PROGRAMA E LIVRO DE RESUMOS

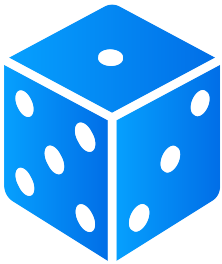
XXVI CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA

SPE 2023

GUIMARÃES, 11 A 14 DE OUTUBRO DE 2023

EDIÇÕES SPE

SOCIEDADE PORTUGUESA DE ESTATÍSTICA



© 2023 Comissão Organizadora do Congresso SPE 2023
UNIVERSIDADE *do* MINHO
Guimarães, Portugal
https://w3.math.uminho.pt/SPE2023
✉: spe2023@cmat.uminho.pt

Fundação
para a Ciência
e a Tecnologia

A organização do XXVI Congresso da Sociedade Portuguesa de Estatística (SPE 2023) foi parcialmente financiada por Fundos Portugueses através da FCT (Fundação para a Ciência e a Tecnologia) no âmbito dos Projetos UIDB/00013/2020 e UIDP/00013/2020.

(The organization of the XXVI Congress of the Portuguese Statistical Society (SPE 2023) 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.)

Ficha Técnica:

PROGRAMA E LIVRO DE RESUMOS

Luís Meira-Machado, Maria Conceição Serra, Marta Ferreira, Raquel Menezes

Editora: Sociedade Portuguesa de Estatística

Capa: Nicolau Moreira (GCI-UMinho)

Impressão: Instituto Nacional de Estatística

Tiragem: 250 exemplares

ISBN: 978-972-8890-49-0

Depósito Legal: 521068/23

COMISSÃO ORGANIZADORA
(ORGANIZING COMMITTEE)

Raquel Menezes (Presidente), UNIVERSIDADE *do* MINHO
Ana Paula Amorim, UNIVERSIDADE *do* MINHO
Arminda Manuela Gonçalves, UNIVERSIDADE *do* MINHO
Carla Moreira, UNIVERSIDADE *do* MINHO
Cecília Castro, UNIVERSIDADE *do* MINHO
Inês Sousa, UNIVERSIDADE *do* MINHO
Luís Meira-Machado, UNIVERSIDADE *do* MINHO
Maria Conceição Serra, UNIVERSIDADE *do* MINHO
Marta Ferreira, UNIVERSIDADE *do* MINHO
Susana Faria, UNIVERSIDADE *do* MINHO

COMISSÃO CIENTÍFICA
(SCIENTIFIC COMMITTEE)

Miguel de Carvalho (Presidente), UNIVERSITY *of* EDINBURGH
Carlos Tenreiro, UNIVERSIDADE *de* COIMBRA
Clara Cordeiro, UNIVERSIDADE *do* ALGARVE
Cláudia Neves, KINGS COLLEGE
Cláudia Nunes Philippart, INSTITUTO SUPERIOR TÉCNICO
Dulce Gomes, UNIVERSIDADE *de* ÉVORA
Inês Sousa, UNIVERSIDADE *do* MINHO
Lígia Henriques Rodrigues, UNIVERSIDADE *de* ÉVORA
Lisete Sousa, UNIVERSIDADE *de* LISBOA
Luís Meira-Machado, UNIVERSIDADE *do* MINHO
Pedro Oliveira, UNIVERSIDADE *do* PORTO
Raquel Menezes, UNIVERSIDADE *do* MINHO
Vera Afreixo, UNIVERSIDADE *de* AVEIRO

Caros participantes, palestrantes e convidados,

É com imenso prazer que a Sociedade Portuguesa de Estatística (SPE) e o Centro de Matemática da Universidade do Minho (CMAT) vos recebem no XXVI Congresso da Sociedade Portuguesa de Estatística (SPE 2023). Este evento decorre de 11 a 14 de outubro de 2023, no emblemático Centro Cultural de Vila Flor em Guimarães, Portugal.

O XXVI Congresso da SPE promete ser uma plataforma imperdível para a partilha de conhecimentos e para o fomento de colaborações entre profissionais e académicos. Destacamos o Minicurso especializado oferecido por Pedro Miranda Afonso, do Erasmus Medical Center Rotterdam. O curso incidirá sobre o pacote R JMbayes2, uma ferramenta essencial para estatísticos e investigadores interessados em análise de dados longitudinais.

As Sessões Plenárias incluem tópicos relevantes como "The Efron-Petrosian estimator" de Jacobo de Uña-Álvarez, abrangendo estatísticas avançadas em astrofísica e biomedicina. Jorge Caiado explorará o agrupamento em grandes volumes de dados temporais, enquanto Maria Eduarda Silva discutirá a ciência de redes em análise de séries temporais. Maria Kateri focará em métodos estatísticos para otimizar experiências em contextos variáveis de stress, relevantes em campos como engenharia e medicina.

O congresso conta também com mais de 150 comunicações orais e pósteres. Com uma gama variada de sessões temáticas, promovidas por instituições como o INE e o Banco de Portugal, bem como áreas mais especializadas como Biometria e Modelos Estocásticos para Dinâmicas Populacionais, este evento representa uma oportunidade única para o aprofundamento técnico e o diálogo interdisciplinar na comunidade estatística. Diversas atividades especiais também estão na agenda, incluindo a sessão temática "Rising Stars", uma iniciativa da SPE/CLAD, que dará palco a jovens investigadores ascendentes na área. Ainda, o IPQ patrocinará a iniciativa "Intersecting Pathways: Statistics and (Inter)national Standardization", com o objetivo de criar sinergias entre a estatística aplicada e os padrões internacionais, sublinhando o papel fulcral da estatística em múltiplas vertentes da governação global.

As áreas temáticas das comunicações cobrem um espetro amplo, desde Análise de Sobrevivência e Ciência de Dados até Métodos Bayesianos. Este leque de temas não só oferece um panorama atualizado da estatística, como também visa facilitar oportunidades valiosas para futuras colaborações e crescimento profissional.

Complementando o rigor académico, o programa inclui também momentos para interação social e cultural. O Jantar do Congresso, na imponente Pousada de Santa Marinha, é uma ocasião perfeita para estreitar laços profissionais, enquanto as visitas guiadas a monumentos históricos de Guimarães oferecem um mergulho na rica tapeçaria cultural da cidade.

Neste evento tão querido para nós, que combina rigor académico com oportunidades para interação social e cultural, encontramos o cenário perfeito para prestar uma homenagem especial ao Professor Daniel Paulino. Um académico que, durante mais de meio século, enriqueceu indelevelmente a estatística portuguesa, particularmente no campo Bayesiano. A sua extensa obra - que inclui uma série de publicações científicas, livros, conferências e cursos - reflete o espírito multifacetado que este congresso aspira representar. Daniel Paulino não foi apenas um investigador e professor, mas também um líder. No seu mandato como Presidente da SPE entre 2012 e 2014, delineou um caminho para a estatística portuguesa no cenário global. A sua visão para os congressos da SPE espelha a alma e a missão desta sociedade: estimular a produção científica, fomentar colaborações multidisciplinares e reforçar a coesão e a afirmação da SPE no âmbito nacional e internacional.

Esta homenagem vai além de um mero tributo; é uma celebração contínua do legado inestimável que o Professor Daniel Paulino nos deixou. Em sintonia com os valores e padrões aos quais ele deu vida através do seu trabalho, este congresso aspira a ser um palco de excelência científica e coesão para a nossa comunidade estatística. Assim, ao recordarmos o Professor Daniel Paulino, reafirmamos o nosso compromisso com esses princípios.

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

Com os melhores cumprimentos,
A Comissão Organizadora Local

PROGRAM

—PROGRAMA—

4ªfeira - 11 de outubro

8:15	**Registo e entrega documentação.** Hall do Auditório			
9:00	**Minicurso:** *Modelação Conjunta de Dados Longitudinais e de Sobrevivência* **Pedro Miranda Afonso** Auditório			
10:30	**Pausa para café**			
11:00	**Minicurso** (cont.)			
12:30	**Pausa para almoço**			
14:00	**Minicurso** (cont.)			
16:00	**Pausa para café**			
16:30	**Sessão de abertura do congresso**. Auditório			
16:45	**Sessão Plenária I**. **Jorge Caiado** *Clustering of Big Data Time Series* Auditório / Moderador: João A. Branco			
17:45	**Comunicações Orais I**			
	Sessão Temática - Banco de Portugal ***Reproducibility and Statistical Disclosure Control*** Organizadora: Rita Sousa / Moderadora: Susana Faria Auditório	**Ciência de Dados I** Moderadora: Maria J. Polidoro Sala 1	**Métodos Não Paramétricos I** Moderador: Carlos Tenreiro Sala 2	
	Perturbation Methods: some application results **Rita Sousa**, Jorge Morais, Susana Faria	*Prasanta Chandra Mahalanobis and his praised Mahalanobis Distance* Ana M. Pires, **João A. Branco**	*Joint modeling of longitudinal binary responses: A nonparametric Bayesian approach using acute and chronic malnutrition data* **André Nunes**, Giovani Silva, Luzia Gonçalves	
	Confidentiality and the statistical data sharing Diogo Barbosa, João Falcão Silva, **Ana Barbara Pinto**	*Perfil de Risco de um Cliente que entra em Incumprimento - Crédito à Habitação* **Sofia Comparada**, Eduardo Severino, Teresa Alpuim	*Simulation Study to Compare the Performance of Signed Klotz and the Signed Mood Weighted Generalized Coefficients* **Sandra M. Aleixo,** Júlia Teles	
	Replication App **Gustavo Iglésias**	*Robust clustering based on trimming and choice of parameters* **Luis Angel García-Escudero**, Christian Hennig, Agustín Mayo-Iscar, Gianluca Morelli, Marco Riani	*Exploring the Mutual Information Rate Decomposition in Situations of Pathological Stress* **Helder Pinto**, Celeste Dias, Ana Paula Rocha	
18:45	**Deslocação (a pé) para o Museu Martins Sarmento**			
19:15	**Receção de boas vindas** Museu Martins Sarmento			

5ªfeira - 12 de outubro

8:15	**Registo e entrega documentação. Hall do Auditório**			
9:00	**Comunicações Orais II**			
	Sessão Temática ***Stochastic Models for Population Dynamics*** Organizadora / Moderadora: Conceição Serra Auditório	**Estatística Multivariada I** Moderadora: Irene Oliveira Sala 1	**Bioestatística e Epidemiologia I** Moderadora: Carla Moreira Sala 2	**Estatística Computacional I** Moderadora: Cecília Castro Sala 3
	General models for harvesting in randomly varying environments: impact of Allee effects **Carlos A. Braumann**, Clara Carlos, Nuno M. Brites	*Classification and Survival Analysis of Multi-omics Data for the Identification of Novel Diagnostic and Prognostic Biomarkers in Glioma* **Francisca G. Vieira**, Regina Bispo, Marta B. Lopes	*"PACE-Gate": statistical lessons learned from a high-profile and controversial clinical trial* **Nuno Sepúlveda**	*Estimação robusta de modelos com dados em painel* **Anabela Rocha,** M. Cristina Miranda
	A class of stochastic models to describe the dynamics of biological populations with migrations **Manuel Mota**, Manuel Molina	*A Study on the Pearson and Mean Deviance Estimators of the Dispersion Parameter in Poisson Regression* **Rui Miranda,** Rita Gaio	*Longitudinal antibody kinetics in kidney transplant recipients who recovered from severe acute respiratory syndrome coronavirus 2 infection* **Maria J. Polidoro**, Ana Pinho, Manuela Bustorff, Natércia Durão, Rui Martins	*Revisitando métodos de reamostragem na estimação de parâmetros de acontecimentos raros* **Dora Prata Gomes,** M. Manuela Neves
	Stochastic models to describe the evolution of two-sex biological populations **Manuel Molina**, Manuel Mota	*Model-based Clustering of Distributional Data - An Application in Official Statistics* **Paula Brito,** A. Pedro Duarte Silva	*Speckle Tracking Echocardiography in detecting myocardial deformation in the left ventricle: Systematic Review and Meta-Analysis* **Helena Cardoso**, Brígida Mónica Faria, Rita Amaral	*Comparative Analysis of the Marginalized LASSO Estimator in Multiple Linear Models: A Comprehensive Evaluation* **Mina Norouzirad,** Filipe J. Marques
	Approximate Bayesian computation approach on the maximal offspring and parameters in controlled branching processes **Carmen Minuesa**, Miguel González, Inés del Puerto	*PCA from compositional and symbolic perspectives in the study of mortality in Portugal* **Marta Maltez**, Adelaide Freitas, Magda Monteiro	*Application of Machine Learning Techniques for a Recommendation System in Pharmacy* **Beatriz Torres**, Alexandra Oliveira, Brígida Mónica Faria, Sandra Alves	*Avaliação de ferramentas de programação em Modelos Lineares Generalizados: estudo de simulação* **Ana Marinho**, Susana Faria, Rita Sousa
	Modeling stochastic introgression in a spatio-temporally varying environment with branching processes **Maria Conceição Serra**		*Fail-safe number: a simulation study* **Vera Afreixo,** Filipa Rocha	
10:40	**Pausa para café e Sessão de Posters I**			
	1 - Use of quantile regression methods in growth curves **Bianca Rafaelle da Silva**, Thiago G. Ramires, Marcelo F. Silva			
	2- Constant effort harvesting with Allee effects in random environments using logistic type models: optimization and an application **Clara Carlos**, Nuno M. Brites, Carlos A. Braumann			
	3 -Semiparametric model applied to crime recidivism data **Felipe António Magro**, Thiago Ramires, Marcelo F. Silva			
	4 - Population Growth and geometrically thinned EVT **Dinis D. Pestana**, Maria de Fátima Brilhante, Ivette Gomes, Sandra Mendonça, Pedro D. Pestana			
	5 - Improvement of COVID-19 symptoms: a survival analysis study from a Portuguese cohort **Leandro Duarte**, Carla Moreira, Luís Meira-Machado, Ana Paula Amorim, Joana Costa, Paula Meireles			

5ªfeira - 12 de outubro

	Pausa para café e Sessão de Posters I (Continuação)
	6 - Perspectives of statistical inference on interval-valued data **Catarina Rodrigues**, Conceição Amado
	7 - Selection of models and thresholds in the Peaks-Over-Threshold (POT) methodology: Application to extreme precipitation values in Madeira and Porto Santo islands **Délia Gouveia-Reis**, Luiz Guerreiro Lopes, Sandra Mendonça
	8 - Os anos de vida saudável perdidos nas doenças não comunicáveis na União Europeia **Margarida Torres**, Alcina Nunes, João Paulo Martins
	9 - Desempenho de metodologias de classificação sexual baseadas em ortopantomografias João Alves, Cristiana Palmela Pereira, **Rui Santos**
	10 - Os desafios do Jornalismo de Dados **Cláudia Silvestre**, Helena Pina, Susana Araújo
	11 - Jackson exponentiality test **Ayana Mateus**, Frederico Caeiro
	12 - Big Data Analysis of Solar Radiation Patterns in the Colombian Caribbean Region **Glória Carrasca**l, Jhonathan Barrios, Jairo Plaza, Flora Ferreira
	13 - Environmental Exposure Index for Early Life Exposure Assessment Tool (ELEAT) **Beatriz Costa**, Lisete Sousa, Célia Rasga, Astrid Vicente
	14 - Failure time in a pulp drying press **Luís Margalho**, Francisco Paiva
	15 - A influência dos Riscos Psicossociais na Qualidade de Vida dos colaboradores **Pedro Pereira**, Maria Mourão, Pedro Carvalho
	16 - Propagação de incertezas e fiabilidade estrutural num modelo mecânico **Luísa Hoffbauer**, Carlos Conceição António
	17 - Count models and randomness patterns **Sílvio Velosa**, Dinis Pestana, Sandra Mendonça
	18 - Desvendando o sucesso escolar: uma jornada através dos Modelos Lineares **Marcos Machado**, Maria Fernanda Diamantino, Luísa Loura
	19 - Regiões portuguesas: os desafios estratégicos e o papel das finanças públicas locais **Irene Oliveira**, Patrícia Martins
	20 - Long-Term Trends and Daily Variations in Global Irradiation in Cabinda, Angola: Implications for Solar Energy Production and Sustainability **Faustino Maciala**, Jhonathan Barrios, Arminda Gonçalves
	21 - Analysis of crew time series absenteeism in the railway sector **Catarina Afonso**, Ricardo Saldanha, Regina Bispo, Gonçalo Matos, Luís Albino
	22 - Application of Statistical Methodologies as a Contribute to Define Disease Control Strategies for a Sustainable Viticulture **Nuno Domingues**, Lisete Sousa, Gonçalo Laureano, Andreia Figueiredo, Marisa Maia
	23 - Use of the random forest model in precision agriculture **Natália Costa Martins**, Thiago G. Ramires, Luiz R. Nakamura
	24 - The profile of the hyper user in an Emergency Department **Loide Ascenso**, Gonçalo Jacinto, Hugo Quintino, Paulo Infante
	25 - Análise estatística temporal da sinistralidade laboral em Portugal de 2009 a 2019 Ricardo Dourado, Marina Almeida-Silva, **Miguel Felgueiras**
	26 - Flexible odds ratio curves for continuous predictors: the flexOR package **Marta Azevedo**, Luís Meira-Machado, Carla Moreira, Artur Araújo
	27 - Multialphabetic hypercubes and disaggregation of sums of squares **Carla Francisco**, Manuela Oliveira, Francisco Carvalho, Tiago Mexia
	28 - Survival Analysis Applied to Extreme Longevity Research **Laetitia Teixeira**, Lia Araújo, Denisa Mendonça, Constança Paúl, Oscar Ribeiro
11:40	**Sessão Plenária II** **Eduarda Silva** *Time Series Analysis via Network Science* Auditório / Moderador: Carlos Brauman
12:40	**Pausa para almoço**
14:30	**Sessão Plenária III** **Maria Kateri** *Step-Stress Models: Statistical Inference and Optimal Design of Experiments* Auditório / Moderador: Miguel de Carvalho
15:35	**Passeio do Congresso**

6ªfeira - 13 de outubro

8:15	**Registo e entrega documentação. Hall do Auditório**			
9:00	**Comunicações Orais III**			
	Sessão Temática - Biometria Organizadoras: Inês Sousa e Mª José Ginzo Villamayor Moderadora: Inês Sousa Auditório	**Bioestatística e Epidemiologia II** Moderadora: Lisete Sousa Sala 1	**Estatística Multivariada II** Moderadora: A. Manuela Gonçalves Sala 2	
	A method for determining groups in cumulative incidence curves **Marta Sestelo**, Luís Meira-Machado, Nora M. Villanueva, Javier Roca-Pardiñas	*Regressing a static response on longitudinal predictors: the case-study of childhood obesity* **Mafalda Oliveira**, Susana Santos, Rita Gaio	*Métodos de Machine Learning para previsão de Dados Longitudinais* **Elsa Soares**, Inês Sousa	
	A retrospective analysis of alcohol-related emergency calls to the ambulance service in Galicia **Mª José Ginzo Villamayor**, Paula Saavedra Nieves, Dominic Royé, Francisco Caamaño Isorna	*Dimensionality reduction in survival models based on gene expression data: an application to brain cancer* **João Brandão**, Marta B. Lopes, Eunice Carrasquinha	*The trace ratio method for robust multigroup classification* **M. Rosário Oliveira**, Giulia Ferrandi, Igor Kravchenko, Michiel E. Hochstenbach	
	Joint model for multiple longitudinal responses with informative time measurements **Inês Sousa**	*O Test Negative Design na avaliação da efetividade de vacinas* **André Martins**, João Paulo Martins, Marlene Santos	*Statistical analysis to identify protein adducts by mass spectrometry: a tool for biomarker investigation* **Filipa Costa**, Conceição Amado, Alexandra M. M. Antunes, Judit Morello	
10:00	**Pausa para café e Sessão de Posters II**			
	29 - Factors that influence energy to water nexus in urban and rural households **A. Manuela Gonçalves**, Cristina Matos			
	30 - Modelos Lineares Generalizados Mistos: uma aplicação a dados de acidentes rodoviários **Susana Faria**, Jair Santos, Elisabete Freitas			
	31 - Classification of compositional data using distributions defined on the hypersphere **Adelaide Figueiredo**			
	32 - Prediction of perceived depression for SHARE survey data in COVID 19 waves Sara Ribeiro Pires, **M. Rosario Ramos**, Paula Vaz-Fernandes			
	33 - An approach to estimate infection by COVID-19 **Manuela Oliveira**, Eugénio Garção			
	34 - Diagnóstico do COVID-19 nos registros de SRAG - Síndrome Respiratória Aguda Grave **Lúcia Barroso**, Gustavo Kanno			
	35 - Regression Modeling of Marine Species Abundance Indicators: Exploring Spatial Distribution and Environmental Correlations André Dias, **Raquel Menezes**, Maria Manuel Angélico			
	36 - Data Mining and Statistical Quality Control **Fernanda Otilia Figueiredo**, Adelaide Figueiredo, Maria Ivette Gomes			
	37 - A hybrid robust-weighted AMMI modeling approach with generalized weighting schemes **Vanda M. Lourenço**, Marcelo B. Fonseca, Paulo C. Rodrigues			
	38 - The use of statistical techniques to evaluate the impact that dierent chocolates have on the sensory perception of three dierent categories of Port wine **Elisete Correia**, Gabriela Santos, Alice Vilela			
	39 - Estilo de vida e bem-estar dos estudantes do Politécnico de Leiria **Daniel Santos**, Rui Santos, Susana Ferreira			
	40 - Risk assessment of vulnerabilities exploitation **Fernando Sequeira**, Maria de Fátima Brilhante, Pedro D. Pestana, Maria Luísa Rocha			
	41 - Probabilistic Procedures for SIR and SIS Epidemic Dynamics on Contact Random Network **J. Leonel Rocha**, Sónia Carvalho, Beatriz Coimbra			
	42 - Multi-state modeling of composite indexes for assessing the economic conditions of firms. A comparative study between energy and non-energy Portuguese firms **Gustavo Soutinho**, Vitor M. Ribeiro, Isabel Soares			

6ªfeira - 13 de outubro

Pausa para café e Sessão de Posters II (Continuação)

43 - Enhancing Gait Analysis through Transformation-Based Multiple Linear Regression Normalization
Jhonathan Barrios, Barbára Araújo, Estela Bicho, Miguel Gago, Wolfram Erlhagen, Flora Ferreira

44 - What statistics can reveal about occupant behavior in diverse building types
Barbara Lumy Noda Nogueira, Celina Pinto Leão, Solange Leder

45 - Exploring the Influence of Various Factors on Salaries: A Regression Analysis Approach
Flora Ferreira, Ana Pedrosa, Ana Borges, José Soares, Pedro Pacheco

46 - Risk analysis for evaluating the water quality of a hydrological basin
Ana Pedra, Arminda Gonçalves, Irene Brito

47 - Seleção de variáveis em misturas de modelos de regressão linear: um estudo de simulação
Ana Moreira, Susana Faria

48 - Microbial diversity and discrimination of Azeitão and Nisa P.D.O. cheeses based on metagenomic data
Carlota Teles, Lisete Sousa, Sílvia Rebouças, Teresa Crespo, Teresa Semedo-Lemsaddek

49 - Unveiling Gene Signatures in Glioma: A Comprehensive Analysis using Regularized Logistic Regression, Dimensionality Reduction, and Outlier Detection
João F. Carrilho, Roberta Coletti, Marta B. Lopes

50 - Desenho e implementação de um questionário: uma avaliação do grau de satisfação dos utilizadores BPLIM
Eduardo Dias, Rita Sousa, Inês Sousa

51 - Diferenças de valores humanos em indivíduos que afirmam pertencer a uma religião e indivíduos que afirmam não pertencer, no período 2002 a 2020, na Europa
Maria Paula Lousão, Cláudia Silvestre, José Casanova

52 - Modelagem logística para ajuste de dados em pacientes diagnosticados com neoplasia
Jorge Alves de Sousa, José Joedson Lima de Sousa, Anselmo Ribeiro Lopes

53 - Construção e análise espacial de um índice de desenvolvimento sustentável
Conceição Ribeiro, Paula Pereira, Sílvia Pedro Rebouças

54 - Tail independence: a comparative analysis of estimation methods
Sandra Dias, Marta Ferreira

55 - A crossinggram for random fields on lattices
Helena Ferreira, **Marta Ferreira**, Luis A. Alexandre

56 - Analysis of M^X/M/c/n systems with impatient customers
Fátima Ferreira, Antonio Pacheco, **Helena Ribeiro**

57 - Extremal behavior of some bivariate integer models
Sandra Dias, **Maria da Graça Temido**

11:00 **Comunicações Orais IV**

Métodos Não Paramétricos II Moderador: Luis Meira-Machado Auditório	**Aplicações em Econometria, Finanças e Gestão** Moderador: Frederico Caeiro Sala 1	**Aplicações em Ambiente, Clima, Geociências e Agricultura I** Moderadora: Isabel Natário Sala 2	**Estatística Computacional II** Moderadora: Marília Antunes Sala 3
Estimation of AUC in logistic regression with missing data: removing the bias **Susana Rafaela Guimarães Martins**, Jacobo Uña-Alvarez, Maria del Carmen Iglésias-Perez	*Innovation and Product Positioning* **Diogo Pereira**, Claudia Nunes, Anne Balter, Peter Kort	*The importance of experimental design principles in agricultural field trials* **Elsa Gonçalves**	*Screening the Discrepancy Function of a Computer Model* **Rui Paulo**, Pierre Barbillon, Anabel Forte
On a Parzen-Rosenblatt type density estimator for circular data **Carlos Tenreiro**	*Modelação estatística do custo em contratos de assistência automóvel* **Bárbara Botellho,** Sandra Ramos	*Modeling landing per unit effort (LPUE) abundance of fish using functional data analysis* **Manuel Oviedo-de la Fuente**, Raquel Menezes, Alexandra A. Silva	*Survapp: a Shiny application for survival data analysis* **Emanuel V.M. da Silva**, Luís Meira-Machado, Gustavo Soutinho
Reduced bias estimation of the residual dependence index – Pareto meets Fréchet **Cláudia Neves**		*Multivariate random fields and systems of stochastic partial differential equations* **Sílvia Guerra**, Fernanda Cipriano, Isabel Natário	*How to define prior information for generalized maximum entropy estimation?* **Jorge Cabral**, Vera Afreixo, Pedro Macedo

12:00 **Prémio Carreira**

13:00 **Pausa para almoço**

6ªfeira - 13 de outubro

14:30	**Sessão Plenária IV** **Jacobo de Uña Álvarez** *The Efron-Petrosian estimator* Auditório / Moderadora: Maria Ivette Gomes			
15:30	**Comunicações Orais V**			
	Métodos Bayesianos Moderador: Giovani Silva Auditório	**Extremos** Moderador: Marta Ferreira Sala 1	**Análise de Sobrevivência** Moderador: Pedro Oliveira Sala 2	**Séries Temporais I** Moderador: Clara Cordeiro Sala 3
	Bayesian prediction of football outcomes - Application to the Portuguese 1st League **Rui Martins**, Daniel Andrade	*A new class of conditional tail expectation estimators* **Lígia Henriques-Rodrigues**, M. Ivette Gomes, Fernanda Figueiredo, Frederico Caeiro	*An additive shared frailty model for recurrent gap time data in the presence of zero-recurrence subjects* **Ivo Sousa-Ferreira**, Ana Maria Abreu, Cristina Rocha	*State-space models: fitting, modeling, and calibration* **Marco Costa**
	Bayesian approach for treating missing data in count time series **Isabel Silva**, Maria Eduarda Silva, Isabel Pereira	*Location invariant estimation of the Weibull tail coefficient* **Maria Ivette Gomes**, Frederico Caeiro, Lígia Henriques-Rodrigues	*Comparative Analysis of Cox Proportional Hazard Model and Machine Learning Approaches for Predicting Financial Distress in SMEs* **Ana Borges**, Mariana Reimão Carvalho	*Regression models for count time series: an application to health care indicators* Rui Soares, **Magda Monteiro**, Isabel Pereira
	Bayesian modeling count time series with structural breaks **Isabel Pereira**, Betty Nakyambadde, Cláudia Santos	*A partially reduced bias Hill estimator of the Extreme Value Index* **Frederico Caeiro**, Ivette Gomes, Lígia Henriques-Rodrigues	*NIV treatment effect estimation for ALS patients using Propensity Score methodology* **Luis Garcez,** Sara Madeira, Helena Mouriño	*Automatic Event Diagnosis in Water Consumption* **Rita Leite**, Conceição Amado, Margarida Azeitona
	Bayesian Smoothing for Time-Varying Joint Extremes **Miguel de Carvalho**	*A direct approach in extremal index estimation* **M. Cristina Miranda**, Manuela Souto De Miranda, M. Ivette Gomes	*Parametrizações do modelo conjunto para dados longitudinais e de sobrevivência* **Maria Helena Oliveira**, Isolde Previdelli	*Improving parameter estimation and prediction accuracy by handling outliers in state-space modeling* **F. Catarina Pereira**, A. Manuela Gonçalves, Marco Costa
16:50	**Pausa para café**			
17:10	**Comunicações Orais VI**			
	Sessão Temática - INE ***Inovações Recentes em Estatísticas Oficiais*** Organizadores: Carlos Marcelo, Pedro Campos Moderador: Carlos Marcelo Auditório	**Aplicações em Ambiente, Clima, Geociências e Agricultura II** Moderador: Conceição Ribeiro Sala 1	**Controlo Estatístico da Qualidade, Fiabilidade e Risco** Moderador: Fernanda Otília Figueiredo Sala 2	**Bioestatística e Epidemiologia III** Moderador: Nuno Sepúlveda Sala 3
	Classification of CPP - Application of a Multilayer Neural Network **M. Conceição Ferreira**, Almiro Moreira, Ana Carmona, David Santos, Rui Alves	*Air Quality Data Analysis with Symbolic Principal Components* **Catarina P. Loureiro**, M. Rosário Oliveira, Paula Brito, Lina Oliveira	*Green exchange-traded fund performance evaluation using the EU-EV risk model* **Irene Brito**, Ana Isabel Azevedo, José Azevedo	*An Application of Cluster Analysis with Mortality Rates of Non-communicable Diseases* **Ana Paula Nascimento**, Cristina Prudêncio, Brígida Mónica Faria, Mónica Vieira, Helena Bacelar-Nicolau
	Estimation of free riding in plastic package waste using put-on-market and business turnover information **João S. Lopes**, Filipa Chambel, Nuno Romão	*Spatio-temporal modelling of commercial fish species distribution* **Daniela Silva**, Raquel Menezes, Susana Garrido	*Análise e avaliação de falhas em estações maregráficas* **Dora Carinhas**, Paulo Infante, António Martinho	*The magnitude and stability of protection against Omicron SARS-CoV-2 acquired by hybrid immunity* **João Malato,** Ruy M. Ribeiro, Eugénia Fernandes, Pedro Pinto Leite, Pedro Casaca, Carlos Antunes, Válter R. Fonseca, Manuel Carmo Gomes, Luís Graça
	Application of a Statistical Disclosure Control Algorithm for Data Protection of Portuguese Census 2021 **Inês Rodrigues de Sá,** Pedro Campos		*From sums of unequal size samples to the mean and standard deviation* Miguel Casquilho, **Cecília Castro**, Jorge Buescu	*Adjustment methods for confounding variables: a comparative study* Inês Fortuna, Luís Antunes, Cláudia Vieira, **Brígida Mónica Faria**
18:10	**Assembleia Geral da SPE**			
20:00	**Jantar do Congresso**			

sábado - 14 de outubro

9:00	**Comunicações Orais VII**			
	Sessão Temática - SPE/CLAD ***Rising Stars*** Organizadoras: M. Rosário Oliveira e Adelaide Figueiredo Auditório	**Estatística em Ciências Sociais e Educação / Estatísticas Oficiais** Moderador: Rui Martins Sala 1	**Ciência de Dados II** Moderador: Cláudia Nunes Philippart Sala 2	**Estatística Espacial** Moderador: Rita Gaio Sala 3
	A utilização de funções de penalização na seleção de variáveis em misturas de modelos de regressão com efeitos aleatórios **Luísa Novais**, Susana Faria	*Item Response Theory and Confirmatory Factor Analysis of emotional distress in women from a Breast Cancer Screening Program* **Ana Meireles**, Bruno de Sousa, Isabel Natário, Vitor Rodrigues	*Identifying Consumption Profiles in Load Data Analysis* **Lucas Henriques**, Cecília Castro, Felipe Lima	*Modelação espacial da probabilidade de persistência de cadáveres de aves em estudos de monitorização da mortalidade em linhas elétricas* **Ema Biscaia**, Joana Bernardino, Regina Bispo
	An interactive R-Shiny app for the identification of atypical clusters **Ana Helena Tavares**, Vera Afreixo, Diana Lucas, Paula Brito	*How to move towards an inclusive education* **Bruno de Sousa**	*Optimizing Retail Sentiment Analysis with SentiLex-PT and Machine Learning* **Catarina Almeida**, Cecília Castro, Ana Cristina Braga, Ana Freitas	*Using a constructed covariate that accounts for preferential sampling* **Andreia Monteiro**, Isabel Natário, Maria Lucília Carvalho, Ivone Figueiredo, Paula Simões
	Multivariate Time Analysis via Multilayer Quantile Graphs **Vanessa Freitas Silva**, Maria Eduarda Silva, Pedro Ribeiro, Fernando Silva	*A study on the development of analytical skills of computer science students at Polytechnic of Porto* **Eliana Costa e Silva**, Cristovão Sousa	*Modelling uncertainty regarding the location of Fire Stations: a case study applied to Porto region* **Tiago Ribeiro**, Maria Isabel Gomes, Regina Bispo	*Geostatistical Models for Identifying Juvenile Fish Hotspots in Marine Conservation* **Francisco Gonçalves**, Raquel Menezes, Daniela Silva, Inês Dias, Alexandra A. Silva
	The Role of Translation Quality Estimation at Unbabel **José Maria Pombal**, André F. T. Martins	*The evolution of immigrant groups in Luxembourg - What are the different pathways in the labour market?* **Catarina Campos Silva**, Paula Brito, Pedro Campos	*Aprendizagem Automática vs Modelação Estatística* **Ricardo Coelho**, Isabel Natário	*Density Surface Model vs. Spatial Models with INLA for animal abundance estimation* **Iúri J. F. Correia**, Tiago A. Marques, Christine Cuyler, Soraia Pereira
10:20	**Pausa para café**			
10:40	**Homenagem Daniel Paulino**			

sábado - 14 de outubro

12:00	**Comunicações Orais VIII**			
	Sessão Temática SPE - IPQ|DRAPN ***Intersecting Pathways: Statistics and (Inter)national Standardization*** Organizadoras: Maria J. Polidoro, Conceição Amado Auditório	**Séries Temporais II** Moderador: Isabel Pereira Sala 1	**Ciência de Dados III** Moderador: Bruno de Sousa Sala 2	**Bioestatística e Epidemiologia IV** Moderador: Luzia Gonçalves Sala 3
	Intersecting Pathways: Statistics and (Inter)national Standardization **Maria J. Polidoro**, **Mafalda T. Costa**, **Rui Pereira**, **Miguel de Carvalho**	*Space-time autoregressive models for time series of counts* **Ana Martins**, Manuel G. Scotto, Christian H. Weiss, Sónia Gouveia	*Classifying distributional data into more than two groups* Ana Santos, **Sónia Dias**, Paula Brito, Paula Amaral	*Analysis of Hospitalization Patterns Using Custom Dissimilarities* Daniel Cordeiro, Ana Azevedo, Bárbara Peleteiro, Lucybell Moreira, Elsa Guimarães, Raquel Cadilhe, **Rita Gaio**
		Comparing the effects of concurrent promotions over demand with interpretable deep learning **Micael Gomes**, Alexandra Oliveira, Luís Paulo Reis	*Linear regression for symbolic density-valued data* **Rui Nunes**, Paula Brito, Sónia Dias	*A Joint Model for Multiple (Un)Bounded Longitudinal Outcomes, Competing Risks, and Recurrent Events* **Pedro Miranda Afonso**, Dimitris Rizopoulos, Anushka Palipana, John P. Clancy, Rhonda D. Szczesniak, Eleni-Rosalina Andrinopoulou
		Previsão horária da descarga de um aproveitamento a fio de água Joana Seabra-Silva, **Paula Milheiro-Oliveira,** Paulo Avilez-Valente	*Rules for predicting lab-grown diamonds prices: a comparative analysis* **Margarida G. M. S. Cardoso**, Luís Chambel	*Comparative study on the performance of different classification algorithms, combined with pre- and post-processing techniques to handle imbalanced data, in the diagnosis of adult patients with familial hypercholesterolemia* **João Albuquerque**, Ana Margarida Medeiros, Ana Catarina Alves, Mafalda Bourbon, Marília Antunes
13:00	**Pausa para almoço**			
14:30	**Prémio SPE 2022**			
15:00	**Prémio SPE 2023**			
15:30	**Encerramento do Congresso**			

BOOK OF ABSTRACTS

—LIVRO DE RESUMOS—

Table of Contents
(Índice)

Mini Course
(Minicurso) 1

Plenary Sessions
(Sessões Plenárias) 5

Thematic Sessions
(Sessões Temáticas) 11

Intersecting Pathways: Statistics and (Inter)national Standardization
— SPE/IPQ/DRAPN —
(Pontos de Interseção entre Estatística e Normalização Inter(nacional) - - SPE/IPQ/DRAPN) 13

Stochastic Models for Population Dynamics
(Modelos Estocásticos para a Dinâmica de Populações) 17

Biometrics
— SPE Section & SGAPEIO —
(Biometria - Secção da SPE & SGAPEIO) 25

Recent Innovations in Official Statistics
— Statistics Portugal —
(Inovações Recentes em Estatísticas Oficiais - Instituto Nacional de Estatística) 31

Reproducibility and Statistical Disclosure Control
— Portuguese Central Bank —
(Reprodutibilidade e Controlo de Confidencialidade - Banco de Portugal) 37

Rising Stars
— SPE/CLAD —
(Estrelas em Ascensão - SPE/CLAD) 43

Oral Sessions
(Comunicações Orais) **49**

Posters
(Pósteres) **132**

Authors / Autores **193**

Mini Course

—Minicurso—

Modelos Conjuntos para Dados Longitudinais e de Sobrevivência

Pedro Miranda Afonso [a]
p.mirandaafonso@erasmusmc.nl

[a] *Erasmus Medical Center, The Netherlands*

Abstract: Habitualmente, em estudos longitudinais, são recolhidas múltiplas variáveis resposta de vários tipos. Estas incluem biomarcadores longitudinais e o tempo até que um evento de interesse ocorra (e.g. morte ou recorrência de cancro). Estes resultados são comumente analisados separadamente, mas a sua modelação conjunta deve ser utilizada quando o interesse científico recai na inter-relação dessas respostas. O desenvolvimento de modelos conjuntos para dados longitudinais e tempo-até-evento tem sido motivado essencialmente por duas aplicações. Primeiro, quando o interesse recai no tempo-atéevento e se quer incluir o efeito de covariáveis endógenas dependentes do tempo medidas com erro. Segundo, quando o interesse recai no biomarcador longitudinal e é necessário corrigir para abandono não aleatório.
Este curso fornecerá uma introdução abrangente a modelos conjuntos. Será explicado em que situações estes modelos devem ser preferenciais, os seus pressupostos e como podem ser utilizados para extrair informação relevante dos dados. Serão também apresentados alguns exemplos práticos usando a biblioteca `R JMbayes2` . `JMbayes2` (https://drizopoulos.github.io/JMbayes2/) é uma poderosa biblioteca `R` que visa facilitar a análise conjunta de dados longitudinais e de sobrevivência por parte de investigadores no seu dia-a-dia. Permite ao utilizador:

- incluir múltiplas respostas longitudinais com diferentes distribuições de probabilidade;
- incluir diferentes processos de tempo-até-evento, tais como riscos competitivos;
- associar as respostas longitudinais ao modelo de risco através de diferentes formas;
- estimar predições dinâmicas individualizadas a partir dos modelos ajustados.

A biblioteca inclui também funções para resumir e visualizar resultados, assim como realizar diagnóstico de regressão.
No final do curso, os participantes serão capazes de identificar situações em que um modelo conjunto é necessário, definir um modelo apropriado para responder à sua questão de investigação, ajustar um modelo com a biblioteca JMbayes2 e interpretar corretamente os resultados.

Programa:

1. Introdução
2. Modelo Misto Linear Generalizado
3. Modelo de Riscos Proporcionais
4. Modelo Conjunto Básico
5. Extensões de Modelos Conjuntos
6. Predições Dinâmicas

Pré-requisitos:
Este curso pressupõe o conhecimento de conceitos básicos de inferência estatística e de modelos de regressão. Conhecimentos básicos de `R` serão vantajosos, mas não são essenciais. Os participantes devem trazer os seus computadores portáteis com a bateria totalmente carregada. Antes do curso, serão fornecidas instruções para a instalação do software necessário.

Plenary Sessions

—Sessões Plenárias—

Step-Stress Models: Statistical Inference and Optimal Design of Experiments

Maria Kateri [a]
maria.kateri@rwth-aachen.de

[a] *Aachen University, Germany*

Abstract: Accelerated life testing (ALT) is widely used in reliability analysis with applications in diverse fields, ranging from material sciences and quality control to biomedical sciences and ecology statistics. Step-stress models form an essential part of ALT. Under a step-stress ALT (SSALT) model, the test units are exposed to stress levels that gradually increase at intermediate time points of the experiment. Statistical inference is then developed for, e.g., the mean lifetime under each tested stress level. The estimation of the mean lifetime under normal (not tested) operating conditions is possible by means of estimating the parameters of an appropriate link function that connects the stress level to the associated mean lifetime. The assumptions made about the time points of stress level change, the termination point of the experiment, the underlying lifetime distributions, the type of censoring, if present, and the way of monitoring, lead to respective models.

We discuss SSALT models and their options for flexible modelling. We focus on a model that considers a general scale family of distributions, which allows for flexible modelling and leads to explicit expressions for the maximum likelihood estimators of the scale parameters of the underlying lifetime distributions. The approach is presented for Type-I censored experiments under continuous as well as interval monitoring of the test items. Statistical inference, frequentist and Bayes, is considered while the issue of optimal designing an SSALT experiment is also discussed. Finally, we deal with SSALT modelling of heterogeneous populations, when, based on their aging behaviour, the test items are split in groups. In this case, heterogeneity is captured through a mixture model approach.

Clustering of Big Data Time Series

Jorge Caiado [a]
jcaiado@iseg.ulisboa.pt

[a] *Instituto Superior de Economia e Gestão (ISEG), Portugal*

Abstract: The current revolution in Artificial Intelligence, Machine Learning, and Big Data presents new possibilities and challenges for researchers and analysts alike. This is particularly true in the case of time series, as for many domains, analysts now have access to collections of very long time series in many areas of interest, such as astronomy, geophysics, medicine, social media, economics and finance. These researchers and analysts are challenged when they are required to compare and cluster such long and diverse time series. On the whole, it is not usually possible to use traditional methods of analysis for these tasks, such as estimating models and comparing features, as these methods imply lengthy computations, including the inversion of extremely large matrices.

Caiado, Crato, and Poncela (2020) proposed a spectral method of synthesizing and comparing time series characteristics which is nonparametric and is focused on the data's cyclical features. Instead of using all the information available from the data, which is computationally very timeconsuming, this procedure uses regularization rules to select and summarize the most relevant information for clustering purposes. This method does not imply the computation of the full periodograms, but only that of the periodogram components related to frequencies of interest, comparing and clustering the respective periodogram ordinates for the various time series using common clustering methods. They named this approach the 'fragmented-periodogram method'.

More recently, Albino, Caiado and Crato (2023) proposed a new approach for clustering big data time series which can be considered to be an alternative to the periodogram method in the case of time series: the 'fragmented-autocorrelation based method'. Essentially, these authors suggest using the autocorrelation function of time series which is only computed for the lags of greatest interest. In a large Monte Carlo simulation study, they explore whether this procedure is able to condense the relevant information of the time series. This method is illustrated in an empirical study of the evolution of various stock market indices.

The Efron-Petrosian estimator

Jacobo de Uña-Álvarez [a]
jacobo@uvigo.es

[a] *Vigo University, Spain*

Abstract: The seminal paper by Bradley Efron and Vahe Petrosian on nonparametric methods for doubly truncated data, published by the Journal of the American Statistical Association in September 1999, will have its 25th birthday by next year. Many contributions on the topic have appeared along all this time, being relevant to Astronomy, Epidemiology, Engineering or Economics, among other fields. In this talk I will revisit some features of the Efron-Petrosian estimator, including: maximum-likelihood and self-consistency properties; inverse-probability weighting representation; numerical algorithms for practical computation; large-sample behaviour; and applications to smooth curve estimation, two-sample problems, regression analysis, goodnessof-fit tests and multi-state models. Some open questions will be presented too. I will defend the need to include methods for doubly truncated data in textbooks on Survival Analysis due to the impact they have in estimation. I will also discuss the importance of performing specification tests for the sampling bias in the doubly truncated setting. Real data examples will be used for illustration purposes. The existing R packages to deal with doubly truncated data will be reviewed.

Time Series Analysis via Network Science

Maria Eduarda Silva [a]
mesilva@fep.up.pt

[a] *Universidade do Porto, Portugal*

Abstract: Large amounts of data index by time are becoming increasingly common in many organizations. Feature-based approaches have become common for exploring and understanding structures and patterns and to identify unusual observations in these large sets of time series. For univariate time series, there are well-define sets of time series features in the literature. This work introduces a new set of features for multivariate time series based on complex networks. First, we provide an overview of mapping approaches to represent univariate time series as single layer complex networks. Then, we introduce two new mapping methods appropriate for multivariate time series: the multilayer horizontal visibility graph that is based on the new concept of cross-horizontal and a quantile-based transition mapping. The topological measures extracted from the resulting multilayer networks constitute a set of multivariate time series features. The proposed mappings and topological measures are parameter-free, do not require data pre-processing and are applicable to any multivariate time series dataset. The features are evaluated and validate the proposed mappings. The results indicate that these features capture useful characteristics of multivariate time series.

Thematic Sessions

—Sessões Temáticas—

Intersecting Pathways: Statistics and (Inter)national Standardization

— SPE/IPQ/DRAPN —

Pontos de Interseção entre Estatística e Normalização Inter(nacional) - SPE/IPQ/DRAPN

Maria J. Polidoro
Instituto Politécnico do Porto e CEAUL, mjp@estg.ipp.pt
Conceição Amado
CEMAT, IST-ULisboa

Intersecting Pathways: Statistics and (Inter)national Standardization

Maria J. Polidoro [a,b], **Mafalda T. Costa** [c], **Rui Pereira** [d], **Miguel de Carvalho** [b,e]

mjp@estg.ipp.pt, anamafalda.costa@drapnorte.gov.pt, rpereira@ipq.pt, Miguel.deCarvalho@ed.ac.uk

[a] *Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Porto (ESTG-IPP), Felgueiras, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *Direção de Serviços de Controlo e Estatística, Direção Regional de Agricultura e Pescas do Norte, Portugal*
[d] *Unidade de Gestão Operacional de Normalização, Departamento de Normalização, Instituto Português da Qualidade, Portugal*
[e] *School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK*

Keywords: IPQ/CT 225, ISO/TC 69, standardization in the applications of statistical methods

Abstract: This session will provide a gentle introduction to the intersection of Statistics and International Standardization, showcasing some of the ongoing developments and open opportunities in this field in Portugal.

