologias estatísticas mais frequentemente aplicadas ao estudo da relação entre microbiota e fatores dietéticos, com especial atenção a dados moleculares obtidos a partir de amostras uterinas humanas. A análise estatística neste contexto enfrenta desafios particulares: as amostras são difíceis de obter, pois a biópsia endometrial é um procedimento doloroso e invasivo, o que limita o número de participantes por questões éticas. Além disso, as análises moleculares, como a sequenciação 16S, têm custos elevados, restringindo a escala dos estudos, especialmente em grupos com menos recursos. Apresentamos ainda dados preliminares de um estudo realizado com mulheres seguidas no Hospital Universitário Virgen de la Arrixaca (Murcia, Espanha), e discutimos a adequação de testes não paramétricos, modelos de regressão e medidas de diversidade microbiana neste tipo de investigação. Pretende-se com este trabalho contribuir para uma escolha mais informada de estratégias estatísticas em contextos biomédicos com elevada restrição amostral e forte variabilidade individual.

Keywords: Dieta mediterrânica · Estatística · Microbioma uterino

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UID/00006/2025 e UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020.

An Exploratory Comparison of Metaheuristic Algorithms for Threshold Selection in Extreme Value Analysis

Beatriz Leça Pereira [1], Luiz Guerreiro Lopes [1,2[0000-0002-6145-8520]] and Délia Gouveia Reis [1,3[0000-0002-5087-3120]]

2005422@student.uma.pt, lopes@staff.uma.pt, delia.reis@staff.uma.pt

[1] *Faculty of Exact Sciences and Engineering, University of Madeira*
[2] *NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)*
[3] *CEAUL – Centre of Statistics and its Applications, University of Lisbon*

Abstract: Threshold selection is a crucial step in extreme value analysis, as it has a direct impact on the statistical reliability of inferences about rare events. Traditional approaches, such as graphical diagnostics, heuristic rules, and empirical criteria [1], are commonly used to determine an appropriate threshold. However, these methods often involve subjectivity and may lack consistency. The application of metaheuristic optimization approaches, such as evolutionary, swarm-based, or metaphor-less optimization algorithms, holds significant potential for the systematic automation of threshold selection. The use of metaheuristic optimization algorithms seeks to balance goodness of fit with the retention of sufficient exceedance data, thereby enhancing the robustness and reproducibility of extreme value modeling. In this study, two simple metaheuristic algorithms, Particle Swarm Optimization (PSO) [2] and Jaya [3], were employed to investigate their potential for automating threshold selection in extreme value modeling. These metaheuristic methods were applied to environmental time series, focusing on extreme rainfall data, to explore their potential and assess their strengths and limitations in identifying suitable thresholds. The results obtained in this study offer preliminary insight for future applications of metaheuristic optimization algorithms in extreme value analysis.

Keywords: Extreme values · Metaheuristics · Threshold selection

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2025 and UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020.

References

[1] Caeiro, F., Gomes, M. I.: Threshold selection in extreme value analysis. In: Dey, D. K., Yan, J. (eds.) Extreme Value Modeling and Risk Analysis: Methods and Applications, pp. 69–86, Chapman and Hall, NewYork (2016). DOI:10.1201/b19721

[2] Freitas, D., Lopes, L. G., Morgado-Dias, F.: Particle swarm optimisation: A historical review up to the current developments, Entropy, Vol. 22, No. 3, Art. 362 (2020) DOI:10.3390/e22030362

[3] Silva, B., Lopes, L. G., Mendonça, F.: Parallel GPU-acceleration of metaphor-less optimization algorithms: Application for solving large-scale nonlinear equation systems, Applied Sciences, Vol. 14, No. 12, Art. 5349 (2024) DOI:10.3390/app14125349

Misturas de Distribuições Gaussianas e Cotação de Criptomoedas

Susana Ferreira [1][0000-0002-1077-5980], Rui Santos [1,2][0000-0002-7371-363X] e Miguel Felgueiras [1,2,3][0000-0001-5450-7374]
susfer@ipleiria.pt, rui.santos@ipleiria.pt, mfelg@ipleiria.pt