Stochastic Models for Population Dynamics

Modelos Estocásticos para a Dinâmica de Populações

Maria Conceição Serra
Centro de Matemática da Universidade do Minho, mcserra@math.uminho.pt

General models for harvesting in randomly varying environments: impact of Allee effects

Carlos A. Braumann [a,b], Clara Carlos [a,c], Nuno M. Brites [d]
braumann@uevora.pt, clara.carlos@estbarreiro.ips.pt, nbrites@iseg.ulisboa.pt

[a] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora*
[b] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora*
[c] *Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal*
[d] *ISEG/UL - Universidade de Lisboa, Department of Mathematics; REM - Research in Economics and Mathematics, CEMAPRE*

Keywords: Allee effects, extinction, harvesting models, stationary distribution, stochastic differential equations

Abstract: We consider general autonomous stochastic differential equation models for the growth of a harvested population with Allee effects living in a randomly varying environment. Besides mild regularity conditions, the model only requires general assumptions dictated by biological properties. We show that, if the *per capita* net growth rate (difference between the geometric average natural growth rate and the harvesting mortality rate) is positive when population size is very small, there is a stochastic equilibrium with a stationary density. If, however, that rate is negative (overharvesting), the population becomes extinct. These results extend the ones obtained in [1, 1] for general models with harvesting but without Allee effects and in [3] for general models with Allee effects but without harvesting. In [2], particular cases were used to assess the impact of Allee effects.

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

References

[1] Braumann, C.A. Variable effort fishing models in random environments. *Mathematical Biosciences*, 156, 1–19, 1999. https://doi.org/10.1016/S0025-5564(98)10058-5

[2] Braumann, C.A. Variable effort harvesting models in random environments: generalization to density dependent noise intensities. *Mathematical Biosciences*, 177 & 178, 229–245, 2002. https://doi.org/10.1016/S0025-5564(01)00110-9

[3] Brites, N.M., Braumann, C.A. Profit optimization of stochastically fluctuating populations under harvesting: the effects of Allee effects. *Optimization*, 71:11, 3277–3293, 2022. https://doi.org/10.1080/02331934.2022.2031191

[4] Carlos, C., Braumann, C.A. General population growth models with Allee effects in a random environment. *Ecological Complexity*, 30, 26–33, 2017. http://dx.doi.org/10.1016/j.ecocom.2016.09.003

A class of stochastic models to describe the dynamics of biological populations with migrations

Manuel Mota [a,b], Manuel Molina [a,b]
mota@unex.es, mmolina@unex.es

[a] *Department of Mathematics, University of Extremadura, Spain*
[b] *Institute of Advanced Scientific Computation, University of Extremadura*

Keywords: branching processes, biological systems, mathematical modeling, multitype populations, population dynamics

Abstract: This research deals with mathematical modeling in complex biological systems in which several types of individuals coexist in various populations. Migratory phenomena among the involved populations are allowed. We propose a class of stochastic models to describe the demographic dynamics of these type of complex biological systems. The probability model is mathematically described through a sequence of random matrices in which rows and columns represent the various populations and the several types of individuals, respectively. We prove that this stochastic sequence can be studied under the general setting provided by the multitype branching process theory. By considering a methodology based on such a theory, see e.g. [1], several probabilistic properties are established and some limiting results are derived. The results obtained have interest in biology and ecology, especially in studies about population dynamics of species. As application, we present an illustrative example about the population dynamics of biological systems formed by long-lived raptor colonies. For more details concerning this research line see [2].

Acknowledgements: We thank to the Ministerio de Ciencia, Innovación y Universidades of Spain, Grant PID2019-108211GB-I00/AEI/10.13039/501100011033, for the financial support given to this research.

References

[1] Mode, C.J. *Multitype Branching Processes. Theory and Applications*. American Elsevier Publishing Co., Inc., New York, USA, 1971.

[2] Molina, M., Mota, M. Demographic dynamics in multiple populations with migrations. *Mathematics*, 9, 246, 2021. https://doi.org/10.3390/math9030246.

Stochastic models to describe the evolution of two-sex biological populations

Manuel Molina [a,b], Manuel Mota [a,b]
mmolina@unex.es, mota@unex.es

[a] *Department of Mathematics, University of Extremadura, Spain*
[b] *Institute of Advanced Scientific Computation, University of Extremadura*

Keywords: branching processes, mathematical modeling, population dynamics, two-sex processes

Abstract: In the general context of mathematical modeling, we continue the research line considered in [1] and [2] about the class of two-sex branching processes with various mating and reproduction strategies. By considering the possibility of immigration of female and male individuals from external populations, we present several probabilistic and statistical contributions. The results are illustrated through simulated examples. The class of stochastic models under consideration has particular interest to mathematically describe the population dynamics of semelparous species, namely, biological species with a single reproductive episode before dying. Semelparity occurs in a wide variety of biological species including amphibians, arachnids, insects, reptiles, etc., see [3].

Acknowledgements: We thank to the Ministerio de Ciencia, Innovación y Universidades of Spain, Grant PID2019-108211GB-I00/AEI/10.13039/501100011033, for the financial support given to this research.

References

[1] Molina, M., Mota, M., Ramos, A. Estimation of parameters in biological species with several mating and reproduction alternatives. *Mathematical Biosciences*, 329, 108471, 2020. https://doi.org/10.1016/j.mbs.2020.108471.

[2] Molina, M., Mota, M. Some contributions to the class of branching processes with several mating and reproduction strategies. *Mathematics*, 10, 2061, 2022. https://doi.org/10.3390/math10122061.

[3] Ranta, E., Tesar, D., Kaitala, V. Environmental variability and semelparity vs. iteroparity as life histories. *Journal of Theoretical Biology*, 217, 391-398, 2002.

Approximate Bayesian computation approach on the maximal offspring and parameters in controlled branching processes

Carmen Minuesa [a], Miguel González [a], Inés del Puerto [a]
cminuesaa@unex.es, mvelasco@unex.es, idelpuerto@unex.es

[a] *Department of Mathematics, University of Extremadura, Badajoz, Spain*

Keywords: approximate Bayesian computation, Bayesian inference, controlled branching processes, logistic growth population model

Abstract: In a Bayesian framework, the purpose of this work is to estimate the posterior distribution of the parameters of interest of controlled branching processes without a prior knowledge of the maximum number of children that an individual can give birth and without explicit likelihood calculations.
To that end, we have developed approximate Bayesian computation (ABC) methods in the context of branching processes. More precisely, we present a rejection ABC algorithm for model choice to select a reasonable maximum number of offspring per individual, and which is based on the comparison with the raw data observed. In a second step, we estimate the posterior distributions of the parameters of interest by applying a tolerance-rejection algorithm with a post-sampling correction method making used of a suitable summary statistic.
We illustrate the accuracy of the proposed methods by means of a simulated example developed with the statistical software R. We also apply our results to a real dataset of harbour seals.

Acknowledgements: This research has been supported by grant PID2019-108211GB-I00 funded by MCIN/AEI/10.13039/501100011033, by "ERDF A way of making Europe"

References

[1] González, M., Minuesa, C., del Puerto, I. Approximate Bayesian computation approach on the maximal offspring and parameters in controlled branching processes. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 116(4), 147, 2022. doi:10.1007/s13398-022-01290-w

Modeling stochastic introgression in a spatio-temporally varying environment with branching processes

Maria Conceição Serra [a]
mcserra@math.uminho.pt

[a] *Centro de Matemática da Universidade do Minho*

Keywords: branching processes, introgression, invasion dynamics, multitype

Abstract: Invasion dynamics are important in many contexts, such as evolution, epidemics, metapopulation dynamics and environmental management. Recently, interest in the study of invasions has increased, because human activities such as trade, travel and agriculture, and processes like climate change, increase the spread of exotic species into new habitats, and enhance the risk of genetic introgression (the permanent incorporation of genes of one population or species into the genome of another).
In this work, we use multitype branching processes to model the evolution of populations that are exposed to introgression events. In particular, we consider a situation where invasion attempts occur repeatedly over time from a source population into a receptor population. The receptor population contains two locations that confer different offspring distribution to invaders. Besides the demographic stochasticity, expressed through different offspring probability distributions, we consider also temporal variation which leads to variations in the fitnesses of local invaders. We investigate how the combination of temporal environmental stochasticity, spatial patterns and demographic stochasticity affects establishment success in situations with repeated invasions as described above.

References

[1] Ghosh, A., Serra, M.C., Haccou, P. Quantifying time-inhomogeneous stochastic introgression processes with hazard rates. *Theoretical Population Biology*, Volume 81, Issue 1, 253–263, 2012. https://doi.org/10.1016/j.tpb.2011.11.006

[2] Ghosh, A., Serra, M.C., Haccou, P. Quantifying stochastic introgression processes in random environments with hazard rates. *Theoretical Population Biology*, Volume 100, 1–5, 2015. https://doi.org/10.1016/j.tpb.2014.11.005

[3] Haccou, P., Serra, M.C. Establishment versus population growth in spatio-temporally varying environments. *Proceedings of the Royal Society B: Biological Sciences*, 288:20202009, 2021. https://doi.org/10.1098/rspb.2020.2009

Biometrics

— SPE Section & SGAPEIO —

Biometria — Secção da SPE & SGAPEIO

Inês Sousa
CMAT, Escola de Ciências, Universidade do Minho, isousa@math.uminho.pt
Mª José Ginzo Villamayor
Universidade de Santiago de Compostela, mariajose.ginzo@usc.es

A method for determining groups in cumulative incidence curves

Marta Sestelo [a,b], Luís Meira-Machado [c], Nora M. Villanueva [b,d], Javier Roca-Pardiñas [a,b]
sestelo@uvigo.es, lmachado@math.uminho.pt, nmvillanueva@uvigo.es, roca@uvigo.es

[a] *CITMAga, CP.15782, Santiago de Compostela, Spain*
[b] *Dep. Statistics and O.R. & SiDOR Group, University of Vigo, Vigo, Spain*
[c] *Centre of Mathematics and Department of Mathematics, University of Minho - School of Sciences, Campus de Azurém, Guimarães, Portugal*
[d] *CINBIO, University of Vigo, Vigo, Spain*

Keywords: clustering, compering risk, cumulative incidence function, multiple curves, survival analysis

Abstract: Survival Analysis is the standard method for analyzing data when the variable under study represents the time from a well-defined initial moment to the occurrence of a single event of interest. If more than one type of endpoint is present, these endpoints are called competing risks since the observation of a particular event prevents an individual from observing any other type of event. The cumulative incidence function is the standard method for estimating the marginal probability of a given event in the presence of competing risks. One basic but important goal in the analysis of competing risk data is the comparison of these curves, for which limited literature exists. We proposed a new procedure that lets us not only test the equality of these curves but also group them if they are not equal. The proposed method allows determining the composition of the groups as well as an automatic selection of their number. Simulation studies show the good numerical behavior of the proposed methods for finite sample size. Finally, the applicability of the proposed method is illustrated using real data.

Acknowledgements: Marta Sestelo and Javier Roca-Roca-Pardiñas acknowledges financial support from Grant PID2020-118101GB-I00 funded by Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033). Luís Meira-Machado acknowledges financial support from Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the project UIDB/00013/2020, UIDP/00013/2020

A retrospective analysis of alcohol-related emergency calls to the ambulance service in Galicia

Mª José Ginzo Villamayor [a], Paula Saavedra Nieves [a], Dominic Royé [b], Francisco Caamaño Isorna [c]
mariajose.ginzo@usc.es, paula.saavedra@usc.es, dominic.roye@ficlima.org, francisco.caamano@usc.es

[a] *Department of Statistics, Mathematical Analysis and Optimization (USC) and Galician Centre for Mathematical Research and Technology (CITMAga)*
[b] *Climate Research Foundation (FIC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)*
[c] *Department of Public Health (USC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)*

Keywords: alcohol, bayesian hierarchical models, Galicia, nonparametric level set estimation

Abstract: This work will be focused on the introduction of statistical methods for data processing and modeling in society, specifically, on alcohol consumption and abuse in Galicia. Dataset is available from a retrospective cohort study based on the telephone calls to the Galicia-061 Public Health Emergency Foundation after alcohol consumption from 1 January 2007 to 4 February 2018. Bayesian hierarchical models and nonparametric level set estimation techniques will applied.
The main objective is modeling spatial and spatio-temporal patterns of emergency calls to the department ethyl poisoning in this region. By fixing administrative areas, for example, municipalities, spatial and spatio-temporal methods for counting data can be considered in this setting. This approach allows to allow to study the evolution of callings patterns. Specifically, hierarchical modeling, through Besag York Molliè (BYM) method will be used to meet this goal. Integrated Nested Laplace Approximation will be considered in order to fit this kind of models. The analysis will be performed by using covariates such as age, gender, study level, Gini index, incomes, number of bars and regulations/sanctions.
Nonparametric level set estimation techniques will be applied in order to identify the hot-spots of emergency calls. Significant covariates detected from hierachical models fittings will be taken into account. In particular, differences between patterns by gender will be studied.

Acknowledgements: Mª José Ginzo and Paula Saavedra acknowledge financial support from Grant PID2020-116587GB-I00 funded by Agencia Estatal de Investigación (AEI) del Ministerio de Ciencia e Innovación.

References

[1] Besag, J., York, J., Molliè, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20, 1991.

[2] Rue, H., Held, L. *Gaussian Markov Random Fields*. London: Chapman and Hall, CRC Press, 2005.

Joint model for multiple longitudinal responses with informative time measurements

Inês Sousa [a,b]
isousa@math.uminho.pt

[a] *Centro de Matemática, Escola de Ciências, Universidade do Minho*
[b] *Departamento de Matemática, Escola de Ciências, Universidade do Minho*

Keywords: longitudinal, joint models

Abstract: In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristic should allow for an association between longitudinal and time measurements processes. In this work we consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In this case the follow-up time process should be considered dependent on the longitudinal outcome process. Results are presented showing that, ignoring the latent process of time measurements brings bias results when the collected time points are associated with the observed process.

References

[1] McKeigue, P. Fitting joint models of longitudinal observations and time to event by sequential Bayesian updating, *Statistical Methods in Medical Research*, 31 (10), 1934–1941, 2022.

[2] Asar, O., Bolin, D., Diggle, P.J., Wallin, J. Linear mixed effects models for non-Gaussian continuous repeated measurement data, *Journal of the Royal Statistical Society Series C-Applied*, 69 (5), 1015-1065, 2020.

[3] Szczesniak, R.D., Su, W., Brokamp, C., Keogh, R.H., Pestian, J.P., Seid, M., Diggle, P.J., Clancy, J.P. Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression, *Statistics in Medicine*, 39 (6), 740-756, 2020.

Recent Innovations in Official Statistics

— Statistics Portugal —

Inovações Recentes em Estatísticas Oficiais—Instituto Nacional de Estatística

Carlos Marcelo
Statistics Portugal, carlos.marcelo@ine.pt
Pedro Campos
Statistics Portugal, pedro.campos@ine.pt

Application of a statistical disclosure control algorithm for data protection of portuguese Census 2021

Inês Rodrigues de Sá [a], **Pedro Campos** [a]
ines.rodrigues@ine.pt, pedro.campos@ine.pt

[a] *Instituto Nacional de Estatística, DMSI/ME*

Keywords: Census 2021, statistical disclosure control, TRS

Abstract: The protection of the confidentiality of data published in the context of the 2021 Census is particularly important when we consider the growing awareness and concern of citizens regarding the privacy of their personal data. As a result of the work carried out in the pursuit of the harmonisation of methods for protecting the confidentiality of Census data in the European Statistical System, two methods were proposed to protect the confidentiality of these data: targeted record swapping (TRS) and cell key method (CKM). Both methods recommended at European level have important advantages and limitations. It appears, however, that the limitations associated with TRS do not cause difficulties that could jeopardise its application. On the contrary, the limitations associated with the CKM raise significant questions about the feasibility of its implementation. In particular, the loss of additivity of the tables caused by this method may not be well accepted by users. Additivity after application of CKM can only be achieved at the expense of loss of consistency of results - and it is recognised that this is an equally important feature for users. As such, and adding the fact that the implementation of the CKM is much more demanding in operational terms, it has been considered more relevant to use TRS as a method of protecting the confidentiality of aggregated Census data. TRS is a pre-tabular algorithm where every individual and household is assessed for uniqueness or rarity over several characteristics. Households and their individuals that are unique or rare on one or more of those characteristics are highlighted as ”risky records”, and all these households would be swapped. Similar households that match on some basic characteristics are sought from other areas to be used as ”swaps”, to preserve data quality. These characteristics included household size, so that the numbers of individuals and numbers of households in each area are preserved. The TRS was applied to the 2011 Census microdata and, from a risk and utility analysis based on six different tables, it was possible to conclude that the TRS does not lead to a significant disruption of the data, allowing to decrease the risk mainly by increasing the uncertainty associated with any attempt to disclose confidential information from the released data.

Classification of CPP - Application of a multilayer neural network

M. Conceição Ferreira[a], Almiro Moreira [b], Ana Carmona [c], David Santos [a], Rui Alves [d]
maria.ferreira@ine.pt, almiro.moreira@ine.pt, ana.carmona@ine.pt, david.santos@ine.pt, rui.alves@ine.pt

[a] *Instituto Nacional de Estatística, DMSI/II*
[b] *Instituto Nacional de Estatística, DRGD*
[c] *Instituto Nacional de Estatística, DCN/CNA*
[d] *Instituto Nacional de Estatística, DRGD/IP*

Keywords: classification, CPP, neural network, text mining

Abstract: One of the most challenging problems in dealing with Census data is the classification of open answers, such as the job classification. The 2011 Portuguese Census data regarding individual jobs were limited to web answers collected through an open answer corresponding to more than 2.5 million cases. The rest of data were collected on paper and processed with OCR (Optical Character Recognition) and, due to their particular characteristics, were left aside in this work. The Standard Occupational Classification (SOC) or CPP classification (in Portuguese) is used as a standard to classify jobs in categories. The CCP is the set of all professions existing in Portugal and their respective functional description, aggregated by professional groups. It is a fundamental instrument for statistics on occupations, both in terms of observation, analysis, consolidation of series and statistical technical coordination, and for statistical comparability at European and international level at all these common levels. The classification of occupations is relevant not only for the Census but also for other more regular statistical operations such as the Employment Survey (IE) or the Living Conditions and Income Survey (ICOR), for example. In this work we use a 1-digit classification of the CPP of the 2011 Census (Large Group levels) in a multiclass classification problem (10 classes) by applying a multilayer neural network. Word Embeddings have been used, as a type of word representation that allows words with similar meaning to have a similar representation. Roughly speaking, word embedding, transforms text into numbers. Therefore, a technique like word embedding is used to map words or phrases from a vocabulary to a corresponding vector of real numbers. The algorithm used to learn word embedding was Embedding Layer. Results show that that after evaluating the classes predicted in the test data, we find out that this model has an accuracy of 90

Estimation of free riding in plastic package waste using put-on-market and business turnover information

Joao S Lopes [a], Filipa Chambel [a], Nuno Romão [a]
joao.lopes@ine.pt, filipa.chambel@ine.pt, nuno.romao@ine.pt

[a] *Statistics Portugal*

Keywords: business turnover, economic activity, free riding, plastic packages, put-on-market

Abstract: Our work aims to calculate the annual amount of undeclared plastic packages put-on-market (POM) (i.e. free riding). Firstly, we cross information on annual business turnover (TO) with reported POM to characterize the relation k between TO and POM. Using this information, we estimate total plastic packages POM for both declarant and non-declarant businesses. This methodology relies on a few assumptions, the biggest being ratio k can successfully capture business characteristics. Accepting this assumption requires grouping businesses in moderately homogeneous clusters and calculate k for each one. This was done in a two-step procedure: a) defining an initial set of clusters; and b) choosing clusters where most of the non-declarant businesses are free riders. For the first task, we consider two characteristics of businesses, i.e. number of employees and primary economic activity, and used an *ad hoc* approach relying on previous knowledge by stakeholders and on data exploration. For the second task, we considered a coefficient c that evaluates the relative amount of free riding compared to the reported POM, and suggest to only consider clusters in which reported goods POM account for at least 20% of total goods POM. This work should be considered work in progress, providing a base estimate for future studies. Several aspects of the methodology should be further studied: using robust summary statistics for cluster characterization; considering further categories for cluster disaggregation; implementing a statistics-based algorithm to obtain an initial set of clusters; performing sensitivity analysis on threshold for cluster discarding.

Reproducibility and Statistical Disclosure Control

—Portuguese Central Bank —

Reprodutibilidade e Controlo de Confidencialidade — Banco de Portugal

Rita Sousa
Portuguese Central Bank, rcsousa@bportugal.pt

Perturbation methods: some application results

Rita Sousa [a], Jorge Morais [b], Susana Faria [b,c]
rcsousa@bportugal.pt, jogerogerio98@hotmail.com, sfaria@math.uminho.pt

[a] *Banco de Portugal, Centro de Matemática e Aplicações - FCT/UNL*
[b] *Universidade do Minho*
[c] *Centro de Matemática da Universidade do Minho*

Keywords: data perturbation, data utility, disclosure risk, package *sdcMicro*, statistical disclosure control (SDC)

Abstract: The demand for data access has been growing a lot in recent years. The compromise between the utility of the information provided and the protection of confidentiality is increasingly important. Statistical Disclosure Control (SDC) techniques suggest methods for modifying data so that they can be published without revealing confidential information that can be linked to specific respondents[2][4].
In this study, we describe and compare different perturbation approaches based mainly on linear and non-linear models. We show the advantages and disadvantages of each method for data perturbation. We also present several measures for data utility and disclosure risk to evaluate the method's performance. This study illustrates an application of these perturbation methods using functions from *sdcMicro* package in R [5].
In the literature, it is clear that Exact General Additive Data Perturbation (EGADP) and Data Shuffling produce the lowest disclosure risk and the highest data utility [3]. These conclusions are supported by the application to a real micro dataset [1]. Nevertheless, we also point out other perturbation methods as quite efficient in terms of data utility.

References

[1] Banco de Portugal Microdata Research Laboratory (BPLIM). Incentives Systems Data. *Banco de Portugal*, 2021. doi:10.17900/SI.APR2021.V1

[2] Benschop, T., Machingauta, C., Welch, M. Statistical Disclosure Control: A Practice Guide. *The World Bank*, 2021.

[3] Rao, C.R., Chakraborty, R., Sen, P. K. Handbook of Statistics. *Bioinformatics in Human Health and Heredity, 1st. North Holland IFIP*, Vol. XXVIII, 2012.

[4] Matthias, T. *Statistical Disclosure Control for Microdata: Methods and Applications in R.* Springer International Publishing, Eng. 1st Ed., Vol I., 2017.

[5] Matthias, T., Bernhard, M., Alexander, K. Package'sd-cMicro'. *Technical Report*, 2021.

Replication App

Gustavo Iglésias [a]
giglesias@bportugal.pt

[a] *Banco de Portugal, BPLIM*

Keywords: containers, replicability, reproducibility

Abstract: Reproducibility of scientific research papers is increasingly gaining traction in several fields, Economics being one of them. Top scientific journals now require access to authors' code and data, as well as good documentation on both, in order to publish research papers. Data editors make sure that the provided code and data produce the results of submitted papers. The use of confidential data by researchers makes the job of the data editor more difficult, since access to the original data is not always possible. Coordination between data editors and data centers is paramount in such cases. In recent years, BPLIM - through workshops and close contact with researchers and data editors - has been advocating for the importance of research reproducibility, trying to get researchers that work with BPLIM data to adhere to the best practices available. For researchers working with perturbed data, it is mandatory that they use an application developed by BPLIM. In this presentation, we show this application - its usage, purpose and main advantages, as well as its use of controled software environments. Although the application is only mandatory for users that work with modified data, our goal is to convince every researcher working with BPLIM datasets to use it. We hope that researchers outside of BPLIM use it as well, since there is a public version under development.

Acknowledgements: BPLIM Team

Confidentiality and the statistical data sharing

Diogo Barbosa [a], **João Falcão Silva** [a]
dfbarbosa@bportugal.pt, jmfsilva@bportugal.pt

[a] *Banco de Portugal*

Keywords: non-financial corporations, statistical confidentiality

Abstract: Data sharing is crucial for the compilation process and uses of the statistical data. From the compiler's perspective, it is essential to ensure that the statistical information is provided accurate, timely delivered and complete. Nevertheless, the information that is published must preserve individual confidentiality, ensuring that no reporting or economic agent is identified. In addition, it is very important that the users of the statistical data, are aware and understand the limitations on the data sharing, namely the confidentiality criteria that is used. For the purpose of the economic analyses, these limitations can be addressed in advance through the implementation of robust analyses processes, that can make the best use of the existing information.

This article aims to identify statistical confidentiality rules and the rationale for their existence. Furthermore, these rules are implemented in the analysis carried out on data published by Banco de Portugal on non-financial corporations. Finally, the effects of statistical confidentiality on the different dimensions published data will be considered.

Acknowledgements: The authors are thankful to the Central Balance Sheet Division colleagues of Banco de Portugal, for their contributions.

References

[1] D′Aguiar, L., Menezes, P. Impact of benefits of micro-databases' integration on the statistics of the Banco de Portugal. *Proceedings of the 59th World Statistics Congress Hong Kong*, 2013.

[2] Duncan, G., Elliot, M., Salazar-González, J.-J. *Statistics for Social and Behavioral Sciences*. Springer, New York, New York 2011. doi:https://doi.10.1007/978-1-4419-7802-8_1

[3] ECB Statistical confidentiality protection in the european system of central banks. *Confidentiality Report 2017 (Summary)*, 2018.

Rising Stars

— SPE/CLAD —

Estrelas em Ascensão — SPE/CLAD

M. Rosário Oliveira
CEMAT, Instituto Superior Técnico, rosario.oliveira@tecnico.ulisboa.pt
Adelaide Figueiredo
Universidade do Porto e LIAAD INESC TEC Porto, adelaide@fep.up.pt

A utilização de funções de penalização na seleção de variáveis em misturas de modelos de regressão com efeitos aleatórios

Luísa Novais[a], Susana Faria[a]
luisa_novais92@hotmail.com, sfaria@math.uminho.pt

[a] *Universidade do Minho*

Keywords: algoritmo CEM, algoritmo EM, máxima verosimilhança penalizada, modelos de mistura, seleção de variáveis

Abstract: Ao longo das últimas décadas, os modelos de mistura têm sido amplamente utilizados na modelação de dados provenientes de populações heterogéneas. Contudo, os avanços tecnológicos dos últimos anos provocaram a existência de conjuntos de dados contendo um elevado número de observações e/ou um elevado número de variáveis, pelo que a seleção de variáveis possui um papel fundamental no estudo de modelos de mistura, envolvendo a procura de um modelo o mais simples possível, mas que descreva adequadamente os dados observados.
No entanto, os métodos de seleção de variáveis clássicos requerem intensa computação, em particular em modelos de mistura, mesmo na modelação com um número moderado de variáveis explicativas. Como tal, dada a grande complexidade da generalidade das bases de dados atuais, surgiu a necessidade de desenvolver novas metodologias mais eficientes e que permitam acomodar a complexidade computacional existente, como é o caso dos métodos baseados em funções de penalização. Neste trabalho investigam-se diferentes métodos de seleção de variáveis baseados em funções de penalização que atuam sobre os coeficientes das variáveis, em particular os métodos *Least Absolute Shrinkage and Selection Operator* (LASSO), *Adaptive Least Absolute Shrinkage and Selection Operator* (ALASSO), HARD e *Smoothly Clipped Absolute Deviation* (SCAD), comparando-se o seu desempenho na identificação do subconjunto de variáveis mais relevantes em misturas de modelos de regressão com efeitos aleatórios, recorrendo aos algoritmos *Expectation-Maximization* (EM) e *Classification Expectation-Maximization* (CEM).

Acknowledgements: O trabalho de L. Novais foi financiado pela FCT - Fundação para a Ciência e a Tecnologia, através da bolsa de doutoramento com a referência SFRH/BD/139121/2018.

References

[1] Celeux, G., Govaert, G. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3), 315–332, 1992. doi:10.1016/0167-9473(92)90042-E

[2] Dempster, A.P., Laird, N.M., Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodol.)*, 39(1), 1–22, 1977. doi:10.1111/j.2517-6161.1977.tb01600.x

[3] Novais, L., Faria, S. Variable selection using the EM and CEM algorithms in mixtures of linear mixed models. *Journal of Statistical Computation and Simulation*, 1–36, 2023. doi:10.1080/00949655.2023.2176503

An interactive R-Shiny app for the identification of atypical clusters

Ana Helena Tavares[a,c], Vera Afreixo[b,c], Diana Lucas[c], Paula Brito[d]
ahtavares@ua.pt, vera@ua.pt, dianalucas@ua.pt, mpbrito@fep.up.pt

[a] *Águeda School of Technology and Managment, University of Aveiro*
[b] *Department of Mathematics, University of Aveiro*
[c] *Center for Research & Development in Mathematics and Applications (CIDMA)*
[d] *FEP, University of Porto & LIAAD INESC TEC*

Keywords: clustering, distributional data, outlyingness, simulation

Abstract: In this work, we explore a procedure for the identification of atypical group of observations. By atypical group, we mean a cluster of observations whose 'mean' pattern stands out from the majority of the 'mean' patterns of the remaining clusters. Our work focus on data whose elements are discrete distributions. The key idea of our proposal is to combine a clustering method with a functional outlyingness criterion to capture atypical class prototypes. To identify a partition of the distributional data we iteratively combine two steps: the first creates a partition, while the second flags atypical curves within each cluster, based on a measure of functional outlyingness [1]. Clusters with atypical curves, are forwarded for (sub)clustering, and the procedure is repeated until no outlying curves are identified in such clusters. Once the final partition is obtained, each cluster is represented by a class prototype, whose outlyingness is evaluated according to the same functional approach. Clusters with an atypical class prototype are pointed as atypical.

We developed an R-Shiny app designed specifically for implementing the proposed method and to perform a simulation study. The procedure is applied to investigate clusters of genomic words in human DNA by studying their inter-word lag distributions [2]. These experiments demonstrate the potential of the new method for identifying clusters of distributions with outlying patterns.

Acknowledgements: This work was partially supported by the CIDMA through the Portuguese Foundation for Science and Technology (FCT), references UIDB/04106/2020 and UIDP/04106/2020. P. Brito acknowledge support by the Portuguese funding agency, FCT, within project LA/P/0063/2020.

References

[1] Rousseeuw, P. J., Raymaekers, J., Hubert, M. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27:2, 345–359, 2018. doi:10.1080/10618600.2017.1366912

[2] Tavares, A.H., Raymaekers, J., Rousseeuw, P.J. et al. Clustering genomic words in human DNA using peaks and trends of distributions. *Advances in Data Analysis and Classification*, 14, 57–76 (2020). doi:10.1007/s11634-019-00362-x

Multivariate time analysis via multilayer quantile graphs

Vanessa Freitas Silva [a], Maria Eduarda Silva [b], Pedro Ribeiro [a], Fernando Silva [a]
vanessa.silva@fc.up.pt, mesilva@fep.up.pt, pribeiro@fc.up.pt, fmsilva@fc.up.pt

[a] *CRACS-INESC TEC, Faculdade de Ciências, Universidade do Porto*
[b] *LIAAD-INESC TEC, Faculdade de Economia, Universidade do Porto*

Keywords: dimensionality reduction, multilayer networks, multivariate time series, quantile graphs

Abstract: Multi-dimensional temporal data have become, in recent years, predominant in the most varied scientific fields. Such data can measure several different variables for a long time which induces huge volumes of data to be analyzed. Issues related to serial and cross-dependency and the high-dimensionality of data are fundamental and open problems in time series analysis. The network-based approach is a complementary and very promising way of analyzing time series data [1, 2] whose focus in recent years has turned to multivariate time series data. In the univariate context, time series transition properties can be captured by *Quantile Graphs* that allow data dimensionality reduction by mapping the observations for a reduced number of sample quantiles. Thinking about the problem of the dimensionality curse that is more and more prominent, we extend the quantile graphs to a multivariate version for multivariate time series analysis, the *Multilayer Quantile Graphs*. In this mapping method, each time series is replaced with the corresponding quantile graph and the contemporaneous quantiles of a pair-wise time series are associated by inter-layer connections that map the cross-dimensions of contemporary transitions. The resulting multilayer network has a much smaller dimension than the original multivariate time series and allows us to characterize and analyze the serial and cross-dimension dynamic transitions of the data via network analysis. In this work, we present and illustrate this mapping method on multivariate real-world context datasets, we explore the topological structures of the resulting network, and we analyze the topological features for the analysis of the original datasets.

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

References

[1] Silva, V.F., Silva, M.E., Ribeiro, P., Silva, F. Time series analysis via network science: Concepts and algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(11), e1404, 2021. doi:10.1002/widm.1404

[2] Silva, V.F., Silva, M.E., Ribeiro, P., Silva, F. Novel features for time series analysis: a complex networks approach. *Data Mining and Knowledge Discovery*, 36, 1062–1101, 2022. doi:10.1007/s10618-022-00826-3

The role of translation quality estimation at Unbabel

José Maria Pombal [a], André F. T. Martins [a]
jose.pombal@unbabel.com, andre.martins@unbabel.com

[a] *Unbabel*

Keywords: large language models, quality estimation, translation

Abstract: Unbabel's translation pipeline can be summarized in three steps: translation, quality estimation, and post-edition (if necessary). The quality estimation phase plays a critical role in monitoring the quality of machine-generated translations across various languages and domains. It ensures that translations meet client standards and align with specific business requirements. Additionally, it effectively identifies imperfect outputs that require post-edition. In high-stakes translation domains, such as medical reports, where even the smallest error in a quantity or terminology can have disastrous consequences quality estimation becomes increasingly vital.

In this talk, we will provide an overview on Unbabel's translation pipeline, highlighting the approaches employed at each step. Our focus will be on how quality estimation is powered by artificial intelligence (AI), and how it can be leveraged to guarantee the correctness and robustness of translations across domains. Furthermore, we will delve into the role of large language models in the ongoing paradigm shift of quality estimation. Traditionally, quality estimation was performed at the segment level, assigning a single quality score to each segment in a document, with word-level analysis receiving less attention. With the advent of large language models, we are witnessing a shift in focus towards word-level quality estimation, which enables more comprehensive and informative quality reports.

Oral Sessions

—Comunicações Orais—

Robust clustering based on trimming and choice of parameters

Luis Angel García-Escudero [a], Christian Hennig [b], Agustín Mayo-Iscar [a], Gianluca Morelli [c], Marco Riani [c]
lagarcia@uva.es, christian.hennig@unibo.it, agustin.mayo.iscar@uva.es, gianluca.morelli@unipr.it, marco.riani@unipr.it

[a] *University of Valladolid*
[b] *University of Bologna*
[c] *University of Parma*

Keywords: clustering, outliers, robustness, trimming

Abstract: Outliers can have a very negative impact on various statistical procedures, and thus, it is recommended to employ robust alternatives. With this idea in mind, robust clustering methods have been developed to better handle outlying observations in Cluster Analysis. The TCLUST method, introduced in [1], is a robust clustering approach based on "impartial" trimming, which involves specifying the number of clusters k and the trimming level α. The term "impartial" means that the data set itself determines the fraction α of observations to be trimmed. TCLUST includes an eigenvalues-ratio constraint that prevents the detection of "spurious" (non-interesting) clusters.
In [2], a graphical procedure was proposed to select sensible values for k and α. This procedure relies on the visual examination of the so-called "classification trimmed likelihood" curves. Some theoretical background will be provide on this graphical tool and the underlying elements involved are analyzed. Furthermore, a parametric bootstrap method will be presented to generate a reduced list of sensible selections for the k and α parameters. This automated approach eliminates the need for visual inspection of the curves and minimizes subjective decisions associated with sample variability. Using this reduced list, users can select the robust cluster partition that best aligns with their specific clustering objectives by utilizing standard cluster validation tools.

Acknowledgements: Research partially supported by Spanish Ministerio de Ciencia e Innovaciòn, grant PID2021-128314NB-I00.

References

[1] García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar. A. A general trimming approach to robust cluster analysis, *Annals of Statistics*, 36, 1324–1345, 2008. doi:10.1214/07-AOS515

[2] García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar. A. Exploring the number of groups in robust model-based clustering, *Statistics and Computing*, 21, 585–599, 2011. doi:10.1007/s11222-010-9194-z

Prasanta Chandra Mahalanobis and his praised Mahalanobis Distance

João A. Branco [a], Ana M. Pires [a]
jbranco1802@gmail.com, ana.maria.n.pires@gmail.com

[a] *Department of Mathematics, IST, University of Lisbon*

Keywords: high dimension low sample size, Mahalanobis, Mahalanobis distance, multivariate analysis

Abstract: Mahalanobis was one of the most eminent statisticians of all times. In brief we highlight his attitude towards the value of science and how its advancement should be conducted. Major contributions all along his intense statistical activity will be mentioned to reaffirm the diversity of his work and its importance in the development of statistics. Since 1936 onwards Mahalanobis Distance (MD) has been a crucial statistical tool used in many applied fields where multivariate analysis is at stake. However, as other traditional statistical methods and concepts, it faces the challenge brought about by the advent of real modern data. We look at the case where the number of variables, p, approaches de number of observations, n, and assist to a progressive degradation of MD until we observe its complete degeneration when $p \geq n - 1$. The consequences are surprisingly devastating and should be known by all who attempt to use MD and related concepts under these conditions.

Perfil de risco de um cliente que entra em incumprimento - crédito à habitação

Sofia Comparada [a], Eduardo Severino [a,b], Teresa Alpuim [a,b]
sofia.reganha@hotmail.com, jeseverino@fc.ul.pt, mtalpuim@fc.ul.pt

[a] *Faculdade de Ciências da Universidade de Lisboa, Portugal*
[b] *CEAUL,Faculdade de Ciências da Universidade de Lisboa, Portugal*

Keywords: árvores de decisão, clusters, incumprimento, regressão logística, risco

Abstract: Com o objetivo de identificar quais os fatores mais determinantes para a entrada em incumprimento, foi analisada e tratada uma amostra de clientes composta por características associadas a cada cliente e ao crédito concedido e, tendo em conta o estado de incumprimento daquele, ajustaram-se três modelos de regressão logística. O primeiro modelo apenas continha as variáveis que representam as características inerentes aos clientes e não conduziu a bons resultados (AIC e AUC). Os outros dois modelos incluíram também variáveis macroeconómicas de duas formas distintas: uma em que foram incluídas individualmente e a outra em que se incluíram componentes principais associadas às mesmas, apresentando resultados idênticos. Seguidamente, e como se pretendia segmentar os clientes de acordo com os seus perfis, as variáveis que se revelaram mais significativas em cada um dos modelos e que eram numéricas foram categorizadas e incluídas desta forma. Ajustaram-se dois novos modelos (com as variáveis macroeconómicas e com as suas componentes principais) com as variáveis nesta forma. Ambos conduziram a resultados melhores e semelhantes. Finalmente, com base nas variáveis mais relevantes em cada um dos modelos construíram-se duas árvores de decisão, com o objetivo de mais facilmente discriminar um cliente enquanto incumpridor, ou não. Em suma, independentemente do modelo analisado, concluiu-se que a antiguidade da conta, a situação profissional e a profissão do cliente, a finalidade e o número de intervenientes do crédito e o rácio entre o montante do empréstimo e do imóvel são as variáveis essenciais para apoiar na tomada de decisão do Banco.

Joint modeling of longitudinal binary responses: A nonparametric Bayesian approach using acute and chronic malnutrition data

André Nunes [a,b], Giovani Silva [a,b], Luzia Gonçalves [b,c]
andre.b.nunes@tecnico.ulisboa.pt, giovani.silva@tecnico.ulisboa.pt, LuziaG@ihmt.unl.pt

[a] *Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[c] *Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa*

Keywords: bayesian nonparametric, longitudinal data, mixed model, MCMC methods

Abstract: In the analysis of longitudinal data, Gaussian random effects are usually adopted in order to control the heterogeneity of the individuals observed over time. On this work is questioned the advantage of making that Gaussian assumption for analysing binary outcomes that are jointly distributed and dependent according to Bayesian Nonparametric (BNP) random effects. Notice that conditional on the random effects the joint binary response variables are considered independent, facilitating the construction of the likelihood function for later inference about the model parameters. Statistical data analysis has traditionally used parametric methods such as Gaussian distribution. Nonparametric and semiparametric approaches have the advantage of not imposing a given distribution for either responses or random effects. We assume BNP random effects and keep Bernoulli distribution for binary responses. Regarding the BNP, Dirichlet process (DP), stick-breaking process, Pitman-Yor process, and Polya's tree are performed and are implemented in R package NIMBLE. Markov Chain Monte Carlo (MCMC) methods for making inference on the model parameters are also employed. Finally, our motivation is due to a binary data study from a longitudinal four-arm randomized parallel trial conducted in Bengo Province to explore the effects of four interventions on wasting and stunting response variables. A total of 121 children with intestinal parasitic infections received baseline treatment, and were allocated to the four arms.

Acknowledgements: The authors namely thank Carolina Gasparinho do Centro de Investigação em Saúde de Angola (CISA).

References

[1] Gasparinho, C., Gonçalves, M.H., Chissaque, A., Silva, G.L., Fortes, F., Gonçalves, L. Wasting, stunting, and anemia in angolan children after deworming with albendazole or a test-and-treat approach for intestinal parasites: binary longitudinal models with temporal structure in a four-arm randomized trial. *Nutrients,* 24(14), 2185, 2022. doi:10.3390/nu14112185

Simulation study to compare the performance of signed klotz and the signed mood weighted generalized coefficients

Sandra M. Aleixo [a,b], Júlia Teles [c,d]
sandra.aleixo@isel.pt, jteles@fmh.ulisboa.pt

[a] *CEAUL – Centro der Estatística e Aplicações da Universidade de Lisboa*
[b] *ISEL – Instituto Superior de Engenharia de Lisboa / IPL – Instituto Politécnico de Lisboa*
[c] *CIPER – Centro Interdisciplinar de Performance Humana*
[d] *FMH – Faculdade de Motricidade Humana / UL – Universidade de Lisboa*

Keywords: monte carlo simulation, signed klotz coefficient, signed mood coefficient, van der waerden coefficient, weighted coefficients

Abstract: In this work a Monte Carlo simulation study was carried out to compare the performance of three weighted coefficients that emphasized the top and bottom ranks at the same time, namely the signed Klotz and the signed Mood weighted coefficients previously proposed by the authors [2] and the van der Waerden weighted coefficient [3], with the Kendall's coefficient that assigns equal weights to all rankings. As the main result of the simulation study, we highlight the best performance of Klotz coefficient in detecting concordance in situations where the agreement is located in a lower proportion of extreme ranks, contrary to the case where the agreement is located in a higher proportion of extreme ranks, in which the Signed Mood and van der Waerden coefficients have best performance.

Acknowledgements: Sandra M. Aleixo was partly supported by the Fundação para a Ciência e Tecnologia, under Grants UIDB/00006/2020 and UIDP/00006/2020 to CEAUL - Centre of Statistics and its Applications. Júlia Teles was partly supported by the Fundação para a Ciência e Tecnologia, under Grant UIDB/00447/2020 to CIPER - Centro Interdisciplinar para o Estudo da Performance Humana (unit 447).

References

[1] Aleixo, S. M., Teles, J. Weighting Lower and Upper Ranks Simultaneously Through Rank-Order Correlation Coefficients. In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2018. ICCSA 2018. *Lecture Notes in Computer Science*, Vol 10961, 318–334, Springer, Cham, 2018.

[2] Aleixo, S. M., Teles, J. Weighted Coefficients to Measure Agreement among Several Sets of Ranks Emphasizing Top and Bottom Ranks at the Same Time. In S. Misra et al. (eds) Computational Science and Its Applications – ICCSA 2019. ICCSA 2019. *Lecture Notes in Computer Science*, Vol 11620 (Part II) 23–33, Springer, Cham, 2019.

[3] Hájek, J. , Šidák, Z. *Theory of Rank Tests. Academic Press*, New York, 1972.

Exploring the mutual information rate decomposition in situations of pathological stress

Helder Pinto [a], Celeste Dias [b], Ana Paula Rocha [a]
helder.pinto@fc.up.pt, mcdias@med.up.pt, aprocha@fc.up.pt

[a] *Centro de Matemática de Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto*
[b] *Centro Hospitalar de São João, Faculdade de Medicina, Universidade do Porto*

Keywords: coupling, entropy, information decompositions, information theory

Abstract: In information theory, the coupling between two processes can be assessed using Mutual Information Rate (MIR) decomposition of the entropy rate or the conditional mutual information [1]. We have applied a non-parametric approach based on nearest-neighbours to compute these decompositions in the study of cardiorespiratory interactions [2]. Plateau waves (PWs) are a specific pattern of Intracranial Pressure (ICP) changes observed in patients with severe traumatic brain injuries (TBI), characterized by a sudden sustained increase in ICP. They are often associated with relevant changes in Heart Rate Variability (HRV), reflecting an Automonic Nervous System (ANS) dysfunction [3]. In this work,we explore the coupling between ICP and HRV by decomposing the Mutual Information Rate. This framework is first tested on simulations of linear and non-linear bivariate systems, then it is applied to data consisting of Heart Period and Intracranial Pressure time series measured in TBI patients with PW occurrence. The obtained results evidence that MIR decompositions are able to highlight the interdependence of HRV and ICP in PWs episodes, suggesting that these critical phenomenon are associated with autonomic stress.

Acknowledgements: Work supported by CMUP, financed by FCT – Fundação para a Ciência e Tecnologia, I.P., projects UIDB/00144/2020, UIDP/00144/2020. H.P. thanks FCT, for the Ph.D. Grant 2022.11423.BD.

References

[1] Miao, H., Zhang, F., Tao, R. Mutual information rate of nonstationary statistical signals. *Signal Processing*, 171, 107531, 2020. doi:10.1016/j.sigpro.2020.107531.

[2] Pinto, H., Antonacci, Y., Pernice, R., Barà, C., Javorka, M., Faes, L., Rocha, A.P. Decomposing the Mutual Information Rate of Heart Period and Respiration Variability Series to Assess Cardiorespiratory Interactions, *Annu Int Conf IEEE Eng Med Biol Soc. 2023 Jul. In press.*

[3] Pinto, H., Dias, C., Rocha, A.P. Multiscale Information Decomposition of Long Memory Processes: Application to Plateau Waves of Intracranial Pressure. *Annu Int Conf IEEE Eng Med Biol Soc. 2022 Jul*; 2022:1753-1756. doi:10.1109/EMBC48229.2022.9870925.

Classification and survival analysis of multi-omics data for the identification of novel diagnostic and prognostic biomarkers in glioma

Francisca G. Vieira [a], Regina Bispo [a,b], Marta B. Lopes [a,b,c]
fmg.vieira@campus.fct.unl.pt, r.bispo@fct.unl.pt, marta.lopes@fct.unl.pt

[a] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Portugal*
[b] *Department of Mathematics, NOVA School of Science and Technology, Portugal*
[c] *UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Portugal*

Keywords: canonical correlation analysis, classification, glioma, multi-omics, survival analysis

Abstract: Glioma is one of the most common types of primary brain cancer. Given its high level of heterogeneity, many efforts have been made to classify the type of glioma in each patient, which is critical to improve early diagnosis and survival. Recent sequencing technologies allow the extraction of an increasing quantity of biological data, with multiple omics modalities collected from the same individuals. Given the high-dimensional nature of this data, its multivariate analysis is particularly challenging and variable selection is required, since only a smaller subset of variables is desired for interpretation. In contrast to the conventional canonical correlation analysis, which application fails in a high-dimensional context and includes all variables in the fitted vectors, a recent method DIABLO [1] extends it to a sparse variant applicable to more than two datasets. Using TCGA data, along with the reclassified patient labels, this study explores this method for the classification of different glioma types, while integrating multiple omics layers. We further investigate the effect of the selected variables (with non-zero coefficients in the canonical variates) in survival probability, using the non-parametric Kaplan-Meier estimator. Hence, this research uses DIABLO to (1) find consistent patterns across datasets that vary between the disease types, and (2) identify novel biomarkers to each group with impact on patient survival.

Acknowledgements: This work was supported by the FCT - Fundação para a Ciência e a Tecnologia, I.P., with references PTDC/CCI-BIO/4180/2020 (MONET project), CEECINST/00042/2021, UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI).

References

[1] Singh, A., Shannon, C.P., Gautier, B., et al. "DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays". *Bioinformatics*, 35.17, 3055–3062, 2019. doi:10.1093/bioinformatics/bty1054

A study on the Pearson and mean deviance estimators of the dispersion parameter in Poisson regression

Rui Miranda [a], Rita Gaio [a,b]
up201804962@fc.up.pt, argaio@fc.up.pt

[a] *Department of Mathematics, Faculty of Sciences, University of Porto, Portugal*
[b] *Center of Mathematics of the University of Porto, Portugal*

Keywords: dispersion parameter, mean deviance estimator, pearson estimator, poisson regression

Abstract: In Poisson regression, the dispersion parameter ϕ of the exponential family of distributions is equal to 1, due to the equality between mean and variance. The accuracy of the estimation of ϕ is crucial for the identification of overdispersion, a common problem in count data that leads to invalid inferences, as a consequence of incorrect standard errors for the estimation of the regression coefficients. Current estimators of ϕ include the Pearson estimator $\hat{\phi}_P$ and the Mean Deviance Estimator $\hat{\phi}_D$, defined in the framework of Poisson fixed-effects regression, which are advertised to keep their properties when applied to Poisson mixed models, where their use is also widely spread. In the present work, we study the properties of these estimators. For the Poisson mixed model consisting of a fixed and a random intercept, and with a *small* conditional mean, we implemented a simulation procedure to study the mean and variance of $\hat{\phi}_D$ and $\hat{\phi}_P$, and their dependence on the fixed and random components. Our results suggest that these estimators are, in fact, sensitive to the presence of random effects. In particular, the greater the variance of the random effect, the greater the deviation from the true value of the dispersion parameter, especially if the fixed component of the linear predictor is *small*.

Acknowledgements: Rita Gaio was partially supported by CMUP, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project with reference UIDB/00144/2020.

References

[1] Dunn, P. K., Smyth, G. K. *Generalized Linear Models With Examples in R*. Springer, New York, 2018. doi:10.1007/978-1-4419-0118-7

[2] Payne, E. H. et al. An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data, *Communications in Statistics - Simulation and Computation*, 47:6, 1722–1738, 2018. doi:10.1080/03610918.2017.1323223

Model-based clustering of distributional data
An application in Official Statistics

Paula Brito [a,b], A. Pedro Duarte Silva [c]
mpbrito@fep.up.pt, psilva@ucp.pt

[a] *Faculdade de Economia, Universidade do Porto*
[b] *LIAAD - INESC TEC*
[c] *Católica Porto Business School & CEGE, Universidade Católica Portuguesa*

Keywords: clustering, finite mixture model, histogram data, symbolic data analysis

Abstract: The classical data representation model, where for each statistical unit a single value is recorded for each variable, is too restrictive when the data to be analysed are not real numbers or single categories but comprise variability. Symbolic Data Analysis (see e.g. [1]) provides a framework for the representation and analysis of such data. In this talk, we consider numerical distributional data [1], i.e., data where units are described by histogram or interval-valued variables, representing intrinsic variability of the corresponding observations. Parametric probabilistic models are introduced, which are based on the representation of each distribution by a location measure and interquantile ranges, extending the model for interval-valued data presented in [3]. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. This framework then allows for multivariate parametric analysis of distributional data. Leveraging on the proposed model, we propose a model-based approach for clustering numerical distributional data, following the approach in [4]. The algorithm relies on a suitable adaptation of the EM algorithm to the likelihood maximization, for the different covariance configurations. Applications to official data illustrate the proposed approach, putting in evidence its added value in this context.

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

References

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

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

[3] Brito, P., Duarte Silva, A.P. Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, 39(1), 3–20, 2012.

[4] Brito, P., Duarte Silva, A.P., Dias, J.G. Probabilistic clustering of interval data. *Intelligent Data Analysis*, 293-313, 2015.

PCA from compositional and symbolic perspectives in the study of mortality in Portugal

Marta Maltez [a,b], Adelaide Freitas [a,b], Magda Monteiro [a,c]
martamaltez@ua.pt, adelaide@ua.pt, msvm@ua.pt

[a] *CIDMA, University of Aveiro*
[b] *Department of Mathematics, University of Aveiro*
[c] *Águeda School of Technology and Management, University of Aveiro*

Keywords: compositional vector, principal component analysis, symbolic data, symbolic histogram variable

Abstract: Principal component analysis is a useful tool for identifying the dominant direction of variation of multivariate data. Compositional data are constrained positive data reflecting an overall composition such as histograms of categorical variables or percentages of parts within a whole. An observation defined by a composition of p D-compositional variables (i.e., p variables each one with D-part compositional components) is a compositional data. This type of multivariate observation is referred to as a (p-dimensional) compositional data vector. Symbolic data, such as histograms of numerical variables, emphasizes the distribution of values taking into account the internal variability of each class. The computation of the symbolic frequency histogram involves a count of the number of individual descriptions that match a certain logical dependency in the data. It is possible to extend the concept of histogram in symbolic data to the p-dimensional captures relationships and interactions between categorical variables. Compositional data and symbolic data represent then types of complex data structures that can describe phenomena that conventional data is unable to explain.
The aim of this work is to apply PCA on both compositional vectors and histogram symbolic data on a set of data related to the cause of death in Portugal in 2020. The data set summarizes deaths by cause of death, age group and region. In addition, this works aspires to show the results and conclusions from two different perspectives.

"PACE-Gate": statistical lessons learned from a high-profile and controversial clinical trial

Nuno Sepúlveda [a,b]
N.Sepulveda@mini.pw.edu.pl

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

Keywords: power calculation, primary endpoint, ROC curve, sample size determination

Abstract: In 2011, the high-profile medical journal *The Lancet* published the large and controversial PACE trial for the treatment of Chronic Fatigue Syndrome [1]. The main finding was that cognitive behavioural and graded exercise therapies were both able to reduce fatigue and physical function in the enrolled patients more than specialised medical care and adaptive pacing therapy. Both patients and research community reacted to this finding with skepticism and subsequent independent analyses showed evidence for a flawed trial (though, vehemently refuted by the trial's authors). This case is now known as the "PACE-Gate" [2] alluding to eventual similarities with the famous Watergate scandal during the Richard Nixon presidency. In this talk, I will discuss this trial from a pure statistical standpoint by simply analysing the published protocol and outcomes reported in the original paper. This discussion will include topics such as sample size determination, the intention-to-treat analysis, pre- versus post-power calculations, and the adjustment of trial's primary endpoints for hypothetical placebo (or nocebo) effects. Finally, I will provide some practical recommendations to improve trial's protocol and the respective reporting under the 2010 CONSORT guidelines.

References

[1] White, P. D., Goldsmith, K. A., Johnson, A. L., Potts, L., Walwyn, R., DeCesare, J. C., Baber, H. L., Burgess, M., Clark, L. V., Cox, D. L., Bavinton, J., Angus, B. J., Murphy, G., Murphy, M., O'Dowd, H., Wilks, D., McCrone, P., Chalder, T., Sharpe, M. & PACE trial management group (2011). Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial. *Lancet*, 377, 823–836, 2011. doi:10.1016/S0140-6736(11)60096-2

[2] Geraghty K. J. 'PACE-Gate': When clinical trial evidence meets open data access. *Journal of Health Psychology*, 22(9), 1106-1112, 2017.

A joint model for multiple (un)bounded longitudinal outcomes, competing risks, and recurrent events

Pedro Miranda Afonso[a,b], Dimitris Rizopoulos[a,b], Anushka Palipana[c], John P. Clancy[d], Rhonda D. Szczesniak[c], Eleni-Rosalina Andrinopoulou[a,b]
p.mirandaafonso@erasmusmc.nl

[a] *Department of Biostatistics, Erasmus University Medical Center*
[b] *Department of Epidemiology, Erasmus University Medical Center*
[c] *Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center*
[d] *Cystic Fibrosis Foundation*

Keywords: bounded outcome, competing risks, joint model, multivariate longitudinal data, recurrent events

Abstract: Cystic fibrosis (CF) is a genetic disease that affects the lungs and digestive system. Clinicians are interested in the associations between lung function ($ppFEV_1$) decline, changes in nutritional status (BMI), and the risks of recurrent acute pulmonary exacerbations (PEx), lung transplantation, and death. Previous analyses have been limited to the first PEx, disregarding subsequent occurrences and neglecting informative censoring due to transplantation and death. Furthermore, despite $ppFEV_1$ being bounded, it has previously been modeled using a Gaussian distribution, leading to predictions outside the feasible range. We present a Bayesian shared-parameter joint model for recurrent events, competing risks, and multiple longitudinal markers using all available US CF Foundation Patient Registry data. We model $ppFEV_1$ and BMI assuming beta and Gaussian distributions, respectively. We allow the specification of various functional forms to link the longitudinal and time-to-event processes. Our model accommodates discontinuous risk intervals and both the gap and calendar timescales. The model is available in the R package `JMbayes2`. A highlight of our results is the finding that a ten-unit increase in the rate of $ppFEV_1$ decline increases the hazard of PEx by 14.69% (95%CI 13.09–14.69%). The incidence of PEx is positively associated with transplantation and death, with a one-standard-deviation increase in the frailty term increasing the hazard by 290.74% (95%CI 264.96–317.43%) and 229.95% (95%CI 211.98–247.93%), respectively. Our model allows for improved estimates of the risks posed by PEx and can support monitoring strategies to reduce the frequency of episodes.

References

[1] Andrinopoulou, E.-R., Clancy, J.P., Szczesniak, R.D. Multivariate joint modeling to identify markers of growth and lung function decline that predict cystic fibrosis pulmonary exacerbation onset. *BMC Pulmonary Medicine*, 20, 1–11, 2020. doi:10.1186/s12890-020-1177-z

[2] Rizopoulos, D., Papageorgio, G., Afonso, P.M. JMbayes2: Extended joint models for longitudinal and time-to-event data. 2023. https://drizopoulos.github.io/JMbayes2/

Fail-safe number: a simulation study

Vera Afreixo [a], Filipa Rocha [a]

[a] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*

Keywords: fail-safe number, meta-analysis, publication bias, simulation

Abstract: The fail-safe number (FSN) refers to a statistical concept used to assess the robustness of the rejection of the null hypothesis through the overall meta-analytic effect. The fail-safe number estimates the number of hypothetical studies with null or reverse results that would need to be added to the analysis to nullify or reverse the estimated overall effect. If the fail-safe number is large, a substantial number of unpublished studies with null findings would be needed to negate the observed effect, thus indicating a relatively robust result. On the other hand, if the fail-safe number is small, it raises concerns about the possibility of the "file drawer problem", where studies with non-significant results are less likely to be published. The FSN is a statistical tool that provides an estimate and should be interpreted with caution. Other factors, such as study quality, heterogeneity, and the overall body of evidence, should also be considered when evaluating the credibility and generalization of meta-analytic results.

In this study, we propose the incorporation of the Iyengar and Greenhouse modification [1] into the Rosenberg FSN approach [2]. We present a comprehensive simulation that evaluates and contrasts various FSN estimates (including Rosenthal FSN, Orwin FSN, and Rosenberg FSN) with and without the implementing the Iyengar and Greenhouse modification. We also discuss the implications of these results within the context of concrete meta-analysis studies.

Acknowledgements: This research was funded by Portuguese funds through CIDMA, The Center for Research and Development in Mathematics and Applications of University of Aveiro, and the Portuguese Foundation for Science and Technology (FCT–Fundação para a Ciência e a Tecnologia), within projects UIDB/04106/2020 and UIDP/04106/2020.

References

[1] Iyengar, S., Greenhouse, J.B. Selection models and the file drawer problem. *Statistical Science*, 109–117, 1988. doi:10.1037/0033-2909.86.3.638

[2] Rosenberg, M.S. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, 59(2),464–468, 2005. doi:10.1554/04-602

Speckle Tracking Echocardiography in detecting myocardial deformation in the left ventricle: Systematic Review and Meta-Analysis

Helena Cardoso [a], Brigida Monica Faria [a,b], Rita Amaral [a, c]
10160273@ess.ipp.pt, monica.faria@ess.ipp.pt, rml@ess.ipp.pt
[a] *ESS, Polytechnic of Porto (ESS-P.PORTO), Porto, Portugal*
[b] *Laboratory of Artificial Intelligence and Computer Science (LIACC), University of Porto, Porto, Portugal*
[c] *Research Center in Health Technologies and Services (CINTESIS) – University of Porto, Porto, Portugal*

Keywords: left ventricular strains, myocardial deformation, speckle tracking echocardiography

Abstract: Transthoracic Echocardiogram studies the mobility of the left ventricular wall through the ejection fraction. However, Speckle Tracking is a recent modality that allows the assessment of this function through strains that assess the relative changes in the extent/thickness of the myocardium to the original shape. This paper presents a Systematic Review to analyze the specificity of Speckle Tracking in kinetic changes, compared to conventional Transthoracic Echocardiography. Observational studies were searched in three electronic databases (PubMed, B-on, MDPI) to determine which method is most accurate for evaluating patients with kinetic or ischemic changes. Reviewers independently extracted data and assessed the quality of evidence with GRADE (Grading of Recommendations Assessment, Development and Evaluation). Pooled studies were analyzed using a random effects model and results were presented as standardized mean differences. Fifty-six articles met the inclusion and review criteria; 10 articles were grouped in meta-analysis. The combined mean of ejection fraction and strains was lower in patients with cardiomyopathy, myocardial infarction, and coronary artery disease when compared to control groups. However, there was no difference in patients with left ventricular hypertrophy. In conclusion, Speckle Tracking can be used to assess left ventricular dysfunction in patients with cardiomyopathies, myocardial infarction, and coronary artery disease. As global longitudinal strain was the most studied, it was found to be a powerful marker of myocardial dysfunction in these heart diseases.

Acknowledgements: The authors would like to express their gratitude to ESS-P.PORTO. This work was partially financially by Base Funding - UIDB/00027/2020 of the Artificial Intelligence and Computer Science Laboratory – LIACC - funded by national funds through the FCT/MCTES (PIDDAC).

References

[1] Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (editors) *Cochrane Handbook for Systematic Reviews of Interventions version 6.3*. Cochrane, 2022. Available:www.training.cochrane.org/handbook

[2] Collier, P., Phelan, D., Klein, A. A Test in Context: Myocardial Strain Measured by Speckle-Tracking Echocardiography. *J Am Coll Cardiol.*, 69(8), 1043-1056, 2017. doi:10.1016/j.jacc.2016.12.012

Application of machine learning techniques for a recommendation system in Pharmacy

Beatriz Torres [a], Alexandra Oliveira [a,b,c], Brígida Mónica Faria [a,b], Sandra Alves [a]
10170145@ess.ipp.pt, aao@ess.ipp.pt, monica.faria@ess.ipp.pt, salves@ess.ipp.pt
[a] *ESS, Polytechnic of Porto (ESS-P.PORTO), Porto, Portugal*
[b] *Artificial Intelligence and Computer Science Laboratory (LIACC), University of Porto, Porto, Portugal*
[c] *Retail Consult, Portugal*

Keywords: community pharmacy, non-prescription products, products clustering, products interactions, recommendation system

Abstract: Community Pharmacy (CP) plays a crucial role in the population, improving patients' quality of life and minimising medication risks. In Portugal, CPs dispense prescription and non-prescription products. Professionals have an added responsibility of advising them and should pay attention to self-medication and possible interactions, especially in polymedicated patients. Although software solutions for CP in Portugal are valuable, they have limited information on these products. It is a general system not aligned with individualized healthcare that requires patient-specific recommendations based on their characteristics and product classification. A product recommendation system that incorporates relevant information about the products can support a more informed recommendation by the professional. This work aims to develop a conceptual pharmaceutical product recommendation framework. Identify relevant groups of products (over-the-counter medication, homoeopathic medication and dermocosmetics) according to their characteristics (Active Substance, Pharmaceutical Form, Indication, Age, Adverse Effects by Categories, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding and Pharmacotherapeutic Group) by applying machine learning techniques to public databases of non-prescription products and comparing the techniques used. It aims to define and evaluate a distance function capable of creating groups of clinically relevant products for pharmaceutical counselling. Communicate the results adjusted to healthcare professionals. For this purpose, hierarchical and non-hierarchical clustering techniques were applied and evaluated. As a result, there is a database of non-prescription products with scientific information and groups of similar products formed. K-means was the most effective clustering approach forming pharmacologically homogeneous groups based mainly on safe use during pregnancy and breastfeeding and pharmacotherapeutic group. Clustering techniques are used to reduce data sparsity, and increase the generation of recommendation speed, and accuracy. While this method provided more homogeneous groups, further refinement is still necessary to obtain more valuable groups in CPs during counselling.

Acknowledgements: We would like to express our gratitude to ESS-P.Porto and Retail Consult for their support and assistance.

Estimação robusta de modelos com dados em painel

Anabela Rocha [a,b], M. Cristina Miranda [a,b]
anabela.rocha@ua.pt, cristina.miranda@ua.pt

[a] *ISCA, University of Aveiro*
[b] *CIDMA, University of Aveiro*

Keywords: dados em painel, estimação robusta, simulação

Abstract: Modelação de dados em painel é um problema que se coloca de forma transversal em diversas áreas, nomeadamente em demografia, economia, finanças, biologia, climatologia e ambiente. Em estudos empíricos, a estimação dos parâmetros destes modelos é usualmente feita com base em métodos de estimação clássicos. A presença de outliers em dados de painel é uma situação frequente em conjuntos de dados reais em diferentes áreas e pode afetar drasticamente as estimativas obtidas por métodos de estimação clássicos. O objetivo deste trabalho é desenvolver técnicas de estimação para modelos com dados em painel que sejam robustas, isto é, que originem melhores estimativas do que os estimadores clássicos, quando os dados não verificam os pressupostos necessários para a validação das propriedades dos estimadores clássicos. As propriedades de robustez dos procedimentos propostos são investigadas por meio de estudos de simulação, sob diferentes cenários. Para ilustrar o desempenho dos métodos robustos propostos, é ainda apresentado um exemplo, baseado num conjunto de dados económicos em painel reais, conhecido da literatura.

Acknowledgements: Este trabalho foi apoiado pelo Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA) através da Fundação para a Ciência e Tecnologia (FCT), com os projetos UIDB/04106/2020 e UIDP/04106/2020.

References

[1] Baltagi, B. H. *Econometric Analysis of Panel Data*. John Wiley, New York, 2001.

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

[3] Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data 2e*. MIT Press, 2010.

Revisitando métodos de reamostragem na estimação de parâmetros de acontecimentos raros

Dora Prata Gomes [a], M. Manuela Neves [b]

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

Keywords: estatística de valores extremos, métodos de reamostragem, parâmetros de acontecimentos raros

Abstract: O principal objectivo da *Estatística de Valores Extremos* é a estimação da probabilidade de ocorrência de acontecimentos geralmente para além do intervalo de dados disponíveis. Na análise de valores extremos existem alguns parâmetros de particular interesse, entre os quais o *índice de valores extremos*, ξ, relacionado com o peso da cauda da distribuição. Este parâmetro é a base para a estimação de todos os outros parâmetros de acontecimentos extremos, excepto para o *índice extremal*, θ. Este último parâmetro tem interesse real para amostras dependentes (situação comum na prática) e pode ser definido grosseiramente como o recíproco da duração esperada de valores acima de um nível elevado. A maioria dos estimadores semi-paramétricos destes parâmetros apresenta o mesmo tipo de comportamento: boas propriedades assintóticas, mas uma elevada variância para valores pequenos de k, o número de estatísticas de ordem superior usadas no cálculo das estimativas, um viés alto para valores grandes de k. Há portanto necessidade real de escolha adequada de k. Após uma breve introdução de alguns estimadores dos parâmetros mencionados e suas propriedades assintóticas, propomos o uso dos métodos de reamostragem *bootstrap* e *jackknife* e um algoritmo para a escolha de k e a estimação adaptativa de ξ e θ. Foi realizado um estudo de simulação e são apresentadas aplicações a dados reais.

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 e UIDP/00297/2020 (Centro de Matemática e Aplicações) e do projeto UIDB/00006/2020 (CEAUL).

References

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

Comparative analysis of the marginalized LASSO estimator in multiple linear models: A comprehensive evaluation

Mina Norouzirad [a], Filipe J. Marques [a,b]
m.norouzirad@fct.unl.pt, fjm@fct.unl.pt

[a] *Center for Mathematics and Applications (NOVA Math), NOVA SST, Caparica, Portugal*
[b] *Department of Mathematics, NOVA SST, Caparica, Portugal*

Keywords: linear regression models, Marginalized LASSO, prediction accuracy, ROC curve, variable selection

Abstract: Variable selection is a crucial task in multiple linear models, where the commonly used LASSO estimator lacks a closed form solution. This paper focuses on the marginalized LASSO estimator, a novel approach based on marginal theory. The paper conducts a comprehensive comparative analysis of the LASSO and Marginalized LASSO estimators, emphasizing their performance in prediction accuracy using the mean squared error (MSE) metric and their effectiveness in variable selection using evaluation criteria such as ROC curves. Practical insights are provided through the application of both estimators to a real dataset, highlighting their performance and discussing their advantages and disadvantages. By emphasizing the Marginalized LASSO estimator's features, this paper offers valuable insights into its potential as an alternative approach for variable selection in multiple linear models.

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. http://www.jstor.org/stable/2346178

[2] Saleh, A. K. Md. E., Arashi, M., Kibria, B. M. G. *Theory of ridge regression estimation with applications*, Wiley, 2019.

[3] Saleh, A. K. Md. E., Arashi, M., Saleh, R. A., Norouzirad, M. *Rank-based methods for shrinkage and selection: With application to machine learning*, Wiley, 2022.

Avaliação de ferramentas de programação em Modelos Lineares Generalizados: estudo de simulação

Ana Marinho [a], Susana Faria [a,b], Rita Sousa [c]
anamarinho2000@gmail.com, sfaria@math.uminho.pt, rcsousa@bportugal.pt

[a] *Departamento de Matemática, Universidade do Minho*
[b] *Centro de Matemática, Universidade do Minho*
[c] *Banco de Portugal; Centro de Matemática e Aplicações - FCT/UNL*

Keywords: estudo de simulação, modelos lineares generalizados

Abstract: O modelo normal linear é utilizado na descrição de fenómenos aleatórios e pressupõe a normalidade dos erros e variância constante. Quando o fenómeno sob estudo não apresenta uma resposta que permita garantir estes pressupostos, opta-se por transformar as variáveis de forma a se aproximarem da normalidade. Contudo, nem todos as situações permitem alcançar os pressupostos pretendidos e, por isso, os modelos não lineares ou não normais são uma alternativa nessas situações. Assim, os Modelos Lineares Generalizados (MLG), definidos por Nelder e Wedderburn em 1972, vieram unificar esses modelos [1].
Com o intuito de avaliar o desempenho dos MLG, programaram-se estes modelos com diferentes linguagens de programação, em R [2], Stata [3] e Phyton [4], de forma a avaliar e comparar a o desempenho de diferentes ferramentas. Para cada uma das linguagens realizaram-se estudos de simulação de forma a perceber, dentro das funções existentes, qual seria a mais adequada no uso destes modelos. Para tal, foram analisados diferentes critérios como o desempenho computacional, através do número de iterações e do tempo necessário para a estimação do modelo, o erro quadrático médio e o viés do coeficiente associado a cada variável explicativa. Estudou-se também a capacidade de previsão que estes modelos apresentam para diferentes dimensões da população e da amostra. Estes resultados foram analisados e aplicados a uma base de microdados reais fornecida pelo Laboratório de Investigação e Microdados do Banco de Portugal (BPLIM).

References

[1] Nelder, J. A., Wedderburn, R. W. M. Generalized linear models *Journal of the Royal Statistical Society*, Series A, 370-384, 1972. doi:https://doi.org/10.2307/2344614

[2] Dunn, P. K., Smyth, G. K. *Generalized linear models with examples in R*, Springer, New York, 2017. doi:https://doi.org/10.1007/978-1-4419-0118-7

[3] Hardin, J. W., Hilbe, J. M. *Generalized linear models and extensions*. Stata Press, 3rd ed., 2012.

[4] McKinne, W. *Python for Data Analysis*, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2013.

Regressing a static response on longitudinal predictors: the case-study of childhood obesity

Mafalda Oliveira [a,b,c], Susana Santos [b,c], Rita Gaio [a,d]
up201805422@fc.up.pt, susana.m.santos@ispup.up.pt, argaio@fc.up.pt

[a] *Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Porto, Portugal*
[b] *EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal*
[c] *Laboratório para a Investigação Integrativa e Translacional em Saúde Populacional (ITR), Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal*
[d] *Centro de Matemática da Universidade do Porto, Porto, Portugal*

Keywords: elastic net, longitudinal predictors, mixture of regressions, obesity

Abstract: Childhood obesity is a complex public health issue that requires early identification of at-risk individuals. This study aims to build a predictive model for obesity at age 13 through static and longitudinal predictors measured at 4,7, and 10 years old. Due to the cross-sectional nature of the response variable, traditional longitudinal models are not applicable. Therefore, we applied two different approaches to provide a solution to the problem. First, we fitted four separate logistic regression models, one for each time point, to analyze the associations between exposures at that time point and the outcome. Secondly, we applied a dynamic model considering all exposures in a penalized regression method with elastic net (ENET). The models' performances were evaluated by AUC (Area Under the ROC Curve), Specificity, Sensitivity, and Prediction Error. Additionally, we fit a finite mixture of regressions model and evaluated the results using the mean squared error, mean absolute error, and mean absolute percentage error. The analysis was conducted on R using data from cohort *Generation XXI*, comprising 4246 observations and 55 variables. The four static models' results suggest that including the most recent predictors improves the performance measures. The dynamic model with ENET regression presented satisfactory performance results, although slightly lower than those achieved by the 10 years-old model. The mixture of regressions outperformed the previous models, with a mean absolute percentage error of only 5%.

Acknowledgements: G21 was funded by Programa Operacional de Saúde – Saúde XXI, Quadro Comunitário de Apoio III and Administração Regional de Saúde Norte (Regional Department of Ministry of Health). It has support from the Portuguese Foundation for Science and Technology and from Calouste Gulbenkian. This study was supported by the European Union Horizon 2020 Research and Innovation Programme under Grant Agreement 824989 (EUCAN-Connect) and 874583 (ATHLETE). Rita Gaio was partially supported by CMUP, which is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the project with reference UIDB/00144/2020.

Dimensionality reduction in survival models based on gene expression data: an application to brain cancer

João Brandão [a], Marta B. Lopes [b,c], Eunice Carrasquinha [a,d]
fc53139@alunos.ciencias.ulisboa.pt, marta.lopes@fct.unl.pt,
eitrigueirao@ciencias.ulisboa.pt

[a] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa*
[b] *NOVA Math, NOVA School of Science and Technology, Portugal*
[c] *UNIDEMI, NOVA School of Science and Technology, Portugal*
[d] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: gene expression, high-dimensional data, network-based regularization, regularized optimization, survival analysis

Abstract: Glioblastoma is the most malignant type of brain cancer, with great heterogeneities in prognosis, clinicopathological features, immune landscapes, and immunotherapeutic responses. As molecular sciences evolve and our understanding of disease mechanisms increases, the possibility of developing personalized medicine approaches becomes increasingly feasible. The identification of long or short-term survivors and associated gene expression markers open promising avenues for the development of novel personalized therapies.

The analysis of gene expression data, however, is particularly challenging due to the high dimensionality of the data. The main objective of this study consists in the application of dimensionality reduction techniques on survival data, paired with gene expression data from tumors of glioblastoma patients, with the ultimate goal of obtaining a list of possible outlier observations, whose survival time is much greater or smaller than expected. The study of these outliers might give us information about important genes in the survival time of these patients.

Elastic Net and network-based regularization techniques (Hub Cox and Orphan Cox) were the methods used for dimensionality reduction. The martingale residuals were then obtained for the different models by the feature reduction methods and the Rank Product Test [1] was used to create a consensus between the models. This method allowed us to identify the observations whose residuals were systematically high amongst the different models and, therefore, identify consensual outliers between different models.

Acknowledgements: Funded by the Portuguese Foundation for Science & Technology with references PTDC/CCIBIO/4180/2020, UIDB/00006/2020 (CEAUL), UIDB/00297/2020 and UIDP/00297/2020 (NOVA MATH), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and CEECINST/00042/2021.

References

[1] Carrasquinha, E., Veríssimo, A., Lopes, M.B., Vinga, S. Identification of influential observations in high-dimensional cancer survival data through the rank product test *BioData Mining*, 2018. doi:10.1186/s13040-018-0162-z

O *test negative design* na avaliação da efetividade de vacinas

André Martins [a], João Paulo Martins [b,c], Marlene Santos [a,c]
andre20.martins@gmail.com, jom@ess.pp.pt, mes@ess.ipp.pt

[a] *Centro de Investigação em Saúde e Ambiente, Instituto Politécnico do Porto, Rua Dr. António Bernardino de Almeida, 4200 -072 Porto, Portugal*
[b] *Escola Superior de Saúde, Instituto Politécnico do Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal*
[c] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: efetividade, ensaios clínicos aleatorizados, gripe, test negative design, vacina da gripe

Abstract: Os estudos observacionais são uma ferramenta muito útil e, cada vez mais utilizada para a avaliação da efetividade de vacinas (e outros medicamentos). O tipo de estudo *test negative design* (TND) é uma derivação do estudo caso controlo e a forma mais comum de avaliar a efetividade da vacina da gripe sazonal. Pessoas com sintomas mais ou menos severos acorrem ao hospital ou outro local de prestação de cuidados de saúde e é-lhes realizado um teste para confirmar a presença do vírus [1]. Indivíduos com resultado positivo serão os casos, os negativos os controlos. Para avaliar a qualidade deste tipo de estudo na avaliação da efetividade da vacina da gripe sazonal foi desenvolvida uma revisão sistemática e meta-análise que compara os resultados obtidos através desse tipo de estudo com os estudos de referência, os ensaios clínicos aleatorizados (RCT). As estimativas de efetividade encontradas nos estudos de referência não são significativamente mais elevadas do que aquelas que se verificam nos estudos TND quando as estirpes de vírus incluídas na vacina e as mais prevalentes em circulação são concordantes (p-value=0.27, I^2=17.4%). Neste trabalho, conclui-se que os TND são uma alternativa fiável, mais rápida e económica para avaliar a efetividade da vacina da gripe sazonal.

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] Zhang, L., Wei, M., Jin, P., Li, J., Zhu, F. An evaluation of a test-negative design for EV-71 vaccine from a randomized controlled trial. *Hum Vaccin Immunother*, 17, 2101–2106, 2021. doi:10.1080/21645515.2020.1859900

Statistical analysis to identify protein adducts by mass spectrometry: a tool for biomarker investigation

Filipa Costa [a], Conceição Amado [a], Alexandra M. M. Antunes [b], Judit Morello [c]
filipabcosta@tecnico.ulisboa.pt,
conceicao.amado@tecnico.ulisboa.pt

[a] *CEMAT and IST-ULisboa*
[b] *CQE-Institute of Molecular Sciences and IST-ULisboa*
[c] *iNova4Health and NOVA Medical School-UNL*

Keywords: acrylamide, adductomics, classical PCA, mass spectrometry, robust PCA

Abstract: Exposure to reactive chemical agents can lead to the formation of protein covalent adducts, which have significant implications for human health. Identifying these adducts offers opportunities to understand disease mechanisms, develop biomarkers for diagnosis/prognosis, and accurately assess chemical toxicant exposure. Mass spectrometry (MS)-based methodologies are ideal for analyzing protein covalent modifications, providing accurate, sensitive, and unbiased identification with quantitative information. However, detecting low-abundance adducts remains challenging despite technological advancements.
Rodents histones exposed to the chemical carcinogen acrylamide were analyzed. To identify the adducts, a statistical analysis is performed after LC-MS data preprocessing with MZmine. In the pre-processing step, the selection of potential adducts is done based on m/z increments corresponding to Acrylamide and glycidamide incorporation. Then a statistical analysis is applied to identify adducts differently present in exposed and non-exposed cells. Results were compared to standard proteomics methodology, aiming to detect previously unnoticed adducts.
In addition to improving our understanding of protein epigenetic changes, this approach also identifies adducts from both endogenous and exogenous chemical exposure. This strategy holds promise for identifying elusive adducts and expanding our knowledge of the molecular events underlying diseases.

Acknowledgements: This work is partially funded by National Funds through the FCT - Fundação para a Ciência e Tecnologia, under the scope of the projects 2022.06292.PTDC, UIDB/04621/2020, UIDP/04621/2020 of CEMAT, UIBD/00100/2020 and UIDP/00100/2020 of CQE, LA/P/0056/2020 of IMS.

Métodos de Machine Learning para previsão de dados longitudinais

Elsa Soares [a], Inês Sousa [a]
elpombo98@hotmail.com, isousa@math.uminho.pt

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

Keywords: análise de dados longitudinais, dados de alta dimensão, métodos de machine learning, métodos estatísticos, previsões dinâmicas

Abstract: Dados longitudinais são originados quando diferentes indivíduos são medidos repetidamente ao longo do tempo para alguma variável resposta de interesse. Geralmente, os modelos longitudinais descrevem o processo subjacente aos dados observados, permitindo diferentes fontes de variabilidade nos dados. Após a análise longitudinal dos dados, é de interesse fazer previsões para trajetórias individuais futuras da variável resposta com base no modelo utilizado [1]. A qualquer momento, será interessante fazer previsões sobre o resultado longitudinal, considerando toda a história passada, e isso é chamado de previsões dinâmicas. Este tipo de ferramenta de previsão é útil em contextos onde é necessário ter mecanismos de tomada de decisão em tempo real e baseado em dados sempre atualizados.
Já está bem estabelecido na literatura o uso de métodos estatísticos clássicos, como métodos bayesianos ou métodos de inferência de verossimilhança [2]. Neste trabalho, pretende-se desenvolver investigação sobre métodos de *machine learning* para a análise de dados longitudinais, em particular desenvolver métodos estatísticos para previsões individuais. Também pretendemos comparar métodos tradicionais de verossimilhança e métodos de *machine learning* para previsões de dados longitudinais. A hipótese é que um melhor desempenho será alcançado com métodos de *machine learning* à medida que a dimensão dos dados aumenta [3] [1].
Este trabalho pretende apresentar uma revisão bibliográfica das bases teóricas de métodos de *machine learning* para previsões individuais no contexto de modelos longitudinais e apresentar uma aplicação a um conjunto de dados reais implementando previsões dinâmicas.

References

[1] Lim, D.K.E., Boyd, J.H., Thomas, E., Chakera, A., Tippaya, S., Irish, A., Manuel, J., et al. Prediction models used in the progression of chronic kidney disease: A scoping review. *PLOS ONE*, 17(7), 2022. doi:10.1371/journal.pone.0271619

[2] Szczesniak, R.D., Su, W., Brokamp, C., Pestian, J.P., Seid, M., Diggle, P.J., et al. Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression. *Statistics in Medicine*, 39(6), 740-756, 2020. doi:10.1002/sim.8443

[3] Chen, S., Grant, E., Wu, T.T., Bowman, F.D.B. Statistical Learning Methods for Longitudinal High-dimensional Data. *Review Computation Statistics*, 6(1), 10-18, 2014. doi:10.1002/wics.1282

[4] Lin, J., Luo, S. Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine*, 41(15), 2894-2907, 2022. doi:10.1002/sim.9392

The trace ratio method for robust multigroup classification

M. Rosário Oliveira[a], Giulia Ferrandi[b], Igor Kravchenko[a],
Michiel E. Hochstenbach[b]
rosario.oliveira@tecnico.ulisboa.pt, g.ferrandi@tue.nl,
igor.kravchenko@tecnico.ulisboa.pt, M.E.Hochstenbach@tue.nl

[a] *CEMAT and Department of Mathematics, Instituto Superior Técnico, University of Lisbon, Portugal*
[b] *Department of Mathematics and Computer Science, TU Eindhoven, The Netherlands*

Keywords: Fisher's discriminant analysis, linear dimensionality reduction, minimum covariance determinant, multigroup classification, trace ratio method

Abstract: Classification is an important statistical task where an observation is assigned to one of the non-overlapping known groups, based on the statistical properties of the data characterizing these groups. Recently, trace ratio (TR) optimization has gained in popularity due to its computational efficiency, as well as the occasionally better classification results. Like Fisher's discriminant analysis (FDA), TR uses linear dimensionality reduction strategies for the multigroup classification problem. However, a statistical understanding is still incomplete.
In this work, we propose a robust TR method, obtained by exploiting Mininum Covariance Determinant estimates family of the within and between covariance matrices. The method can deal with high-dimensional data since it uses regularized MCD estimates.
We compare TR and FDA on synthetic and real datasets. Synthetic scenarios consider cases where one method performs better than the others. In this case, FDA and TR are used as classifiers and are compared with two different criteria. While the first one is based on the performance of the associated classification rules, the second criterion is related to the proximity between the true solution of one method and the estimated one. Real datasets have been chosen from the UCI and KEEL platforms, and they illustrate the performance of FDA and TR as dimensionality reduction methods, used before the construction of a classifier. In many of these datasets, FDA is as good as or better than TR. Moreover, robust TR shows clear improvements compared to classical TR in several datasets.

Acknowledgements: This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 812912. It has also received support from Fundação para a Ciência e Tecnologia, Portugal, through the project UIDB/04621/2020.

References

[1] Ferrandi, G., Kravchenko, I.V., Hochstenbach, M.E., Oliveira, M.R. On the trace ratio method and Fisher's discriminant analysis for robust multigroup classification. *arXiv*, 2211.08120, 2022. https://doi.org/10.48550/arXiv.2211.08120

Estimation of AUC in logistic regression with missing data: removing the bias

Susana Rafaela Guimarães Martins [a,c], Jacobo de Uña-Alvarez [b,c], Maria del Carmen Iglesias-Perez [b,c]
srgm@estg.ipvc.pt, jacobo@uvigo.es, mcigles@uvigo.es

[a] *Escola Superior de Desporto e Lazer, Instituto Politecnico de Viana do Castelo*
[b] *Department of Statistics and OR, Universidade de Vigo*
[c] *CINBIO, Universidade de Vigo*

Keywords: missing data, optimistic AUC, prediction, ROC curve

Abstract:
Logistic regression is a well-known approach to predict a binary outcome given covariates. To evaluate the predictive capacity of a regression model, the Area Under the Curve (AUC) is often used.
In this work we investigate the issue of estimating the AUC in the presence of missing data, both in the variable of interest and in the covariables. The covariates may be continuous or discrete. For the construction of the predictive models, different missing data methodologies were applied: Complete Case Analysis, Inverse probability Weight and Multiple Imputation, and the apparent AUC was estimated for each one of them. With a simulation study we evaluate de performance of the several estimators for the AUC; in particular, the Monte Carlo bias and the mean squared error of the estimators are obtained. The Bias is defined as the difference between AUC apparent and AUC out-of-sample, where AUC out-of-sample approximates the true AUC, that is, the AUC of population.
Traditionally, the apparent AUC overestimates the true AUC. In this work we consider several approaches to correct for this overestimation: Split-Sample, K-fold and Leave-one-out, adapted to missing data. We carried out a simulation study to evaluate the performance of the correction methods in the presence of missing data. Furthermore, we apply these methods to a set of real data.

Acknowledgements: Work supported by the Grant PID2020-118101GB-I00, Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033).

References

[1] Iparraguirre, A., Irantzu, B., Rodríguez-Álvarez, M.X. On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. *SORT-Statistics and Operations Research Transactions*, 1, 145-162, 2019. doi:https://raco.cat/index.php/SORT/article/view/356185

[2] Li, P., Taylor, J.M.G., Spratt, D.E., Karnes, R.J., Schipper, M.J. Evaluation of predictive model performance of an existing model in the presence of missing data *Statistics in Medicine*, 40, 3477–3498, 2021.

On a Parzen-Rosenblatt type density estimator for circular data

Carlos Tenreiro [a]
tenreiro@mat.uc.pt

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

Keywords: bandwidth selection, circular data, kernel density estimation

Abstract: Given an independent and identically distributed sample of angles $X_1, \ldots, X_n \in [0, 2\pi[$ from some absolutely continuous random variable X with unknown probability density function f, the standard kernel density estimator is defined, for $\theta \in [0, 2\pi[$, by

$$\tilde{f}_n(\theta; g) = \frac{c_g(L)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \cos(\theta - X_i)}{g^2}\right),$$

where $L : [0, \infty[\to \mathbb{R}$ is a bounded function satisfying some additional conditions, $g = g_n$ is a sequence of positive numbers such that $g_n \to 0$, as $n \to \infty$, and $c_g(L)$, depending on the kernel L and the bandwidth g, is chosen so that $\tilde{f}_n(\cdot; g)$ integrates to unity. In this work we consider an alternative estimator of f which is closer in spirit to the Parzen-Rosenblatt estimator for linear data. For $\theta \in [0, 2\pi[$, it is defined by

$$\hat{f}_n(\theta; h) = \frac{d_h(K)}{n} \sum_{i=1}^{n} K_h(\theta - X_i),$$

where $h = h_n > 0$ is a sequence of strictly positive real numbers converging to zero as n tends to infinity, K_h is a real-valued periodic function on $\mathbb{R}$, with period 2π, such that $K_h(\theta) = K(\theta/h)/h$, for $\theta \in [-\pi, \pi[$, with K a kernel on $\mathbb{R}$, that is, an integrable real-valued function on $\mathbb{R}$ with $\int_{\mathbb{R}} K(u)du > 0$, and $d_h(K)$ is a normalizing constant depending on the kernel K and the bandwidth h which is chosen so that $\hat{f}_n(\cdot; h)$ integrates to unity. The connection between $\tilde{f}_n(\cdot; g)$ and $\hat{f}_n(\cdot; h)$, the consistency of the Parzen-Rosenblatt type estimator $\hat{f}_n(\cdot; h)$, and the automatic selection of the bandwidth for $\hat{f}_n(\cdot; h)$ are, among others, issues that will be addressed in this presentation.

Acknowledgements: Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324/2019, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

[1] 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. doi:10.1080/10485252.2022.2057974

[2] Tenreiro, C. A Parzen-Rosenblatt type density estimator for circular data: exact and asymptotic optimal bandwidths. *DMUC Preprint*, 23-10, 2023. https://www.mat.uc.pt/preprints/ps/p2310.pdf

Reduced bias estimation of the residual tail dependence index: Pareto meets Fréchet

Cláudia Neves [a]
claudia.neves@kcl.ac.uk

[a] *King's College London*

Keywords: asymptotic independence, extreme value theory, regular variation, semi-parametric inference, tail dependence

Abstract: Unlike univariate extremes, multivariate extreme value distributions cannot be specified through a finite-dimensional parameter family of distributions. Instead, the many facets of multivariate extremes are mirrored in the inherent dependence structure of component-wise maxima which must be dissociated from their marginal distributions. Statistical procedures for eliciting extremal dependence typically rely on standardisation of the unknown marginal distributions, through which process pseudo-observations of either Pareto or Fréchet distribution are often considered. The relative merits of either of these choices for transformation of marginals have been discussed in the literature, particularly in the context of domains of attraction of a multivariate extreme value distribution. This talk is set within this context as we will introduce a class of reduced-bias estimators for the residual dependence index that eschews consideration of this choice of marginals. The proposed unifying class of reduced bias estimators includes but is not limited to variants of the Hill estimator. Adapted conditions of regular variation lay the groundwork for obtaining their asymptotic properties, whose effectiveness is borne by a simulation study. The leading application is aimed at discerning asymptotic independence from monsoon-related rainfall occurrences at several locations in Ghana. This is joint work with Emily Black, Jennifer Israelsson and David Walshaw.

Innovation and product positioning

Diogo Pereira [a], Anne Balter [b], Cláudia Nunes [a], Peter Kort [b]
dia.pereira@campus.fct.unl.pt, A.G.Balter@tilburguniversity.edu,
cnunes@math.tecnico.ulisboa.pt, kort@uvt.nl

[a] *CEMAT and IST*
[b] *Tilburg University*

Keywords: Geometric Brownian Motion, optimal stopping problem, vertical and horizontal differentiation

Abstract: In today's economy, firms need to be innovative to survive. To improve their strategic position, a firm can work along two dimensions. First, a firm can make its product such that it differs from the products of its competitors. In the literature, this is called horizontal differentiation. A way to model this is the linear city or Hotelling line. Second, a firm could try to improve its product in such a way that the resulting product quality exceeds that of products of other firms. This is called vertical differentiation. This entails costs: the firm needs to invest in R&D costs, and needs to pay to introduce this new product.
In this paper, we study the impact of the introduction of a new product when an existing one is already in the market. The firm is currently producing the existing product at a certain location (in the Hotelling line) and needs to decide how much should invest in R&D in order to have a new product, when and where to invest in the new one. For both products the demand follows a Geometric Brownian motion and the customers are distributed uniformly in the Hotelling line. The horizontal differentiation dimension corresponds to the location of the new product on the line. The problem can be stated as a maximization/optimal stopping time. Using the Hamilton-Jacobi-Bellman (HJB) equations, we can characterize the optimal strategy in terms of the level of demand that triggers the investment decision in the new product, how much the firm should invest in R&D, where to locate the new product and if the firm keeps producing both products or if (and when) to replace the existing product by the new one.
The analytical results are then illustrated with some examples, where one specifies the parameters of the demand (notably the drift and the volatility), as well as the costs associated with the production of the existing and the new product.