[1] *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Portugal*
[2] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Universidade de Aveiro, Portugal*

Abstract: Em finanças, é comummente avaliada a evolução dos retornos logarítmicos (log-retornos), ou seja, $x_t = \ln(X_t) - \ln(X_{t-1})$ onde X_t pode representar o valor de fecho de um determinado índice da bolsa, ou a cotação de uma criptomoeda. O ajustamento de modelos a este tipo de dados é uma questão relevante, e diversos modelos têm sido considerados apropriados, tais como misturas de distribuições gaussianas, distribuições estáveis com cauda paretiana, distribuições t-Student, distribuições hiperbólicas generalizadas, entre outros.
Atendendo à sua flexibilidade e ao poder atual da computação, as misturas de distribuições gaussianas podem ser utilizadas com bastante sucesso, pelo que importa analisar a qualidade do seu ajuste. Neste contexto, simplificações habituais como igualdade de médias e de variâncias devem ser analisadas.
Neste trabalho serão utilizadas diferentes misturas de distribuições gaussianas de forma a modelar a distribuição dos log-retornos de algumas criptomoedas.

Keywords: Criptomoedas · Log-retornos · Misturas de Gaussianas

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/00006/2025 e UIDB/00006/2020. https://doi.org/10.54499/UIDB/00006/2020

References

[1] Behr, A., Potter, U.: Alternatives to the normal model of stock returns: gaussian mixture, generalized logF and generalized hyperbolic models. Annals of Finance **5**, 49–68 (2009). https://doi.org/10.1007/s10436-007-0089-8
[2] Felgueiras, M., Martins, J., Santos, R.: Gaussian Scale Mixtures. Journal of Numerical Analysis, Industrial and Applied Mathematics **11**(1-2), 1-10 (2017). https://jnaiam.org/2024/02/gaussian-scale-mixtures/
[3] Felgueiras, M., Martins, J., Santos, R.: Gaussian Mixtures with common Variance. WSEAS Transactions on Mathematics **23**, 276-281 (2024). https://doi.org/10.37394/23206.2024.23.30

Multivariate Empirical Bayes Analysis for Time-Resolved Omics: A Grapevine — Pathogen Infection Case Study

Nuno Domingues [1], **Lisete Sousa** [1,2], Gonçalo Laureano [3], Vincent Carré [4], Jasmine Hertzog [4], Andreia Figueiredo [3] and Marisa Maia [3]
fc52631@alunos.ciencias.ulisboa.pt, lmsousa@ciencias.ulisboa.pt, gmlaureano@ciencias.ulisboa.pt, vincent.carre@univ-lorraine.fr, jasmine.hertzog@univ-lorraine.fr, aafigueiredo@ciencias.ulisboa.pt, mrmaia@ciencias.ulisboa.pt

[1] *DEIO, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal*
[2] *CEAUL, Universidade de Lisboa, Lisboa, Portugal*
[3] *Grapevine Pathogen Systems Lab (GPS Lab), Biosystems & Integrative Sciences Institute (BioISI), Departamento de Biologia Vegetal, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal*
[4] *Université de Lorraine, LCP-A2MC, F-57000 Metz, France*

Abstract: The development of high-throughput omics technologies has contributed to increasing advancements within biological knowledge, allowing researchers to better understand the mechanisms underlying life across its multiple levels. Despite these advancements, omics data entails some significant challenges in data analysis, primarily due to its high dimensionality. Moreover, omics studies may involve complex experimental designs with numerous experimental factors measured at different time points, resulting in datasets containing vast amounts of information, but challenging to interpret. Such complexity may hinder the extraction of meaningful insights. Conventional dimensionality reduction techniques, such as Principal Component Analysis and Partial Least Squares Discriminant Analysis, though efficient in dealing with high-dimensional data, do not explicitly incorporate time as an experimental factor. To address this limitation, we highlight the application of Multivariate Empirical Bayes Analysis (MEBA), which ranks variables according to their differential behaviour over time, thus allowing the identification of key temporal features. The illustration of MEBA will be performed on data regarding grapevine untargeted metabolomics in the context of its susceptibility to pathogens such as *Plasmopara viticola*. In the context of metabolomics studies, such as the above, MEBA enables the extraction of biologically meaningful insights that would likely be missed by conventional approaches, allowing for a more comprehensive analysis of time-resolved data.