Modelação estatística do custo em contratos de assistência automóvel

Bárbara Botelho [a], Sandra Ramos [a,b]
1180676@isep.ipp.pt, sfr@isep.ipp.pt

[a] *Instituto Superior de Engenharia do Instituto Politécnico do Porto*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa*

Keywords: contratos de assistência automóvel, modelação de custos, processos pontuais não homogéneos, simulação estocástica

Abstract: Os contratos de assistência têm vindo a ganhar força no mercado automóvel, já que contribuem para a manutenção de um relacionamento duradouro e satisfatório com o cliente, oferecendo um atendimento de alta qualidade com foco na lealdade e fidelização. Geralmente, estes contratos cobrem todos os custos de manutenção e grande parte dos custos de reparação, durante um horizonte predeterminado de tempo ou de quilómetros, em troca de um prémio mensal fixo e isentam os clientes de custos incertos. Para garantir a rentabilidade, os prémios devem cobrir, no mínimo, os custos esperados para o horizonte do contrato.
Apesar do custo do contrato incluir uma fração de quatificação simples (associada a manutenções recomendadas pelos fabricantes), existe uma parte, associada com eventos não previstos, de mais difícil estimação e à qual estão, geralmente, associados custos elevados. Um problema de interesse prático é a modelação destes custos imprevistos.
Neste trabalho, apresentam-se resultados da modelação do custo total de contratos de assistência, em função do tempo/quilometragem do contrado, através da aplicação de processos estocásticos pontuais não homogéneos para modelar grupos de eventos aleatórios associados com intervenções imprevistas. Para além da produção de estimativas pontuais e intervalares para o custo total, o modelo desenvolvido permite simular tempos de ocorrências de necessidades de intervenção.

The importance of experimental design principles in agricultural field trials

Elsa Gonçalves [a]
elsagoncalves@isa.ulisboa.pt

[a] *LEAF—Linking Landscape, Environment, Agriculture and Food Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal*

Keywords: blocking, grapevine field trials, linear mixed models, randomisation, repetitions

Abstract: Experiments begin long before data analysis, with the previous design of the research strategy. Agriculture has a long tradition in the development of experimental designs to establish rigorous field trials, mainly in plant breeding research. The overall planning of the field trial is a key point to guarantee the success of the experimental process. It is well-known that many of the important principles of experimental design were developed in the 1920s and 1930s, in particular by R.A. Fisher: randomisation and replication of treatments and as well as blocking to control extraneous variation. While the importance of these principles is continuously recognized in many texts about experimental designs in agriculture, forests, and related sciences, they are not well understood and they are not well implemented in practice by some researchers. For example, randomization ensures that all experimental units are equally likely to receive any treatment, minimizing systematic errors or bias induced by the experimenter, therefore, it is essential to provide a valid estimation of the error variance. However, notwithstanding randomization is a basic principle in research, its importance is not always fully respected. A recurrent problem in many experiments is distinguishing between true repetitions (independent repetitions) and pseudo-repetitions. The objective of this work is to reinforce the importance of these principles (replication, randomizations, and blocking) when conducting agricultural experiments. The importance of respecting these principles is highlighted and demonstrated using real data obtained in grapevine field trials and fitting linear mixed models.

Acknowledgements: This research was funded by the projects "GrapeVision" (PTDC/BIA-FBT/2389/2020) and "Save the intra-varietal diversity of autochthonous grapevine varieties" (PRR-C005-i03–000016).

Spatio-temporal modelling of commercial fish species distribution

Daniela Silva[a], Raquel Menezes[b], Susana Garrido[c]
danyelasylva2@gmail.com, rmenezes@math.uminho.pt,
susana.garrido@ipma.pt

[a] *Centre of Mathematics (CMAT), Minho University, Braga*
[b] *Centre of Mathematics (CMAT), Minho University, Guimarães*
[c] *Division of Modelling and Management of Fishery Resources, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisboa*

Keywords: geostatistics, joint modelling, preferential sampling, *sardina pilchardus*, spatio-temporal species distribution model

Abstract: Scientific tools capable of identifying the species distribution patterns are important as they contribute to improve knowledge of causes of species fluctuations. Species distribution data often implies residual spatial autocorrelation and temporal variability, so both components are important to study the evolution of species distribution from an ecological point of view. Fishery data can arise from two main sources, fishery-independent data that are often derived from commercial fleets and fishery-independent data that typically rely on research surveys. Research surveys occur once or twice a year over a larger spatial region and smaller number of spatial locations are often sampled in a standardized sampling design. Data from commercial fleets often benefit from a higher occurrence where more locations are sampled in a smaller region since there is a preferential selection of these locations. The two data sources may provide different but important information, can been used as complementary. However, joint modelling the two sources will require an approach capable of dealing with the different sampling designs, as classical tools are able to deal with standardized sampling designs but not with the preferential nature from commercial data. In this context, we present a hierarchical spatio-temporal model for sardine (*Sardina pilchardus*) absense/presence, which integrates both data sources while considering the preferential sampling from fishery-dependent data.

Acknowledgements: The authors acknowledge the FCT Foundation for funding their research through grant PD/BD/150535/2019, as well as projects PTDC/MAT-STA/28243/2017, UIDB/00013/2020 and UIDP/00013/2020. They also express their appreciation to MAR2020 for funding the SARDINHA2020 project (MAR-01.04.02-FEAMP-0009) and to all colleagues involved in this study.

References

[1] Alglave, B., Rivot, E., Etienne, M., Woillez, M., Thorson, J. T., Vermard, Y. Combining scientific survey and commercial catch data to map fish distribution, *ICES Journal of Marine Science*, 79 (4), 1133–1149, 2022. https://doi.org/10.1093/icesjms/fsac032

Multivariate random fields and systems of stochastic partial differential equations

Sílvia Guerra [a], Fernanda Cipriano [b], Isabel Natário [b]
silvia.guerra@novasbe.pt, mfsm@fct.unl.pt, icn@fct.unl.pt

[a] *NOVA School of Business and Economics, Carcavelos, Portugal*
[b] *NOVA MATH & NOVA School of Science and Technology, Caparica, Portugal*

Keywords: bayesian inference, Gaussian Markov random fields, Gaussian random fields, stochastic partial differential equations

Abstract: In Spatial Statistics, Gaussian fields (GF) have been used to represent and model spatial variability, as it has been noted that the solutions of certain stochastic partial differential equations (SPDE), being GF, have convenient covariance function for modelling spatial phenomena. To overcome the computational difficulties of modelling space dependence, Gaussian Markov random fields (GMRF) are considered, corresponding to the numerical approximations of the solutions of the SPDE, obtained through the finite element method and allowing estimation under a Bayesian statistical framework.
Some problems call for a multivariate spatial model, if one is interesting in modelling two or more quantities that depicts spatial dependency. That can be addressed through systems of SPDE, which is considered in this work.
The main goal of this study is to estimate the wind intensity and velocity, considering a system of SPDE, and applying Bayesian inference, through the INLA methodology. We present theoretical results and calculations, not explicitly presented in the literature, that supports INLA methodology, central for doing approximate Bayesian inference.
The results are encouraging, and open new lines of investigation, such as applying statistical methods to study the solution of stochastic partial differential equations.

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] Hu, X., Steinsland, I. Spatial modeling with system of stochastic partial differential equations. *WIREs Comput Stat*, 8:112–125, 2016. doi:10.1002/wics.1378

[2] Hu, X. et al. Multivariate Gaussian random fields using systems of stochastic partial differential equations. *In: arXiv preprint arXiv:1307.1379*, (cit. on p. 84), 2013.

[3] Bolin, D., Lindgren, F. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, Vol 5, No. 1, 523–550, 2011. doi:10.1214/10-AOAS383

How to define prior information for generalized maximum entropy estimation?

Jorge Cabral [a,b], Vera Afreixo [a,b], Pedro Macedo [a,b]
jorgecabral@ua.pt, vera@ua.pt, pmacedo@ua.pt

[a] *Department of Mathematics, University of Aveiro, Aveiro, Portugal*
[b] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal*

Keywords: generalized maximum entropy, M estimation, ordinary least squares, ridge, support space

Abstract: The maximum entropy (ME) formalism provides an approach for solving ill-posed problems [1]. The generalized ME (GME) selects the most conservative or noncommittal solution to the linear model and coefficients that are maximally informative [2]. Its implementation begins with the choice of the sets of discrete points (support spaces) based on prior information about the coefficients to be estimated. However, this information is usually unknown. We propose a user-independent approach for choosing the magnitude of the support spaces based on a first GME estimation of the coefficients on standardized data which will define the extremes of a zero centred symmetric support with equally spaced points for each parameter. Coefficients are then repeatedly re-estimated for support spaces with decreasing ranges keeping the ones that produce the lowest k-fold cross-validation root mean square error (CV-RMSE). We test our approach against other estimation methods in a simulation study with 100 replications and 100 observations. A specific vector of parameters, a different number of independent variables drawn from a standard normal distribution, random correlation matrices and errors that follow normal distributions with zero-mean and different standard deviations are used to generate data. The supports are defined with different number of points. The GME estimator with our proposal for the selection of the supports generally returned the lowest median 5-fold CV-RMSE. Generally using supports with 7 points resulted in a neglectable decrease in error at the expense of a significant increase in time of computation.

Acknowledgements: The authors are supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT), project UIDB/04106/2020. Cabral is grateful to the PhD fellowship at CIDMA-DMat-UA, reference UIDP/04106/2020.

References

[1] Jaynes, E.T. Information Theory and Statistical Mechanics. *Physical Review*, 106, 620–630, 1957. doi:10.1103/PhysRev.106.620

[2] Golan, A., Judge, G.G., Miller, D. *Maximum entropy econometrics: robust estimation with limited data*. Wiley, Chichester, New York, 1996.

Screening the discrepancy function of a computer model

Rui Paulo[a], Pierre Barbillon[b], Anabel Forte[c]
rui@iseg.ulisboa.pt, pierre.barbillon@agroparistech.fr, anabel.forte@uv.es

[a] *CEMAPRE/REM and Lisbon School of Economics and Management, Universidade de Lisboa*
[b] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France*
[c] *Universitat de Valencia*

Keywords: Bayesian variable selection, calibration, Gaussian processes, uncertainty quantification

Abstract: Traditionally, screening refers to the problem of detecting active inputs in the computer model. We develop methodology that applies to screening, but the main focus is on detecting active inputs not in the computer model itself but rather on the discrepancy function that is introduced to account for model inadequacy when linking the computer model with field observations. We contend this is an important problem as it informs the modeler which are the inputs that are potentially being mishandled in the model, but also along which directions it may be less recommendable to use the model for prediction. The methodology is Bayesian and is inspired by the continuous spike and slab prior popularized by the literature on Bayesian variable selection. In our approach, and in contrast with previous proposals, a single MCMC sample from the full model allows us to compute the posterior probabilities of all the competing models, resulting in a methodology that is computationally very fast. The approach hinges on the ability to obtain posterior inclusion probabilities of the inputs, which are easy to interpret quantities, as the basis for selecting active inputs. For that reason, we name the methodology PIPS — posterior inclusion probability screening.

Survapp: a Shiny application for survival data analysis

Emanuel V. M. da Silva [a], Luís Meira-Machado [a], Gustavo Soutinho [b]
emanuelvieira1111@gmail.com, lmachado@math.uminho.pt,
gdsoutinho@gmail.com

[a] *Centre of Mathematics, University of Minho*
[b] *EPIUnit - University of Porto*

Keywords: application development, multi-state models, statistical analysis, survival analysis, user-friendly graphical interfaces

Abstract: There is a significant demand for user-friendly graphical interfaces that enable professionals with limited programming knowledge to perform statistical analysis. Although the R software is widely used for statistical analysis, it lacks a sufficiently intuitive graphical interface for individuals without statistical and programming skills. This project aimed to address this gap by developing an application called Survapp that allows users, regardless of their computational knowledge, to conduct survival analysis. The development used R software, RStudio, and the Shiny package to create an interactive web app. Several existing shiny applications, such as MSM.app [1], SmulTCan [2] and MEPHAS [3], were evaluated during the development of Survapp. While Survapp shares some characteristics with these existing applications, it distinguishes itself by employing supervised algorithms. Survapp incorporates diverse methodologies for analyzing survival data, including the Kaplan-Meier, log-rank test, Cox models, parametric accelerated failure time models, decision trees, random forests, and competitive risk analysis (a particular case of multi-state models). Survapp enables users to analyze survival data, offering example databases for various methodologies within the application. However, the primary objective is to allow users to import their own data and conduct their respective analyses. Overall, Survapp proves to be a highly valuable tool for survival data analysis, catering to users' needs and providing a user-friendly interface with a wide range of statistical analysis methods. The Shiny app is available at the Shiny Apps repository: Survapp.

References

[1] Soutinho, G., Meira-Machado, L. F. Analysis of Complex Survival Data: a tutorial using the Shiny MSM.app application. *arXiv:2202.09160*, 2022. doi:http://dx.doi.org/10.2139/ssrn.3996850

[2] Ozhan, A., Tombaz, M., Konu, O. SmulTCan: A Shiny application for multivariable survival analysis of TCGA data with gene sets. *Computers in Biology and Medicine*, 137, 104793, 2021. doi:https://doi.org/10.1016/j.compbiomed.2021.104793

[3] Zhou, Y., Leung, Sw, Mizutani, S. et al. MEPHAS: an interactive graphical user interface for medical and pharmaceutical statistical analysis with R and Shiny. *BMC bioinformatics*, 21, 183, 2020. doi:https://doi.org/10.1186/s12859-020-3494-x

Bayesian prediction of football outcomes – Application to the Portuguese 1st League

Rui Martins [a,b], Daniel Andrade [a]
rmmartins@fc.ul.pt, daniel.andrade14@gmail.com

[a] *Faculdade de Ciências da Universidade de Lisboa (FCUL)*
[b] *Centro de Estatística e aplicações da Universidade de Lisboa (CEAUL)*

Keywords: count data, de finetti measure, zero-modified Poisson distribution

Abstract: Statistical modeling of sports results has become trendy. Different types of models have been proposed to model these data depending on the objectives: from predicting the outcome of a game or team rankings in national championships to the identification of player characteristics that can improve their performance.
Our work shows that Multinomial model is a reasonable approach when it comes to predict the result of a game in terms of win, tie or loss. Whereas if one wants to predict the goals scored by each team on a match a modified–Poisson distribution might be the best way to go. The Bayesian context facilitates the predictions for a new game because they are naturally accommodated in terms of the posterior predictive distribution.
We do not account for possible information embedded in covariates and from this point of view the models are simple and not dependent on context features. The approach is underpinned in terms of random-effects that inform about the attack and defense abilities of the teams enrolled. There is also one parameter that represents the home advantage. From our understanding of the game a home-team has generally some unobserved advantage. This particular feature is also referred throughout the literature.
Several standard metrics (scoring rules) were used for models assessment, namely the proportion of interest outcomes incorrectly predicted by the models which were applied and assessed on a data set related to the 2014–2015 season of the Portuguese first league.

Acknowledgements: This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through the project UIDB/00006/2020.

References

[1] Baio, G., Blangiardo, M. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistic*, 37(2), 253-264, 2010. doi:10.1080/02664760802684177

[2] Diniz, M.A., Izbicki, R., Lopes, D., Salasar, L.E. Comparing probabilistic predictive models applied to football. *Journal of the Operational Research Society*. 70(5), 770-782, 2019. doi:10.1080/01605682.2018.1457485

[3] Conceição, K.S., Suzuki, A.K., Andrade, M.G., Louzada, F. A Bayesian approach for a zero modified Poisson model to predict match outcomes applied to the 2012–13 La Liga season. *Brazilian Journal of Probability and Statistics*, 31(4), 746-764, 2017. doi:10.1214/17-BJPS379

Bayesian approach for treating missing data in count time series

Isabel Silva [a], Maria Eduarda Silva [b], Isabel Pereira [c]
ims@fe.up.pt, mesilva@fep.up.pt, isabel.pereira@ua.pt

[a] *Faculdade de Engenharia, Universidade do Porto and CIDMA*
[b] *Faculdade de Economia, Universidade do Porto and LIADD-INESC TEC*
[c] *Departamento de Matemática, Universidade de Aveiro and CIDMA*

Keywords: approximate Bayesian computation, Bayesian estimation, Gibbs sampler with data augmentation, INAR(1) model, missing data

Abstract: Missing data can arise for several reasons, such as equipment failure, measurement errors, or simply when data is not available for certain time points. If not appropriately addressed, missing data can lead to biased parameter estimates, reduced efficiency, and inaccurate predictions that can affect the validity and reliability of the statistical analysis.
The aim of this work is to contribute to the modelling of time series of counts in the presence of missing data, based on first-order integer-valued autoregressive (INAR) models. Dealing with missing data in time series of counts presents unique challenges because the temporal dependence and discrete nature of the data need to be taken into account. There are few works in the literature for treating missing data in integer-valued autoregressive processes (see for instance [1] and [3]).
Here, the problem of estimating INAR(1) models in the presence of missing data is approached from a Bayesian perspective with Approximate Bayesian Computation (ABC) and Gibbs sampler with Data Augmentation (GDA) algorithms (see [2] and references therein). The methodologies are illustrated with synthetic and real data and the results indicate that the estimates are consistent and present less bias than those obtained when the missing data is neglected.

Acknowledgements: The first and third authors were partially supported by CIDMA through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. The second author was partially supported by Portuguese National Funds through the FCT as part of project LA/P/0063/2020.

References

[1] Andersson, J., Karlis, D. Treating missing values in INAR(1) models: An application to syndromic surveillance data. *Journal of Time Series Analysis*, 31, 12–19, 2010. doi:10.1111/j.1467-9892.2009.00636.x

[2] Silva, I., Silva, M.E., Pereira, I., McCabe, B. Time Series of Counts under Censoring: A Bayesian Approach. *Entropy*, 25, 549, 2023. doi:10.3390/e25040549

[3] Xiong, W. , Wang, D., Wang, X. Imputation-based semiparametric estimation for INAR(1) processes with missing data. *Hacettepe Journal of Mathematics and Statistics*, 49, 1843-1864, 2020. doi:10.15672/hujms.643081

Bayesian modeling count time series with structural breaks

Isabel Pereira [a], Betty Nakyambadde [b], Cláudia Santos [c]
isabel.pereira@ua.pt, betty@ua.pt, csps@ua.pt

[a] *Departamento de Matemática, Universidade de Aveiro and CIDMA*
[b] *Departamento de Matemática, Universidade de Aveiro*
[c] *ESAC, Instituto Politécnico de Coimbra and CIDMA*

Keywords: Gibbs sampling, INAR models, MCMC, structural break

Abstract: Time series of count are prevalent in various scientific fields. For instance, they are used to track the daily number of patients admitted to a hospital, the minute-by-minute transactions of a specific stock, or the monthly count of car accidents in a particular region. Current research focuses on integer autoregressive models, known as INAR(p), where p represents the autoregressive order. These models assume that the parameters remain constant over time, but this may not hold true in practice. Typically, the daily count of affected cases starts low in the initial phase of an epidemic, increases, and eventually decreases.
There are several techniques and model frameworks available for detecting breakpoints in count processes. Considering INAR model with structural breaks, it is applied the proposal of [1] in order to estimate the parameters of the process and the break points through Markov Chain Monte Carlo (MCMC) and Gibbs sampling. Additionally, it is addressed the prediction of future observations. The proposed methodology is applied considering real data sets in health indicators context.

Acknowledgements: The first and third authors were partially supported by CIDMA through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

References

[1] Kashikar, A.S., Rohan, N., Ramanathan, T.V. Integer autoregressive models with structural breaks. *Journal of Applied Statistics*, 40:12, 2653-2669, 2013. doi: 10.1080/02664763.2013.823920

Bayesian smoothing for time-varying joint extremes

Miguel de Carvalho [a]
Miguel.deCarvalho@ed.ac.uk

[a] *Universidade de Edimburgo*

Keywords: asymptotic (in)dependence, Bayesian P-splines, extreme value theory, international equity markets, nonstationary extremes

Abstract: In this talk, I will propose a Bayesian time-varying model that learns about the dynamics governing joint extreme values over time. Our model relies on dual measures of time-varying extremal dependence, that are modelled via a suitable class of generalized linear models conditional on a large threshold. The simulation study indicates that the proposed methods perform well in a variety of scenarios. The application of the proposed methods to some of the world's most important stock markets reveals complex patterns of extremal dependence over the last 30 years, including passages from asymptotic dependence to asymptotic independence. Joint work with J. Lee, A. Rua, and J. Avila.

Acknowledgements: I thank, without implicating, I. Papastathopoulos, R. Huser, L. Mhalla, and V. Inácio de Carvalho for a variety of helpful comments and feedback. The research was partially supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal) through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2020.

References

[1] Lee, J., de Carvalho, M., Rua, A., Avila, J. Bayesian smoothing for time-varying extremal dependence. *Journal of the Royal Statistical Society, Ser.* C, conditionally accepted.

A new class of conditional tail expectation estimators

Lígia Henriques-Rodrigues[a], M. Ivette Gomes[b],
Fernanda Otília Figueiredo[c], Frederico Caeiro[d]
ligiahr@uevora.pt, migomes@ciencias.ulisboa.pt, otilia@fep.up.pt,
fac@fct.unl.pt

[a] *Universidade de Évora and Centro de Investigação em Matemática e Aplicações (CIMA)*
[b] *Universidade de Lisboa and Centro de Estatística e Aplicações (CEAUL)*
[c] *Faculdade de Economia da Universidade do Porto and Centro de Estatística e Aplicações (CEAUL)*
[d] *Universidade Nova de Lisboa and Centro de Matemática e Aplicações (CMA)*

Keywords: conditional tail expectation, generalized means, Monte-Carlo simulation, semi-parametric estimation, statistics of extremes

Abstract: Extreme value theory as shown to be an important tool in finance and risk management to assess the tail risk of a distribution or portfolio. From the several risk measures available, the conditional tail expectation (CTE) will be considered as it is regarded as more informative than the value at risk at a level q (VaR_q), which represents the upper (1-q)-quantile of the loss function, being defined as $\text{CTE}_q = E(X|X > \text{VaR}_q)$. We consider a Pareto tail for the right-tail function and work with heavy tailed models, i.e., models with a positive extreme value index (EVI), quite common in finance. For these models, the classical EVI estimator is the Hill estimator [1]. Among the several classes of CTE estimators, we will consider one particular class proposed by [2]. The link between the estimation of both EVI and CTE allows for the utilization of the class of EVI estimators introduced in [3] and based on the mean of order p of the log-excesses in CTE estimation. Under adequate conditions consistency of the new class of CTE estimators will be derived. To assess the behaviour of this class in finite samples, Monte Carlo simulation experiments will be conducted.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, within the projects UIDB/04674/2020 (CIMA), UIDB/00006/2020 (CEA/UL) and UIDB/MAT/0297/2020 (CMA/UNL).

References

[1] Hill, B.M. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3, 1163–1174, 1975. doi:10.1214/aos/1176343247

[2] Necir, A., Rassoul, A., Zitikis, R. Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, 2010, Article ID 596839, 17 pages, 2010. doi:10.1155/2010/596839

[3] Gomes, M.I., Martins, M.J. Generalizations of the Hill estimator – asymptotic versus finite sample behaviour. *Journal of Statistical Planning and Inference*, 93, 161–180, 2001. doi:10.1016/S0378-3758(00)00201-9

Location invariant estimation of the Weibull tail coefficient

M. Ivette Gomes[a], Frederico Caeiro[b], Lígia Henriques-Rodrigues[c], migomes@ciencias.ulisboa.pt, fac@unl.pt, ligiahr@uevora.pt

[a] *Universidade de Lisboa and Centro de Estatística e Aplicações (CEA/UL)*
[b] *Universidade Nova de Lisboa and Centro de Matemática e Aplicações (CMA)*
[c] Universidade de Évora and Centro de Investigação em Matemática e Aplicações (CIMA)

Keywords: semi-parametric estimation, statistics of extremes, Weibull tail-coefficient

Abstract: The *Weibull tail-coefficient* (WTC) is the index of regular variation in a regularly varying cumulative hazard function $H(x) = -\log(1 - F(x))$. Due to the specificity of the WTC, and its deep and explicit link to a positive *extreme value index* (EVI), any estimator of a positive EVI, like all *generalised means* (GMs) (see [1], among others), generalizing the classical Hill estimator in [2], can be used for the estimation of the WTC (see [3, 4]). Contrarily to the EVI and the WTC, these estimators are scale invariant but not location invariant. With PORT standing for *peaks over random threshold*, new classes of PORT WTC-estimators are now introduced. These classes are dependent on an extra tuning parameter s, $0 \leq s < 1$, and they are both location and scale invariant. The asymptotic normal behaviour of those PORT classes is derived. These EVI-estimators are further studied for finite samples, through a Monte-Carlo simulation study. An adequate choice of the tuning parameters under play is put forward, and some concluding remarks are provided.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, within the projects UIDB/04674/2020 (CIMA), UIDB/00006/2020 (CEA/UL) and UIDB/MAT/0297/2020 (CMA/UNL).

References

[1] Caeiro, F., Gomes, M.I., Beirlant, J., de Wet, T. Mean-of-order-p reduced-bias extreme value index estimation under a third-order framework. *Extremes*, 19:4, 561–589, 2016. doi:10.1007/s10687-016-0261-5

[2] Hill, B.M. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3, 1163–1174, 1975. doi:10.1214/aos/1176343247

[3] Caeiro, F., Gomes, M.I., Henriques-Rodrigues, L. Estimation of the Weibull Tail Coefficient through the Power Mean-of-Order-p. In Bispo R, Henriques-Rodrigues L, Alpizar-Jara R and De Carvalho M (Eds.), *Recent Developments in Statistics and Data Science*, Springer Proceedings in Mathematics and Statistics, vol 398, Springer, Cham, Chapter 4, pp. 41–53, 2022. doi:10.1007/978-3-031-12766-3_4

[4] 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*, Article ID 7290822, 12 pages, 2022. doi:10.1155/2022/7290822

A partially reduced bias Hill estimator of the Extreme Value Index

Frederico Caeiro[a], M. Ivette Gomes [b], Lígia Henriques-Rodrigues[c]
fac@fct.unl.pt, migomes@ciencias.ulisboa.pt, ligiahr@uevora.pt

[a] *Universidade Nova de Lisboa and Centro de Matemática e Aplicações (CMA)*
[b] *Universidade de Lisboa and Centro de Estatística e Aplicações (CEA/UL)*
[c] *Universidade de Évora and Centro de Investigação em Matemática e Aplicações (CIMA)*

Keywords: bias reduction, extreme value index, semi-parametric estimation, statistics of extremes

Abstract: The estimation of the extreme value index (EVI) plays a crucial role in modelling and predicting extreme events, such as floods, earthquakes, and heat waves. The Hill estimator [2], defined as the average of the log-excesses of a high threshold is a popular choice for estimating the EVI, primarily due to its simplicity. However, the Hill estimator is known to suffer from bias, particularly when the estimation is based on a large number of log-excesses. In this paper, we propose a partial bias corrected Hill estimator that addresses this issue and provides more accurate estimates than the Hill estimator. A comparison with other reduced bias estimators, from the literature (see [1]), is also provided.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, within the projects UIDB/04674/2020 (CIMA), UIDB/00006/2020 (CEA/UL) and UIDB/MAT/0297/2020 (CMA/UNL).

References

[1] Caeiro, F., Gomes, M.I., Henriques Rodrigues, L. Reduced-bias tail index estimators under a third order framework. *Communications in Statistics - Theory and Methods*, 38, 1019–1040, 2009. doi:10.1080/03610920802361415

[2] Hill, B.M. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3, 1163–1174, 1975. doi:10.1214/aos/1176343247

A direct approach in extremal index estimation

M. Cristina Miranda [a,b,c], M. Souto de Miranda [a], M. Ivette Gomes [b]
cristina.miranda@ua.pt, manuela.souto@univ.pt,
migomes@ciencias.ulisboa.pt

[a] *CIDMA, University of Aveiro*
[b] *CEAUL, University of Lisbon*
[c] *ISCA, University of Aveiro*

Keywords: extremal index, logistic model, robust estimation

Abstract: It is known that the limit distribution of the maxima of stationary sequences exists under specific conditions, even in the presence of some dependence structures. Dealing with sequences of maxima, the degree of dependence between those observations can be studied through a parameter of the Extreme Value Distribution, named the extremal index (*EI*). That parameter is theoretically known for some particular models and might be interpreted in different contexts, namely, like the inverse of the mean size of clusters of exceedances, or as the multiplicity of a compound Poisson point process (see, e.g., Moloney et al. (2019)). Generally, *EI* estimation methods are focused on the mean clusters size and their properties. Then, the numerical inverse of the mean estimate provides the extremal index estimate. In this study we present a method that is based on direct estimation of the parameter itself. The procedure takes into account the distribution of the inter-exceedances times as derived in Ferro and Segers (2003) and considers one of the most used robust estimators for the logistic regression, as proposed by Bianco and Yohai (1997).

Acknowledgements: Present research was supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT), reference UIDP/04106/2020.

References

[1] Bianco, A., Yohai, V. Robust estimation in the logistic regression model. *Robust Statistics, Data Analysis and Computer Intensive Methods*, Rieder, H. (ed), 17–34. New York. Springer-Verlag, 1997.

[2] Ferro, C., Segers, J. Inference for Clusters of Extreme Values. *J. R. Statist. Soc. B* 65, 545–56, 2003.

[3] Moloney, N., Faranda, D., Sato, Y. An overview of the extremal index. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29 (2), 2019.

An additive shared frailty model for recurrent gap time data in the presence of zero-recurrence subjects

Ivo Sousa-Ferreira[a,c], Ana Maria Abreu[a,d], Cristina Rocha[b,c]
ivo.ferreira@staff.uma.pt, abreu@staff.uma.pt, cmrocha@ciencias.ulisboa.pt

[a] *Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal*
[b] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[d] *CIMA – Centro de Investigação em Matemática e Aplicações, Portugal*

Keywords: additive shared frailty, gap times, non-homogeneous Poisson process, recurrent events, zero-recurrence

Abstract: Over the past decade, considerable efforts have been dedicated to the development of survival models for the analysis of recurrent gap time data. Within this scope, Zhao and Zhou [2] proposed an additive semiparametric model with a rate function derived from a non-homogeneous Poisson process (NHPP). However, when it is relevant to obtain an estimate of the recurrence rate, using a parametric model is more advantageous. Furthermore, the within-subject correlation problem triggered by unobserved heterogeneity, which can lead to biased estimators, was not addressed in [2].
Thus, we propose a shared frailty model for recurrent gap time data, assuming that frailty acts additively on a parametric rate function that is derived from a NHPP. The frailty is included with two purposes: to handle within-subject correlation and to accommodate zero-recurrence subjects. Inspired by Rocha [1], we assume that the frailty has a non-central chi-squared distribution with zero degrees of freedom. Since the proposed model is fully specified, the parameter estimation is based on the unconditional likelihood function. An application to real data is presented to illustrate the potential of the new model.

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

References

[1] Rocha, C.S. Survival Models for Heterogeneity Using the Non-Central Chi-Squared Distribution with Zero Degrees of Freedom. In Jewell, N.P., Kimber, A.C., Lee, ML.T., Whitmore, G.A. (eds.) *Lifetime Data: Models in Reliability and Survival Analysis*. Springer, Boston, MA, 1996.

[2] Zhao, X., Zhou, X. Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data Analysis*, 56(2), 370–383, 2012.

Parametrizações do modelo conjunto para dados longitudinais e de sobrevivência

Maria Helena Santos de Oliveira [a], Isolde Previdelli [a, b]
pg403496@uem.br, isoldeprevidelli@gmail.com

[a] *Programa de Pós Graduação em Bioestatística, Universidade Estadual de Maringá, Brasil*
[b] *Departamento de Estatística, Universidade Estadual de Maringá, Brasil*

Keywords: análise de sobrevivência, COVID-19, dados longitudinais, modelos conjuntos

Abstract: Modelos conjuntos para dados longitudinais e de tempo até o evento são uma ferramenta valiosa para a investigação de associações entre desfechos longitudinais e de sobrevivência. O *framework* de modelagem conjuntos permite que dados longitudinais truncados por morte sejam modelados quando a morte é um mecanismo de perda não aleatória, assim como que processos de sobrevivência sejam modelados na presença de covariáveis tempo-dependentes endógenas. A conexão entre os dois processos pode ser especificada de maneiras diferentes, que levam à interpretações práticas variadas. O objetivo deste estudo é a aplicação de três diferentes parametrizações do modelo conjunto à dados de 58 pacientes infectados pela doença do Coronavírus 2019 internados em Unidade de Tratamento Intensivo (UTI) em um hospital na Arábia Saudita. Estes pacientes foram submetidos a exames laboratoriais diários até a sua alta ou óbito, e suas medidas de cloro foram modeladas utilizando modelos lineares de efeitos mistos em conjunto com regressão de riscos proporcionais para o seu tempo de sobrevivência. Os resultados obtidos indicam que, apesar de os valores absolutos de cloro ao longo do tempo não serem associados à mortalidade, a inclinação da trajetória longitudinal das medidas dos pacientes é, indicando que existe informação importante nas mudanças longitudinais dos níveis de cloro ao longo do tempo, especialmente no desenvolvimento de hipocloremia durante o internamento em UTI. Estes achados evidenciam a importância de explorar diferentes parametrizações ao utilizar a modelagem conjunta de desfechos longitudinais e de sobrevivência, e compreender as diferentes considerações práticas que podem ser derivadas de cada modelo.

Acknowledgements: Agradecemos a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - Brasil) pelo apoio financeiro.

Comparative analysis of Cox proportional Hazard Model and machine learning approaches for predicting financial distress in SMEs

Ana Borges [a], Mariana Reimão Carvalho [a]
aib@estg.ipp.pt, mrc@estg.ipp.pt

[a] *CIICESI, ESTG, Politécnico do Porto*

Keywords: Cox proportional hazard model, financial distress, random survival forests, SMEs, survival trees, time-to-event data

Abstract: A high rate of corporate bankruptcy can have severe repercussions on the economic and entrepreneurial ecosystem, particularly in countries with a significant concentration of Micro, Small, and Medium-Sized Enterprises (SMEs) [1]. Traditional static models often overlook the longitudinal factor of time when predicting corporate bankruptcy. In contrast, this study employs the Cox proportional hazard model (CPHM) to predict the probability of SMEs experiencing financial distress using historical data from SABI. To assess the performance of the CPHM, the study compares its predictive capability with machine learning models such as survival trees and random survival forests. These models consider the temporal aspect of time-to-event data, enabling a more nuanced understanding of the factors influencing financial distress. Several studies have explored the application of different modeling approaches for financial distress prediction in SMEs. For example, [2] compared a CPHM and non-parametric CART decision trees and found good prediction accuracy for both. And also, [3] employed several machine learn ing techniques, identifying Random Forest as the most accurate method for financial distress prediction when incorporating additional factors. For the analysis, a sampleof 18,140 Portuguese SMEs is considered. Results suggest that both the CPHM and machine learning approaches can be utilized to predict financial distress in SMEs. The choice of the most appropriate model may depend on the specific situation and the available data.

Acknowledgements: This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through project UIDB/04728/2020.

References

[1] Borges, A. , Machado, M., Duarte, F. Survival analysis of Portuguese SMEs: A preliminary approach. *AIP Conference Proceedings*, Vol 2186, No. 1, p. 090002, 2019. doi:10.1063/1.5137998

[2] Gepp, A., Kumar, K. Predicting financial distress: A comparison of survival nalysis and decision tree techniques. *Procedia Computer Science*, 54, 396–404, 2015. doi:10.1016/j.procs.2015.06.046

[3] Malakauskas, A., Lakštutienė, A. Financial distress prediction for small and medium enterprises using machine learning techniques. *Engineering Economics*, 32(1), 4–14, 2021. doi:10.5755/j01.ee.32.1.27382

NIV treatment effect estimation for ALS patients using Propensity Score methodology

Luís Garcez [a], Sara Madeira [b], Helena Mouriño [a]
luis_garcez_ferreira@ciencias.ulisboa.pt
[a] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*

Keywords: amyotrophic lateral sclerosis, non-invasive ventilation, prognosis, propensity score, survival analysis

Abstract: Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease, which leads to death usually from respiratory failure. One recommended treatment intervention is non-invasive ventilation (NIV) for ALS patients with respiratory impairment, which is known to increase survival. Propensity score (PS) methodology offers a way to access survival benefits in an intuitive way. This methodology matches treated and non-treated patients regarding a set of confounding variables, reducing the impact of confounders; therefore, an unbiased treatment effect estimate can be calculated. Firstly, this work aims to select the matching framework that best balances the treated and non-treated patients, by testing different combinations of matching and PS estimation methods. Then, after the best matching strategy is found, the matching is performed on different subsets of patients, in order to find the survival effect in each subgroup. The subsets include patients with severe, moderate and mild/no bulbar dysfunction, and patients with slow, average and fast disease functional decline. The survival benefits were significant in all subgroups, but more marked in patients with slow disease progression (with a median survival benefit of 315 days), and in patients with mild/no bulbar dysfunction (with a median survival benefit of 274 days). In the total ALS population, NIV has a marginal Hazard Ratio of 0.540 (0.468-0.623), and therefore reduces the death hazard by 46%, representing a median survival extension of 205 days (6.8 months). This work highlights the importance of NIV as a beneficial symptomatic treatment and that its early use can prolong the patient's survival.

References

[1] Dorst, J., Ludolph, A. C. Non-invasive ventilation in amyotrophic lateral sclerosis. *Ther. Adv. Neurol. Disord.*, 12, 2019.

[2] Rubin, D. B., Rosenbaum, P. R. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, vol. 70 SRC-, no. 1, pp. 41–55, 1983.

State-space models: fitting, modeling, and calibration

Marco Costa [a,b]

[a] *ESTGA-UA – Águeda School of Technology and Management, University of Aveiro, Portugal*
[b] *CIDMA – Research & Development in Mathematics and Applications, University of Aveiro, Portugal*

Keywords: calibration, environmental data, Kalman filter, state space model, time series

Abstract: State space models and the Kalman filter are widely used tools in environmental and climate time series modeling and analysis. They play an important role in the estimation and prediction of environmental variables, such as temperature, precipitation, humidity, sea level, gas concentration, among others. The relevance of these models lies in their flexible structure, the possibility of incorporating uncertainties from various sources, the updating of the predictions through the Kalman filter. In this work, we discuss some aspects related to the fitting of state space models, namely the estimation of unknown parameters and its implications on the optimality of Kalman filter predictors, [1]. In addition, two approaches to the application of state space models are presented: as an instrument for the modelling of time series essentially in a generalization perspective of linear regression models, [2], and as statistical models that allow the combination or assimilation of data, in cases where we have observations of the same phenomenon, but from different sources and, possibly, with different uncertainties, [3].

Acknowledgements: This work was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT), references UIDB/04106/2020 and UIDP/04106/2020.

References

[1] Auger-Méthé, M., Field, C., Albertsen, C., et al. State-space models' dirty little secrets: Even simple linear gaussian models can have estimation problems. *Sci Rep*, 6:26677, 2016. https://doi.org/10.1038/srep26677

[2] Monteiro, M., Costa, M. A time series models comparison for monitoring and forecasting water quality variables. *Hydrology*, 5 (3), 37, 2018. https://doi.org/10.3390/hydrology5030037

[3] Pereira, F.C., Gonçalves, A.M., Costa, M. Short-term forecast improvement of maximum temperature by state-space model approach: the study case of the TO CHAIR project. *Stoch Environ Res Risk Assess*, 37:219–231, 2023. https://doi.org/10.1007/s00477-022-02290-3

Regression models for count time series: an application to health care indicators

Rui Soares [a], **Magda Monteiro** [b,c], Isabel Pereira [a,c]
rui.soares@ua.pt, msvm@ua.pt, isabel.pereira@ua.pt

[a] *Mathematics Department, University of Aveiro*
[b] *ESTGA - Águeda School of Technology and Management, University of Aveiro*
[c] *CIDMA, University of Aveiro*

Keywords: count time series, generalized linear models, partial likelihood, regression models, zero-inflation

Abstract: Count time series naturally arises in several areas from health, to agriculture, finance, among many others. These series often exhibit non-negative autocorrelations and overdispersion and/or a high presence of zeros. In this last situation, the use of conditional distributions such as the Poisson and negative binomial distributions may result in erroneous inferences. To accommodate the specificities of these series several approaches have been proposed, of which we highlight the generalized linear models for the time series approach, also called regression models [3], which includes a feedback mechanism. The response Y at time t, given the information up to time t, has a known distribution with mean λ_t, which is regressed (or some function of it) on past values of the response and/or on covariates. In the absence of excess zeros, the most usual distributions are Poisson and negative binomial, the latter having proved to be useful in the presence of overdispersion [1]. When there is an excess of zeros, the use of distributions such as the zero inflated Poisson or the zero inflated negative binomial should be considered. The inference for such models is performed for instance under a partial likelihood framework [2]. The aim of this work is to apply generalized linear models for the count series in the context of health indicators that can present overdispersion and excess of zeros. The different models are compared and prediction of future values are also evaluated.

Acknowledgements: This research was partially supported by the Center for Research and Development in Mathematics and Applications through the Portuguese Foundation for Science and Technology, grant number UIDB/04106/2020.

References

[1] Liboschik, T., Fokianos, K., Fried, R. tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, 82(5), 1–51, 2017. https://doi.org/10.18637/jss.v082.i05

[2] Yang, M., Zamba, G., Cavanaugh, J.E. Markov regression models for count time series with excess zeros: A partial likelihood approach, *Statistical Methodology*, 14, 26–38, 2013. https://doi.org/10.1016/j.stamet.2013.02.001

[3] Zeileis, A., Kleiber, C., Jackman, S. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8), 1–25, 2008. https://doi.org/10.18637/jss.v027.i08

Automatic event diagnosis in water consumption

Rita Leite [a], Conceição Amado [b] Margarida Azeitona [c]
ritassleite@tecnico.lisboa.pt, conceicao.amado@tecnico.ulisboa.pt,
margarida.azeitona@baseform.com

[a] *IST-ULisboa*
[b] *CEMAT and IST-ULisboa*
[c] *Baseform*

Keywords: event detection, multi-class classification, time series classification, urban water distribution networks, water demand patterns

Abstract: Monitoring water demand is extremely helpful in the early detection of issues and malfunctions in water distribution networks. Baseform develops forward-thinking software for networked water infrastructures that assists water utilities in monitoring the entire network in its daily management. Several methodologies are continuously applied by Baseform to analyze and accurately predict water demand patterns (see [1], [2]) and to establish whether the consumption recorded during a specific period is normal or may reveal abnormal events. These include pipe bursts or leaks, a Legitimate change in consumer behavior, and Irrigation systems. However, it is often found that the appropriate post-processing that the identified events require is laborious, and it is, therefore, essential to develop mechanisms that automate many of the decisions in events' processing and reclassification.
Our approach starts with defining a set of new variables that try to capture the behavior of several types of events and then developing and comparing several statistical and machine learning methods to continuously classify those events.
Through the analysis carried out, it was possible to extract and better understand what features are relevant in the discrimination between different event categories. The first developed classifier still does not possess reliable enough performance to distinguish rare events, such as those related to pipe bursts. In fact, it will be necessary to conduct a more thorough examination of the extracted variables to make it clearer which further pre-processing steps, like event pruning, might be required to lessen the amount of noise in the data and promote a better classifier performance.

References

[1] Azeitona, M., Coelho, S., Vitorino, D. A Parametrização Automática e Dinâmica De Padrões De Consumo Em Abastecimento De Água. 17º ENASB, Sep 14-16 Guimarães, Portugal 2016.

[2] Vitorino, D., Loureiro, D., Alegre, H., Coelho, S., Mamade, A. In defense of the demand pattern, a software approach. *16th Water Distribution System Analysis Conference, WDSA2014*, 89, 982-989, 2014, doi:10.1016/j.proeng.2014.11.215

Improving parameter estimation and prediction accuracy by handling outliers in state-space modeling

F. Catarina Pereira [a], A. Manuela Gonçalves [b], Marco Costa [c]
id9976@alunos.uminho.pt, mneves@math.uminho.pt, marco@ua.pt

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

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

Abstract: Most real-time series present challenges in selecting and specifying an appropriate model. Particularly in time series modeling and forecasting, the existence of outliers poses a significant and common challenge, for these outliers have the potential to impact various aspects, including parameter estimation, forecasting accuracy, and inference results. The main objective of this study is to discuss competitive methods for outlier detection and treatment in time series analysis through state-space modeling. In this study, outlier detection is performed using time series data and the standardized residuals obtained after the model's adjustment. The outlier treatment methodologies proposed in this research include the linear interpolation (LI), an iterative method based on the robustified Kalman filter (RKF), and another iterative method employing Kalman filter predictors (naKF). To evaluate the performance of the proposed outlier detection and treatment methods, a simulation study is presented considering a simplified time-invariant model with a state-space representation and Gaussian errors with different parameter combinations and sample sizes.

Acknowledgements: F. Catarina Pereira 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. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. Marco Costa was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA-UA) through the Portuguese Foundation for Science and Technology – FCT, references UIDB/04106/2020 and UIDP/04106/2020.

References

[1] Rose, F.Z.C., Ismail, M.T., Tumin, M. Outliers detection in state-space model using indicator saturation approach. *Indonesian Journal of Electrical Engineering and Computer Science*, 22, 1688–1696, 2021. doi:10.11591/ijeecs.v22.i3.pp1688-1696

Air quality data analysis with Symbolic Principal Components

Catarina P. Loureiro [a,b], M. Rosário Oliveira [a,b],
Paula Brito [c,d], Lina Oliveira [b,e]
catarinapadrela@tecnico.ulisboa.pt, rosario.oliveira@tecnico.ulisboa.pt,
mpbrito@fep.up.pt, lina.oliveira@tecnico.ulisboa.pt

[a] *CEMAT*
[b] *Dep. Mathematics, Instituto Superior Técnico*
[c] *LIAAD-INESC TEC*
[d] *Fac. Economia, Universidade do Porto*
[e] *CAMGSD*

Keywords: air quality, control chart, interval principal component analysis, outlier detection, symbolic data analysis

Abstract: Symbolic Data Analysis is a growing area of study that focuses on modelling complex data types, including intervals. In this work, we examine air quality data from a monitoring station in Entrecampos, Lisbon, through the lens of Symbolic Data Analysis. The dataset consists of nine pollutants' concentration measurements recorded hourly over 3 years. We start by aggregating the data into intervals, taking the daily minimum and maximum values. The symbolic mean and variance are estimated for each variable by the method of moments. In order to capture the pairwise dependencies between the interval-valued variables, we employ a bivariate copula to estimate the likelihood function. The estimated covariance matrix is then used to perform symbolic principal component analysis, which is used to fit a generalized extreme value distribution. Finally, a control chart is constructed based on the quantiles of this distribution, and is used to identify outlying observations. Additionally, a comparison is made with conventional principal component analysis based on daily averages. The results show that this approach based on interval-valued data is able to detect outlying maximums, and even outperforms the conventional approach in certain instances. These findings demonstrate the applicability of symbolic principal component analysis for interval-valued data and provide insights into the air quality measurements in Lisbon.

Acknowledgements: Catarina P. Loureiro is supported by FCT - Fundação para a Ciência e Tecnologia, Portugal, through the grant UI/BD/153720/2021. This work was also supported by Fundação para a Ciência e Tecnologia, Portugal, through the projects UIDB/04621/2020, UID/MAT/04459/2020 and LA/P/0063/2020.

References

[1] Lin, L.C., Guo, M., Lee, S. Monitoring photochemical pollutants based on symbolic interval-valued data analysis. *Advances in Data Analysis and Classification*, 2022. https://doi.org/10.1007/s11634-022-00527-1

Modeling landing per unit effort (LPUE) abundance of fish using functional data analysis

Manuel Oviedo-de la Fuente [a], Raquel Menezes [b], Alexandra A. Silva [cd]
manuel.oviedo@udc.es, rmenezes@math.uminho.pt, asilva@ipma.pt

[a] *CITIC, University of A Coruña*
[b] *CMAT, Minho University, Guimarães*
[c] *Portuguese Institute for the Sea and Atmosphere (IPMA), Lisboa*
[d] *MARE / ARNET, Faculty of Sciences, University of Lisbon*

Keywords: data monitoring, functional regression, LPUE, variable selection

Abstract: Forecasting the abundance of landed fish per unit effort (LPUE) is a crucial challenge in competitive fish markets. Previous studies have addressed this issue using various models such as ARIMA, generalized linear models (GLMs), generalized additive models (GAMs), and geostatistical models like continuous Gaussian random fields (GRFs), among others used to model species distribution in fisheries. However, this paper proposes an alternative approach based on functional data analysis (FDA). FDA is a statistical branch that focuses on analyzing data consisting of curves or anything else varying over a continuum. We use sensor data monitoring, such as chlorophyll-a concentrations, intensity of ocean currents, Sea Surface Temperature, wind speed, and wind direction curves. Furthermore, this paper addresses the challenge of variable selection by employing distance correlation to investigate the relationships between environmental curves (including their derivatives) and other sources of information, such as sale prices at landing, calendar variables, and the scalar response (LPUE). The proposed functional approach, specifically a functional generalized additive model (GAM) with variable selection, has demonstrated promising results when applied to a real dataset (LPUE of juvenile sardine along the northern Portuguese coast in 2009-2010). These findings present decision makers with a valuable tool to advance marine sustainability and conservation efforts by enhancing our understanding of the factors influencing LPUE.

Acknowledgements: This research/work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01), all of them through the ERDF. Authors also acknowledge the FCT Foundation for funding their research through projects PTDC/MAT-STA/28243/2017, UIDB/00013/2020 and UIDP/00013/2020 and MAR2020 for funding the SARDINHA2020 project (MAR-01.04.02-FEAMP-0009).

References

[1] Febrero-Bande, M., González-Manteiga, W., Oviedo de la Fuente, M. Variable selection in Functional Additive Regression Models. *Computational Statistics*, 34, 469–487, 2019. https://doi.org/10.1007/s00180-018-0844-5.

Green exchange-traded fund performance evaluation using the EU-EV risk model

Irene Brito [a], Ana Isabel Azevedo [b], José Azevedo[b]
ireneb@math.uminho.pt, aazevedo@iscap.ipp.pt, jazevedo@iscap.ipp.pt

[a] *Center of Mathematics, Department of Mathematics, University of Minho, 4800-045 Guimarães, Portugal*
[b] *CEOS.PP, ISCAP, Polytechnic of Porto*

Keywords: EU–EV risk model, exchange-traded funds, performance evaluation

Abstract:
This work evaluates the performance of green exchange-traded funds (ETFs) using the expected utility, entropy and variance (EU-EV) risk model. Data from a dataset of 14 green ETFs analysed in earlier literature in the in-sample period from January 2008 to December 2010 are used.
The analysis consists first in assessing the green ETFs using the recently proposed EU-EV risk model. The green ETFs are ranked according to their risk, considering the returns' expected utility, entropy and variance, and the best-ranked ETFs are selected to construct equally weighted portfolios. Then, the performance of the green ETFs portfolios is evaluated and compared with those of the S&P500 index. Cumulative returns in in-sample and out-of-sample periods and different performance metrics, such as Sharpe ratio, Sortino ratio, Beta and Alpha, are analysed. The results show that, in general, the equally weighted portfolios formed with half the number of best-ranked ETFs outperform the benchmark index in the in-sample period and for specific time ranges in the out-of-sample periods.

References

[1] Brito, I. A portfolio stock selection model based on expected utility, entropy and variance. *Expert Systems with Applications*, 213, Part A, 118896, 2023. doi: 10.1016/j.eswa.2022.118896

[2] Sabbaghi, O. Do Green Exchange-Traded Funds Outperform the S&P500?. *Journal of Accounting and Finance*, 11, 50–59, 2011.

[3] Tsolas, I.E., Charles, V. Green exchange-traded fund performance appraisal using slacks-based DEA models. *Operational Research*, 15, 51–77, 2015. doi: 10.1007/s12351-015-0169-x

Análise e avaliação de falhas em estações maregráficas

Dora Carinhas[a,b], Paulo Infante[a], António Martinho[c]
dora.carinhas@hidrografico.pt, pinfante@uevora.pt, santos.martinho@marinha.pt

[a] *CIMA/IIFA e DMAT/ECT, Universidade de Évora*
[b] *Instituto Hidrográfico*
[c] *CINAV, Escola Naval, Instituto Universitário Militar*

Keywords: estação maregráfica, fiabilidade, modelos de sobrevivência, previsão

Abstract: A clássica questão da qualidade dos dados maregráficos tornou-se mais importante na última década, pois surgiram novas tecnologias, as redes maregráficas foram modernizadas, factos que obrigam a olhar com atenção o desempenho dos marégrafos. A norma NP EN 13306, de 2007, define fiabilidade como a "aptidão de um bem para cumprir uma função requerida sob determinadas condições, durante um dado intervalo de tempo". A distribuição de Weibull é muito flexivel em termos de modelação de fiabilidade dos vários componentes de uma estação maregráfica, mesmo quando dispomos de um número reduzido de dados [1]. Os sistemas complexos reparáveis, como é o caso das estações maregráficas, são reparados em vez de serem substituídos, quando ocorrem avarias; essas avarias podem ter causas variadas, assim estamos perante modos de falhas misturados e a distribuição de Weibull conduz a um ajustamento com parâmetro de forma igual a 1 [1]. O modelo sugerido é o de Crow-AMSAA, desenvolvido nos anos 70 por Larry Crow [2]. Neste estudo, os dados de falha necessários para análise de fiabilidade foram recolhidos nas estações maregráficas de Leixões, Setúbal-Troia e Sines, durante, aproximadamente, 15 anos. O software R foi utilizado para o cálculo das estimativas de parâmetros e recorreu-se ao teste de Kolmogorov-Smirnov para rejeitar ou aceitar a hipótese do modelo de distribuição.

References

[1] Pereira, F., Sena, F. *Fiabilidade e sua aplicação à manutenção*. Publindústria, Porto, 2012.

[2] Comerford, N. Crow/AMSAA Reliability Growth Plots. *16th Annual Conference 2005 - Rotorua*, 2005.

From sums of unequal size samples to the mean and standard deviation

Miguel Casquilho [a], **Cecília Castro** [b], Jorge Buescu [c]
mcasquilho@tecnico.ulisboa.pt, cecilia@math.uminho.pt, jsbuescu@fc.ul.pt

[a] *IST (Instituto Superior Técnico), University of Lisbon, Lisbon, Portugal and CERENA*
[b] *University of Minho, Braga, Portugal, and CMAT*
[c] *FCUL (Faculdade de Ciências), University of Lisbon, and CMAFCIO*

Keywords: computation, Gaussian distribution, inference, internet, statistical quality control

Abstract: The sums of unequal size samples are data of frequent use in many activities, mainly for control reasons. Typically, the individual item values are not known, such as in the case of a load of bags on a truck. We show how to obtain the point estimation of the mean and standard deviation, and their confidence intervals, for Gaussian items, from those sums alone. The estimations may, namely, help to confirm estimates from statistical process control. For the parameters mean and standard deviation of the distribution we derive point estimates, which lead to weighted statistics, and we derive confidence intervals, leading to a conjecture for the mean, and a proposal for the standard deviation. The results can be efficiently verified by direct computation and by simulation, which can be run on public web pages of ours, namely for possible industrial use.

Acknowledgements: M. C. is at the Department of Chemical Engineering, IST, University of Lisbon, and CERENA, "Centro de Recursos Naturais e Ambiente", under Project UID/04028/2020 of FCT, "Fundação para a Ciência e a Tecnologia" (Portuguese *National Science Foundation*). C. C. is at the Dept. of Mathematics, Univ, of Minho, and CMAT, "Centro de Matemática da Universidade do Minho"). J. B. is at the Dept. of Mathematics, FCUL, Univ. of Lisbon, and CMAFCIO, "Centro de Matemática Aplicada e Fundamentos e Investigação Operacional", under FCT Project UID/MAT/04561/2020. The computing resides at the system of CIIST, "Centro de Informática do IST" (Informatics Centre of IST).

References

[1] Casquilho, M., Buescu, J. Standard deviation estimation from sums of unequal size samples. *Monte Carlo Methods and Applications*, 28:3, 235–253, 2022. https://doi.org/10.1515/mcma-2022-2118

[2] Casquilho, M. http://web.tecnico.ulisboa.pt/~mcasquilho/compute/qc/f-BagsPECI.php

[3] Wilkinson, M. D., et al. Comment: the FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, Article ID 160018, 2016. https://doi.org/10.1038/sdata.2016.18

An application of cluster analysis with mortality rates of non-communicable diseases

Ana Paula Nascimento [a, b, c], Cristina Prudêncio [a, c, d], Brígida Mónica Faria [a, e], Mónica Vieira [a, c, d], Helena Bacelar-Nicolau [f, g]
ananascimento@ess.ipp.pt, cprudencio@ess.ipp.pt, monica.faria@ess.ipp.pt, mav@ess.ipp.pt, hbacelar@psicologia.ulisboa.pt

[a] *ESS, Polytechnic of Porto (ESS-P.PORTO), Porto, Portugal*
[b] *Health and Environment Research Center (CISA), Porto, Portugal*
[c] *Center for Translational Health and Medical Biotechnology Research (TBIO), Porto, Portugal*
[d] *Instituto de Investigação e Inovação em Saúde (i3S), Porto, Portugal*
[e] *Artificial Intelligence and Computer Science Laboratory (LIACC), University of Porto, Porto, Portugal*
[f] *Faculty of Psychology, University of Lisbon (FPUL), Lisboa, Portugal*
[g] *Instituto de Saúde Ambiental (ISAMB/FMUL), Lisboa, Portugal*

Keywords: cluster analysis, non-communicable diseases, Public Health

Abstract: Public Health dedicates to study and preventing diseases, extending lifespans, and enhancing quality of life through organized endeavors and well-informed decision-making. Therefore the analysis of population health factors is essential. Clustering results can help in identifying common causes or risk factors, aiding in the improvement of preventive medicine. This work is a continuation of a previous work presented in [1] with more recent data and applying agglomerative hierarchical cluster analysis of non-communicable diseases [2]. The age-standardized mortality rates are used as variables and the non-communicable diseases separated by gender are used as individuals. Using the Euclidean distance to measure diseases dissimilarity, the obtained hierarchy provides six main clusters of diseases. It is found that these clusters are sequentially formed, with increasing order of disease severity, corresponding to the increasing order of levels. Considering the hierarchy levels from the lowest to the highest, there is a difference between the diseases belonging to the third and fourth clusters compared to the previous work [1], meaning that the severity of the diseases belonging to these groups has changed in recent years. Once again, the results show that cerebrovascular diseases and ischemic heart disease belong to the last group, particularly in the case of males.

References

[1] Nascimento, A. P., Prudêncio, C., Vieira, M., Pimenta, R., Bacelar-Nicolau, H. *A typological study of Portuguese mortality from non-communicable diseases*. Adv. Science, Tech. and Eng. Systems, 5(5), 613–619 Doi:doi.org/10.25046/AJ050575

[2] Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á., Bacelar-Nicolau, L. *Clustering of variables with a three-way approach for health sciences*. TPM in Applied Psychology, 21(4), 435–447 Doi:doi.org/10.4473/TPM21.4.5

The magnitude and stability of protection against Omicron SARS-CoV-2 acquired by hybrid immunity

João Malato[a,b], Ruy M Ribeiro[a,c], Eugénia Fernandes[d], Pedro Pinto Leite[d], Pedro Casaca[d], Carlos Antunes[e,f], Válter R Fonseca[a], Manuel Carmo Gomes[e], Luís Graca[a]
jmalato@medicina.ulisboa.pt, lgraca@medicina.ulisboa.pt

[a]*Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Portugal*
[b]*Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal*
[c]*Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545*
[d]*Direção de Serviços de Informação e Análise, Direção Geral da Saúde, Lisboa, Portugal*
[e]*Faculdade de Ciências, Universidade de Lisboa, Portugal*
[f]*Instituto Dom Luiz, Universidade de Lisboa, Portugal*

Keywords: hybrid immunity, Omicron SARS-CoV-2, reinfection, vaccination

Abstract: By mid-2022, the SARS-CoV-2 Omicron BA.5 subvariant progressively displaced earlier subvariants, BA.1 and BA.2, displaying ability to evade immune responses elicited by prior infection.At that time, understanding the impact of BA.1 infection on the risk of reinfection with BA.5 was crucial, as adapted vaccines under development were based on BA.1.

Using the Portuguese COVID-19 registry, we estimated the relative risk (RR) of BA.5 infection in individuals with a previous infection during periods for distinct variants' predominance (Wuhan-Hu-1, Alpha, Delta, BA.1/BA.2) in relation to individuals without any documented infection for the same periods. We then modelled the stability of protection granted by hybrid immunity (vaccine + infection) over time. We found that a prior SARS-CoV-2 infection reduced the risk for BA.5 infection, with the protection effectiveness ($(1 - RR) \times 100\%$) for a first-time infection being 51.6% (95% CI, 50.6–52.6%), 54.8% (51.1–58.2%), 61.3% (60.3–62.2%), and 75.3% (75.0–75.6%), respectively, for Wuhan-Hu-1, Alpha, Delta, and BA.1/BA.2. Afterwards, when considering the the stability of acquired immunity over time, we estimated that there is a small but gradual decay of protection, as RR for BA.5 increased from 0.06 to 0.35 over a period of 3 to 8 months post BA.1/BA.2 infection, stabilising around 0.37 thereafter.

These results should be interpreted in the context of breakthrough infections within a highly vaccinated population (Portuguese population had >98% with completed primary vaccination series at the time of the study). In conclusion, infection with BA.1/BA.2 reduces the risk for breakthrough infections with BA.5 in a highly vaccinated population. This finding is critical to appraise the ever-evolving epidemiological situation, as well as the development of adapted vaccines, and the need for booster doses.

Acknowledgements: JM acknowledges funding from Fundação para a Ciência e a Tecnologia (grant refs. SFRH/BD/149758/2019 and UIDB/00006/2020).

Adjustment methods for confounding variables: a comparative study

Inês Fortuna [a, b],Luis Antunes [a], Claudia Vieira [b], Brígida Mónica Faria [a, c]
inesfortuna@gmail.com, lba@ess.ipp.pt, claudia.vieira@ipoporto.min-saude.pt, monica.faria@ess.ipp.pt

[a] *ESS, Polytechnic of Porto (ESS-P.PORTO), Porto, Portugal*
[b] *Instituto Português de Oncologia do Porto (IPO-Porto), Porto, Portugal*
[c] *Artificial Intelligence and Computer Science Laboratory (LIACC), University of Porto, Porto, Portugal*

Keywords: breast cancer, confounders, IPTW, neoadjuvant pertuzumab, propensity score, treatment efficacy

Abstract: Observational studies provide relevant evidence, however, they have an inherent lack of balance of baseline variables distribution between the study groups, making it difficult to understand the real treatment effect. There are many methods to balance the confounders. Traditional covariate adjustment is the most used, however, currently, it is also common to apply techniques based on propensity score (PS). One of them is Inverse Probability of Treatment Weighting (IPTW). The application of IPTW involves comparing two groups of samples weighed by inverse probability of treatment. The main advantage of using IPTW, in comparison to other propensity score techniques, is that it preserves all patient data while also enabling the balancing and evaluation of confounders before assessing the outcome. In this study, the effect of two neoadjuvant treatments for HER2-positive breast cancer was analysed. The treatments differed in four additional cycles of pertuzumab. Two methods of balancing the distribution of variables were applied, the IPTW and the traditional regression adjustment methods. The results after the application of both mentioned techniques allowed us to conclude that the therapy with double-block anti-HER2 seems more favourable. Besides, this treatment enabled a greater number of patients with pathologic complete response (pCR). It also allowed a reduction in the number of radical mastectomies. Although there were statistically significant differences in the type of surgery between the study groups, the difference in pCR was not significant. The IPTW methods should continue to be applied in other clinical studies to understand better their impact on the calculation of real treatment effect, including studies with more HER2 positive BC that apply the double block treatment with trastuzumab plus pertuzum1ab. Future studies should include more patients and these should have greater heterogeneity baseline features.

Acknowledgements: The authors would like to express their gratitude to ESS-P.PORTO and IPO-Porto for providing access to their facilities and resources.

The evolution of immigrant groups in Luxembourg What are the different pathways in the labour market?

Catarina Campos Silva [a], Paula Brito [b], Pedro Campos [c]
up201705753@fep.up.pt, mpbrito@fep.up.pt, pcampos@fep.up.pt

[a] *Faculdade de Economia, Universidade do Porto, Portugal*
[b,c] *Faculdade de Economia, Universidade do Porto & LIAAD INESC TEC, Portugal*

Keywords: clustering, immigration, labour force survey, Luxembourg, symbolic data analysis

Abstract: Luxembourg, known for its immigration tradition, attracts immigrants for work. This study examines different immigrant groups in the labour market from 2014 to 2022. The data source is the Labour Force Survey (LFS), and Symbolic Data Analysis (SDA) [1] was used to analyse it.
Microdata was aggregated and 21 symbolic objects were created based on birthplace and length of residence in Luxembourg. The objects were primarily described by 16 modal variables which are multi-valued variables with a frequency attached to each category. Then, clustering algorithms were applied and the hierarchical clustering using complete linkage and the χ^2-divergence dissimilarity, demonstrated the greatest separation between clusters and homogeneity within clusters. The Heuristic Identification of Noisy Variables (HINoV) algorithm [3], suggests that with only six variables the objects may be separated into groups with similar labour market profiles. The Monitoring the Evolution of Clusters (MEC) framework [2], monitors cluster transitions over time by identifying temporal relations between these structures. This was used to track the movement of population groups between clusters.
Results indicate that people from the European Union (EU) and Neighbouring countries have similar profiles while the Portuguese have opposite characteristics. The Luxembourgers are in between. Profiling people from non-EU countries is challenging. Lastly, the MEC framework revealed significant object movements from 2017 to 2018 and in the period 2019-2022.
This work combines LFS and SDA, making it easy to replicate in nations that use the LFS, enabling comparison of results and monitoring to continue in the future.

References

[1] Diday, E., Noirhomme-Fraiture, M. *SYMBOLIC DATA ANALYSIS AND THE SODAS SOFTWARE*. John Wiley & Sons, 2008.

[2] Oliveira, M., Gama, J. Understanding clusters evolution. *Workshop on Ubiquitous Data Mining*, Vol 500, 16–20, 2010.

[3] Walesiak, M., Dudek, A. Identification of noisy variables for nonmetric, symbolic data in cluster analysis. *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation eV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, 85–92, 2008.

Item response theory and confirmatory factor analysis of emotional distress in women from a breast cancer screening program

Ana Meireles [a], Bruno de Sousa [b], Isabel Natário [a], Vitor Rodrigues [c,d]
amc.martins@campus.fct.unl.pt, bruno.desousa@fpce.uc.pt, icn@fct.unl.pt, vrodrigues@netcabo.pt

[a] *Nova University Lisbon, FCT, Portugal*
[b] *Univeristy of Coimbra, Faculty of Psychology and Education Sciences, CINEICC, Portugal*
[c] *Univeristy of Coimbra, Faculty of Medicine, Portugal*
[d] *Liga Portuguesa Contra o Cancro - Núcleo Regional do Centro, Portugal*

Keywords: breast cancer screening program, CFA, IRT, PCQ questionnaire, psychometric properties

Abstract: Breast cancer is a public health problem, with high incidence and mortality, particularly among women. In Portugal, the Portuguese Cancer League (LPCC), with its Breast Cancer Screening Program, has played a key role in the diagnosis and prevention of this disease. Despite the great benefits that these programs bring, it is also important to evaluate the psychological impact that they may cause on the participants. This study presents and discusses the psychometric properties of the Portuguese version of the emotional sub-scale of the questionnaire PCQ-Negative Consequences in Breast Cancer Screening Program. This sub-scale is composed of five items on which women must manifest the frequency of certain feelings/emotions experienced about breast cancer in the week that follows the exam, using an ordinal scale of four points (0 – Not at all, . . . , 3 – Quite a lot of the time). A total of 1437 women, aged 43 to 76, who participated in the breast cancer screening program were randomly selected through a stratified proportional sampling design in 34 Municipalities in central Portugal. Women had a face-to-face interview and, if diagnosed negative for breast cancer, had a follow-up telephone interview Data collection took place from March 2020 through December 2021. The psychometric properties of the emotional dimension of the PCQ-Negative Consequences was evaluated using both Item Response Theory (IRT) and Confirmatory Factor Analysis (CFA). Results from both IRT and CFA show evidence of a good fit of the emotional sub-scale.

Acknowledgements: Funded by Fundação para a Ciência e a Tecnologia (FCT), POCI/01/0145/FEDER/029443 – SHSADReM.

References

[1] Cockburn, J., De Luise, T., Hurley, S., Clover, K. Developement and validation of the PCQ: A Questionnaire to measure the psycological consequences of screening mammography. *Soc. Sci. Med*, 34(10), 1129–1134, 1992 May. doi10.1016/0277-9536(92)90286-y.PMID:1641674

A study on the development of analytical skills of computer science students at Polytechnic of Porto

Eliana Costa e Silva [a,b], Cristovão Sousa [c,d]
eos@estg.ipp.pt, cds@estg.ipp.pt

[a] *CIICESI, ESTG, Politécnico do Porto, Portugal*
[b] *Centre Algoritmi, University of Minho, Portugal*
[c] *ESTG, Politécnico do Porto, Portugal*
[d] *INESC TEC, Portugal*

Keywords: analytical skills, correlation, data exploration, statistical inference, statistical literacy

Abstract: Analytical skills are crucial in today's data-driven world, since they enable us to make sense of complex information and find meaningful insights.
To contribute to the effective development of the analytical skills of the 2nd year students of the degree in Informatics Engineering at ESTG, P.Porto, a transdisciplinary project involving the course units of Software Engineering II (SE) and Computational Mathematics II (MC) was put into practice.
In groups with up to four elements, students had to: 1) formulate two or three research questions related to a set of Software Quality metrics, identified in the scope of a SCRUM-based software project development; 2) collect relevant data in six milestones, and synthesize those according to a specific deliverable template; 3) decide on the statistical techniques, apprehended at CM, that should be applied to answer the research questions; 4) perform data cleaning, exploration, and visualization, and deriving insights; 5) perform the statistical analysis of the data, using R-Studio; 6) produce a written report with the results and their interpretation; 7) produce a 10 minutes video communicating the main insights; 8) defend their work.
In the present work, the analysis between the results obtained by the different artifacts delivered by the students, namely: i) the report; ii) the video; iii) the final discussion of the results and; iv) the written test of MC are presented, in order to assess the ability to effectively use analytical tools, interpret results, and communicate findings in the context of future professional activities to better manage and adapt the software development methodologies.

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

How to move towards an inclusive education

Bruno de Sousa [a]
bruno.desousa@fpce.uc.pt

[a] *University of Coimbra, Faculty of Psychology and Education Sciences, CINEICC, Portugal*

Keywords: education statistics, distance learning, inclusive education, roadmap, universal design

Abstract: The Covid-19 pandemic forced all of us to go into distance learning, urging us to reflect on our teaching practices and finding new ways to reach out to our students. The internationalization of educational programs redefined concepts such as inclusion, where it is no longer restricted to special needs students, but expanded to a much broader concept where different cultures, languages, gender and sexual orientation perspectives need to be taken into consideration. With the enforcement of distance learning due to pandemics, are we actually creating a more inclusive environment for learning or are we enabling learning inequalities all over the World? UNESCO horrifying projections (UNESCO, 2020) that over 24 million individuals from pre-school to tertiary education will not return to school after the closing of the schools during Covid-19 pandemics, strengthening the importance of inclusive education as a way to fight against discriminatory attitudes, towards a society open to real diversity in terms of socioeconomic status, ethnicity, culture, disability or LGBTQ+ individuals. The results of a case study that took place during Covid-19 lock down will be presented here in which learning objects were proposed taken into consideration the principals of Universal Design proposed by architect Ronald Mace (1985). The roadmap created by this project will provide a reflection on the teaching practices and approaches used in order to reach out to all kinds of students.

References

[1] The Center for Universal Design. Environments and Products for All people. 1989. https://design.ncsu.edu/research/center-for-universal-design/

[2] UNESCO. *Towards Inclusion in Education: Status, Trends and Challenges. The UNESCO Salamanca Statement 25 Years on United Nations*. United Nations Educational, Scientific and Cultural Organization: Paris, France, 2020. https://reliefweb.int/sites/reliefweb.int/files/resources/374246eng.pdf

Identifying consumption profiles in load data analysis

Lucas Henriques [a], Cecilia Castro [a], Felipe Lima[b]
lucasdestefano2@hotmail.com, cecilia@math.uminho.pt,
felipepratalima@gmail.com

[a] *Centre of Mathematics, University of Minho, Braga, Portugal*
[b] *IT Department, Federal Institute of Alagoas, Murici, Alagoas, Brazil*

Keywords: hierarchical clustering, k-means, load profile analysis, machine learning techniques, self-organizing maps

Abstract: In this study we embark on an in-depth exploration of residential electrical load data gathered from a wide-ranging selection of households across Brasil. Leveraging the capabilities of unsupervised machine learning techniques, we aim to unearth unique consumption patterns, with a key focus on discerning low, medium, and high consumption clusters within the dataset. The analytical methodology was constructed in a series of stages, initiating with a rigorous preprocessing phase to tackle the inherent challenges of real-world data such as missing values and presence of irrelevant data. The core of our analysis lies in the application of three different clustering techniques: K-means, Agglomerative Hierarchical Clustering (AHC), and Self-Organizing Maps (SOM). Each of these methods brought unique strengths to the table and contributed significantly towards decoding complex consumption behaviors. Notably, all methodologies were successful in distinguishing between the predetermined consumption categories, thus substantiating their individual value in load data study. A comparative assessment of these methods highlighted their complementary nature and suitability under diverse data conditions and analysis requirements. Consequently, the findings position these unsupervised learning techniques as potent instruments for energy consumption analysis, with the choice of method to be guided by the specific research objectives and the nature of the data at hand. We further delineate potential areas for future investigations, including the consideration of additional consumption-influencing factors, the exploration of alternative clustering techniques, a more detailed analysis of temporal consumption patterns, and the practical applications of the identified consumption profiles. The latter holds the promise of crafting tailored energy efficiency initiatives and smarter load management strategies, responding to the pressing need for sustainable energy practices in today's world.

Acknowledgements: This work was partially supported by Portuguese funds through the CMAT - Research Centre of Mathematics of University of Minho - within references UIDB/00013/2020, UIDP/00013/2020.

Aprendizagem Automática *vs* Modelação Estatística

Ricardo Coelho [a], Isabel Natário [a]
rpe.coelho@campus.fct.unl.pt, icn@fct.unl.pt

[a] *CMA & Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Portugal*

Keywords: comparação, modelação estatística, modelos de aprendizagem automática

Abstract: O desenvolvimento tecnológico tem originado conjuntos de dados cada vez maiores e estruturalmente mais complexos. Muitas aplicações que requerem a análise de dados para estabelecer relações e fazer previsões, têm privilegiado o uso de modelos de aprendizagem automática (*machine learning*) sobre os modelos estatísticos tradicionais. Ambas as abordagens são poderosas nesse objetivo e, apesar de poderem resultar em conclusões semelhantes, na realidade são bastante distintas em concepção, pressupostos e aplicação, o que nem sempre é óbvio para os utilizadores. A modelação estatística centra-se em modelos para o processo de geração de dados e na inferência sobre as relações entre as variáveis observadas, baseados em pressupostos fortes sobre a distribuição dos dados, o tipo de relações a considerar e a parcimoniosidade. Tais pressupostos devem de ser validados para que possamos ter confiança nos resultados. Os modelos de aprendizagem automática procuram descortinar padrões e estabelecer relações de forma automática, usando métodos estatísticos e probabilísticos para aprender diretamente dos dados, guiados pela procura de um bom desempenho preditivo. Não dependem de suposições específicas sobre a distribuição dos dados, ganhando flexibilidade. Aliado à necessidade de lidar com grandes quantidades de dados, o seu uso indiscriminado tem sido potenciado, o que faz questionar se não seria preferível a modelação estatística, mais robusta, estabelecendo relações interpretáveis e permitindo a inferência. Comparam-se as abordagens da modelação estatística versus aprendizagem automática, com base nos seus métodos mais comuns, estabelecendo semelhanças e diferenças e elencando as vantagens e as desvantagens de cada um deles.

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 e UIDP/00297/2020 (Centro de Matemática e Aplicações).

References

[1] Bzdok, D., Krzywinski, M., Altman, N. Machine learning: supervised methods. *Nature methods*, 15, 5, 2018.

[2] Dangeti, P. *Statistics for machine learning.*. Packt Publishing, 2017.

[3] Bennett, M., Kleczyk, E.J., Hayes, K., Mehta, R. Evaluating Similarities and Differences between Machine Learning and Traditional Statistical Modeling in Healthcare Analytics. *IntechOpen*, 2022.

Optimizing retail sentiment analysis with SentiLex-PT and Machine Learning

Catarina Almeida [a], Cecilia Castro [b], Ana Cristina Braga [c], Ana Freitas [d]
pg46717@alunos.uminho.pt, cecilia@math.uminho.pt, acb@dps.uminho.pt, ANCFREITAS@mc.pt

[a] *Master in Mathematics and Computer Science, University of Minho, Portugal*
[b] *Centre of Mathematics, University of Minho, Braga, Portugal*
[c] *ALGORITMI Research Centre, LASI, University of Minho, Braga, Portugal*
[d] *MC Sonae, Senhora da Hora, Matosinhos, Portugal*

Keywords: customer feedback, machine learning, Portuguese retail, SentiLex-PT, sentiment analysis

Abstract: In our study, we employ a unique approach combining SentiLex-PT, a comprehensive Portuguese sentiment lexicon, with various machine learning techniques to conduct sentiment analysis of Portuguese retail customer feedback. We focus on multiple levels of analysis, including document, sentence, and phrase level, providing a comprehensive sentiment understanding. Multiple machine learning models, including Naive Bayes, Random Forests, and Multinomial Logistic, were assessed. The Multinomial Logistic model emerged as the top performer, exhibiting strong predictive capabilities for positive and neutral sentiments. Fine-tuning of the model through feature engineering and hyperparameter optimization further improved its performance. Despite promising results with pre-trained models such as Twitter Roberta base Sentiment Latest, our study suggests the Multinomial Logistic model's superiority for sentiment analysis tasks in this context. However, we emphasize the need for further research in refining these models and adapting other pre-trained models for different languages and cultural nuances.

Acknowledgements: This work was partially supported by Portuguese funds through the CMAT - Research Centre of Mathematics of University of Minho - within references UIDB/00013/2020, UIDP/00013/2020.

Modelling uncertainty regarding the location of Fire Stations: a case study applied to Porto region

Tiago Ribeiro [a], Maria Isabel Gomes [b], Regina Bispo [b]
tmb.ribeiro@campus.fct.unl.pt, mirg@fct.unl.pt, r.bispo@fct.unl.pt

[a] *Mestrado em Matemática e Aplicações, Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa, Portugal*
[b] *Centro de Matemática e Aplicações (NOVA Math) e Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade de NOVA de Lisboa, Portugal*

Keywords: fire stations, kernel density estimation, location models, urban fires

Abstract: In light of the increasingly higher number of urban and forest fires in Portugal as well as that of the human and material costs associated, the awareness and interest over this topic has been growing in the scientific community.
In the process of fighting these phenomena, there are several circumstances and factors which hinder the firefighters' ability to intervene. This means that there is a need to analyze and understand whether the actual network of fire stations is suitable, or not, given the past records of fires. For that, a case study focusing on Porto region was performed to understand the suitability of the current fire stations' locations concerning the urban fire fighting response.
The available data included records of urban fires between 2012 and 2020. Given this data, we first estimated the process intensity using kernel density estimation. These results were used as uncertainty parameters in the application of location optimization models.
The results indicated that the current distribution of fire stations may not be optimal for effective urban firefighting. This suggests that improvements can be made to the locations of the existing network. The results of this study provide valuable information for planning and locating new fire stations and improving firefighting capabilities.

Acknowledgements: This work was funded by national funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the project DSAIPA/ DS/ 0088/2019 and research and development units UNIDEMI (project UIDB/ 0067/2020) and NOVAMATH (projects UIDB/ 00297/ 2020 and UIDP/ 00297/2020).

Modelação espacial da probabilidade de persistência de cadáveres de aves em estudos de monitorização da mortalidade em linhas elétricas

Ema Biscaia [a], Joana Bernardino [b], Regina Bispo [c]
e.biscaia@campus.fct.unl.pt, jbernardino@cibio.up.pt, r.bispo@fct.unl.pt

[a] *Mestrado em Matemática e Aplicações, Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa, Portugal*
[b] *CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Portugal e BIOPOLIS Program in Genomics, Biodiversity and Land Planning, Portugal*
[c] *Centro de Matemática e Aplicações (NOVA Math) e Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade de NOVA de Lisboa, Portugal*

Keywords: dados espaciais, INLA, probabilidade de persistência

Abstract: Na monitorização do impacte ambiental causado pelas linhas elétricas sabe-se que a mortalidade de aves observada difere da mortalidade real, uma vez que os cadáveres podem ser removidos por predadores. Os planos de monitorização das linhas contemplam, por isso, a realização de ensaios experimentais, designados por *testes de remoção*, que permitem obter dados sobre o tempo de persistência do cadáver no solo até à sua remoção, podendo estes ser dados censurados. A realização destes ensaios a cada nova linha elétrica implica, no entanto, avultados custos financeiros e logísticos. Em 2022, Bernardino *et al.* [1] compilaram informação sobre 30 ensaios, realizados em linhas de muito alta tensão em Portugal continental, dando origem a dados geo-referenciados. Assim, neste estudo pretende-se construir um modelo que permita estimar, para qualquer ponto geográfico do território nacional, a probabilidade de persistência de um cadáver obviando assim a realização de testes a cada novo projeto. Para responder a este objetivo, foi usado o método INLA (*Integrated Nested Laplace Approximation*) conjugado com a abordagem SPDE (*Stochastic Partial Differential Equations*), o que permitiu modelar a probabilidade de persistência de um cadáver considerando quer efeitos fixos (tamanho do cadáver e época do ano) quer efeitos aleatórios (localização geográfica e projeto). Os resultados permitiram analisar a variação desta probabilidade em todo o território nacional e criar uma ferramenta prática que estima a probabilidade de persistência, numa dada localização, como função das covariáveis consideradas.

References

[1] Bernardino, J., Martins, R. C., Bispo, R., Marques, A. T., Mascarenhas, M., Silva, R., Moreira, F. Ecological and methodological drivers of persistence and detection of bird fatalities at power lines: Insights from multi-project monitoring data. *Environmental Impact Assessment Review*, 93, 106707, 2022. doi:10.1016/j.eiar.2021.106707

Using a constructed covariate that accounts for preferential sampling

Andreia Monteiro [a], Isabel Natário [b], Maria Lucília Carvalho [c], Ivone Figueiredo [c,d], Paula Simões [e]
andreiaforte50@gmail.com, icn@fct.unl.pt, mlucilia.carvalho@gmail.com, ifigueiredo@ipma.pt, pc.simoes@campus.fct.unl.pt

[a] *CIDMA - Center for Research and Development in Mathematics and Applications, Portugal.*
[b] *Department of Mathematics, NOVA School of Science and Technology, NOVA MATH, Portugal.*
[c] *Center of Statistics and its Applications (CEAUL), Faculty of Sciences of the University of Lisbon, Portugal.*
[d] *Portuguese Institute for Sea and Atmosphere (IPMA).*
[e] *Military Academy Research Center - Military University Institute (CINAMIL)*

Keywords: constructed covariates, nearest neighbour distances, preferential sampling

Abstract: In Geostatistics, the common assumption is that the selection of the sampling locations does not depend on the values of the spatial variable of interest. However, dependence can be observed for example in fishery data, where catches are certainly associated with the locations where the fisheries take place, in order to maximize capture effort. Thus, the abundance process under study determines the data-locations and the above mentioned assumption is violated. This phenomenon is coined as preferential sampling and ignoring the preferential nature of the sampling can lead to biased estimates and misleading inferences. We plan to investigate the use of constructed covariates, based on an average value of the distances of nearest neighbors observations, that are able to mitigate preferential sampling. The inclusion of this covariate in the geostatistical model is able to explain the stochastic dependence of sampling locations on the spatial variable. If this dependence is no longer detected after this step, then we can use standard statistical techniques for inference without problems. This approach is assessed in a simulation study and we also discuss issues specific to this approach that arise when several study configurations are accounted in the simulations. The methodology is illustrated using two real data sets, one provided by the Instituto Português do Mar e da Atmosfera and the second concerns biomonitoring of lead pollution in Galicia.

References

[1] Diggle, P., Menezes, R., Su, T. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2), 191–232, 2010. doi:10.1111/j.1467-9876.2009.00701.x

[2] Illian, J. B., Sørbye, S. H., Rue, H. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *The Annals of Applied Statistics*, 1499–1530, 2012. doi:10.1214/11-AOAS530

Geostatistical models for identifying juvenile fish hotspots in marine conservation

Francisco Gonçalves[a], Raquel Menezes[a], Daniela Silva[a], Inês Dias[b], Alexandra A. Silva[bc]
pg46708@alunos.uminho.pt, rmenezes@math.uminho.pt, danyelasylva2@gmail.com, inesc.dias@ipma.pt, asilva@ipma.pt

[a] *Centre of Mathematics (CMAT), Minho University, Guimarães*
[b] *Division of Modelling and Management of Fishery Resources, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisboa*
[c] *MARE - Marine and Environmental Sciences Centre / ARNET - Aquatic Research Network, Faculty of Sciences, University of Lisbon*

Keywords: Bayesian framework, geostatistics, recruitment areas, species distribution models

Abstract: Species distribution models (SDMs) play a crucial role in the management and conservation of commercially important marine species. The growing interest in SDMs stems from the necessity to ensure fisheries' sustainability, supported by geostatistical models that address the specificities inherent to this type of data. This work focuses on investigating multivariate geostatistical models that associate species occurrence or abundance observations with environmental covariates in a limited number of locations, enabling the prediction of species presence and extent in unobserved areas. The primary objectives of this study are to identify hotspots of juvenile richness and map recruitment areas and seasons. Our analysis focuses on the landing per unit of effort (LPUE) of small (size T4) sardine (*Sardina pilchardus*) along the northern Portuguese coast, during a period with fewer administrative fishing restrictions (2007-2011). Adopting a Bayesian-INLA framework, we account for the complexity associated with hierarchical geostatistical models capable of handling temporally collected data. The Integrated Nested Laplace Approximation (INLA) approach is employed to construct multiple models, incorporating a spatial field generated through the Stochastic Partial Differential Equation (SPDE) methods. Regarding model evaluation and comparison, the DIC (Deviance Information Criterion) and CPO (Conditional Predictive Ordinate) metrics based on goodness of fit and complexity are utilized to select the most influential environmental covariates. The outcomes of this study, enhancing our understanding of juvenile sardine distributions and accurately identifying hotspots, will hopefully contribute to the sustainability of marine ecosystems and the preservation of commercially important species.

Acknowledgements: The authors acknowledge the FCT Foundation for funding their research through projects PTDC/MAT- STA/28243/2017, UIDB/00013/2020 and UIDP/00013/2020. They also express their appreciation to MAR2020 for funding the SARDINHA2020 project (MAR-01.04.02-FEAMP-0009) and to all colleagues involved in this study.

Density Surface Model vs. Spatial Models with INLA for animal abundance estimation

Iúri J. F. Correia [a], Tiago A. Marques [a, b, c], Christine Cuyler [d], Soraia Pereira [a]
ijcorreia@fc.ul.pt

[a] *Centro de Estatística e Aplicações – Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa (FCUL), Portugal*
[b] *Centre for Research into Ecological and Environmental Modelling, University of St. Andrews, Scotland*
[c] *Departamento de Biologia Animal, FCUL, Portugal*
[d] *Greenland Institute of Natural Resources, P.O. Box 570, Nuuk, Greenland*

Keywords: abundance estimation, caribou, distance sampling, DSM, inlabru

Abstract: *Rangifer tarandus groenlandicus* is a caribou species native to the Greenland West coast. Its importance to the human population reaches not only cultural traditions and subsistence but harvesting as well. Thus, caribou long-term monitoring is essential to adjust management strategies. This is controlled by Greenland Institute of Natural Resources helicopter-based surveys. Here, two approaches to estimate their spatial density were compared, based on data collected on said surveys. Using Distance Sampling methods, a detection function was fitted to the data to estimate caribou detection probability in the region. Then, a Density Surface Model (DSM) was fitted describing caribou abundance as a function of additional environmental covariates. In parallel, since the caribou data are point-referenced, it is also intuitive to work under the spatial point processes' framework. Thus, a spatial model was adjusted via the `inlabru` package, providing another detailed modelling perspective. `inlabru` operates with Integrated Nested Laplace Approximation methods for complex spatial model fitting to survey data where the detection probability is unknown. Finally, the DSM results were compared with the latter model.

Acknowledgements: Bolsa FCT UI/BD/152236/2021 and CEAUL's strategic project.

References

[1] Cuyler, C., Rosing, M., Molgaard, H., Heinrich, R., Raundrup, K. Status of two West Greenland caribou populations 2010; 1) Kangerlussuaq-Sisimiut, 2) Akia-Maniitsoq. Technical Report 78, GINR, 2011. ISBN:87-91214-60-2

[2] Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., Thomas, L. *Introduction to Distance Sampling - Estimating abundance of biological populations*. Oxford University Press, 2001. ISBN:9780198509271

[3] Miller, D. L., Burt, M. L., Rexstad, E. A., Thomas, L. Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution*, 4(11):1001-1010, 2013. doi:10.1111

[4] Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B. `inlabru`: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10: 760–766, 2019. https://doi.org/10.1111/2041-210X.13168

Space-time autoregressive models for time series of counts

Ana Martins [a], Manuel G. Scotto [b], Christian H. Weiß [c], Sónia Gouveia [a,d]
a.r.martins@ua.pt, manuel.scotto@tecnico.ulisboa.pt, weissc@hsu-hh.de, sonia.gouveia@ua.pt

[a] *Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal*
[b] *Center for Computational and Stochastic Mathematics (CEMAT), Department of Mathematics, IST, University of Lisbon, Lisbon, Portugal.*
[c] *Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany.*
[d] *Intelligent Systems Associate Laboratory (LASI), Portugal.*

Keywords: binomial thinning, space-time, STINAR models, time series of counts

Abstract: The theoretical development of statistical models aiming the analysis of time series of counts is relatively new when compared to those designed for real-valued data. For integer-valued data, the class of Integer Autoregressive and Moving Average (INARMA) models plays a central role. This class was inspired by the continuous ARMA counterpart, where the discrete nature of the process is ensured by replacing the multiplication with the binomial thinning operator (BTO, hereafter), and by assuming a innovation process which follows a discrete distribution. Given the recent developments, it is not surprising that the knowledge regarding multivariate models for count data is still underdeveloped. Hence, this work introduces the novel class of models addressed as Space-Time Integer-valued ARMA (STINARMA), driven by the continuous STARMA class endowing a matrix BTO and a multivariate discrete innovation process with independent marginals. After the preliminary results on the STINMA class, this works focuses specifically on the purely autoregressive component, i.e., the STINAR class. The theoretical properties of the STINAR process are derived, namely first- and second-order moments, and estimation approaches are proposed. The finite-sample performance of the estimation is evaluated through a simulation study. Finally, the applicability of the STINAR model is illustrated with real data concerning the number of daily hospital admissions nearby the Aveiro region in Portugal.

References

[1] Pfeifer, P. E., Deutsch, S. J. A Three-Stage Iterative Procedure for Space-Time Modeling. *Technometrics*, 22, 35-47, 1980.

[2] Franke, J., Subba Rao, T. Multivariate First-Order Integer-Valued Autoregressions. Technical Report, University of Kaiserslaute, 1993.

[3] Martins, A., Scotto, M. G., Weiß, C. H., Gouveia, S. Space-time integer-valued ARMA modelling for time series of counts. *Submitted*.

Comparing the effect of concurrent promotions over demand with interpretable deep learning

Micael Gomes [a,c], Alexandra Oliveira [a,b,c], Luís Paulo Reis [b,c]
up201709390@up.pt, alexandra.oliveira@retail-consult.com, lpreis@fe.up.pt

[a] *Retail Consult, Porto, Portugal*
[b] *Artificial Intelligence and Computer Science Laboratory (LIACC), University of Porto, Porto, Portugal*
[c] *FEUP, Faculty of Engineering from Porto University*

Keywords: deep learning, forecasting, interpretability, promotions, time-series

Abstract: Accurately predicting customer demand is challenging but crucial for businesses and the environment. Numerous factors can impact demand, and promotions are of particular interest as they are known to alter customer behavior. Often, multiple promotional events occur at the same time, making it difficult to assess the individual impact of each one but of critical importance for decision-making.
Current forecasting methods face limitations in this context, especially in interpreting the effects of different promotional events on customer behavior.
Therefore, this paper applies and compares interpretable deep-learning models to a wide range of real-world time series (12,830) representing daily product sales with promotional information. The focus is on improving accuracy and interpretability, identifying the most critical factors in predicting the effects of concurrent promotions.
The employed models comprise SARIMAX, Short-Term Long Memory models, XGBoost, and Temporal Fusion Transformer (TFT), each of which is implemented with a sliding window method, operating on a 14-day window to generate one-step-ahead forecasts. The data processing and analysis were done in Python, and the models' evaluation was done through the use of MAE, MASE, and RMSE.
The TFT model outperforms others across all metrics- with a mean value of 1.83, 2.60, and 0.71 for MAE, RMSE, and MASE, respectively. An interpretability analysis reveals the factors that influenced the target variable. This analysis was performed at a global, category, and item level. Item-level research demonstrates the impact of individual promotions. The findings inform more effective and targeted promotional strategies, suggesting that promotions play a significant role in sales forecasting.