Keywords: Metabolomics · Multivariate data analysis · Time-resolved analysis

Acknowledgements: This research was funded by FCT through: MM research contract (DOI: 10.54499/2022.07433.CEECIND/CP1715/CT0009), exploratory project (DOI: 10.54499/2023.14736.PEX), CEAUL (UID/00006/2025, DOI: 10.54499/UIDB/00006/2020) and BioISI (UID/00100, DOI: 10.54499/UIDB/04046/2020). The MassLor research infrastructure is acknowledge for the mass spectrometry analysis.

Maximum Likelihood Estimation for a Folded Directional Distribution

Adelaide Figueiredo [1] [0000-0002-5734-3851]
and Fernanda Otília Figueiredo [2] [0000-0003-0255-4106]
adelaide@fep.up.pt, otilia@fep.up.pt

[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, Portugal*

Abstract: The field of directional data analysis, which deals with unit vectors on the surface of the hypersphere, has received increasing attention and development in recent years. When directional data fall on the positive orthant of the unit hypersphere, folded directional distributions become more appropriate than standard ones. Such data often arise, for instance, when compositional data are transformed into directional data via the square root transformation. Since the signs of the vector components are unknown after this transformation, the resulting data can be modeled using a folded directional distribution. In this study, we focus on the maximum likelihood estimation for the folded von Mises-Fisher distribution. As analytical solutions for the maximum likelihood estimates are not available, we adopt a numerical approach to solve the likelihood equations. Specifically, we apply an Expectation-Maximization (EM) algorithm to obtain the estimates, and we conduct a simulation study to investigate the properties of the resulting estimators.

Keywords: Directional data · EM algorithm · Folded distributions · Simulation study · Von Mises-Fisher distribution

Acknowledgements: This work is funded by national funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, under the support UID/50014/2023 (https://doi.org/10.54499/UID/50014/2023), UID/00006/2025 and UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Predictor Variable Selection for Mathematics Achievement: A Study Using PISA Data

Susana Faria [1][0000-0001-8014-9902]
sfaria@math.uminho.pt

[1] *Centre of Mathematics, Department of Mathematics, University of Minho*

Abstract: Using data from the 2022 Programme for International Student Assessment (PISA) for Portugal, this study applies penalized regression techniques to identify predictor variables associated with Portuguese students' mathematics scores. Specifically, it examines the influence of students' backgrounds, attitudes toward mathematics, home environment, parental involvement, and school-related factors on their mathematical literacy.
Penalized regression techniques have been recognized as powerful tools for variable selection in model development, particularly when working with high-dimensional datasets in educational research.
The dataset comprises 6793 Portuguese students from 224 schools, as part of the PISA 2022 assessment. Mathematics performance scores were used as the dependent variable, while 44 variables from the student questionnaires served as predictors.
Among the most influential factors associated with students' mathematics performance were their economic, social, and cultural status (ESCS), gender, grade repetition, self-efficacy in formal and applied mathematics, preference for mathematics, access to ICT at home, and family support for self-directed learning.

Keywords: LASSO · Mathematical literacy · PISA data · Variable selection

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] Hastie, T., Tibshirani, R.: Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Statistical Science, **35**, 579–592 (2020). DOI:10.1214/19-STS733

[2] OECD.: PISA 2022 Technical Report. OECD Publishing, Paris (2023). https://www.oecd.org/publications/pisa-2022-technical-report

[3] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), **58**(1), 267–288 (1996). DOI:10.1111/j.2517-6161.1996.tb02080.x

[4] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Methodological), **67**, 301–320 (2005). DOI:10.1111/j.1467-9868.2005.00503.x

A Statistical Approach to Pandemic Impact Analysis: Clustering Time Series Features in the US Retail Sector

José Canoso [1], and **Joana Leite** [1,2,3][0000-0001-6828-9486]
a2022107985@alumni.iscac.pt, jleite@iscac.pt

[1] *Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal*
[2] *CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Portugal*
[3] *Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal*