Acknowledgements: Thanks are extended to Retail Consult for the data provided, enabling the realization and success of this study. Acknowledgment is also due to the Faculty of Engineering at Porto University for their vision in supporting this emergent research area. Special thanks to Dr. Alexandra Oliveira and Professor Luís Paulo Reis, whose contributions significantly enriched the project. Their expertise and guidance were fundamental in achieving the objectives of this study.

Previsão horária da descarga de um aproveitamento a fio de água

Joana Seabra-Silva[a,b], **Paula Milheiro-Oliveira**[c,d], Paulo Avilez-Valente[c,e]
joanassilva0598@gmail.com, poliv@fe.up.pt, pvalente@fe.up.pt

[a] *Sonae SGPS, Maia, Portugal*
[b] *Faculdade de Ciências, Universidade do Porto, Portugal*
[c] *Faculdade de Engenharia, Universidade do Porto, Portugal*
[d] *Centro de Matemática da Universidade do Porto, Portugal*
[e] *Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Portugal*

Keywords: múltiplos períodos, recursos hídricos, séries temporais

Abstract: A gestão dos recursos hídricos desempenha um papel cada vez mais importante na nossa sociedade. A consciência do carácter finito dos recursos naturais tem levado a um interesse crescente pelos estudos e pela investigação neste domínio. A construção de barragens constitui uma fonte de energia hidroelétrica, produzindo energia renovável e não poluente. O planeamento estratégico e eficiente das albufeiras permite a redução das perdas de água e da poluição, ajuda a prevenir inundações e secas e proporciona um acesso adequado das populações à água. Este trabalho tem como objetivo demonstrar como se pode modelar e prever a descarga horária numa barragem, considerando a barragem de Crestuma-Lever, no Douro, como caso de estudo. As tarefas de modelação e previsão exploram um conjunto de modelos alternativos de séries temporais, focando-se nos que permitem lidar com múltiplos períodos, como o modelo ARIMAX com termos de Fourier, mas também nos modelos VAR e ARDL, entre outros. É analisada a sua capacidade de previsão neste caso em concreto. Os resultados mostram que o modelo VAR é o mais adequado para a previsão com um horizonte de 48 h. As previsões baseadas nos modelos SARIMAX e ARIMAX com termo de Fourier apresentaram um bom desempenho para previsões a longo prazo, como a previsão de um ano hidrológico completo. O modelo ARDL aparenta ser o que melhor captura grandes flutuações do caudal, mas produz previsões com atraso e prevê erradamente alguns caudais de ponta. Os modelos de mudança de regime de Markov revelaram-se vantajosos, requerendo menos informação para uma precisão semelhante, reduzindo significativamente o tempo de computação.

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, e pelo projeto EsCo-Ensembles (PTDC/ECI-EGC/30877/2017), cofinanciado pelo NORTE 2020, Portugal 2020 e União Europeia através do FEDER, e pela FCT através de fundos nacionais. O primeiro autor realizou grande parte da investigação enquanto estudante da Faculdade de Ciências da Universidade do Porto.

Classifying distributional data into more than two groups

Ana Santos [a], **Sónia Dias** [b,d], Paula Brito [c,d], Paula Amaral [d]
up202103086@edu.fc.up.pt, sdias@estg.ipvc.pt, mpbrito@fep.up.pt, paca@fct.unl.pt

[a] *Faculdade de Ciências, Universidade do Porto, Portugal*
[b] *ESTG - Instituto Politécnico de Viana do Castelo, Portugal*
[c] *Faculdade de Economia, Universidade do Porto, Portugal*
[d] *LIAAD-INESC TEC, Portugal*
[e] *Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa & CMA, Portugal*

Keywords: classification, histogram data, linear discriminant function, Mallows distance, Symbolic Data Analysis

Abstract: This work addresses multiclass classification of distributional data (see [1]). The proposed method relies on linear discriminant functions to variables whose observations are empirical distributions, represented by the corresponding quantile functions [2]. A discriminant function allows obtaining a score (quantile function) for each unit. For the classification of an unit in one of two groups, the Mallows distance between its score and the score obtained for the barycentric histogram of each class is computed. The observation is then assigned to the closest class. When considering more than two *a priori* classes, two approaches are considered. The first one consists in dividing the multiclass classification problem into several binary classification subproblems. In this case, two well-known multiclass classification techniques may be applied: One-Versus-One (OVO) and One-Versus-All (OVA). The alternative approach, named Consecutive Linear Discriminant Functions, consecutively defines several linear discriminant functions, under the condition that each new discriminant function must be uncorrelated with all previous ones. This leads to several score histogram-valued variables with null symbolic linear correlation coefficient. Classification is then based on a suitable combination of the corresponding obtained scores, using the Mallows distance.

This method is applied to discrimination of network data, described by the distributions, across the network nodes, of four centrality measures. The goal is to identify the network model used to generate the networks.

References

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

[2] Dias, S., Brito, P., Amaral, P. Discriminant analysis of distributional data via fractional programming *EJOR*, 294(1), 206–218, 2021. https://doi.org/10.1016/j.ejor.2021.01.025.

Linear regression for symbolic density-valued data

Rui Nunes [a,d], Paula Brito [b,d], Sónia Dias [c,d]
up201400313@up.pt, mpbrito@fep.up.pt, sdias@estg.ipvc.pt

[a] *Faculdade de Ciências da Universidade do Porto*
[b] *Faculdade de Economia da Universidade do Porto*
[c] *Instituto Politécnico de Viana do Castelo*
[d] *LIAAD-INESC TEC, Portugal*

Keywords: functional data analysis, linear regression, symbolic data

Abstract: The increasing complexity and size of data has led to the development of techniques for analyzing group behavior. Symbolic Data Analysis (SDA)[1] addresses data aggregation challenges while considering the inherent variability of the aggregated data. This study introduces a density-valued variable representation using Functional Data Analysis (FDA)[3] approaches. Since we are working with quantile functions to represent the observations of the variables, we can, on the one hand, apply FDA techniques such as the Concurrent Model for linear regression. On the other hand, we can extend the Symmetric Distribution Model (DSD)[2] to continuous variables. The objective is to minimize the Sum of Squared Errors (SSE) between observed and predicted quantile functions, with regression coefficients estimated under specific constraints. The model is applied to a dataset of 31 European countries' GDP from 1995 to 2022, analyzing the "Import Goods" component in relation to others. The initial findings show a good fit of the model to the data. Penalization has been applied to the model and validation was performed using cross-validation techniques. Future research includes investigating a cluster-wise method for density-valued data. This approach provides a valuable framework for analyzing complex, large-scale data with inherent variability and contributes to the field of Symbolic Data Analysis in understanding group behavior.

References

[1] Diday, E. The symbolic approach in clustering and related methods of data analysis. *Proceedings of IFCS, Classification and Related Methods of Data Analysis(1987)*, 1988.

[2] Dias, S., Brito, P. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(2):75–113, 2015. doi:10.1002/sam.11260

[3] Kokoszka, P., Reimherr, M. *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, 2017. doi:10.1201/9781315117416

Rules for predicting lab-grown diamonds prices: a comparative analysis

Margarida G. M. S. Cardoso [a,b], Luís Chambel [b]
margarida.cardoso@iscte-iul.pt, luischambel@sinese.pt

[a] *ISCTE-IUL, BRU-IUL*
[b] *Sínese*

Keywords: k-nearest neighbors, lab-grown diamonds, price, propositional rules

Abstract: Lab-grown gem-quality diamonds have shown fast market-share growth since the mid-2010s. Several attempts have been made to predict natural diamond prices based on their characteristics, namely, using Machine Learning techniques. There are no similar studies referring to lab-grown diamonds. In the present study, we aim at predicting lab-grown diamond unit prices. Data used were collected from *https://www.1215diamonds.com* and *https://www.miadonna.com* , on 2022, including 44 443 and 18 283 observations (lab-grown diamonds), respectively. These data include the diamond prices and also attributes to be used as predictors: Carat (measured in ct, 0.2 g), Color (7 levels, where the first is the colorless diamond); Clarity (8 levels where the first has no visible inclusions or internal flaws); Cut, including 10 different shapes, the predominant being the "Round". For the task at hand – predicting lab-grown diamonds prices based on physical characteristics- we propose using Propositional Rules (RULES). K-Nearest Neighbors (KNN) will be used as a baseline. We resort to the R packages "FNN" (for KNN) and "Cubist" (for RULES). Metrics estimated on the test set (30% of the data) including R-Squared, MAE-Mean Absolute Error, and MAPE- Mean Absolute Percentage Error, enable evaluation of the predictive capacity of the proposed approach. We conclude that RULES generally produce better predictions than KNN, and also provide easy-to-interpret outputs and useful insights regarding specific observations. Future analysis may include additional predictors such as Cut quality and Certificate.

Acknowledgements: This work was supported by FCT,Grant UIDB/50021/2020.

Analysis of hospitalization patterns using custom dissimilarities

Daniel Cordeiro [a], Ana Azevedo [b,c,d], Bárbara Peleteiro [b,c,d], Lucybell Moreira [b], Elsa Guimarães [b], Raquel Cadilhe [b], Rita Gaio [a,e]
up201506370@edu.fc.up.pt, a.oliveira@chsj.min-saude.pt, barbara.peleteiro@chsj.min-saude.pt, lucybell.moreira@chsj.min-saude.pt, elsa.guimaraes@chsj.min-saude.pt, raquel.cadilhe@chsj.min-saude.pt, argaio@fc.up.pt

[a] *Departamento Matemática, Faculdade de Ciências, Universidade do Porto*
[b] *Centro de Epidemiologia Hospitalar, Centro Hospitalar Universitário de São João*
[c] *Instituto de Saúde Pública da Universidade do Porto*
[d] *Departamento de Ciências da Saúde Pública e Forenses, e Educação Médica Faculdade de Medicina, Universidade do Porto*
[e] *Centro de Matemática da Universidade do Porto*

Keywords: clustering, custom dissimilarity, hospitalization trajectories, partition around medoids

Abstract: This work aims to present an innovative approach to analyzing pediatric hospitalization trajectories at Centro Hospitalar Universitário de São João (CHUSJ). Focusing on data from 1710 hospitalizations of children aged between 3 months and 13 years old, from December 2021 to November 2022, we devise a unique perspective on these hospitalization pathways.

Each hospitalization trajectory is represented by a sequence of, at most, seven key pediatric departments and the associated lengths of stay in each; we see each sequence of departments as a 'word'. These 'words' are then assessed based on a custom dissimilarity measure, capturing discrepancies between the department sequences and lengths of stay.

The novelty of this methodology lies in combining various word and time-based distances and in integrating penalization based on disparities in total length-of-stay. This results in a nuanced, comprehensive understanding of hospitalization patterns, exposing the complex dynamics of patient care. Based on the dissimilarities, we then apply the partition around medoids clustering algorithm. The silhouette plot provided 8 homogeneous groups.

Our approach will provide meaningful and actionable insights for hospital administrators, healthcare providers, and policy makers, enabling them to strategize and improve healthcare delivery. Consequently, this innovative approach is anticipated to illuminate the evolving landscape of pediatric patient care at CHUSJ and inform the refinement of future healthcare strategies, ultimately facilitating more personalized and effective patient care.

Acknowledgements: Rita Gaio was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the projects with reference UIDB/00144/2020 and UIDP/00144/2020.

Longitudinal antibody kinetics in kidney transplant recipients who recovered from severe acute respiratory syndrome coronavirus 2 infection

Maria J. Polidoro [a,b], Ana Pinho [c], Manuela Bustorf [c], Natércia Durão [d], Rui Martins [b,e]
mjp@estg.ipp.pt, ana.pinho@chsj.min-saude.pt,
maria.guerra@chsj.min-saude.pt, natercia@upt.pt, rmmartins@fc.ul.pt

[a] *Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Porto (ESTG-IPP), Felgueiras, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *Departamento de Nefrologia, Centro Hospitalar Universitário de São João, Porto, Portugal*
[d] *Universidade Portucalense, REMIT – Research on Economics, Management and Information Technologies, Porto, Portugal*
[e] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: anti-N IgG , anti-S1 IgG, COVID-19, kidney transplant recipients, longitudinal antibody kinetics

Abstract: Kidney transplant recipients (KTRs) are at an elevated risk of death from coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Evaluating the immune response of KTRs who recover from SARS-CoV-2 infection and the factors that may influence it, namely the role of immunosuppression, is crucial to understanding the immune response's quality and durability to natural infection.
In this work, we report the longitudinal antibody kinetics using two SARS-CoV-2 antigens, namely the nucleocapsid and the S1 domain of spike protein, anti-N, and anti-S1 IgG ratio values, respectively. This is the largest known longitudinal study describing the variability of memory of the immunosuppressive response of KTRs in a state of primary infection in the absence of vaccination.

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. Maria J. Polidoro, Natércia Durão and Rui Martins thank Ana Pinho, Manuela Bustorf and her department who kindly provided the data used in this work.

References

[1] Søfteland J.M., Gisslén M., Liljeqvist J.Å., et al. Longevity of anti-spike and anti-nucleocapsid antibodies after COVID-19 in solid organ transplant recipients compared to immunocompetent controls. *Am J Transplant*, 22(4), 1245-1252, 2022. doi:10.1111/ajt.16909

Comparative study on the performance of different classification algorithms, combined with pre- and post-processing techniques to handle imbalanced data, in the diagnosis of adult patients with familial hypercholesterolemia

João Albuquerque [a], Ana Margarida Medeiros [b], Ana Catarina Alves [b], Mafalda Bourbon [b], Marília Antunes [a]
joaodavid.alb@gmail.com, ana.medeiros@insa.min-saude.pt, catarina.alves@insa.min-saude.pt, mafalda.bourbon@insa.min-saude.pt, marilia.antunes@ciencias.ulisboa.pt

[a] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal*
[b] *Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal*

Keywords: extreme gradient boosting, logistic regression, naive Bayes, random forest, synthetic minority oversampling technique

Abstract: Familial Hypercholesterolemia (FH) is an inherited disorder of cholesterol metabolism. Current criteria for FH diagnosis, like Simon Broome (SB) criteria, lead to high false positive rates. The aim of this work was to explore alternative classification procedures for FH diagnosis, based on different biological and biochemical indicators. For this purpose, logistic regression (LR), naive Bayes classifier (NB), random forest (RF) and extreme gradient boosting (XGB) algorithms were combined with Synthetic Minority Oversampling Technique (SMOTE), or threshold adjustment by maximizing Youden index (YI), and compared.

Data was tested through a 10×10 repeated k-fold cross validation design. The LR model presented an overall better performance, as assessed by the areas under the receiver operating characteristics (AUROC) and precision-recall (AUPRC) curves, and several operating characteristics (OC), regardless of the strategy to cope with class imbalance. When adopting either data processing technique, significantly higher accuracy (Acc), G-mean and F1 score values were found for all classification algorithms, compared to SB criteria ($p < 0.01$), revealing a more balanced predictive ability for both classes, and higher effectiveness in classifying FH patients. Adjustment of the cut-off values through pre or post-processing methods revealed a considerable gain in sensitivity (Sens) values ($p < 0.01$). Although the performance of pre- and post-processing strategies was similar, SMOTE does not cause model's parameters to loose interpretability. These results suggest a LR model combined with SMOTE can be an optimal approach to be used as a widespread screening tool.

Acknowledgements: The current work was supported by the programme Norte2020 [Grant Number NORTE-08-5369-FSE-000018] and by Fundação para a Ciência e Tecnologia (FCT) [Grant Number UID/MAT/00006/2020].

Posters

—**Pósteres**—

Use of quantile regression methods in growth curves

Bianca Rafaelle da Silva [a], Thiago G. Ramires [a], Marcelo F. Silva [a]
biasil@alunos.utfpr.edu.br, thiagogentil@gmail.com,
marcelosilva@utfpr.edu.br

[a] *Federal University of Technology - Paraná*

Keywords: animal growth, GAMLSS, management

Abstract: Animal growth curves (weight/ages) are important tools for activity management and selection of animals with greater weight gain. Growth is directly related to the quantity and quality of meat, the final product of cattle breeding, and can be summarized as an increase in the size or weight of the animal. Such curves are generally analyzed using non-linear models, however, for growth curves that present cyclic behavior, the convergence of such parameters is not trivial. As an alternative, parametric quantile regression methods, pioneered by Cole [1] and Cole and Green [2], can be applied in these cases. Any distribution can be used to fit percentile curves using the generalized additive models for location, scale and shape (GAMLSS) [3]. In this survey, we propose the use of GAMLSS as an alternative for the adjustment of growth curves, using a dataset composed of 55 female Hereford breeds. Based on GAMLSS models, we observed that the average weight of animals grows from birth to 200 days, then decreases from 200 to 400 days at a slower rate, grows from 400 to 500 days, then the weight remains practically constant up to 615 days. This average behavior can be explained due, e.g., to excessive rainfall or food shortages.

Acknowledgements: This work was supported by the Federal University of Technology - Paraná. I also would like to thank the Research and Postgraduate Directorate (DIRPPG-AP).

References

[1] Cole, T.J. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 151, 385–406, 1988. doi:10.2307/2982992

[2] Cole, T.J., Green, P.J. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in medicine*, 11, 1305–1319, 1992. doi:10.1002/sim.4780111005

[3] Stasinopoulos, D.M., Rigby, R.A. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 1–46, 2008. doi:10.18637/jss.v023.i07

Constant effort harvesting with Allee effects in random environments using logistic type models: optimization and an application

Clara Carlos [a,b], Nuno M. Brites [c], Carlos A. Braumann [a,d]
clara.carlos@estbarreiro.ips.pt, nbrites@iseg.ulisboa.pt, braumann@uevora.pt

[a] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora*
[b] *Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal*
[c] *ISEG/UL - Universidade de Lisboa, Department of Mathematics; REM - Research in Economics and Mathematics, CEMAPRE*
[d] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora*

Keywords: constant effort harvesting, logistic-like model with Allee effects, optimization, stationary distribution, stochastic differential equations

Abstract: Conditions for existence of a stochastic equilibrium for general autonomous stochastic differential equation models of population growth with harvesting in a random environment can be seen in [1] or in our other SPE 2023 communication (an extension of [3] to harvested populations), according to whether Allee effects are absent or present. Using those results and results in [2], we present the particular case of constant harvesting effort and constant noise intensity for specific comparable models, namely the logistic (without Allee effects) and the logistic-like Allee effects models, including expressions for stationary densities and expected sustainable profits and yields. We assess the impact of Allee effects by comparing the two models and their optimal profits and yields for the Pacific halibut data in [4].

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

References

[1] Braumann, C.A. Variable effort harvesting models in random environments: generalization to density dependent noise intensities. *Mathematical Biosciences*, 177 & 178, 229–245, 2002. https://doi.org/10.1016/S0025-5564(01)00110-9

[2] Brites, N.M., Braumann, C.A. Profit optimization of stochastically fluctuating populations under harvesting: the effects of Allee effects. *Optimization*, 71:11, 3277–3293, 2022. https://doi.org/10.1080/02331934.2022.2031191

[3] Carlos, C., Braumann, C.A. General population growth models with Allee effects in a random environment. *Ecological Complexity*, 30, 26–33, 2017. http://dx.doi.org/10.1016/j.ecocom.2016.09.003

[4] Hanson, F.B., Ryan, D. Optimal harvesting with both population and price dynamics. *Mathematical Biosciences*, 148(2), 129–146, 1998. https://doi.org/10.1016/S0025-5564(97)10011-6

Semiparametric model applied to crime recidivism data

Felipe Antonio Magro [a], Thiago G. Ramires [a], Marcelo F. Silva [a]
felipe.am2003@hotmail.com, thiagogentil@gmail.com,
marcelosilva@utfpr.edu.br

[a] *Federal University of Technology - Paraná*

Keywords: cure rate models, GAMLSS, splines

Abstract: Crime recidivism is a recurrent problem in several countries, which can be influenced by several factors [1]. To identify factors associated with the theme, we propose a new semiparametric Weibull cure rate model, using the generalized additive models for location, scale and shape (GAMLSS) framework, that can fit all distribution parameters using parametric or nonparametric functions. Using a crime recidivism database from Brazil, we fitted the new model, which it was possible to conclude that only the jail time influences on the average time of the crime recurrence, where the greater jail time, the greater is the average of recidivism time. It was also possible conclude that the estimated probability of non-recurring in crime for parole freedom is greater than definitive freedom which is greater than semi-open freedom; the greater time in jail implies greater probability of recidivism in crime. Finally, the age presents a nonlinear effect in the probability of non-recurring in crime. Further, the probability of not committing a crime again increases up to 45 years, and then begins to decrease.

Acknowledgements: This work was supported by the Federal University of Technology - Paraná. I also would like to thank the Research and Postgraduate Directorate (DIRPPG-AP).

References

[1] Cottle, C.C., Lee, R.J., Heilbrun, K. The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal justice and behavior*, 28, 367–394, 2001. doi:10.1177/0093854801028003

Population growth and geometrically thinned EVT

Dinis D. Pestana[a,b], M. Fátima Brilhante[c,b], Sandra Mendonça[d,b], Pedro D. Pestana[e,f]
ddpestana@ciencias.ulisboa.pt, maria.fa.brilhante@uac.pt,
migomes@ciencias.ulisboa.pt, sandram@staff.uma.pt, pedro.pestana@uab.pt

[a] DEIO, *Faculdade de Ciências da Universidade de Lisboa* (FCUL)
[b] *Centro de Estatística e Aplicações,* UL (CEA/UL)
[c] DME, *Faculdade de Ciências e Tecnologia, Universidade dos Açores*
[d] DM-FCEE, *Universidade da Madeira*
[e] *DCT, Universidade Aberta*
[f] *Centro de Investigação em Ciência e Tecnologia das Artes*

Keywords: extreme value theory, fractional calculus, population growth

Abstract: Order statistics and products of powers of independent uniform *random variables* (RVs) originate families of RVs such as Beta and Logarithmic RVs, with natural parameters, which can be further generalized using fractional parameters (see [1], and references therein). Extensions of the Verhulst model lead to population growth models equilibrium such as the Gompertz model, tied to the Gumbel *extreme value* (EV) model, a max-stable model of high relevance in EV *theory* (EVT), and will be here under discussion. Details on statistical EVT can be found in [2], among other review papers. Aside from the classical *general* EV model, EV models in randomly stopped extreme schemes (see, [3]) are also discussed, in the lines of [4]. Observing that the logistic distribution is max-geo-stable and the Gompertz function is proportional to the Gumbel max-stable distribution, growth models, related to geometrically thinned EVT, are investigated.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, Portugal, project UIDB/00006/2020 (CEAUL) and by HiTEc Cost Action CA21163.

References

[1] Brilhante, M.F., Gomes, M.I., Mendonça, S., Pestana, D.D., Pestana, P.D. Generalized Beta Models and Population Growth, so many routes to chaos. *Fractal Fract.* 7:2, 194, 2023. doi:10.3390/fractalfract7020194

[2] Gomes, M.I., Guillou, A. Extreme value theory and statistics of univariate extremes: A review. *Internat. Statist. Review* 83:2, 263–292, 2015. doi:10.1111/insr.12058

[3] Rachev, S.T., Resnick, S. Max-geometric infinite divisibility and stability. *Communications in Statistics — Stochastic Models* 7, 191–218, 1991. doi:10.1080/15326349108807184

[4] Mendonça, S., Pestana, D.D., Gomes, M.I. Randomly Stopped k-th Order Statistics. In: Kitsos C, Oliveira T, Rigas A, Gulati S (eds.), *Theory and Practice of Risk Assessment.* Springer Proceedings in Mathematics & Statistics, vol 136, Springer, Cham, 249–266, 2015. doi:10.1007/978-3-319-18029-8_19

Improvement of COVID-19 symptoms: a survival analysis study from a Portuguese cohort

Leandro Duarte [a], Carla Moreira [a,b], Luís Meira-Machado [a], Ana Paula Amorim[a], Joana Costa[b], Paula Meireles[b]
pg45191@alunos.uminho.pt, d8434@math.uminho.pt, machado@math.uminho.pt

[a] *Centro de Matemática da Universidade do Minho, Universidade do Minho, 4800-058 Guimarães, Portugal*
[b] *EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas, n○ 135, 4050-600 Porto, Portugal*

Keywords: covid-19, estimation of survival, regression, survival analysis

Abstract: The COVID-19 pandemic has had a profound impact on the world, affecting millions of people and causing widespread illness and death. As the disease continues, it is critical to understand the patterns and predictors of the disease in order to get valuable information that can be used to develop strategies for preventing and managing COVID-19. Survival analysis techniques have been widely used in medical research to analyze longitudinal time-to-event data, such as time from diagnosis to recovery or death. These techniques provide valuable insights into the risk factors and the outcome of the disease. We used a registry of 3481 COVID-19 patients diagnosed at Centro Hospitalar Universitário de São João (CHUSJ) between March 01, 2020 and January 01, 2021. Symptoms of the disease were reported at admission, and its improvement was investigated using phone interviews. Descriptive statistics were performed according to the measurement level of the variable, and some nonparametric localization tests were used to compare groups. For the longitudinal analysis, the product-limit estimator of survival (Kaplan and Meier) was used to describe COVID-19-associated symptom duration. The estimated survival curves were used to compare the improvement of COVID-19 symptoms for categorical predictors, and formal hypothesis tests were used. Simple and multiple regression models were used to estimate the effect of potential predictors on the improvement of COVID-19 symptoms.

Acknowledgements: We would like to thank the Centro Hospitalar Universitário de São João (CHUSJ) for providing the information on which this study is based. Additionally, the authors would like to acknowledge the financial support received from FCT — Fundação Ciência e Tecnologia through the project EXPL/MAT-STA/0956/2021.

References

[1] Kaplan, E.L., Meier, P. Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457—481, 1958.

[2] Klein, J.P., Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

Perspectives of statistical inference on interval-valued data

Catarina Rodrigues [a], Conceição Amado [a]
catarinasaleirorodrigues@tecnico.ulisboa.pt,
conceicao.amado@tecnico.ulisboa.pt

[a] *CEMAT and IST-ULisboa*

Keywords: bootstrap, hypothesis testing, inference, interval-valued, symbolic data

Abstract: In classical statistics, we assume that we know the exact values of the associated quantities. However, in several situations, due to the preservation of confidentiality or availability of data, we have only intervals containing such values. The concept of symbolic data, introduced by [1], is a way to deal with this type of data. Despite the number of approaches introduced to dealing with interval data in symbolic data analysis, such as regression models, principal component analysis, or clustering, few have addressed inference concerns. Following the work of Grzegorzewski and Śpiewak [2], we review recent approaches to hypothesis tests on interval data for the mean from epistemic and ontic perspectives. In this spirit, we develop new hypothesis tests, taking into account the interval centers and also a random choice of a value in the interval. Additionally, we also propose to use bootstrap hypothesis testing for this type of symbolic data. A simulation experiment compares the effectiveness of several strategies.

References

[1] Diday, E. Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolique - Numérique.* CEREMADE, Université Paris, 21-56, 1987.

[2] Grzegorzewski, P., Śpiewak, M. The sign test and the signed rank test for interval-valued data., *International Journal of Intelligent Systems*, 34: 2122- 2150, 2019. doi.org/10.1002/int.22134

[3] Roy, A., Klein, D. Testing of mean interval for interval-valued data. *Communications in Statistics - Theory and Methods*, 49:20, 5028-5044, 2020. doi:10.1080/03610926.2019.1612915 2019.

Selection of models and thresholds in the Peaks-Over-Threshold (POT) methodology: Application to extreme precipitation values in Madeira and Porto Santo islands

Délia Gouveia-Reis [a,b], Luiz Guerreiro Lopes [a], Sandra Mendonça [a,b]
delia.reis@staff.uma.pt, lopes@uma.pt, sandram@staff.uma.pt

[a] *Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira*
[b] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: exponential distribution, generalized Pareto distribution, intense precipitation, POT methodology, threshold selection

Abstract: The study of extremes in the Peaks-Over-Threshold (POT) methodology requires the analysis of observations that exceed a specified threshold. The choice of this threshold, as well as the distribution to be used in the modeling of the resulting extreme values, is a topic of great practical importance. The choice of the threshold involves a balance between the size of the bias of the estimators and their variances, where high threshold values result in small bias and large variances and low values lead to the opposite effect. Such value can be chosen using graphical analysis, heuristic rules, and other methodologies [1, 2]. The conditional distribution function of the difference between values above the threshold and this value can be well approximated by a generalized Pareto distribution. The choice between the exponential distribution and the non-exponential generalized Pareto distribution can be made using large sample theory tests, such as likelihood ratio, goodness-of-fit, and others [2]. A bibliographic review of the scientific literature on these topics is carried out in the present work. To demonstrate the methodologies and tests examined, an application of these to daily precipitation values on the islands of Madeira and Porto Santo from 1961 to 2022 is carried out in this study.

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. The authors are grateful to the Portuguese Institute for Sea and Atmosphere (IPMA) for the data provided for this study.

References

[1] Dey, D.K., Yan, J., eds. *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press, Boca Raton, FL, 2016. doi:10.1201/b19721

[2] Gomes, M.I., Guillou, A. *Extreme value theory and statistics of univariate extremes: A review. International Statistical Review*, 83, 2, 263–292, 2015. doi:10.1111/insr.12058

Os anos de vida saudável perdidos nas doenças não comunicáveis na União Europeia

Margarida Torres [a], Alcina Nunes [b], João Paulo Martins [a,c]
10130257@ess.ipp.pt, alcina@ipb.pt, jom@ess.ipp.pt

[a] *Escola Superior de Saúde, Instituto Politécnico do Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal*
[b] *UNIAG, Instituto Politécnico de Bragança, Portugal*
[c] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: anos de vida saudável perdidos, dados em painel, despesas em saúde, doenças não comunicáveis

Abstract: As doenças não comunicáveis (DNC) apresentam-se como um problema de saúde mundial [1]. Para medir o seu impacto nos sistemas de saúde desenvolveu-se a métrica anos de vida saudável perdidos (DALY). Esta métrica, usada no *Global Burden of Disease* (GBD), permite perceber a carga de uma doença, de forma a definir políticas para os cuidados de saúde [2, 3]. Este estudo analisou a evolução dos DALY para DNC e sua relação com a evolução das despesas de saúde públicas e privadas na União Europeia, entre 2000 e 2019. Observou-se uma diminuição dos DALY para as doenças cardiovasculares e neoplasias que constituem as DNC com DALY mais elevados. Através do ajustamento de modelos parcimoniosos para dados em painel, verificou-se que o aumento das despesas em saúde teve um impacto significativo na diminuição dos DALY das DNC, na generalidade das DNC analisadas. As doenças musculoesqueléticas e mentais, apresentaram uma evolução desfavorável no período em análise em que o aumento da despesa não se tem traduzido numa diminuição dos DALY.

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] GBD 2019 Viewpoint Collaborators. Five insights from the Global Burden of Disease Study 2019. *Lancet*, 396, 1135–1159, 2020. doi:10.1016/S0140-6736(20)31404-5

[2] Dodhia, H., Phillips, K. Measuring burden of disease in two inner London boroughs using disability adjusted life years. *J Public Health (Bangkok)*, 30, 313–321, 2008. doi:10.1093/pubmed/fdn015

[3] Lyttkens, C.H. Time to disable DALYs? On the use of disability-adjusted life years in health policy. *European Journal of Health Economics*, 4, 195–202, 2003. doi:10.1007/s10198-003-0169-2

Desempenho de metodologias de classificação sexual baseadas em ortopantomografias

João Alves[a], Cristiana Palmela Pereira[b,c,d], **Rui Santos**[d,e]
email.do.alves@gmail.com, cpereira@campus.ul.pt, rui.santos@ipleiria.pt

[a] *Estudante do Mestrado em Ciência de Dados da ESTG–Politécnico de Leiria*
[b] *Faculdade de Medicina Dentária da Universidade de Lisboa*
[c] *Faculdade de Medicina da Universidade de Lisboa*
[d] *CEAUL – Centro de Estatística e Aplicações, Universidade de Lisboa*
[e] *Escola Superior de Tecnologia e Gestão, Politécnico de Leiria*

Keywords: análise discriminante, classificação sexual, ortopantomografias, redes neuronais convolucionais, regressão logística

Abstract: As estruturas ósseas craniomandibulares, por serem mais resistentes aos processos de tafonomia, são relevantes na diagnose sexual de esqueletos adultos. Este passo é primordial na vertente reconstrutiva de um cadáver não identificado. Assim, com base numa amostra obtida por estudantes da Faculdade de Medicina Dentária da Universidade de Lisboa através de um conjunto de medições efetuadas em ortopantomografias (radiografias panorâmicas), neste trabalho é avaliado o desempenho de diferentes metodologias de classificação do sexo. Algumas das metodologias avaliadas são baseadas nas medições realizadas, como a regressão logística, a análise discriminante, os k-vizinhos mais próximos, entre outras. É igualmente avaliada a aplicação de redes neuronais pré-treinadas, como a VGG16, a RESNET-50 e a INCEPTION V-3, que concretizam a classificação diretamente das ortopantomografias. A amostra utilizada foi aleatoriamente dividida em 80 por cento para a estimação dos parâmetros de cada metodologia (treino) e as restantes 20 por cento para avaliação do desempenho (teste). A comparação do desempenho foi baseada na matriz de confusão e medidas associadas (acurácia, sensibilidade, especificidade, valores preditivos e F-score) e na área sob a curva ROC.

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] Ionel, V. *Diagnose sexual baseada em parâmetros radiológicos craniomandibulares*. Mestrado em Medicina Dentária da Universidade de Lisboa, 2021.

[2] Sairam, V., Geethamalika, M.V., Kumar, P.B., Naresh, G., Raju, G.P. Determination of sexual dimorphism in humans by measurements of mandible on digital panoramic radiograph. *Contemp Clin Dent*, 7(4), 434–439, 2016. doi:10.4103/0976-237X.194110

[3] Santos, R., Martins, J.P., Felgueiras, M., Ferreira, L. Accuracy Measures for Binary Classification Based on a Quantitative Variable. *REVSTAT — Statistical Journal*, 17(2), 223–244, 2019. doi:10.57805/revstat.v17i2.266

[4] Vila-Blanco, N., Varas-Quintana, P., Aneiros-Ardao, Á., Tomás, I., Carreira, M.J. Automated description of the mandible shape by deep learning. *Int J Comput Assist Radiol Surg*, 16(12), 2215–2224, 2021. doi:10.1007/s11548-021-02474-2

Os desafios do Jornalismo de Dados

Cláudia Silvestre [a,b], Helena Pina [a], Susana Araújo [a]
csilvestre@escs.ipl.pt, hpina@escs.ipl.pt, saraujo@escs.ipl.pt
[a] *Escola Superior de Comunicação Social*
[b] *CEAUL*

Keywords: bases de dados, estatística, jornalismo, visualização

Abstract: Numa sociedade orientada por dados, é cada vez mais premente saber extrair conhecimento a partir de bases de dados de grande dimensão, em particular de informação numérica. O jornalismo, profissão em constante mudança, tem-se adaptado a esta nova realidade. Atualmente nas redações, de grande dimensão, podemos encontrar profissionais na área da infografia e do jornalismo de dados.
O jornalista com a sua capacidade para escrutinar e movido por uma curiosidade inesgotável, tem requisitos fundamentais para dar vida aos números. Mas, para tal precisa ter conhecimentos estatísticos, o que não é muito frequente em profissionais desta área. Esta é uma necessidade premente, dado que a complexidade das bases de dados disponíveis exige competências técnicas, na área da Estatística, para que os dados possam ser examinados com rigor. Só assim o jornalista poderá usar a potencialidade dos dados para compreender melhor as questões sociais, políticas e económicas, bem como comunicar essa informação através duma narrativa clara e rigorosa.
Embora a Estatística esteja presente nos cursos de Comunicação, geralmente os jornalistas de dados consideram-se autodidatas. O que evidencia a necessidade do ensino da Estatística mais direcionado para esta área. Neste trabalho, pretendemos apresentar como a Estatística pode ser usada ao serviço do Jornalismo, apresentando alguns exemplos do trabalho desenvolvido na Escola Superior de Comunicação Social. E o que pretendemos fazer com o objetivo de capacitar os estudantes a investigar, analisar e visualizar dados numéricos de maneira eficaz, possibilitando a identificação de tendências e a criação de narrativas envolventes baseadas em evidências sólidas.

Acknowledgements: Trabalho parcialmente financiado pelo Politécnico de Lisboa ao abrigo do programa IDI&CA com a ref.:
$IPL/IDI\&CA2023/MOOC - JorDt_ESCS$.

References

[1] Bhaskaran, H., Kashyap, G., Mishra, H. Teaching Data Journalism: A Systematic Review. *Journalism Practice*, 68, 1–22, 2022. doi:10.1080/17512786.2022.2044888

[2] Cushion, S., Lewis, J., Callaghan, R. Data Journalism, Impartiality And Statistical Claims. *Journalism Practice*, 11:10, 1198–1215, 2017. doi:10.1080/17512786.2016.1256789

[3] Silvestre, C. Numbers and Journalism during the Covid-19 Pandemic. *Comunicação Pública*, 16:31, 2021. doi:https://doi.org/10.34629/cpublica.245

[4] Spiegelhalter, D. *The Art of Statistics: Learning from Data*. Pelican Books, UK, 2010 9.

Jackson exponentiality test

Ayana Mateus [a], Frederico Caeiro [a]
amf@fct.unl.pt, fac@fct.unl.pt

[a] *Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA SST*

Keywords: exponential distribution, Jackson statistic, Monte Carlo simulation, power of a statistical test

Abstract: The exponential model finds application in a wide range of fields, including, queueing theory, reliability engineering, survival analysis, financial, telecommunications, quality control, machine learning and artificial intelligence, among others. Consequently, testing exponentiality is a relevant research topic in Statistical Analysis. Numerous tests have been put forward in the existing literature to address this issue. This work aims to revisit the Jackson statistic, which is commonly employed to test the hypothesis of exponentiality against a more general alternative. Revisiting the Jackson exponentiality test involves reevaluating and examining the exact and asymptotic properties of the test statistic, as well as studying its performance through Monte Carlo computations. This revisitation contributes to a better understanding of the test's capabilities, limitations, and suitability for various statistical analyses involving the assumption of exponentiality.

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] Jackson, O.A.Y. An analysis of departures from the exponential distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 29, 540–549, 1967.

[2] Balakrishnan, K., Basu, A.P. *Exponential distribution: theory, methods and applications*. Routledge, 2020.
doi:https://doi.org/10.1201/9780203756348

[3] Noughabi, H.A., Arghami, N.R. Testing exponentiality based on characterizations of the exponential distribution. *Journal of Statistical Computation and Simulation*, Vol 81(11), 1641–1651, 2011.
doi:https://doi.org/10.1080/00949655.2010.498373

Big Data analysis of solar radiation patterns in the Colombian Caribbean region.

Gloria Carrascal [a], Jhonathan Barrios[b], Jairo Plaza [a], Flora Ferreira [b]
gpcarrascal@mail.uniatlantico.edu.co, id10605@uminho.pt,
jairoplaza@mail.uniatlantico.edu.co, fjferreira@math.uminho.pt

[a] *Physics Department, Science Faculty, Universidad del Atlántico, Colombia*
[b] *Centre of Mathematics, School of Sciences, University of Minho, Portugal*

Keywords: Big Data, climate variables, solar radiation, statistical analysis, time-series modeling

Abstract: This study focuses on analysing solar radiation in the Colombian Caribbean region, utilizing historical solar radiation data provided by Institute of Hydrology, Meteorology and Environmental Studies (IDEAM) from January 1, 2012, to December 31, 2022. The data, corresponding to solar irradiance measurements, have a temporal granularity per minute and are collected individually for each station located in different parts of the national territory. The main aim of this work is to conduct an exhaustive analysis of these data using various statistical techniques. In the initial exploratory data analysis phase, solar radiation values in the Colombian Caribbean region over time are examined to understand their distribution, identify outlier values, and verify data quality. The work also involved univariate and bivariate statistical analyses to calculate the descriptive statistics of solar radiation to understand its variability, and explore the relationships between solar radiation and other relevant climatic and geographical variables such as precipitation, temperature, altitude, and latitude. Additionally, time-series modelling was implemented to comprehend the trends, seasonality, and cyclical patterns in solar radiation in the region. This process involves decomposing the time series into its main components and applying the ARIMA model to predict future values. Finally, the validation of the entire statistical analysis was carried out through the application of statistical significance tests. This study will provide valuable insight for future steps, particularly for the application of forecasting methods using Artificial Neural Network (ANN). This work has significant potential to advance our understanding of solar radiation patterns in the Colombian Caribbean region, which can have important implications for solar energy development and climate change adaptation in the region.

Acknowledgements: The authors would like to express gratitude to IDEAM in Colombia to supply the data for the work.

Environmental exposure index for Early Life Exposure Assessment Tool (ELEAT)

Beatriz Costa [a], Lisete Sousa [a,b], Célia Rasga [c,d], Astrid Vicente [c,d]
fc53161@alunos.ciencias.ulisboa.pt, lmsousa@ciencias.ulisboa.pt,
celia.rasga@insa.min-saude.pt, astrid.vicente@insa.min-saude.pt

[a] *DEIO, Faculdade de Ciências, Universidade de Lisboa*
[b] *CEAUL, Faculdade de Ciências, Universidade de Lisboa*
[c] *DPS, Instituto Nacional de Saúde Doutor Ricardo Jorge*
[d] *BioISI, Faculdade de Ciências, Universidade de Lisboa*

Keywords: autism spectrum disorder, eleat, exposure index, factor analysis of mixed data, LASSO regression

Abstract: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by a heterogeneous clinical presentation. Genetic factors explain 40% of ASD etiology. Recent studies indicate a substantial effect of environmental factors on the onset of ASD [1].
The ELEAT (Early Life Exposures Assessment Tool) will be administered to a sample of ASD children and typically developing controls, colleted in Hospital Pediátrico de Coimbra and Hospital Garcia de Orta, to create an exposure index algorithm, which will evaluate the level of exposure, to which groups of toxicants, in which developmental period and how exposure occurs. This Index will be developed from a logical sequence of steps [2]. These include the choice of variables used in the basic data set by applying the LASSO regression model and then hierarchical clustering with Gower Distance for mixed data. The variables with the largest inter-cluster gaps will be selected as the most discriminating. Secondly, the extraction from these variables of a set of orthogonal factors which we expect to represent the different exposure dimensions related to ASD. To solve the multidimensional concept exposure, a Factor Analysis of Mixed Data (FAMD) will be conducted on the variables selected as the most discriminating. Subsequently, the conversion of the factors back into weighted original variables in an aggregated index, called, exposure index. Finally the validation of the index scores and removal of redundant variables, using multiple regression analysis of the orthogonal factors from the FAMD. This work is still in development, as is the data collection.

Acknowledgements: CEAUL under the FCT project UIDB/00006/2020.

References

[1] Santos, J.X., Rasga, C., Vicente, A. Exposure to Xenobiotics and Gene-Environment Interactions in Autism Spectrum Disorder: A Systematic Review. In *Autism Spectrum Disorder-Profile, Heterogeneity, Neurobiology and Intervention* (M. Fitzgerald, Ed.), 2021. doi:10.5772/intechopen.95758

[2] Chadee, S., Stoute, V. Development of an urban intensity index to facilitate urban ecosystem studies in Trinidad and Tobago. *Journal of Applied Statistics*, 45:3, 508–527, 2018. doi:10.1080/02664763.2017.1282440

Failure time in a pulp drying press

Luís Margalho [a], Francisco Paiva [a]
lmelo@isec.pt, a2021110526@isec.pt

[a] *Coimbra Engineering Institute / Coimbra Politechnic Institute*

Keywords: maintenance, paper pulp industry, time to failure

Abstract: Given the current competitiveness of the global market, it is essential to ensure the reliability of all assets present in a production unit, maintaining manufacturing stability and meeting the required quality.
Maintenance procedures occur when parts are expected to be worn out, preventing failures and halting the production processes for more time than strictly necessary, focusing on preventing future failures. Predictive maintenance is the most common type of approach, aiming to optimize costs and increase equipment availability [1].
With this work, it is intended to conduct a reliability analysis of a pulp drying press, to evaluate the efficiency of the predictive maintenance methodologies applied. The variable monitored is the electric current intensity, considering data gathered with a sampling period of 1 minute. Data cleaning reveals to be most challenging task. The Kaplan-Meier estimate of the survival function [2] is presented.

References

[1] Mateus, B., Mendes, M., Farinha, J., Assis, R., Cardoso, A. Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press. *Energies*, 14, 2021. https://doi.org/10.3390/en14216958

[2] Kaplan, E., Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53:282, 457-481, 1958.

A influência dos riscos psicossociais na qualidade de vida dos colaboradores de uma IES - uma visão crítica

Pedro Pereira [a], Maria Mourão [b], Pedro Carvalho [b]
pereira1998pedro@gmail.com, fmourao@estg.ipvc.pt, pc@estg.ipvc.pt

[a] *Universidade de Trás os Montes e Alto Douro*
[b] *Instituto Politécnico de Viana do Castelo*

Keywords: conciliação, família, qualidade de vida, riscos picossociais, trabalho

Abstract: Segundo Muchinsky (2000), a atividade em que o ser humano está mais envolvido na sua vida é o trabalho. Assim, cada vez mais se tem verificado o processo de humanização do trabalho, pelo que urge que se estudem as melhores formas de gerir os recursos humanos das organizações do séc. XXI. O objetivo deste trabalho é essencialmente perceber a influência que os riscos psicossociais no trabalho poderão ter na vida pessoal dos colaboradores de uma Instituição de Ensino Superior (IES), por aplicação da versão portuguesa do COPSOQ II. Foi aplicada uma análise estatística exploratória e confirmatória aos resultados com o intuito de perceber se fatores do foro laboral poderiam influenciar fatores da vida pessoal. Concluiu-se que é dada muita importância ao emprego e que fatores do foro profissional podem, realmente, influenciar dimensões da vida pessoal desses colaboradores. Entre outras coisas, verifica-se que, no caso da IES em análise, as exigências cognitivas influenciam diretamente a satisfação no trabalho, que as exigências quantitativas têm efeitos diretos no conflito trabalho-família e que a satisfação no trabalho apresenta influência direta no stress somático e cognitivo. Por outro lado, verifica-se que o conflito trabalho-família é essencialmente influenciado pelas exigências quantitativas, pelo que deverá haver uma boa distribuição da carga de trabalho pelos trabalhadores, definindo-se objetivos concretos.