Abstract: Retail commerce was significantly disrupted by the COVID-19 epidemic in the United States (US), changing consumer behaviour and commercial dynamics across the country. In order to study these effects, a comparative analysis is performed on time series data from the pre- and post-pandemic periods, with the objective of identifying changes and emerging patterns in retail activity. The analysis employs monthly retail trade indicators obtained from the Federal Reserve Economic Data (FRED) database, which includes a diverse range of sub-sectors. Time series feature-based clustering algorithms are used to group retail sub-sectors based on their attributes [1], allowing for the detection of distinct impacts and behavioural similarities. Statistical features, including trend, seasonality, autocorrelation, and entropy, are extracted to capture the underlying dynamics of retail sales over time [2, 3]. The resulting clusters highlight distinct response profiles, enabling a more detailed interpretation of how various segments within the sector adapted or were impacted in the aftermath of the pandemic.

Keywords: Clustering · COVID-19 · FRED database · Retail trade · Time series features

References

[1] Paparrizos, J., Yang, F., Li, H.: Bridging the Gap: A Decade Review of Time-Series Clustering Methods (2024). DOI:10.48550/arXiv.2412.20582

[2] Kang, Y., Hyndman, R. J., Smith-Miles, K.: Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting, **33**(2), 345–358 (2017). DOI:10.1016/j.ijforecast.2016.09.004

[3] Wang, X., Smith, K., Hyndman, R.: Characteristic-Based Clustering for Time Series Data. Data Mining and Knowledge Discovery, **13**(3), 335–364 (2006). DOI:10.1007/s10618-005-0039-x

Bivariate INAR models with zero inflated innovations

Sandra Dias[1][0000-0001-5071-3023], Cristina Martins[2][0000-0003-4606-5953], and **Maria da Graça Temido**[3][0000-0002-5159-0528]
sdias@utad.pt, cmtm@mat.uc.pt, mgtm@mat.uc.pt

[1] *Universidade de Trás-os-Montes e Alto Douro, Dep. de Matemática*
[2] *Universidade de Coimbra, FCTUC*
[3] *Universidade de Coimbra/CMUC, FCTUC*

Abstract: Modelling and predicting the pattern of count time series has attracted the attention of many authors over the last four decades. Apart from academic interest, this attention is mainly due to the increasing availability of discrete data relevant to various fields, including social sciences, industry, finance, economics, medicine, etc. Traditional methods often become inadequate to describe the discreteness of the data. To address this limitation—and to draw a parallel with classical ARMA models—the usual multiplication operation was replaced by the binomial thinning operator introduced by Steutel and van Harn. This interesting operator allowed the development of integer ARMA models (INARMA), starting with the foundational integer-valued autoregressive (INAR) model. The literature on univariate time series for counts is widely developed, while research on multivariate time series models has progressed more slowly and not as detailed.
In this work we study a bivariate INAR(1) model with double zero-inflated innovations, considering three different scenarios for the double distribution of the innovations process. We start with the expected cases of the zero-inflated Poisson and the zero-inflated geometric double distributions. We also propose a mixed case where the innovations have different marginal distributions—one Poisson and one geometric—both modified to incorporate zero (or eventually one) inflation.
In addition to the traditional Yule-Walker and CLS methods, we propose a method for estimating the innovation parameters, driven by the frequent occurrences of zeros and clusters of zeros that characterize these models.

Keywords: Bivariate INAR model · Estimation · Zero-inflated innovations

Exploring Statistical Indices for Weekly Seasonality Analysis

Joana Carvalheiro [1], Joana Leite [1,2,3][0000-0001-6828-9486], and **Clara Viseu** [1,2,3][0000-0001-9906-276X]
a2023113433@alumni.iscac.pt, jleite@iscac.pt, cviseu@iscac.pt

[1] *Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal*
[2] *CEOS.PP Coimbra, Polytechnic University of Coimbra, Coimbra, Portugal*
[3] *Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Coimbra, Portugal*

Abstract: Short-term seasonal patterns are commonly observed in daily time series across various sectors, including transportation, energy, and public health, which makes their understanding important for informed decision-making. Although many statistical indices have been proposed to assess seasonality in monthly data, especially for annual cycles, their relevance to daily data and weekly seasonality is still underexplored in the literature [1]. In this context, the focus of this study is on examining the efficacy of statistical indices in detecting and quantifying weekly seasonality in daily observations. Customisation for this type of data and seasonality includes basic indices, such as the seasonality ratio and the seasonality indicator [2], as well as the Gini and Theil coefficients [3]. The analysis explores how effectively these metrics reveal the presence and intensity of weekly patterns, showing that they capture distinct dimensions of temporal concentration and distributional variability, thereby providing complementary analytical perspectives.

Keywords: Daily data · Indices · Time series · Weekly seasonality

References

[1] Rosselló, J., Sansó, A.: Yearly, monthly and weekly seasonality of tourism demand: A decomposition analysis. Tourism Management, **60**, 379–389 (2017). DOI:10.1016/j.tourman.2016.12.019
[2] Karamustafa, K., Ulama, S.: Measuring the seasonality in tourism with the comparison of different methods. EuroMed Journal of Business, **5**(2), 191–214 (2010). DOI:10.1108/14502191011065509
[3] Yabanci, O.: Seasonality of tourism demand in Turkey: a multi-methodical analysis. Current Issues in Tourism, **27**(12), 1930–1946 (2024). DOI:10.1080/13683500.2023.2217350

Improved estimation of the shape parameter for the shifted log-logistic distribution: Theory and Applications

Ayana Mateus [1[0000-0002-5630-3321]], Frederico Caeiro [1[0000-0001-8628-7281]]
amf@fct.unl.pt, fac@fct.unl.pt

[1] *Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA SST*

Abstract: The log-logistic distribution is commonly described in the literature with two parameters: one related to the shape and the other to the scale. To enhance its flexibility in modeling empirical data, an additional location parameter can be added, resulting in the three-parameter log-logistic distribution, also known as the shifted log-logistic or Pareto Type III distribution. In this research, we propose a reduced-bias estimator for the shape parameter of the three-parameter log-logistic model, derived from the classic Hill estimator. We examine the theoretical properties of the proposed estimator and evaluate its finite-sample performance through Monte Carlo simulations. Furthermore, we demonstrate its practical applicability using real-world data. Both theoretical analysis and simulation results indicate that the proposed method outperforms existing estimators in the literature, particularly with respect to bias and mean squared error.

Keywords: Bias-reduction · Log-logistic distribution ·Shape parameter

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

References

[1] Ahsanullah, M., Alzaatreh, A.: Parameter estimation for the log-logistic distribution based on order statistics. Revstat – Statistical Journal, **16**(4), 429–443 (2018).
[2] Mateus, A., Caeiro, F.: Improved Shape Parameter Estimation for the Three-Parameter Log-Logistic Distribution. Computational and Mathematical Methods, 1–13 (2022). DOI:https://doi.org/10.1155/2022/8400130
[3] Balakrishnan, N., Malik, H.J., and Puthenpura, S.: Best linear unbiased estimation of location and scale parameters of the log-logistic distribution. Communications in Statistics - Theory and Methods, **16**(12), 3477–3495 (1987).

A Comparative Study of Some Parametric and Nonparametric Control Charts

Fernanda Otília Figueiredo [1[0000-0003-0255-4106]],
Adelaide Figueiredo [2[0000-0002-5734-3851]],
and M. Ivette Gomes [3[0000-0002-2903-6993]]
otilia@fep.up.pt, adelaide@fep.up.pt, migomes@fc.ul.pt

[1] *University of Porto, School of Economics and Management, and University of Lisbon, Centre of Statistics and Applications (CEAUL)*
[2] *University of Porto, School of Economics and Management, INESC TEC*
[3] *University of Lisbon, Faculty of Science, Centre of Statistics and Applications (CEAUL)*

Abstract: Control charts are the most powerful tools for statistical process monitoring. However, the assumptions underlying their implementation must be fully verified. In many practical situations involving process monitoring, the data distribution is often non-normal or unknown since justifying a specific distribution for the data is difficult. This underscores the importance of developing nonparametric, or distribution-free, control charts, that are not dependent on the data distribution. In this study, we conduct simulation experiments to compare the performance of parametric and nonparametric control charts commonly used to monitor non-normal data. We also explore possible modifications to control chart statistics to improve their performance. We examine the performance of the charts in terms of in-control and out-of-control properties for data from symmetric non-normal distributions with different tail weights.