Acknowledgements: Este trabalho foi cofinanciado pelo programa COMPETE2020, Portugal2020 e Fundo Europeu de Desenvolvimento Regional - POCI-05-5762-FSE-000328

References

[1] Muchinsky, P.M. Emotions in the workplace: The neglect of organizational behavior. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 21(7), 801–805, 2000.

[2] Kristensen, T.S., Hannerz, H., Høgh, A., Borg, V. The Copenhagen Psychosocial Questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian journal of work, environment & health*, 31(6):438–49. doi:10.5271/sjweh.948

Propagação de incertezas e fiabilidade estrutural num modelo mecânico

Luísa Hoffbauer [a], Carlos Conceição António [b]
lnh@isep.ipp.pt, cantonio@fe.up.pt

[a] *Instituto Politécnico do Porto, Instituto Superior de Engenharia, LAETA*
[b] *Universidade do Porto, Faculdade de Engenharia, LAETA*

Keywords: compósito laminado, fiabilidade estrutural, propagação de incertezas

Abstract: O recurso a materiais compósitos laminados em todos os tipos de estruturas tem aumentado nos últimos anos, sobretudo nas indústrias aeroespacial, automóvel e naval.
As variações nos parâmetros de fabrico e a heterogeneidade dos materiais são fatores que introduzem incertezas. O estudo da propagação de incertezas decompõe-se em três fases: a definição do modelo matemático representando o comportamento físico do sistema considerado, a caracterização probabilística das incertezas nos parâmetros de entrada e a propagação das mesmas incertezas ao longo do modelo.
É, então, fundamental obter informação sobre a aleatoriedade da resposta estrutural devida à aleatoriedade da entrada. Essa informação pode ser obtida utilizando métodos que fornecem uma estimativa dos dois primeiros momentos da resposta, que dão informação sobre a cauda da função densidade de probabilidade da resposta (probabilidade de falha) ou que pretendam representar a aleatoriedade completa da resposta.
Neste trabalho, a estrutura considerada é uma casca cilíndrica encastrada em compósito laminado. As variáveis aleatórias de entrada são as propriedades mecânicas (o módulo elástico longitudinal, o módulo elástico transversal, a resistência transversal em tração e a resistência ao corte); as variáveis-resposta são o deslocamento critico e o número de Tsai critico. São utilizados métodos que fornecem a média e a variância destas variáveis. A informação sobre a cauda da distribuição é obtida usando o índice de fiabilidade.

References

[1] Sudret, B. Uncertainty propagation and sensitivity analysis in mechanical models - Contributions to structural reliability and stochastic spectral methods, 2007.

Count models and randomness patterns

Sílvio Velosa[a,b], Dinis D. Pestana [c,b,d], Sandra Mendonça[a,b]
silviov@staff.uma.pt, ddpestana@ciencias.ulisboa.pt, sandram@staff.uma.pt

[a] DM-FCEE, *Universidade da Madeira*
[b] *Centro de Estatística e Aplicações,* UL (CEA/UL)
[c] DEIO, *Faculdade de Ciências da Universidade de Lisboa* (FCUL)
[d] *Instituto de Investigação Científica Bento da Rocha Cabral*

Keywords: count models, extra-pair paternity, randomness patterns

Abstract: The formal mathematical organization of discrete models, such as Kempf's [2] use of hypergeometric functions or Hess *et al.* [1], [5] and references therein, use of Panjer's type recursive probability mass functions, doesn't in general enhance the randomness patterns underlying phenomena such as extra-pair paternity [3]. We focus on randomness patterns and how they influence statistical modelling, discussing the relevance of count models [4]. We also discuss truncation, and how support constriction may substantially expand the parameters space.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, Portugal, project UIDB/00006/2020 (CEA/UL) and by HiTEc Cost Action CA21163.

References

[1] Hess, K.T., Liewald, A., Schmidt, K.D. An extension of Panjer's recursion. *ASTIN Bull.* 32, 283–297, 2002. doi:10.2143/AST.32.2.1030

[2] Johnson, N.L, Kemp, A.W, Kotz, S. *Univariate Discrete Distributions.* Wiley, Hoboken, New Jersey, 2005. doi:10.1002/0471715816

[3] Marques, T.A., Pestana, D., Velosa, S. Count data models in biometry and randomness patterns in birds extra-pair paternity. *Listy Biometryczne Biometrical Letters*, 42, 81–112, 2005. https://sparrow.up.poznan.pl/biometrical.letters/full/BL%2042%202%201.pdf

[4] McCabe, B.P.M., Skeels, C.L. Distributions you can count on... But what's the point? *Econometrics*, 8(1):9, 2020. doi:10.3390/econometrics8010009

[5] Mendonça, S., Pestana, D., Velosa, S. Panjer Count Models and Aggregate Claims, *Notas e Comunicações do CEAUL*, 2023.

Desvendando o sucesso escolar: uma jornada através dos Modelos Lineares Hierárquicos no contexto dos estudantes portugueses

Marcos Machado[a,b]
fc55403@alunos.ciencias.ulisboa.pt
Maria Fernanda Diamantino[a,b]
mfdiamantino@ciencias.ulisboa.pt
Luísa Loura[a]
ldloura@ciencias.ulisboa.pt

[a] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *DEIO, Faculdade de Ciências da Universidade de Lisboa*

Keywords: avaliação educativa, Modelo Linear Hierárquico, teoria da resposta ao item

Abstract: O desempenho escolar dos estudantes é um tema de grande relevância para a sociedade, uma vez que está diretamente relacionado com o sucesso educativo e o desenvolvimento futuro dos estudantes. Neste contexto, é indiscutível a necessidade de um entendimento profundo dos fatores que interferem nesse desempenho, permitindo assim o planeamento de estratégias educativas eficazes e ajustadas à realidade de cada aluno. Este estudo visou desvendar os fatores que influenciam o desempenho a matemática dos alunos do 9.º ano em Portugal, analisando dados fornecidos pelo Ministério da Educação, através de Modelos Lineares Hierárquicos de três níveis [1]. Esta abordagem foi útil na descrição da estrutura hierárquica dos dados, ponderando não só efeitos individuais dos estudantes, mas também os impactos da turma e da escola em que se inserem. Durante a investigação, e após revisões extensivas de literatura, constatou-se que a variável nível socioeconómico das escolas se revela como um preditor relevante do desempenho escolar. Foi, por isso, sob esta ótica e através da Teoria da Resposta ao Item (TRI), que foi criado um indicador para avaliar o nível socioeconómico das escolas portuguesas. Apesar da existência de uma ampla literatura dedicada a estudos que utilizam modelos lineares hierárquicos de dois níveis - focados, principalmente, nos efeitos escolares ou de sala de aula sobre o desempenho dos estudantes -, os estudos dedicados a três níveis (aluno-turma-escola) continuam a ser escassos. Desta forma, ambicionamos contribuir para uma compreensão mais completa e multifacetada do rendimento a matemática dos alunos em Portugal.

References

[1] Raudenbush, S.W., Bryk, A.S. *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Advanced quantitative techniques in the social sciences series: Vol. 1. Sage Publications, 2002.

Regiões portuguesas: os desafios estratégicos e o papel das finanças públicas locais

Irene Oliveira [a], Patrícia Martins [b]
ioliveir@utad.pt, smartins@utad.pt

[a] *Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro e CEMAT - Centro de Matemática Computacional e Estocástica*
[b] *Departamento de Economia, Sociologia e Gestão -UTAD e CETRAD*

Keywords: análise classificatória, análise regional, finanças locais

Abstract: Este trabalho aborda dois desafios estratégicos para a economia portuguesa: a demografia e as desigualdades e coesão social. Assim, a caracterização económica das regiões portuguesas é complementada pelas dimensões populacional e social, para avaliar as escolhas políticas passadas relativas à despesa pública social e para aferir o possível impacto das entidades intermunicipais na coesão socioeconómica.
Os dados referem-se a 2009 e 2019, com o propósito de identificar a situação das regiões (NUTS III) aquando da crise económica e sua evolução ao longo de uma década, antes da crise pandémica. Foram utilizadas variáveis normalizadas "min-max" e posterior análise da estrutura correlacional entre variáveis, examinando a significância estatística e a dinâmica da estrutura de significâncias de 2009 para 2019.
Realizaram-se análises classificatórias hierárquicas, testando vários métodos de ligação entre clusters, e não hierárquicas. De seguida implementaram-se testes não paramétricos para investigar as diferenças significativas entre os grupos, considerando as dimensões económica, populacional e social, finanças locais e participação eleitoral. A análise principal exclui as áreas metropolitanas de Lisboa e Porto, outliers nas três dimensões analisadas. Os resultados obtidos para as 21 regiões NUTS III de Portugal Continental, utilizando o método de ligação de Ward, sugerem uma divisão do país em:

- regiões pobres e envelhecidas, que dependem mais das transferências do Estado e têm recursos financeiros limitados para executar níveis mais elevados de despesas;
- regiões pobres com alta densidade empresarial e populacional, que apresentaram maior crescimento do PIB per capita entre 2009 e 2019;
- regiões mais ricas, com excepção das áreas metropolitanas, com saldos migratórios superiores à média nacional.

Acknowledgements: Este trabalho é financiado por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, no âmbito dos projetos UIDB/04011/2020 e UID/MULTI/04621/2020

Long-term trends and daily variations in global irradiation in Cabinda, Angola: Implications for solar energy production and sustainability

Faustino Maciala [a], Jhonathan Barrios [a], A. Manuela Gonçalves [b]
id10604@uminho.pt, id10605@uminho.pt, mneves@math.uminho.pt

[a] *Centre of Mathematics, School of Sciences, University of Minho, Braga, Portugal*
[b] *Centre of Mathematics, School of Sciences, University of Minho, Guimarães, Portugal*

Keywords: daily variations, global irradiation, long-term trends, seasonal patterns, solar energy

Abstract: In global climate change and the urgent need for sustainable energy solutions, harnessing solar energy effectively is crucial. Understanding the irradiation patterns is fundamental to optimizing solar energy production. This work was conducted with data of Cabinda, Angola, a region with considerable solar energy potential. We analysed seasonal patterns, daily variations, and long-term trends in global irradiation and their implications for solar energy production and sustainability. Global irradiation data derived from the CAMS v4.5 radiation service, based on satellite data, was utilized. An exploratory approach was employed to identify patterns, trends, and anomalous behaviours in the data. The average global irradiation in Cabinda was found to be approximately $138.944 Wh/m^2$, exhibiting a noticeable seasonality following a 12-month pattern. A decomposition of the time series revealed both a long-term trend and a 12-month seasonality in the data. Using criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), we selected the `SARIMA(0,1,1)(1,0,1)[12]` model as the most suitable for explaining the data's seasonality. This model demonstrated a good fit and a low mean squared error (MSE). Daily variations in global irradiation were also analysed, showing a daily change rate of approximately 0.7%, which may be influenced by the time of day and season. Lastly, we identified a long-term trend in global irradiation in Cabinda. While this trend was not evident in the initial analysis, a more thorough examination confirmed its existence through a regression line. These findings are crucial in understanding the seasonal patterns and daily variations in global irradiation in Cabinda, providing key insights for solar energy production and promoting sustainability in the region.

Acknowledgements: F. Maciala would like to express gratitude to INAGBE - Angola (National Institute for Administration and Management of Scholarships) for the sponsorship provided to this research and the financial support.

References

[1] Puindi, A. C. Structural Models: A Computational Look for Signal Extraction and Forecasting Seasonal Time Series. *SN Computer Science*, 2, 96, 2021. doi.org/10.1007/s42979-021-00474-2

[2] Boland, J. Time Series Modelling of Solar Radiation. *Modeling Solar Radiation at the Earth's Surface*, Vol I, 283–312, 2014. doi.org/10.1007/978-3-540-77455-6

Analysis of crew time series absenteeism in the railway sector

Catarina Afonso [a,b], Gonçalo Matos [b], Luís Albino [b], Ricardo Saldanha [b], Regina Bispo [c,d]
cmr.afonso@campus.fct.unl.pt, goncalo.matos@siscog.pt, lmalbino@siscog.pt, rsaldanha@siscog.pt, r.bispo@fct.unl.pt

[a] *MSc. Analytics and Big Data Engineering, Departments of Computer Science and Mathmatics, NOVA School of Science and Technology, Portugal*
[b] *SISCOG - Sistemas Cognitivos, SA*
[c] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Portugal*
[d] *Department of Mathematics, NOVA School of Science and Technology, Portugal*

Keywords: absenteeism, forecasting methods, machine learning, planning, railway operators

Abstract: Crew absenteeism is a problem for railway operators because it can have repercussions, such as cancellation of train trips, additional expenses, among others. To maintain the normal operation of transport, railway operators carry out operational planning to manage the railway resources. However, if an unexpected event occurs, like unplanned absence of the crew, the operational planning must be readjusted. As some types of absences of crew members, particularly those caused by illness or family emergencies, are not planned, the railway operators seek for tools that can help mitigate the impacts caused by absenteeism.
The aim of this study is to develop a prediction model for unscheduled absences, that can estimate the number of crew members absent, segmented by job category (drivers or guards) and by operational base, for a given time period based on historical data. This prediction model can help planners make better decisions, for instance, in creating reserve plans with increased cost-effectiveness.
Several datasets were selected and exported from a railway operator in Northern Europe, a customer of SISCOG[1]. The data was converted to time series to enable the analysis of the evolution of absenteeism over time, in this case, daily data on the percentage of absence by job category and operational base. We looked for patterns and analysed which factors significantly influence absenteeism, and we used Prophet [1] algorithm to perform predictions.

References

[1] Taylor, S.J., Letham, B. Forecasting at scale. *The American Statistician*, 72.1, 37—45, 2018. doi:https://doi.org/10.1080/00031305.2017.1380080

[1]The company that hosts this research and develops decision-support tools for the scheduling and management of railway operations, such as crew operation

Application of statistical methodologies as a contribute to define disease control strategies for a sustainable viticulture

Nuno Domingues[a], Lisete Sousa[a], Gonçalo Laureano[b], Andreia Figueiredo[b], Marisa Maia[b]
fc52631@alunos.ciencias.ulisboa.pt, lmsousa@ciencias.ulisboa.pt, gmlaureano@ciencias.ulisboa.pt, aafigueiredo@ciencias.ulisboa.pt, mrmaia@ciencias.ulisboa.pt

[a]*Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa and CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal*
[b]*Departamento de Biologia Vegetal, Faculdade de Ciências da Universidade de Lisboa and Gravepine Pathogen Systems Lab, BioISI - Biosystems and Integrative Sciences Institute, Universidade de Lisboa, Lisboa, Portugal*

Keywords: metabolomics, multivariate data analysis, sustainable viticulture, time-resolved analysis

Abstract: Elicitation is a technique able to trigger the biosynthesis of phytochemical compounds in plants, inducing physiological changes [1]. This study aims to evaluate two selected molecules, Eicosapentaenoic Acid and Jasmonic Acid, in their ability as elicitors agents on triggering defence responses in gravepine against *Plasmopara viticola*, the pathogen responsible for downy mildew in gravepine. To do so, one of the approaches was metabolomics analysis of elicited plants, after inoculation with the pathogen, at different time points. Up until now, a Multivariate Data Analysis was performed on the metabolomic data, where methods such as Principal Component Analysis (PCA) and Partial Least Squares-Discriminant Analysis (PLS-DA) were used. A time-resolved analysis is now being applied by using ANOVA-Simultaneous Component Analysis (ASCA) [2]. With this method, time is taken into account as an effect, allowing the assessment of how the selected molecules interact with time and, ultimately, the evaluation of how important that interaction is to the global variation of metabolites between the different groups. Therefore, a deeper and more dynamic understanding of the biological information is obtained when compared to only using methods such as PCA and PLS-DA.

Acknowledgements: This work is supported by National Funds through FCT under projects UIDP/00006/2020 (BII to N.D.), UIDB/00006/2020 (CEAUL), UIDB/04046/2020 (BIOISI) and 2022.07433.CEECIND (RC to M.M.).

References

[1] Baenas, N., et al. Elicitation: a tool for enriching the bioactive composition of foods. *Molecules*, 19(9), 13541–13563, 2014. doi:10.3390/molecules190913541

[2] Bertinetto, C., et al. ANOVA simultaneous component analysis: A tutorial review. *Anal Chim Acta X*, 6, 100061, 2020. doi:10.1016/j.acax.2020.100061

Classification of compositional data using distributions defined on the hypersphere

Adelaide Figueiredo
adelaide@fep.up.pt

Faculdade de Economia da Universidade do Porto e LIAAD - INESC TEC Porto

Keywords: compositional data, directional data, hypersphere

Abstract: Compositional data are vectors whose components are the proportions or percentages of some whole. Their sum is constrained to be some constant, equal to 1 for proportions or 100 for percentages. Compositional data arise in many areas, including Geology (mineral compositions of rocks, sediment compositions such as sand, silt, clay compositions), Economics (household budget compositions, portfolio compositions), Environment (pollutant compositions), Geography (US state ethnic compositions, urban-rural compositions, land use composition), etc. This type of data need to be transformed, before applying the standard statistical techniques designed for the Euclidean space. As only the relative variation of the components is of interest, the methods based on the log-ratios of the components are the natural ways of analyzing compositional data (see [1]). But other transformations of the compositional data have been used in the literature, for example the square-root to transform compositional data into directional data (unit vectors on the surface of the hypersphere). The statistics of directional data has been extensively studied (see for example, [2]).

In this paper the square-root transformation is applied to the compositional data to obtain directional data. Then, the identification of a mixture of distributions on the hypersphere is applied to the directional data for clustering the compositional data and the discriminant analysis for a directional distribution is applied for the classification of the compositional data. Finally these methods are illustrated with several sets of compositional data given in the literature and the results obtained are compared with those obtained using the log-ratio transformations.

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

References

[1] Aitchison, J. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 44 (2), 139–177, 1982.

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

Data mining and statistical quality control

Fernanda Otília Figueiredo [a], Adelaide Figueiredo [b], M. Ivette Gomes [c]
otilia@fep.up.pt, adelaide@fep.up.pt, ivette.gomes@fc.ul.pt

[a] *Faculdade de Economia da Universidade do Porto and CEAUL*
[b] *Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC Porto*
[c] *DEIO and CEAUL, Universidade de Lisboa*

Keywords: big data, data mining, statistical quality control

Abstract: In the era of data mining and big data all organizations have available large volume and variety of data, with detailed information, to be processed and monitored. At the same time the organizations are faced with the challenge of measuring the data quality.
In this work we present a brief review of some statistical techniques and data mining tools that can be used in the context of big data as well as some adaptations of the traditional control charts for use in this context. Some examples of control charts suggested in the literature to monitor such type of data are, among others, control charts for attributes and mixed data types, multivariate control charts for profile monitoring and control charts for multiple streams of correlated data. Different types of machine learning algorithms have also been proposed in the literature for quality improvement.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, within the projects UIDB/00006/2020 (CEA/UL) and LA/P/0063/2020.

The profile of the hyper user in an emergency department

Loide Ascenso [a], Gonçalo Jacinto [b], Hugo Quintino [c], Paulo Infante [b]
loide.ascenso@gmail.com , gjcj@uevora.pt , hquintino@hevora.min-saude.pt, pinfante@uevora.pt

[a] *PDMAT/IIFA, Universidade de Évora*
[b] *DMAT/ECT e CIMAA/IIFA, Universidade de Évora*
[c] *Hospital do Espírito Santo EPE*

Keywords: emergency department, exploratory data analysis, generalized additive linear models, generalized linear models

Abstract: Characterizing the profile of frequent users of an emergency service is essential for the definition of health management policies, looking for an efficient meeting of patients' needs and offering a better quality of services to users of the National Health Service. Therefore, it is reasonable that several research studies developed in this area focus on the very frequent users of hospital emergency services, usually referred to as hyper users. Their definition varies according to the studies, but in general, a hype user is a patient with 4 or more visits to the emergency department per year. This study analyzes the inflow of users to the emergency department of the Hospital do Espírito Santo de Évora (HESE), by type of user, in a period between January 2018 and May 2021. Based on Generalized Linear Models (GLM) and Generalized Additive Models (GAM) it is attempted to identify some factors that allow to outline the profile of the hyper user of the HESE emergency service. The results obtained with several models are compared and some considerations are made about the modelling limitations taking into account the type of data and the context of the collection period. The profile of a frequent user of the HESE emergency services includes the triage code indicator, the user residence, age and gender, the admission shift, the waiting time and the physician's diagnosis.

Acknowledgements: This research was partially supported by the Portuguese funding agency, FCT Fundação para a Ciência e a Tecnologia, Portugal, project number UIDB/04674/2020 (CIMA).

Análise estatística temporal da sinistralidade laboral em Portugal de 2009 a 2019

Ricardo Dourado [a], Marina Almeida-Silva [a,b], **Miguel Felgueiras** [c,d]
ricardo.dourado.maria@gmail.com, marina.silva@estesl.ipl.pt, mfelg@ipleiria.pt

[a] *H&TRC–Health & Technology Research Center, ESTeSL – Escola Superior de Tecnologia e Saúde, Instituto Politécnico de Lisboa.*
[b] *OSEAN–Outermost Regions Sustainable Ecosystem for Entrepreneurship and Innovation, 9000–039 Funchal, Portugal.*
[c] *ESTG, Politécnico de Leiria.*
[d] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal.*

Keywords: acidentes de trabalho, correlação, inspetor do trabalho

Abstract: Este trabalho pretende proporcionar um melhor conhecimento da distribuição geográfica e temporal dos acidentes de trabalho com vista ao desenvolvimento e implementação de orientações a nível organizacional da Autoridade para as Condições do Trabalho. Foi observada a tendência dos acidentes de trabalho e analisado o impacto da atividade inspetiva sobre a sinistralidade laboral entre 2009 e 2019. O número de acidentes de trabalho foi correlacionado com o número de inspetores do trabalho e com o número de trabalhadores, numa perspetiva distrital. Dos dados apurados, fica clarificado que os grandes centros urbanos (sobretudo devido ao número de trabalhadores alocados a estas regiões) e áreas com forte implementação industrial (por exemplo, os distritos de Aveiro ou Leiria) necessitam de um claro reforço de meios, pelo menos para acompanhar um maior índice de sinistralidade que evidencia uma maior debilidade das organizações em conseguir proporcionar condições de trabalho com um grau de segurança dentro da média.

Acknowledgements: Research supported by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through projects UIDB/00006/2020, UIDB/04106/2020, UIDP/04106/2020, UIDB/05608/2020 and UIDP/05608/2020. H&TRC authors also acknowledge Instituto Politécnico de Lisboa for funding the project PLASCONGEN (IPL/2022/PLASCONGEN_ESTeSL) under its IDI&CA Program.

References

[1] Alegre, C. *Regime jurídico dos acidente de trabalho e das doenças profissionais*, 2.ª Edição, Coimbra, Almedina, 2000.

[2] Fabela, S., Sousa, J. *Os impactos sócio-económicos no âmbito dos acidentes de trabalho. Representações, práticas e desafios à gestão das organizações de trabalho*, Vila do Conde, Civeri Publishing, 2012.

[3] GEP–Estatísticas de acidentes de trabalho. http://www.gep.mtsss.gov.pt/sinteses

[4] Leitão, L. *Direito do trabalho*, 7ª Edição, Almedina, Coimbra, 2021.

Flexible odds ratio curves for continuous predictors: the flexOR package

Marta Azevedo [a], Luís Meira-Machado [a], Artur Araújo [b] Carla Moreira [a]
marta.vasconcelos4@gmail.com, lmachado@math.uminho.pt,
artur.stat@gmail.com, carlamgmm@gmail.com

[a] *Centre of Mathematics, University of Minho, Portugal*
[b] *University of Vigo, Spain*

Keywords: logistic regression, odds ratio, reference value, smoothing

Abstract: The analysis of odds ratio curves is a valuable tool in understanding the relationship between continuous predictors and binary outcomes. Traditional parametric regression approaches often assume specific functional forms, limiting their flexibility and applicability to complex data. To address this limitation and introduce more flexibility, several smoothing methods may be applied, and approaches based on splines are the most frequently considered in this context. To better understand the effects that each continuous covariate has on the outcome, results can be expressed in terms of splines-based odds ratio (OR) curves, taking a specific covariate value as reference. In this paper, we introduce an R package, flexOR, which provides a comprehensive framework for pointwise nonparametric estimation of odds ratio curves for continuous predictors. The package can be used to estimate odds ratio curves without imposing rigid assumptions on the underlying functional form while a specific reference covariate value. The package offers various options for choosing automatically the degrees of freedom in multivariable Cox models. It also includes visualization functions to aid in the interpretation and presentation of the estimated odds ratio curves. flexOR offers a user-friendly interface, making it accessible to researchers and practitioners without extensive statistical backgrounds.

References

[1] Hastie, T., Tibshirani, R. (1990) *Generalized Additive Models.*, London: Chapman and Hall.

[2] Venables, W. N., Ripley, B. D. (2002) *Modern Applied Statistics with S.*, New York: Springer.

Multialphabetic hypercubes and disaggregation of sums of squares

Carla Francisco [a,b], Manuela Oliveira [a,b], Francisco Carvalho [a,b], Tiago Mexia [c]
carlafrancisco@edu.ulisboa.pt, mmo@uevora.pt, fpcarvalho@ipt.pt, jtm@fct.unl.pt

[a] *Escola de Ciências e Tecnologia, Departamento de Matemática, Universidade de Évora, Portugal*
[b] *Center for Mathematics and Applications, Faculty of Science and Technology, New University of Lisbon, Caparica, Portugal*
[c] *Mathematics and Physics Departmental Unit, Polytechnic Institute of Tomar, Portugal*

Keywords: experimental designs, galois fields, greco-latin squares, hypercubes, multialphabetic

Abstract: This research focuses on developing a complete disaggregation of squares sums for treatments in experimental designs through the use of multialphabetic hypercubes constructed on vector spaces of Galois Fields, [1]. Multialphabetic hypercubes also known as Vigenère squares, are a type of cryptogram based on polyalphabetic letter substitution. They were invented by the French cryptographer Blaise de Vigenère in the 16th century, [2]. These hypercubes are an extension of Greco-Latin squares to higher dimensions and are used for randomized systematic sampling on continuous media. The dimension of a hypercube is determined by the number of factors considered to have prime levels. Algorithms were also developed to decompose the sum of squares for treatments into the effects and interactions of the factors. By using hypercubes, structured series of designs are obtained through nesting, with treatments defined by the crossing of v factors with p prime levels and u external factors with p levels. This approach allows for the study of the effects and interactions of both internal and external factors, unlike traditional designs that only consider internal factors. The use of factors with prime levels enables multilevel nesting, which can facilitate the implementation of designs in the network.

Acknowledgements: This talk is partially supported by Centro de Investigação em Matemática e Aplicações, through project UIDB/04674/2020 of FCT - Fundação para a Ciência e a Tecnologia, Portugal.

References

[1] Lidl, R., Niederreiter, H. Finite Fields Cambridge University Press, 1997. doi: https://doi.org/10.1017/CBO9780511525926

[2] Vigenère, B. Traicté des chiffres ou secrètes manières d'écrire. Chez Blaise de Vigenère, 1586. ark:/12148/bpt6k73371g

Survival analysis applied to extreme longevity research

Laetitia Teixeira [a,b], Lia Araújo [b,c], Denisa Mendonça [a], Constança Paúl [a,b], Oscar Ribeiro [a,d,e]
lcteixeira@icbas.up.pt, liajaraujo@esev.ipv.pt, dvmendon@icbas.up.pt, paul@icbas.up.pt, oribeiro@ua.pt

[a] *Institute of Biomedical Sciences Abel Salazar, University of Porto*
[b] *CINTESIS.ICBAS-UP, University of Porto*
[c] *School of Education, Polytechnic Institute of Viseu*
[d] *CINTESIS.UA, University of Aveiro*
[e] *Department of Education and Psychology, University of Aveiro*

Keywords: extreme longevity, health, left-truncated, survival analysis

Abstract: The world population is ageing, and older persons are increasing in number and in representativity. Portugal observed this pattern and is considered one of the oldest countries of Europe [1]. The population aged 80+ is rapidly growing, and within this group the number of centenarians has risen exponentially. The main objective of this study is to identify sociodemographic and health-related factors that may be associated with survival after 100 years old in a population-based sample of Portuguese centenarians (N=140 centenarians). Follow-up was considered as the time (in months) between the 100th anniversary and death (event of interest) or the last telephone contact (June 2019). The centenarians lost in follow-up or that were still alive at the end of the study were considered as censored. Given that this is a prevalent-study design, we had left-truncated (or late entering) and right-censored data. Univariable and multivariable Cox proportional hazards regression models, with age 100 as time scale and the entry age to the study as left truncation time, were performed to identify potential predictive factors of survival [2, 3]. Proportional hazards assumption was assessed using test based on the Schoenfeld partial residuals. Multivariable analysis revealed that longer survival was associated with acute disease, functional status and physical fatigue. this study adds to the available knowledge on the health related factors that predict longevity in centenarians.

References

[1] Teixeira, L., Araújo, L., Paúl, C., Ribeiro, O. Centenarians: a European Overview. *Springer International Publishing*, 2020. Doi: 10.1007/978-3-030-52090-8

[2] Canchola, A. J., Stewart, S. L., Bernstein, L., West, D. W., Ross, R. K., Deapen, D., ..., Peel, D.. Cox regression using different time-scales. *Western Users of SAS Software*. San Francisco, California, 2019.

[3] Kim, M., Paik, M. C., Jang, J., Cheung, Y. K., Willey, J., Elkind, M. S., Sacco, R. L. Cox proportional hazards models with left truncation and time-varying coefficient: Application of age at event as outcome in cohort studies. *Biometrical Journal,* 59(3), 405-419, 2017.

Factors that influence energy to water nexus in urban and rural households

A. Manuela Gonçalves [a], Cristina Matos[b,c]
mneves@math.uminho.pt, crismato@utad.pt

[a] *University of Minho, Department of Mathematics and Centre of Mathematics, Portugal*
[b] *University of Trás-os-Montes e Alto Douro, School of Science and Technology, Portugal*
[c] *Interdisciplinary Centre of Marine and Environmental Research of the University of Porto, Portugal*

Keywords: influence on consumption of the differentiating factors, ordinal regression models, rural and urban environments, survey

Abstract: Water and energy are two linked resources of primordial importance to humanity. Water consumption has been raising every day and the factors that are influencing this situation are not well studied yet. There is a wealth of research that highlights a large difference between urban and rural household water consumption patterns. One of the factors mentioned in the literature is the type of use; however, characterizing the consumption pattern is not done via the type of environment. It is empirically known that rural environments are characterized as featuring mostly primary activity (agricultural), so the main water consumption (and consequently energy consumption) may be linked to irrigation. The main question resides in assessing how the differences between rural and urban households affect water and energy consumption. Knowing these differences may turn help predict household consumption, which is critical to perform an efficient and effective assessment of the supply and demand balance at household scale (domestic consumption of water, electricity, and natural gas). This study primarily involves a survey regarding the period from December 2016 to January 2017, and its main goal is to reveal the major underlying causes of water and energy consumption in rural and urban households located in Vila Real County, in Northern Portugal. In fact, the main aim of this study is to distinguish which factors support the differences found between consumption in both environments.

Acknowledgements: A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

References

[1] Matos, C., Bentes, I., Pereira, S., Gonçalves, A.M., Faria, D., Briga-Sá, A. Which are the factors that may explain the differences in water and energy consumptions in urban and rural environments? *Science of The Total Environment*, 642, 421–435, 2018. doi:10.1016/j.scitotenv.2018.06.062

Modelos Lineares Generalizados Mistos: uma aplicação a dados de acidentes rodoviários

Susana Faria [a], Jair Santos [b], Elisabete Freitas [c]
sfaria@math.uminho.pt, jaircv57@gmail.com, efreitas@civil.uminho.pt

[a] *Centro de Matemática, Universidade do Minho*
[b] *Departamento de Matemática, Universidade do Minho*
[c] *Departamento de Engenharia Civil, Universidade do Minho*

Keywords: dados de contagem, modelos lineares generalizados mistos, segurança rodoviária

Abstract: A Sinistralidade Rodoviária é um problema de saúde pública com consequências humanas e materiais relevantes. Embora nos países desenvolvidos o número de acidentes tenha diminuído nas últimas décadas, o conhecimento dos fatores que os determinam, visando reduzir a gravidade dos danos causados às vítimas, ainda é incompleto.
O objetivo deste trabalho é desenvolver modelos estatísticos para dados de contagem, que permitam identificar fatores (humanos, infraestruturais e ambientais) associados à ocorrência de acidentes rodoviários. Trata-se de um problema de elevada importância no contexto da segurança rodoviária, pois permite que as empresas responsáveis adotem medidas adequadas para melhorar a segurança nas estradas.
O conjunto de dados utilizado neste trabalho refere-se a acidentes rodoviários ocorridos na autoestrada A4 da região do Grande Porto, entre 2014 e 2021. Cada observação corresponde a um segmento da autoestrada, com uma extensão de 200m. Na construção dos modelos lineares generalizados mistos, são utilizados como variável resposta o número de acidentes de viação, e como variáveis explicativas, são consideradas variáveis relacionadas ao tráfego, ao estado do pavimento e às características geométricas da via (tanto em planta como em perfil).

Acknowledgements: Este trabalho foi apoiado pela Fundação para a Ciência e Tecnologia (FCT) de Portugal no âmbito do Financiamento Estratégico
UIDB/00013/2020 e UIDP/00013/2020 do CMAT-UM.

References

[1] Hedeker, D. Generalized linear mixed models. Encyclopedia of Statistics in Behavioral Science. John Wiley and Sons, New York, 2005.

[2] Faraway, J. J. *Extending the Linear Model with R: Generalized Linear Mixed Effects and Nonparametric Regression Models.* Chapman and Hall/CRC, 2006.

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

Use of the random forest model in precision agriculture

Natália Costa Martins [a], Thiago G. Ramires [a], Luiz R. Nakamura [b]
nataliamartins@alunos.utfpr.edu.br, thiagogentil@gmail.com,
lr.nakamura@gmail.com

[a] *Federal University of Technology - Paraná*
[b] *Departamento de Estatística. Federal University of Lavras*

Keywords: agriculture 4.0, artificial inteligence, classification

Abstract: Cultivation of sugarcane has become the focus in several countries, due to diversity of use, with Brazil being the world's largest producer. One of the biggest problems in sugarcane production, that affects crop productivity, is the invasion of weeds, e.g., invasion of Brachiaria decumbens and Panicum were responsible for the loss of 40% of production [1]. With the development of technology, which generated agriculture 4.0, the management of weeds started to be carried out via crop mapping, which allows producers to apply fertilizers and herbicides locally, increasing productivity and competitiveness [2]. One of the main techniques for mapping cultivars is classification models. In this survey we propose a random forest model, to predict weed in the field, using four color spectra as input, which were obtained by a multispectral camera adapted to an unmanned aerial vehicle. Results show that in the experimental field it was possible to reduce the use of herbicides by 57%.

Acknowledgements: This work was supported by the Federal University of Technology - Paraná. I also would like to thank the Research and Postgraduate Directorate (DIRPPG-AP).

References

[1] Kuva, M., et al. Períodos de interferência das plantas daninhas na cultura da cana-de-açúcar: I-Tiririca. *Planta Daninha*, 18, 241–251, 2000. doi:10.1590/S0100-83582000000200006

[2] Shiratsuchi, L.S., Fontes, J.R.A., Resende, A.V. Correlação da distribuição espacial do banco de sementes de plantas daninhas com a fertilidade dos solos. *Planta daninha*, 23, 429–436, 2005. doi:10.1590/S0100-83582005000300006

Prediction of perceived depression for SHARE survey data in COVID 19 waves

Sara Ribeiro Pires [a], M. Rosário Ramos [a,b], Paula Vaz-Fernandes [a,c]
sararipires@gmail.com, mariar.ramos@uab.pt.pt, paulavaz@uab.pt

[a] *Universidade Aberta, Portugal*
[b] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *Centre for Public Administration and Public Policies, Institute of Social and Political Sciences, Lisbon University, Lisbon, Portugal, 1300-663*

Keywords: classification, logistic regression, self perceived depression, SHARE project

Abstract: The elderly face several challenges resulting from physical and psychological changes associated with the natural aging process. With the expected increase of life expectancy, it is feared that the incidence of mental health problems will increase in the elderly and adults. SHARE (Survey of Health, Ageing and Retirement) is a European project whose aim is to respond to the call made by the European Commission aiming to promote research and study the impact of health, social, economic and environmental policies throughout the lives of European citizens. This study seeks to model the self-perceived depression among adults aged 50 years and older, taking a selection of physical and mental health, as well as daily activities variables. Data are from the special wave of SHARE administered in 2020 and 2021 regarding COVID-19. Binary logistic regression is applied with different link functions, individual effects and interactions, included to examine whether a moderator, like sex for example, will change the strength of the relationship between the independent and dependent variables. Finnaly, Artificial Neural Network is also aplied to data and the performance of both tecnhiques is compared. The self-perception of health status, and feeling lonely in recent weeks, as well as having suffered a hip fracture are among the variables that help to predict the perception of feeling sad or depressed, during the period considered.

Acknowledgements: M.Rosário Ramos and Paula Vaz-Fernandes are partially financed by FCT – Fundação para a Ciência e a Tecnologia under projects UIDB/00006/2020 and UIDB/00713/2020 respectively.

References

[1] Santini, Z. I. et al. Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP):a longitudinal mediation analysis. *The Lancet. Public health*, 5(1), e62–e70,2020. https://doi.org/10.1016/S2468-2667(19)30230-0

[2] Wester, C. et al. Longitudinal changes in mental health following the COVID-19 lockdown: Results from the Survey of Health, Ageing and Retirement in Europe. *Annals of Epidemiology*, 5(1), e62–e70, 2022. https://doi.org/10.1016/j.annepidem.2022.05.010

An approach to estimate infection by COVID-19

Manuela Oliveira [a], Eugénio Garção [b]
mmo@uevora.pt, jesg@uevora.pt

[a] *Departamento de Matemática e CIMA, Universidade de Évora, Évora, Por- tugal*
[b] *Departamento de Engenaria Mecatrónica, Escola de Ciências e Tecnologia, Universidade de Évora, Évora, Portugal*

Keywords: COVID-19, immune, number of asymptomatic, numbers of symptomatic, stratified sampling

Abstract: Given that individuals in a certain population are different (among other things they have a different immune system), it is possible that some are infected with the known virus COVID-19 and are asymptomatic and therefore not diagnosed with the disease. Thus, estimates of the number of infected and dead with COVID-19 may not correspond to reality. This study seeks to indicate a procedure to estimate the number of individuals in the infected population that are asymptomatic (not diagnosed, but possible transmitters of the disease), based on the number of infected individuals (already diagnosed). We showed how with data available (numbers of symptomatic, symptomatic in the hospital and deceased) on the evolution of the pandemic in five regions of mainland Portugal it is possible to estimate the number of asymptomatic and immune individuals in the population.

Acknowledgements: project ref. UIDB/04674/2020 do CIMA and project ref. H2020-MSCA-RISE-2020/101007950, "DecisionES - Decision Support for the Supply of Ecosystem Services under Global Change", funded by the Marie Curie International Staff Exchange Scheme.

Diagnóstico do COVID-19 nos registros de SRAG - Síndrome Respiratória Aguda Grave

Lúcia Pereira Barroso [a], Gustavo de Oliveira Kanno [a]
lbarroso@ime.usp.br

[a] *Instituto de Matemática e Estatística - Universidade de São Paulo*

Keywords: COVID, mortality, surveillance

Abstract: O presente estudo teve como objetivo explorar métodos estatísticos de classificação para aplicação em casos de óbito por síndrome respiratória aguda grave, com agente etiológico não especificado, diferenciando-os entre causados ou não pela COVID-19. Métodos e Procedimentos: Para a classificação dos casos como Covid ou Não-Covid, primeiramente foram utilizados os casos cuja causa de óbito já estava especificada. Foi feita uma análise descritiva dos dados seguida da aplicação das técnicas estatísticas (Hastie et al., 2017): Regressão Logística, Árvores de Classificação, Random Forest, KNN. Além dessas, foi feita uma adaptação do algoritmo COVID-19 Rapid Mortalitity Surveillance (CRMS), inspirado por Duarte-Neto et al. 2021. Na segunda fase, as técnicas de classificação foram aplicadas ao conjunto de dados de SRAG cuja causa de óbito não era especificada. Resultados: Os diferentes modelos e algoritmos foram comparados por meio de um conjunto de métricas, em específico a acurácia, a sensibilidade, a especificidade, o valor preditivo positivo (VPP) e o valor preditivo negativo (VPN). Os maiores valores de acurácia (0,82), sensibilidade (0,93) e valor preditivo negativo (0,63) foram obtidos para a Random Forest com dados balanceados. Já os maiores valores para especificidade (0,99) e valor preditivo positivo (0,99) foram obtidos dessa vez através da Regressão Logística. A adaptação do algoritmo CRMS permitiu sua aplicação a grandes bancos de dados e estimar o grau de subnotificação das mortes por COVID-19 de pacientes com SRAG.

References

[1] Duarte-Neto, A.N., Marinho, M.d.F., Barroso, L.P., Saldiva de André, C.D., da Silva, L.F.F., Dolhnikoff, M., Afonso de André, P., Minto, C.M., de Moura, C.S., Leite, T.F., Filho, J.T., Monteiro, R.A.d.A., Setel, P., Bratschi, M.W., Mswia, R., Saldiva, P.H.N., Bierrenbach, A.L. Rapid mortality surveillance of COVID-19 using verbal autopsy. *International Journal of Public Health*, 66, 1604249, 2021. doi:10.3389/ijph.2021.1604249

[2] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2017.

Regression modeling of marine species abundance indicators: Exploring spatial distribution and environmental correlations

André Dias[a], **Raquel Menezes**[a], Maria Manuel Angélico[b]
a98501@alunos.uminho.pt, rmenezes@math.uminho.pt, mmangelico@ipma.pt

[a] *Centre of Mathematics (CMAT), Minho University, Guimarães*
[b] *Division of Oceanography and Marine Environment, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisboa*

Keywords: environmental predictors, marine species, regression models, spatial distribution

Abstract: The accurate assessment of marine species abundance is vital for understanding and managing the health of marine ecosystems. In this study, we focus on employing regression models to investigate the relationship between the egg count of a target species, serving as an abundance indicator, and various environmental factors in a specific location. The objective is to gain insights into the spatial distribution of the species and its correlation with essential environmental variables, such as sea surface temperature, salinity, and chlorophyll concentrations.
Conducted in collaboration with the Division of Oceanography and Marine Environment at the Portuguese Institute of the Sea and Atmosphere (IPMA), this research aims to contribute valuable knowledge to both the statistical and ecological communities. By using regression modeling, we endeavor to provide a comprehensive understanding of the techniques' efficacy, the interpretation of results, and their practical relevance in addressing intricate ecological challenges.
To achieve our objectives, we will analyze data collected during scientific surveys from targeted marine species, covering diverse environmental conditions. We will explore linear, generalized, and potentially non-linear regression models to predict species abundance based on egg counts. Moreover, we will integrate spatial analysis techniques to account for geographic dependencies and explore the influence of environmental variables across different regions.
By revealing correlations between egg counts and environmental factors, our study can inform resource management decisions and marine conservation efforts. It bridges statistical methodologies and marine ecology, fostering a deeper appreciation for intricate relationships within marine ecosystems, aiding in sustainable practices and biodiversity protection.

Acknowledgements: The authors acknowledge the FCT Foundation for funding their research through projects UIDB/00013/2020 and UIDP/00013/2020. They also express their appreciation to MAR2020 for funding the SARDINHA2020 project (MAR-01.04.02-FEAMP-0009). The survey data analysed was collected under the framework programme PNAB: Portuguese Marine Surveying Programme - P03M02 (EU Data Collection Framework EU- DCF, FEAMP).

A hybrid robust-weighted AMMI modeling approach with generalized weighting schemes

Vanda M. Lourenço [a], Marcelo B. Fonseca [b], Paulo C. Rodrigues [b]
vmml@fct.unl.pt, marcelo_seiya@hotmail.com , paulocanas@gmail.com

[a] *Center for Mathematics and Applications (CMA), FCT NOVA and Department of Mathematics, FCT NOVA, Caparica, Portugal*
[b] *Department of Statistics, Federal University of Bahia, Salvador, Brazil*

Keywords: AMMI model, $G \times E$ interactions, plant breeding, robust modeling, weighting schemes

Abstract: The AMMI model and its variations have proven to be excellent in identifying genotypes with specific adaptability and stability under certain environmental conditions, making it a valuable tool in crop improvement breeding programs. However, the presence of atypical data points in crop data, which can be a result of various sources, such as measurement errors, genotype characteristics, diseases, or climate phenomena, may seriously undermine the performance of the AMMI model, as these data points most times interfere with the underlying assumptions of the model (e.g., the violation of the normality assumption) [1, 2]. It is, therefore, crucial that the AMMI model is equipped with statistical tools that enable it to provide reliable inferential results even when small departures from the model's assumptions occur, so that the AMMI model can maintain its effectiveness in supporting decisions related to crop improvement. This work proposes a hybrid AMMI modeling framework (RW-AMMI) that combines robust and weighted algorithms to model the genotype by environment interaction. Additionally, we introduce a comprehensive set of nine weighting schemes for the weighted (W-AMMI; [1]), robust (R-AMMI; [2]), and robust-weighted AMMI (RW-AMMI) models. To evaluate the performance of our proposed approach, we conduct a Monte Carlo simulation considering both contaminated and uncontaminated data with and without heterogeneous error variance, and compare the proposed method against the AMMI, W-AMMI, and R-AMMI models while using the nine weighting schemes. Furthermore, the effectiveness of our approach is validated via a real crop data application.