Keywords: ARL performance · Control charts · Statistical process monitoring

Acknowledgements: This work is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the projects UID/00006/2025 and UIDB/00006/2020 (DOI:10.54499/UIDB/00006/2020), and LA/P/0063/2020 (DOI:10.54499/LA/P/0063/2020).

References

[1] Surname, N., Surname, N.M.: Paper title. Journal title, **Num. Vol.**(num.), pages (Year). DOI:xxxxx
[2] Surname, N.: Paper title. In: Proceedings of the Symposium ..., pp. pages, Year. DOI:xxxxx
[3] Surname, N.: Book title. Publisher(s) (Year). DOI:xxxxx

Root Cause Analysis in Surface Mount Technology through Explainable AI

Ana Marinho [1][0009-0009-6205-5828], Luís Araújo [2][0009-0009-9613-5578]
b14125@math.uminho.pt, Luis.Araujo3@pt.bosch.com

[1] *Centre of Mathematics CMAT, Minho University, Campus de Azurém,*
[2] *BOSCH Car Multimédia Portugal, Braga, Portugal*

Abstract: High-accuracy classification models can routinely predict "pass" or "fail" outcomes in Surface Mount Technology (SMT) production lines, but the process variables that drive those predictions are rarely uncovered in an automated, systematic way. This study investigates how to leverage the predictive capacity of complex or black-box classifiers to explain why specific cases are classified correctly, turning raw accuracy into actionable insight. We introduce an agnostic, two-stage explanation framework that first applies Shapley additive explanations (SHAP) to obtain a global ranking of influential features [1]. Then, we deployed local interpreters — LIME [2] for case-level attribution and partial-dependences' plots (PDP) for pattern discovery —, to map these features to parameter ranges associated with abnormal behaviour [3].
The method is demonstrated on two practical challenges: (i) identifying which Automated Optical Inspection (AOI) settings most often trigger false calls and (ii) determining which SMT process parameters are correlated with specific component defects. In both scenarios, the framework identifies a concise set of dominant variables and highlights relevant thresholds, enabling targeted root-cause investigations. Since the approach is model- and process-agnostic, it can be extended to any classifier or defect type encountered on the production line.

Keywords: Machine Learning · SMT · XAI

Acknowledgements: This work is supported by national funds, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 179826; Funding Reference: SIFN-01-9999-FN-179826].

References

[1] Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 4768–4777, 2017. DOI:10.48550/arXiv.1705.07874
[2] Ribeiro, M. T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In:Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135-1144, 2016. DOI:10.1145/2939672.2939778
[3] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E.: Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics, **24**(1), 44-65 (2015). DOI:10.1080/10618600.2014.907095

Holistic Defect Detection in Surface-Mount Production Lines with a Machine Learning approach

Gabriel Quartin [1[0000-0003-2232-7656]], Luís Araújo [2[0009-0009-9613-5578]]
b14777@uminho.pt, Luis.Araujo3@pt.bosch.com

[1] *LIP – Laboratory of Instrumentation and Experimental Particle Physics and University of Minho*
[2] *BOSCH Car Multimedia Portugal, Braga, Portugal*

Abstract: Automated Optical Inspection (AOI) systems are widely employed to identify solder joint and component defects in Surface Mount Technology (SMT) assembly lines. However, these systems typically operate as final quality gates, without leveraging upstream process data such as solder paste printing, component placement, reflow profiles, or material traceability. In this study we investigate whether integrating such line-wide tabular features with Machine Learning (ML) classifiers can enhance defect detection. Specifically, we (i) measure how much full-line data improves AOI re-classification accuracy and (ii) examine to what extent post-reflow defects can be predicted without any AOI input at all.

Keywords: Anomaly detection · Machine learning · Surface-mount technology

Acknowledgements: This work is supported by national funds, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 179826; Funding Reference: SIFN-01-9999-FN-179826].

References

[1] Reshadat, V. and Kapteijns, A.R.: Improving the Performance of Automated Optical Inspection (AOI) Using Machine Learning Classifiers. In: Proceedings of the 2021 International Conference on Data and Software Engineering (ICoDSE), pp. 1-5, 2021. DOI:10.1109/ICoDSE53690.2021.9648445

[2] Thielen, N., Werner, D., Schmidt, K., Seidel R., Reinhardt, A. and Franke, J.: A Machine Learning Based Approach to Detect False Calls in SMT Manufacturing. In: Proceedings of the 43rd International Spring Seminar on Electronics Technology, pp. 1-5, 2020. DOI:10.1109/ISSE49702.2020.9121044

Economic Inequality in Europe: Interplay Between Income and Wealth

Kamila Trzcińska [1[0000-0002-4714-4074]], Elżbieta Zalewska [1[0000-0003-1544-300X]]
kamila.trzcinska@uni.lodz.pl, elzbieta.zalewska@uni.lodz.pl

[1] *Department of Statistical Methods, University of Lodz, Poland*

Abstract: Household income represents one of the fundamental economic categories, as it allows for theesenting the accumulated assets of households. Unlike income, wealth is characterized by a much higher level of concentration, making it a particularly interesting subject of analysis. Economi assessment of the material situation of different social groups. Equally important is wealth, reprc literature emphasizes that wealth inequality, in addition to income inequality, can lead to increased social tensions, limit social cohesion, and generate economic instability. For this reason, examining inequalities in the distribution of both income and wealth is an important element in the analysis of the condition of national economies and the well-being of societies. Although the issue of income inequality is relatively well documented, there is still a lack of comprehensive research on the distribution of wealth and the factors determining its concentration. The aim of this study is to analyze income and wealth inequalities among households in selected European countries. The evaluation is conducted using selected concentration measures, including the Zenga coefficient, Gini coefficient and its decomposition, with computations performed in R. The empirical data originate from the European Central Bank databases (2021), ensuring comparability across countries. In addition, a comparative analysis is carried out that takes into account household characteristics such as gender and education. Particular attention is devoted to differences in income and wealth levels between men and women as well as between groups with varying educational attainment. This approach allows for a better understanding of the mechanisms shaping inequalities in Europe.

Keywords: Inequality · Wealth group · Decomposition · Gini coefficient · Gender

Acknowledgements: This work was carried out within the research grant Excellence Initiative – Research University (IDUB; B23112001000193.07)

References

[1] Lerman, Robert I., Yitzhaki, S.: A note on the calculation and interpretation of the Gini index. Economics Letters, **15**(3-4), 363-368 (1984).
[2] Piketty, T., Zucman, G.: Wealth and inheritance in the long run. In Handbook of income distribution, **2**, 1303-1368 (2015).

Goodness-of-fit in multivariate ARMA-based models

Ana Martins [1,2] [0000-0003-4860-7795], and Sónia Gouveia [1,2] [0000-0002-0375-7610]
a.r.martins@ua.pt, sonia.gouveia@ua.pt

[1] *Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal*
[2] *Intelligent Systems Associate Laboratory (LASI), Portugal*

Abstract: The coefficient of determination (R^2) can be interpreted as a measure of the relative predictability of a time series given its history. Regarding univariate time series, Nelson (1976) [1] provided an interpretation of R^2 in autoregressive and moving average (ARMA) models, linking it to the model parameters and the autocorrelation structure. For an AR(p) model, the coefficient is expressed as $R_p^2 = \boldsymbol{\rho}_p^T C_p^{-1} \boldsymbol{\rho}_p$ where $\boldsymbol{\rho}_p = (\rho_1, \ldots, \rho_p)^T$ denotes the $(p \times 1)$ autocorrelation vector and C_p denotes the $(p \times p)$ correlation matrix with entries $[c_{ij}] = corr(z_{t-i}, z_{t-j})$, $i, j = 1, \ldots, p$. For an AR(1), this simplifies to $R_1^2 = \rho_1^2 = \phi_1^2$. For ARMA models, the relation remains valid under the assumption of invertibility, which enables reformulating them as a higher-order AR model.
In spatio-temporal modeling, which inherently involves multivariate time series, the assessment of goodness-of-fit has received limited attention. This work extends the concept of R^2 to the multivariate setting by deriving an analogous formulation for vector autoregressive (VAR) models of order one [2]. Given that any VAR(p) or VARMA(p, q) model can be expressed in a state-space form as a VAR(1), the result applies more broadly. Furthermore, the extension of these results to multivariate integer-valued AR(p) models, MINAR(p), and spatio-temporal variants, e.g., STINAR(p) [3], is straightforward, owing to their autocovariance structure being analogous to that of continuous VAR models. The use and interpretation of R^2 in a spatio-temporal context is illustrated with an application of STINAR(p) models to the time series of the daily number of hospital admissions in 18 Portuguese districts.

Keywords: ARMA · INARMA · Goodness-of-fit · Multivariate

Acknowledgements: This work was funded by the Fundação para a Ciência e Tecnologia, Portugal (FCT, https://www.fct.pt) under unit 00127-IEETA (https://www.ieeta.pt).

References

[1] Nelson, C. R.: The interpretation of R^2 in autoregressive-moving average time series models. The American Statistician, **4:30**, 175–180 (1976).
[2] Lütkepohl, Helmut.: New introduction to multiple time series analysis. Springer Science & Business Media (2005).
[3] Martins, A. Scotto M. G., Weiß, C. H., Gouveia, S.: Space-time integer-valued ARMA modelling for time series of counts. Electronic Journal of Statistics, **17:2**, 3472–3511 (2023).

Factors Influencing Student Engagement in Different Forms of Academic Misconduct

Cristina Veríssimo [1,2][0000-0002-1325-1410], Joselina Barbosa [1][0000-0002-1854-1572], Milton Severo [2,3][0000-0002-5787-4871], Paula Mena Matos [4,5][0000-0001-5763-8070], **Pedro Oliveira** [2,3][0000-0002-2470-0795] and Laura Ribeiro [1,6][0000-0003-1181-9217]

averissimo@med.up.pt, joselina@med.up.pt, msevero@icbas.up.pt, pmmatos@fpce.up.pt, pnoliveira@icbas.up.pt, lribeiro@med.up.pt

[1] *Departamento de Ciências da Saúde Pública e Forenses, e Educação Médica - Unidade de Educação Médica, Faculdade de Medicina da Universidade do Porto*
[2] *Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto*
[3] *EPIUnit – Instituto de Saúde Pública da Universidade do Porto*
[4] *Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto*
[5] *Centro de Psicologia da Universidade do Porto*
[6] *i3S-Instituto de Investigação e Inovação em Saúde da Universidade do Porto*

Abstract: Multiple instruments have been used to assess academic misconduct, yet robust psychometric evidence has been reported only for a few. This study aims to determine the validity and dimensionality of a novel Academic Misconduct Questionnaire (AMQ) and to explore differences between students who engage in distinct misbehaviours. The questionnaire showed good validity and reliability. Health students scored higher in most misbehaviours, especially compared to Economics/Law, Social Sciences and Arts/Humanities, while the latter two disclosed higher Signature Forgery. This study proposes a valid instrument to assess academic misconduct in university students. The predictive models helped to better understand differences between students who engaged in distinct misbehaviours, enabling more targeted interventions.

Keywords: Academic integrity · Academic misconduct questionnaire · Validation study

References

[1] Veríssimo, A.C., Conrado, G.A., Barbosa, J., et al.: Machiavellian Medical Students Report More Academic Misconduct: A Cocktail Fuelled by Psychological and Contextual Factors. Psychol Res Behav Manag., **15**, 2097-2105 (2022). DOI:10.2147/PRBM.S370402

[2] Veríssimo, A.C., Barbosa, J., Severo, M., Matos, P.M., Oliveira, P., Ribeiro, L.: Validation of the academic misconduct questionnaire: exploring predictors of student misconduct. Med Educ Online, **30**(1), (2025). DOI:10.1080/10872981.2025.2506739

Index