References

[1] Rodrigues, P.C., Malosetti, M., Gauch Jr, H.G., van Eeuwijk, F.A. A Weighted AMMI Algorithm to Study Genotype-by-Environment Interaction and QTL-by-Environment Interaction. *Crop Science*, 54, 1555–1570, 2014. DOI:10.2135/cropsci2013.07.0462

[2] Rodrigues, P.C., Monteiro, A., Lourenço, V.M. A Robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. *Bioinformatics*, 32, 58–66, 2016. DOI:10.1093/bioinformatics/btv533

The use of statistical techniques to evaluate the impact that different chocolates have on the sensory perception of three different categories of Port wine

Elisete Correia [a], Gabriela Santos [b], Alice Vilela [c]
ecorreia@utad.pt, gabifreitasantos@gmail.com, avimoura@utad.pt

[a] *CEMAT, Dep. of Mathematics, UTAD, 5001-801 Vila Real, Portugal*
[b] *Student of Enology Master, UTAD, 5001 801 Vila Real, Portugal*
[c] *CQ-VR, Dep. of Biology and Environment, UTAD, 5001-801 Vila Real, Portugal*

Keywords: chocolate, factor analysis, multivariate analysis of variance, Port wine, sensory evaluation

Abstract: Port wine is a Portuguese fortified wine produced exclusively in the Douro Valley. The wine produced is fortified by the addition of aguardente, which helps stop the fermentation leaving residual sugar in the wine, as well as increasing the alcohol content. Before being bottled, it is stored and aged in wood anywhere from 2 years to many decades, often in barrels stored in a cellar at low temperatures and a high degree of humidity. To evaluate, through Quantitative Descriptive Analysis (ADQ) and Temporal Dominance of Sensations (TDS), the impact that different chocolates have on the sensory perception of assorted styles of Port wines a panel of tasters performed the taste analysis of three Port wines (white reserve Port, 20-year-old Tawny Port and LBV 2015 Port) before and after tasting the chocolates (chocolate with fleur de sel, chocolate with almonds and cranberries and 70% cocoa chocolate). To characterize the sensory profile of the three Port wines after tasting the chocolates statistical techniques such as factor analysis and multivariate analysis of variance were used. The results obtained by the ADQ method showed that chocolate had significant effects on the flavor of some of the wines, namely on the bitterness, alcohol, floral, and honey attributes, the 20-year-old Tawny Port wine was the only one in which it did not show statistically significant effects. The results obtained by the TDS method showed that this method was sensitive to assessing the impact that different chocolates have on the sensory perception of Port wines.

Acknowledgements: The authors would like to thank CQ-VR UIDB/00616/2020 and UIDP/00616/2020, CEMAT/IST-ID UIDB/04621/2020 and UIDP/04621/2020 for its financial support and the financial support provided by national funds through FCT-Portuguese Foundation for Science and Technology. The authors also like to acknowledge the tasting panels that participated in the work. This study was funded by the project e-Flavor POCI-01-0247-FEDER-049337, funded by the European Regional Development Fund, through the COMPETE2020.

Estilo de vida e bem-estar dos estudantes do Politécnico de Leiria

Daniel Santos[a], Rui Santos[b,c], Susana Ferreira[b]
danielaubesa@gmail.com, rui.santos@ipleiria.pt, susfer@ipleiria.pt

[a] *Estudante do Mestrado em Ciência de Dados da ESTG–Politécnico de Leiria*
[b] *Escola Superior de Tecnologia e Gestão, Politécnico de Leiria*
[c] *CEAUL – Centro de Estatística e Aplicações, Universidade de Lisboa*

Keywords: ansiedade, classificação, depressão, inquérito, validação

Abstract: A entrada no Ensino Superior traduz-se num ponto de mudança na vida de um estudante. Frequentemente associado ao afastamento do círculo familiar e de amigos, bem como a uma maior exigência das suas tarefas, acompanhado de um aumento da autonomia nas escolhas sobre a sua própria vida. Todos estes fatores podem afetar atitudes e comportamentos relacionados com o estilo de vida e o bem-estar, nomeadamente em termos de alimentação, hábitos de desporto, consumo de álcool e/ou estupefacientes, comportamento sexual, entre outros.

Após os confinamentos associados à pandemia, têm sido reportados diversos estudos realizados nestas idades e os resultados têm sido alarmantes. Com o objetivo de perceber e caraterizar o estilo de vida e bem-estar dos estudantes do Politécnico de Leiria, foi aplicado um inquérito que inclui três conhecidas escalas de avaliação, o PHQ-2 (*Patient Health Questionnaire*) para avaliar sintomas depressivos, o GAD-7 (*Generalized Anxiety Disorder*) para avaliar o nível de ansiedade e o SMILE (*Short Multidimensional Inventory Lifestyle Evaluation*) para avaliar o estilo de vida. Esta última escala surgiu do trabalho da equipa de Balanzá-Martínez, tendo sido aplicado a várias populações nomeadamente nas espanhola e brasileira.

Neste trabalho são apresentadas as principais conclusões da análise aos resultados obtidos, em termos da descrição do estilo de vida e bem-estar dos estudantes do Politécnico de Leiria, das diferenças entre categorias definidas pelo género, pela Escola ou tipologia de curso (utilizando testes t, ANOVA, Mann-Whitney e Kruskal-Wallis), da validação do questionário (Alfa de Cronbach), bem como da aplicação de metodologias de classificação (regressão logística e árvores de decisão) de forma a identificar problemas de depressão e/ou de ansiedade com base nas respostas ao inquérito SMILE. Por fim, a fiabilidade das classificações obtidas são comparadas (matriz de confusão, acurácia, sensibilidade, especificidade, valores preditivos e área sob a curva ROC).

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] Simjanoski M., Ballester P.L., da Mota, J.C., De Boni, R.B., Balanzá-Martínez, V., Atienza-Carbonell, B., *et al.* Lifestyle predictors of depression and anxiety during COVID-19: a machine learning approach, *Trends Psychiatry Psychother*, 44:e20210365, 2022. doi:10.47626/2237-6089-2021-0365

Risk assessment of vulnerabilities exploitation

Fernando Sequeira[a,b], M. Fátima Brilhante[b,c], Pedro D. Pestana[d,e], M.L. Rocha[f,g]
fjsequeira@ciencias.ulisboa.pt, maria.fa.brilhante@uac.pt, pedro.pestana@uab.pt, maria.ls.rocha@uac.pt

[a] *DEIO, Faculdade de Ciências, Universidade de Lisboa*
[b] *Centro de Estatística e Aplicações, Universidade de Lisboa (*CEA/UL*)*
[c] *Faculdade de Ciências e Tecnologia, Universidade dos Açores*
[d] *DCT, Universidade Aberta*
[e] *Centro de Investigação em Ciência e Tecnologia das Artes* (CITAR)
[f] *Faculdade de Economia e Gestão, Universidade dos Açores*
[g] *Centro de Estudos de Economia Aplicada do Atlântico* (CEEAplA)

Keywords: CVSS, EPSS, risk assessment, scoring systems, vulnerabilities

Abstract: Vulnerabilities are weaknesses that can be exploited by cybercriminals to gain unauthorized access to a computer system, eventually running malicious code, installing malware, and stealing sensitive data. Nowadays more than 20,000 new vulnerabilities are reported yearly. CVSS (Common Vulnerabilities Scoring System) [1], using Base, Temporal and Environmental metrics, is a useful scoring system to prioritize the urgency of patching or working around vulnerabilities, since extreme economic impact may result from exploitation [2]. It is however recognized that the mathematical treatment is overly complicated and counterintuitive, leading to insufficient granularity (in practice, 99 possible CVSS scores), and that scores published by vendors are too often High or Critical. Moreover, the CVSS Temporal metrics do not impact in practice the final CVSS score, which is static, while it is obvious that risk evolves with time. In order to shift from static to dynamic scores, we use variables from the life cycle of vulnerabilities, whose larger values may come from extreme value models, eventually subject to geometric thinning; this suggests fitting models such as the General Extreme Value, the Generalized Pareto or Loglogistic distributions, although the traditional Lognormal or other heavy tailed models with paretian tails should also be considered. The ultimate goal is to devise methodologies, whenever possible with machine learning implementation, to alter the CVSS score dynamically.

Acknowledgements: Research partially supported by National Funds through FCT, Fundação para a Ciência e a Tecnologia, Portugal, project UIDB/00006/2020 (CEAUL) and by HiTEc Cost Action CA21163.

References

[1] CVSS — *Common Vulnerability Scoring System* version 3.1: Specification Document: https://www.first.org/cvss/specification-document.

[2] Eling, M., Elvedi, M., Falco, G. The Economic Impact of Extreme Cyber Risk Scenarios. *North American Actuarial Journal*, 1–15, 2022.

Probabilistic procedures for SIR and SIS epidemic dynamics on contact random networks

J. Leonel Rocha [a,b], S. Carvalho [a,b], B. Coimbra [a,b]
jose.rocha@isel.pt; sonia.carvalho@isel.pt; a47549@alunos.isel.pt

[a] *CEAUL-Centre of Statistics and its Applications*
[b] *Department of Mathematics, ISEL-IPL*

Keywords: epidemic threshold, Erdös-Rényi networks, infectious disease, probabilistic procedure, SIR and SIS models

Abstract: Dynamic processes analysis in random contact networks explains the evolution of real propagation and diffusion phenomena, namely ideas, information, influence and epidemics. Understanding the effect of connections in a given population, allows us to comprehend and identify how diseases are likely to spread from one individual to another, concerning several different aspects. In this sense, social contact networks can be modelled using complex random networks, where epidemic phenomena are simulated through the SIS and SIR models. In this work we study the epidemic threshold dynamics for the SIR and SIS models, where some properties are proved, over Erdös-Rényi contact random networks. Probabilistic procedures for these epidemic models are established. Aiming the analysis of the influence maximization problem, some probabilities regarding the triggering of the infectious state are computed.

Acknowledgements: This research was funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020 (CEAUL) and ISEL. B. Coimbra is funded by a research grant from CEAUL, Center of Statistics and Applications of the University of Lisbon, for Masters Students in FCiências.ID - Association for Research and Development of Sciences, within the scope of the Project UIDP/00006/2020.

References

[1] Barabási, A-L. *Network Science*. Cambridge University Press, Cambridge, UK, 2016. doi:10.1177/0094306116681814

[2] Bollobás, B. *The Evolution of Random Graphs, The Giant Component. In Random Graphs*. Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2001. doi:10.1017/CBO9780511814068.008

[3] Erdös, P., Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5, 17–61, 1960. doi:10.2307/1999405

[4] Pastor-Satorras, R., Castellano, C., Mieghem, P.V., Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87, 925–979, 2015. doi:10.1103/RevModPhys.87.925

[5] Rocha, J.L., Carvalho, S., Coimbra, B. Probabilistic Procedures for SIR and SIS Epidemic Dynamics on Contact Random Networks. submitted.

Multi-state modeling of composite indexes for assessing the economic conditions of firms. A comparative study between energy and non-energy Portuguese firms

Gustavo Soutinho [a], Vítor M. Ribeiro [a], Isabel Soares [a]
gdsoutinho@gmail.com, vsribeiro@fep.up.pt, isoares@fep.up.p

[a] *Faculty of Economics, University of Porto*

Keywords: composite index, economic shocks, energy, multi-state modeling

Abstract: Over the last two decades, the economic activities of Portuguese firms were affected by several socioeconomic, financial, and political crises, such as the mortgage crisis period, the troika intervention period, or the pandemic crisis, which contributed to the permanent turmoil of the national economy. Nevertheless, there is a lack of studies addressing the impacts of these economic shocks on the performance of firms. To tackle this, our study aims to investigate the progression of the health economic conditions regarding economic indicators such as economic returns, liquidity, debt or profit margin. To this end, we used the Iberian Balance Sheet Analysis System (SABI) dataset, which includes financial information of more than 900,000 Portuguese firms. Methodology: we started to derive composite indexes using the principal component analysis and some adaptations of the Laspeyres and Paasche indexes making use of the economic indicators for the data aggregation. Afterwards, for each year since 2002, the values of the index were split into four intervals, each one representing the health economic state of the firms. Finally, several inference methods in multi-state models were implemented to analyze the multiple events (e.g. regression models and transition probabilities). Results confirmed that firms belonging to the energy sector were characterized by intermediate positive conditions that lasted for long periods of time. Nevertheless, these energy firms also presented an accentuated fall of performance in short periods of times that led to a low positive or even negative economic conditions. Disaggregating by the sections of the CAE 3 Rev. that comprise the energy sector, this work also revealed that the positive economic conditions over time were mainly explained by electricity, gas or vapor firms (Section D). Quick movements that occurred that led energy firms to low or negative economic conditions were fundamentally due to energy industry firms (Section B).

References

[1] Booysen, F. An Overview and Evaluation of Composite Indices of Development. *Social Indicators Research*, 59, No. 2, 115–151, 2002. DOI: 10.1023/A:1016275505152.

[2] Everitt, B.S., Dun, G. Applied Multivariate Data Analysis. Edward Arnold, London, 97–126, 1991. DOI: 10.1177/096228029300200109.

[3] Meira-Machado, L., Sestelo, M. Estimation in the progresive illness-death model: A nonexhaustive review. *Biometrical Journal*, 61, No. 2, 245–263, 2009. DOI: 10.1002/bimj.201700200.

Enhancing gait analysis through transformation-based multiple linear regression normalization

Jhonathan Barrios [a], Bárbara Araújo[b], Estela Bicho[b], Miguel F. Gago [c,d], Wolfram Erlhagen[a], Flora Ferreira [a]
id10605@uminho.pt, pg47037@alunos.uminho.pt, estela.bicho@dei.uminho.pt, miguelgago@hospitaldeguimaraes.min-saude.pt, wolfram.erlhagen@math.uminho.pt, fjferreira@math.uminho.pt.

[a] *Centre of Mathematics, School of Sciences, University of Minho*
[b] *Algoritmi Centre, School of Engineering, University of Minho*
[c] *Neurology Department, Hospital da Senhora da Oliveira*
[d] *ICVS, School of Medicine, University of Minho*

Keywords: gait analysis, multiple linear regression, nonlinear dynamics, normalization, transformation

Abstract: Gait analysis is increasingly recognized as a critical clinical tool that provides objective assessments of gait impairment and aids in monitoring disease progression and the efficacy of therapeutic interventions. Yet, achieving precise comparisons across individuals remains challenging due to the influence of personal physical characteristics on gait measures. Multiple linear regression (MLR) normalization has been employed to mitigate this interference [1], but its limited ability to capture non-linear effects can impact normalization accuracy. This work aims to explore the potential of transforming independent variables using various mathematical functions to improve MLR models for gait normalization. Gait data from a cohort of 36 healthy adults were utilized. MLR models were constructed using gait variables as independent predictors, and mathematical transformations such as logarithm, square root, square, and cube were applied to enhance normalization performance. Model fit was assessed using Akaike's information criterion (AIC), adjusted R^2, and variance inflation factors (VIFs). Further, we evaluated the correlation of physical properties and gait features using Spearman's rank-order correlation coefficient and point biserial correlation. Finally, Principal Component Analysis (PCA) and k-means clustering were employed to discern distinct clusters of subjects based on normalized gait characteristics from both the MLR models described in [1] and our new MLR models. Preliminary results suggest that the transformed MLR models significantly improve gait normalization, effectively reducing correlations between subject-specific physical characteristics and gait features, further enhancing the potential for accurate comparisons and objective gait assessments.

References

[1] Fernandes, C., Ferreira, F., et al. Discrimination of idiopathic Parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect, *Journal of Biomechanics*, vol. 125, pp. 110214,2021,doi.org/10.1016/j.jbiomech.2020.110214..

What statistics can reveal about occupant behavior in diverse building types

Lumy Noda [a], Celina Pinto Leão, [b], Solange Leder [a]
barbara.lumy@academico.ufpb.br, cpl@dps.uminho.pt,
solangeleder@yahoo.com.br

[a] *Postgraduate Program in Architecture and Urbanism, Federal University of Paraíba, Brazil*
[b] *Centro Algoritmi/LASI, School of Engineering University of Minho, Portugal*

Keywords: buildings, occupant behavior, office building, thermal comfort, work-from-home

Abstract: There has been a growing interest in investigating the relationship between occupant behavior and building energy performance as it plays a crucial role in energy consumption and affects energy estimates deviations. Despite extensive research, accurately predicting occupant energy behavior remains challenging, and data transferability across different buildings remains understudied [1]. To address this, a study on adaptive thermal comfort was conducted in a hot and humid climate city in Brazil. Qualitative and quantitative data were collected from a field survey. The occupant behavior was compared in an office building (n=183) and residential buildings with teleworking participants (n=130), all engaged in the same working activity. The added value lies in using data from 50 common participants in both building typologies. In addition to the descriptive statistical analysis, the study employed Pearson's correlation coefficient and the chi-square test of independence to validate significant differences in the behavior of occupants concerning the actions performed and their combinations, depending on the type of building, with a p-value less than the significance level of 0.05. The study highlights that occupants' thermal comfort-related behavior in office buildings cannot be generalized to residential settings, due to the influence of indoor air conditioning and adaptive control measures on occupants' motivations and actions. This research sheds light on the importance of understanding occupants' interactions in different building typologies to enhance energy efficiency strategies and building design.

Acknowledgements: Grant CAPES PDSE, Brazil, process 88887.717394/2022-00, and FCT – RD Units Project Scope: UIDB/00319/2020.

References

[1] Xu, X., Yu, H., Sun, Q., Tam, V.W.Y. A critical review of occupant energy consumption behavior in buildings: How we got here, where we are, and where we are headed. *Renewable and Sustainable Energy Reviews*, 182, 113396, 2023. doi:10.1016/j.rser.2023.11339

Exploring the influence of various factors on salaries: A regression analysis approach

Flora Ferreira [a,b], Ana Pedrosa [c], Ana Borges [c], José Maria Soares[a], Pedro Pacheco[a]
fjferreira@math.uminho.pt, 8180052@estg.ipp.pt, aib@estg.ipp.pt,
jose.soares@berd.eu, pedro.pacheco@berd.eu

[a] *Centre of Mathematics, University of Minho, Guimarães, Portugal*
[b] *BERD - Bridge Engineering Research and Design, Matosinhos, Portugal*
[c] *CIICESI, ESTG, Polytechnic of Porto, 4610-156 Felgueiras, Portugal*

Keywords: correlation analysis, multiple linear regression, salary estimation

Abstract: Salaries are intricately influenced by various factors including the worker experience and the economic situation of a country. This study focuses on the application of regression analysis to examine the influence of a comprehensive set of factors on the salaries of different job categories. The selection of factors was based on their accessibility and likelihood of impacting salary levels. Chosen factors include average wage, years of experience, and economic variables such as Gross Domestic Product (GDP) per capita, unemployment rate, and inflation rate. For the analysis, a dataset of 8316 salary values for 9 job categories across 14 countries from 2017 to 2021 was used. This dataset is a subset of a larger dataset available on Kaggle, titled "Salary by Profession and Country Over Time" by Kelly Garrett[2]. To explore the relationships between the factors and salaries, correlation analysis techniques such as Pearson correlation and Variance Inflation Factor (VIF) were employed to detect multicollinearity. Multiple Linear Regression models were constructed for each job category, considering all combinations of independent variables that were not correlated and exhibited a linear relationship with the dependent variable. The selection of the best-fit models was based on Akaike's information criterion (AIC) and adjusted R^2 metrics, leading to the identification of three top-performing models. These models individually and combined were evaluated to estimate salary values. By employing multiple linear regression, this study establishes an approach to accurately estimate salary values using easily accessible factors. The findings of this research have the potential to enhance our understanding of the factors influencing salaries and offer valuable insights for decision-making processes related to salary determination in various contexts.

Acknowledgements: Supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within the projects UIDB/00013/2020, UIDP/00013/2020 and UIDB/04728/2020.

[2]https://www.kaggle.com/datasets/thedevastator/uncovering-global-data-professional-salary-trend [Accessed: 6-May-2023]

Risk analysis for evaluating the water quality of a hydrological basin

Ana Cristina Pedra [a], A. Manuela Gonçalves [a], Irene Brito [a]
pg46704@alunos.uminho.pt, mneves@math.uminho.pt, ireneb@math.uminho.pt

[a] *Department of Mathematics, Centre of Mathematics, University of Minho*

Keywords: clustering, Douro River basin, risk measures, time series, water quality

Abstract: Water is a limited, irreplaceable and indispensable natural resource. Statistical methods are important tools for controlling and forecasting changes in the management process of water quality. The main objective of this study is to develop new methodologies, combining concepts from risk theory and times series modelling approaches, for the forecasting of environmental variables in the context of water quality assessment, and to propose new risk indexes of pollution levels. The methodologies are illustrated using a data set of the Douro River basin (in Portugal) in terms of environmental water quality variables, measured monthly in 18 water quality sampling stations and recorded in the period from January 2002 to December 2013. Risk measures, such as entropy, value at risk, probability of excess, will be determined for the considered variables in order to assess the risk of water pollution taking into account the monthly nature of the data. A clustering analysis will be performed to group similar sampling stations considering the different environmental variables, and by taking into account the seasonal variations. The risk measures will be used to classify the clusters. A time series analysis with approaches of Box-Jenkins models and their extensions, the classical decomposition time series associated with multiple linear regression models with correlated errors, and the Holt-Winters method, of the environmental data, corresponding to the specific clusters, will be carried out and predictions for future months will be obtained. The most adequate risk measures will be identified for the forecasting of pollution risk.

Acknowledgements: The authors thank support from FCT through the Projects UIDB/00013/2020 and UIDP/00013/2020. Ana Cristina Pedra thanks support from CMAT through the grant UMINHO/BIM/2022/100.

References

[1] Brachinger, H.W., Weber, M. Risk as a primitive: A survey of measures of perceived risk. *Operations-Research-Spektrum*, 19, 235-250, 1997. doi:https://doi.org/10.1007/BF01539781

[2] Ganoulis, J. *Risk Analysis of Water Pollution*. Wiley, 2009.

[3] Gonçalves, A.M., Costa, M. Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research & Risk Assessment*, 27(5), 1021-1038, 2013. doi:https://doi.org/10.1007/s00477-012-0640-7

[4] Sistema Nacional de Informação de Recursos Hídricos - SNIRH, 2023. https://snirh.apambiente.pt/

Seleção de variáveis em misturas de modelos de regressão linear: um estudo de simulação

Ana Moreira [a], Susana Faria [a]
id10866@uminho.pt, sfaria@math.uminho.pt

[a] *Departamento de Matemática, Universidade do Minho, Portugal*

Keywords: algoritmo CEM, algoritmo EM, estimação por máxima verosimilhança penalizada, estudo de simulação, misturas de modelos de regressão linear

Abstract: A seleção de variáveis constitui uma etapa crucial na construção de um modelo de regressão. Neste trabalho estuda-se o problema da seleção de variáveis em misturas de modelos de regressão linear recorrendo à estimação por máxima verosimilhança penalizada considerando os algoritmos *Expectation-Maximization*(EM) e *Classification Expectation-Maximization* (CEM) e realizando um estudo de simulação.

Os modelos de regressão de misturas finitas fornecem uma ferramenta flexível para modelar dados que surgem de uma população heterogénea. Nas aplicações desses modelos, um grande número de variáveis explicativas é frequentemente considerado, por este motivo, a seleção de variáveis assume uma grande relevância para os modelos de mistura.
Neste trabalho analisa-se o problema da seleção de variáveis em misturas de modelos de regressão linear na presença de um grande número de variáveis explicativas, recorrendo a uma abordagem de verosimilhança penalizada e usa-ndo os algoritmos EM e CEM ([1]) na estimação dos parâmetros. Estudam-se diferentes métodos de seleção baseados em funções de penalização, em particular, o método *Adaptive Least Absolute Shrinkage and Selection Operator* (ALASSO) ([2]) e o método *elastic Net* ([3]), comparando o seu desempenho na seleção de variáveis explicativas. Através de um extenso extudo de si-mulação, conclui-se que diferentes cenários simulados influenciam o desempenho dos diferentes algoritmos e das funções de penalização aplicadas.

Acknowledgements: Este trabalho foi financiado pela FCT - Fundação para a Ciência e a Tecnologia, através da bolsa de doutoramento com a referência 2022.12256.BD.

References

[1] Celeux, G., Govaert, G. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3), 315–332, 1992. doi:10.1016/0167-9473(92)90042-E

[2] Zou, H. The Adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429, 2006. doi:10.1198/016214506000000735

[3] Zou, H., Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320, 2005. doi:10.1111/j.1467-9868.2005.00503.x

Microbial diversity and discrimination of Azeitão and Nisa PDO cheeses based on metagenomic data

Carlota Teles [a,b], Lisete Sousa [a,b], Sílvia Rebouças [c, d],
Maria Teresa Barreto Crespo [e,f], Teresa Semedo-Lemsaddek [g,h]
fc51511@alunos.ciencias.fc.ul.pt, lmsousa@ciencias.ulisboa.pt, smdpedro@gmail.com, tcrespo@ibet.pt, tlemsaddek@fmv.ulisboa.pt

[a] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências*
[b] *FCUL – Faculdade de Ciências, Universidade de Lisboa*
[c] *ISMAT - Instituto Superior Manuel Teixeira Gomes*
[d] *COPELABS - Cognitive and People-centric Computing R&D Unit*
[e] *Instituto de Tecnologia Química e Biológica António Xavier*
[f] *iBET, Instituto de Biologia Experimental e Tecnológica*
[g] *CIISA – Centre for Interdisciplinary Research in Animal Health*
[h] *Associate Laboratory for Animal and Veterinary Sciences (AL4AnimalS)*

Keywords: metagenomics, microbiota, portuguese traditional cheeses

Abstract: Studies focusing on the analysis of the microbiota present in traditional cheeses are crucial steps towards understanding their organoleptic properties, quality, and safety. The data in this study consists of 24 samples of traditional Protected Designation of origin (PDO) cheeses, collected from two PDO cheese producers in Nisa and four producers in Azeitão, Portugal, in 2021. One objective is to compare the microbiota of the PDO cheeses based on their production regions: Principal Coordinate Analysis (PCoA) and Non-Metric Multidimensional Scaling (NMDS) were employed. Additionally, statistical tests such as Analysis of Similarities (ANOSIM) and Permutation Multivariate Analysis of Variance (PERMANOVA) were applied to determine significant differences in the composition of the cheeses from the two regions. The other main objective is to identify the bacteria that play a role in differentiating the microbiota of the PDO cheeses. This will be accomplished by employing classification techniques, such as Linear Discriminant Analysis Effect Size (LEfSe), Lasso regression and Random Forests. Both ANOSIM ($p = 0.003$) and PERMANOVA ($p = 0.035$) tests confirmed significant differences between cheeses from the two regions. Based on the histogram of LDA scores provided by LEfSe, *Lacticaseibacillus, Serratia*, and *Pseudomonas* were found to be more abundant in PDO cheeses produced in Nisa, while *Enterococcus, Carnobacterium, Leuconostoc, Hafnia, Corynebacterium*, and *Latilactobacillus* were characteristic of PDO cheeses produced in Azeitão.

Acknowledgements: This study was supported by FCT—Fundação para a Ciência e Tecnologia IP Portugal, through projects PTDC/OCE-ETA/1785/2020 [EMOTION], UIDB/00276/2020 (CIISA) and LA/P/0059/2020-AL4ANIMALS (AL4AnimalS). Teresa Semedo-Lemsaddek is financially supported by national funds through FCT under the Transitional Standard–DL57/2016/CP1438/CT0004.

Unveiling gene signatures in glioma: A comprehensive analysis using regularized logistic regression, dimensionality reduction, and outlier detection

João F. Carrilho [a,b], Roberta Coletti [b], Marta B. Lopes [a,b,c]
jf.carrilho@campus.fct.unl.pt, roberta.coletti@fct.unl.pt,
marta.lopes@fct.unl.pt

[a] *Department of Mathematics, NOVA School of Science and Technology, Portugal*
[b] *Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Portugal*
[c] *UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Portugal*

Keywords: elastic net, feature selection, glioma, outlier detection, robust multinomial logistic regression

Abstract: Gliomas are among the most aggressive tumors, showing a large intertumoral heterogeneity, making it crucial the search for novel biomarkers and targeted therapies. The knowledge about this disease is rapidly evolving, leading to recent updates in the glioma classification guidelines by the World Health Organization (WHO). The objective of this work was to identify transcriptomic and methylomic biomarkers of glioma heterogeneity, aiming at contributing to the understanding and ultimately to the management of this disease.
The analysis considered transcriptomics and methylomics datasets obtained from The Cancer Genome Atlas (TCGA), with patients' diagnoses updated according to the 2016 and 2021 WHO guidelines. Sparse and robust multinomial logistic regression models revealed a noticeable improvement in class separability for the 2021 update, identifying as outliers patients that changed their diagnostic labels in 2021. A cellwise inspection of the outlying patients showed that those verify a generally diverging gene expression profile from the class they belong to. Dimensionality reduction via the Uniform Manifold Approximation and Projection (UMAP) technique also provided a better distinction between glioma types with the 2021 classification, both before and after feature selection. Classification was also used to explore a set of features selected by a network-based methodology based on transcriptomics data, to assess if these allow the glioma-type distinction. Our results uncovered relevant biological information regarding glioma heterogeneity with high accuracy, with information extracted from the transcriptomics and methylomics patient profiles better supporting 2021 WHO glioma classification guidelines.

Acknowledgements: This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references CEECINST/00042/2021, UIDB/00297/2020 and UIDP/00297/2020 (NOVA MATH), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and developed within the project "MONET: Multiomic networks in gliomas" (PTDC/CCIBIO/4180/2020).

Desenho e implementação de um questionário: uma avaliação do grau de satisfação dos utilizadores BPLIM

Eduardo Dias [a], Rita Sousa [b], Inês Sousa [a]
a95467@alunos.uminho.pt, rcsousa@bportugal.pt, isousa@math.uminho.pt

[a] *Universidade do Minho*
[b] *Banco de Portugal*

Keywords: amostragem aleatória estratificada, bplim, microdados, questionário

Abstract: Nos dias de hoje já não é possível compreender o mundo complexo no qual vivemos analisando apenas dados agregados. É necessário estudar os dados estatísticos individuais sobre pessoas singulares e coletivas, os microdados, que se podem obter a partir de sondagens, censos ou mesmo sistemas administrativos. O Laboratório de Investigação em Microdados do Banco de Portugal é uma unidade autónoma do Banco de Portugal. Este fornece acesso às suas bases de microda dos (anonimizadas) a investigadores internos e externos, bem como assistência e orientação para a sua análise.
Na primeira parte do trabalho, foi criado um questionário eletrónico com a intenção de averiguar o grau de satisfação dos investigadores que utilizaram os serviços do BPLIM. Foi realizada uma amostragem aleatória estratificada de forma a dividir a população em estratos, em função do género e do tipo de investigador (interno ou externo). Seguidamente, selecionam-se aleatoriamente elementos de cada estrato de forma a obter dois grupos representativos da população. Posteriormente, foi enviado, via e-mail, um questionário para ser preenchido de forma anónima aos elementos de um dos grupos e um não anónimo aos elementos do outro grupo.
Na segunda parte, foram realizadas análises e inferências estatísticas aos resultados obtidos, de forma a determinar quais os pontos a melhorar no serviço prestado pelo BPLIM, bem como, para averiguar se o facto do questionário ser anónimo levaria a uma maior taxa de resposta e a avaliações superiores às obtidas através dos questionários não anónimos. Por fim, foram também aplicados diversos modelos de regressão logística para tentar compreender o que caracteriza um investigador que se considera totalmente satisfeito pelo serviço prestado pelo BPLIM.

References

[1] Hilbe, J. *Logistic Regression Models*. Chapman & Hall/CRC, New York, 2009. doi:10.1201/9781420075779

[2] Cochran, W. Sampling Techniques. John Wiley & Sons, New York, 1977.

[3] Hosmer, D., Lemeshow, S. *Applied Logistic Regression*. John Wiley & Sons, 2000. doi:10.1002/0471722146

As diferenças de valores humanos em função da religiosidade, na Europa

Maria Paula Lousão [a,b,c], Cláudia Silvestre [a,d], José Luís Casanova [b,c]
mlousao@sp.ipl.pt, csilvestre@escs.ipl.pt, jose.casanova@iscte-iul.pt

[a] *Politécnico de Lisboa - Escola Superior de Comunicação Social (IPL-ESCS)*
[b] *Centro de Investigação e Estudos de Sociologia (CIES-IUL)*
[c] *Instituto Universitário de Lisboa (ISCTE-IUL)*
[d] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*

Keywords: análise classificatória hierárquica, análise de correspondências múltiplas, informação mútua, religiosidade, valores humanos.

Abstract: O objetivo desta comunicação é avaliar a perceção dos Valores Humanos (VH) e como a religião poderá ser um fator influenciador. Para tal usou-se os dados do European Social Survey, que é um inquérito transnacional de cariz académico realizado na Europa de dois em dois anos, desde 2001. Um dos objetivos deste inquérito é avaliar a mudança na estrutura social, nas condições e atitudes da população europeia. Neste estudo centrado na perceção dos VH usou-se apenas uma parte das variáveis disponíveis na base de dados. Começou-se por explorar associações entre o tipo de relação com a religião, as várias práticas religiosas e o pertencer ou não a uma religião, através duma análise de correspondências múltiplas.
Os VH foram obtidos pela escala desenvolvida por Schwartz. Consideraram-se os 10 tipos motivacionais que podem ser classificados em quatro tipos motivacionais de segunda ordem, organizados em dois eixos concetuais básicos: Autotranscendência versus Autopromoção e Abertura à Mudança versus Conservadorismo. Com o objetivo de avaliar a (dis)semelhança entre as religiões, face à sua posição relativamente aos VH, optou-se pela análise de agrupamento hierárquico. Foram utilizados 3 métodos de agregação: o método de Ward, o método dos k vizinhos mais próximos e o método da ligação média. Os segmentos obtidos através destes métodos foram semelhantes. Com o intuito de perceber se existiam diferenças entre a perceção dos VH por parte das pessoas que são religiosas e as que não são, calculou-se a informação mútua entre esses dois grupos de pessoas e os VH depois de previamente agrupados em 3 classes.

References

[1] Carneiro, A., Sousa, H.F.P., Dinis, M.A.P., Leite, A. Human Values and Religion: Evidence from the European Social. *Social Science*, 10(2), 75, 2021. https://doi.org/10.3390/socsci10020075

[2] Lebart, L., Morineau, A., Piron, M. *Statistique Exploratoire Multidimensionelle*. Dunod, Ed., Paris, 1995.

[3] Le Roux, B., Rouanet, H. *Multivariate Correspondence Analysis*. Sage Publications, London, 2010.

Modelagem logística para ajuste de dados em pacientes diagnosticados com neoplasia

Jorge Alves de Sousa [a], Jossé Joedson Lima de Sousa [a], Anselmo Ribeiro Lopes [a]
jorge.alves@professor.ufcg.edu.br, joedson.studs@gmail.com,
anselmo.ribeiro@professor.ufcg.edu.br

[a] *Universidade Federal de Campina Grande*

Keywords: modelos lineares generalizados, neoplasia maligna, regressão logística, variáveis independentes

Abstract: Em estudos na área de saúde, há interesse em se verificar, por exemplo, quais variáveis são fatores de risco para um paciente apresentar ou não uma doença, ou se determinado tratamento produziu ou não o efeito esperado, também podem ser avaliadas a influência de algumas variáveis em doenças como o câncer, cariciando dois diagnósticos, neoplasias malignas ou benignas. Nesses casos, a variável resposta é do tipo binária. Os fatores de risco são as variáveis independentes ou variáveis explicativas. Para modelar essa relação, entre a resposta e as variáveis independentes, existe uma classe de modelo, dentro dos modelos lineares generalizados (MLGs), denominado modelo de regressão logística. O presente estudo, teve por objetivo modelar a partir de uma base não identificada de pacientes diagnosticados com neoplasia maligna e benigna em função das variáveis: sexo, idade, tempo de tratamento e tipo de tratamento. Os dados foram ajustados a distribuição Binomial utilizando a função de ligação de Cauchy pelo menor valor de AIC (Bozdongan, 1987), apenas às variáveis sexo e tipo de tratamento foram probabilisticamente significativas, com a bondade de ajuste do modelo sendo comprovada pelo teste Hosmer-Lemesshowe, 1985. Uma calibração entre sensitividade a especificidade foi realizada utilizando a curva ROC, observemos que a sensitividade obtida com esta escolha foi de 78% e a especificidade é de 79%.

References

[1] Bozdongan. H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 68, 257–270, 2014.

[2] Hosmer, D.W., Lemeshow, S. *Applied logistic regression.* 2nd ed. John Wiley Sons, p. 156-64, New York, 1985.

Construção e análise espacial de um índice de desenvolvimento sustentável para os países da União Europeia

Conceição Ribeiro [a,b], Paula Pereira [a,c], Sílvia Pedro Rebouças [a,d,e,f]
cribeiro@ualg.pt, paula.pereira@estsetubal.ips.pt, smreboucas@ualg.pt

[a] *Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *Instituto Superior de Engenharia, Universidade do Algarve, Portugal*
[c] *Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, Portugal*
[d] *Escola Superior de Gestão Hotelaria e Turismo, Universidade do Algarve, Portugal*
[e] *ISMAT-Instituto Superior Manuel Teixeira Gomes, Portugal*
[f] *COPELABS-Centro de Investigação em Computação Centrada nas Pessoas e Cognição, Portugal*

Keywords: análise espacial, desenvolvimento sustentável, índice

Abstract: Pretende-se analisar a existência de diferenças, em termos de desenvolvimento sustentável, nos países da União Europeia, de acordo com os Objetivos de Desenvolvimento Sustentável (ODS). Neste sentido, a partir da construção de um índice de desenvolvimento sustentável, será realizada uma análise espacial com vista ao mapeamento das desigualdades entre os vários países em estudo. A construção do índice será feita com recurso à análise fatorial e para o mapeamento das diferenças será usada a modelação espacial.

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] Morrison, D.F. *Multivariate Statistical Methods*. 3rd edition, McGraw-Hill, N.Y, 1990.

[2] Jolliffe, J. T. *Principal component analysis*. SpringerVerlag,New York, 1986.

[3] Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. *Multivariate Data Analysis: With Readings*. London: Prentice Hall International, 1995.

[4] Banerjee, S., Carlin, B., Gelfand, A. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press. 2nd ed., 2014.

[5] Rebouças, S., Araripe-Silva, J., Ribeiro, C., Abreu, M. (2018). Building a sustainable development index and spacial. *Revista de Administração Pública*, 68, 257–270, 2014.

Tail independence: a comparative analysis of estimation methods

Sandra Dias [a], Marta Ferreira [b]
sdias@utad.pt, msferreira@math.uminho.pt

[a] *Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro*
[b] *Centro de Matemática, Universidade do Minho*

Keywords: coefficient of tail independence, extreme value theory, tail index estimation

Abstract: Extreme value theory is focused on developing methods for tail inference where data are scarce. It is thus not surprising that measures that are usually applied to evaluate the central part of data are not adequate in an analysis of extreme values. Such is the case of correlation to assess linear association between two variables. For example, in the context of the Gaussian model, no matter how strong the correlation is, the variables always show an asymptotic tail independence. In this work we address the tail independence coefficient η of Ledford and Tawn ([4], 1996), a measure to assess the presence of an extremal residual dependence. We will see that η can be estimated as a regular variation index, for which several estimators and inferential methods already exist. One of the problems involved in this topic is the selection of the optimal sample fraction to be considered in the estimation (Hall [3], 1990; Danielsson *et al.* [2], 2001; Caeiro and Gomes [1], 2016). Based on a simulation study, we will make a comparative analysis of different methodologies adapted to the estimation of η. We will finish with an illustration in real data.

References

[1] Caeiro, F., Gomes, M. I. Threshold Selection in Extreme Value Analysis. *Extreme Value Modeling and Risk Analysis: Methods and Applications*, Chapman and Hall/CRC, 69–86, 2016. doi:org/10.1201/b19721-5

[2] Danielsson, J., Haan, L., Peng, L., Vries, C.G. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2), 226–248, 2001. doi:org/10.1006/jmva.2000.1903

[3] Hall, P. Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. *Journal of Multivariate Analysis*, 32(2), 177–203, 1990. doi:org/10.1016/0047-259X(90)90080-2

[4] Ledford, A., Tawn, J. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1), 169-187, 1996. doi:org/10.1093/biomet/83.1.169

A crossinggram for random fields on lattices

Helena Ferreira [a,b], **Marta Ferreira** [a,b], Luís A. Alexandre [c]
helenaf@ubi.pt, msferreira@math.uminho.pt, lfbaa@di.ubi.pt

[a] *Universidade da Beira Interior, Centro de Matemática e Aplicações (CMA-UBI), Avenida Marquês d'Avila e Bolama, 6200-001 Covilhã, Portugal*
[b] *Centro de Matemática, Universidade do Minho*
[c] *Departamento de Informática and NOVA LINCS Universidade da Beira Interior, Covilhã, Portugal*

Keywords: extremal coefficients, extreme values, tail dependence coefficients, upcrossings

Abstract: The modeling of risk situations that occur in a space framework can be done using max-stable random fields on lattices. Although the summary coefficients for the spatial behaviour do not characterize the finite-dimensional distributions of the random field, they have the advantage of being immediate to interpret and easier to estimate. The coefficients that we propose give us information about the tendency of a random field for local oscillations of its values in relation to real valued high levels. It is not the magnitude of the oscillations that is being evaluated, but rather the greater or lesser number of oscillations, that is, the tendency of the trajectories to oscillate. We can observe smoother surface trajectories over a region according to a higher crossinggram value. It ranges in $[0, 1]$ and increases with the concordance of the variables of the random field.

Analysis of $M^X/M/c/n$ systems with impatient customers

Fátima Ferreira [a,d], António Pacheco [b,d], Helena Ribeiro [c,d]
mmferrei@utad.pt, apacheco@math.tecnico.ulisboa.pt,
helena.ribeiro@ipleiria.pt

[a] *University of Trás-os-Montes and Alto Douro, Vila Real*
[b] *Instituto Superior Técnico, Universidade de Lisboa, Lisbon*
[c] *Escola Superior de Tecnologia e Gestão, Polytechnic of Leiria, Leiria*
[d] *CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Lisbon*

Keywords: balking, batch arrival, busy-period, queueing system, reneging

Abstract: Customer impatient behaviour is a common phenomena in service provider organisations that deal with queueing systems. Due to perception of a large number of customers waiting in queue before them, incoming customers may either decide not to join the queue or, after joining it, they may decide to abandon the system before getting service, after a certain amount of time elapses. Such behaviours, known in the literature as balking and reneging, may impact the profits of services providers due to undesirable loss of customers. In order to minimise such customer losses, while maintaining a reasonable number of services per period of continuous occupancy, a properly dimensioning of the system is required.
Aiming to guide service providers in such decision making, we compute in this work the joint probability function of the numbers of customers served and lost in busy-periods of $M^X/M/c/n$ systems with balking and reneging. These are multi-server finite capacity queues, in which customers arrive in batches according to a compound-Poisson process. The batch sizes are independent and identically distributed. The customer service times are independent and identically exponentially distributed random variables, and are independent of the arrival process. The computation is accomplished using a recursive procedure that evaluates probability-generating functions of the numbers of customers served and lost in busy-periods. We analyse the sensitivity of the joint probability function under different customer blocking and reneging policies (partial/total and preemptive/non-preemptive) and different key parameters (system capacity; number of servers; batch size and balking probabilities; and arrival, service, and impatience rates).

Acknowledgements: This work was supported by FCT (Fundação para a Ciência e a Tecnologia) under projects UIDB/04621/2020 and UIDP/04621/2020.

Extremal behavior of some bivariate integer models

Sandra Dias [a], **M. G. Temido** [b]
sdias@utad.pt, mgtm@mat.uc.pt

[a] *Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro*
[b] *Universidade de Coimbra, CMUC, Departamento de Matemática, FCTUC*

Keywords: bivariate time series, extreme values, integer models

Abstract: Integer-valued time series have received increasing attention in the statistical literature over the past decades. In the univariate case there is a wide variety of models to describe count data. However, in real applications such as epidemiology, researchers often have to face multivariate time series data. Suitable classes of models to cope with the discreetness of the data are the ones based on the so-called binomial thinning operator. This operator led to the development of a wide variety of interesting INARMA models. Despite the large number of integer-valued models proposed in the literature, little is known about its extremal properties. Anderson [1] gave an important contribution to this limitation by obtaining upper and lower bounds for the limiting distribution of the maximum term of i.i.d. sequences with integer distributions exhibiting an exponentially decaying tail. Based on Anderson's work several stationary models, with different dependence structures, have been studied with respect to the extremal behaviour. Mainly, INAR and INMA models and their variants. In what concerns the multivariate case little has been done so far with respect to extreme values of integer-valued data. Hüsler *et al.* [2] establishes asymptotics for the distribution of the bivariate maximum of infinite INMA sequences by introducing a special class of distributions for the innovations, that, marginally includes the one of Anderson [1]. Dias and Temido [3] study the extremal behaviour of a bivariate max-INAR model with innovations distributed according to a double geometric law.
In this work, we study the extremal behaviour of some integer-valued bivariate time series. Concretely, assuming that the distributional behaviour of the innovations is the one introduced in Hüsler *et al.* [2], we establish asymptotics for the distribution of the bivariate maxima of two max-BINAR models and of a BINMM model.

References

[1] Anderson, C. W. Extreme value theory for a class of discrete distribution with applications to some stochastic processes. *J. Appl. Probab.*, 7, 99–113, 1970. doi:org/10.2307/3212152

[2] Hüsler, J., Temido, M. G., Valente-Freitas, A. On the maximum term of a bivariate infinite MA model with integer innovations. *Methodol Comput Appl Probab*, 24, 2373–2402, 2022. doi:org/10.1007/s11009-021-09920-3

[3] Dias, S., Temido, M. G. On the Maximum of a Bivariate Max-INAR(1) Process. *Recent Developments in Statistics and Data Science. SPE 2021. Springer Proceedings in Mathematics & Statistics*, Vol 398, 55–69, 2022. doi:org/10.1007/978-3-031-12766-3_5

Index

Abreu, Ana Maria, 95
Afonso, Catarina, 155
Afonso, Pedro Miranda, 3, 62
Afreixo, Vera, 46, 63, 84
Albino, Luís, 155
Albuquerque, João, 131
Aleixo, Sandra M., 55
Alexandre, Luís A., 189
Almeida, Catarina, 117
Almeida-Silva, Marina, 160
Alpuim, Teresa, 53
Alves, Ana Catarina, 131
Alves, João, 143
Alves, Rui, 34
Alves, Sandra, 65
Amado, Conceição, 73, 101, 140
Amaral, Paula, 126
Amaral, Rita, 64
Amorim, Ana Paula, 139
Andrade, Daniel, 87
Andrinopoulou, Eleni-Rosalina, 62
Angélico, Maria Manuel, 170
António, Carlos Conceição, 150
Antunes, Alexandra M. M., 73
Antunes, Carlos, 109
Antunes, Luís, 110
Antunes, Marília, 131
Araújo, Artur, 161
Araújo, Bárbara, 177
Araújo, Lia, 163
Araújo, Susana, 144
Ascenso, Loide, 159
Avilez-Valente, Paulo, 125
Azeitona, Margarida, 101
Azevedo, Ana, 129
Azevedo, Ana Isabel, 105
Azevedo, José, 105
Azevedo, Marta, 161

Bacelar-Nicolau, Helena, 108
Balter, Anne, 79
Barbillon, Pierre, 85
Barbosa, Diogo, 41
Barrios, Jhonathan, 146, 154, 177
Barroso, Lúcia Pereira, 169
Bernardino, Joana, 119
Bicho, Estela, 177
Biscaia, Ema, 119
Bispo, Regina, 57, 118, 119, 155
Borges, Ana, 97, 179
Botelho, Bárbara, 80
Bourbon, Mafalda, 131
Braga, Ana Cristina, 117
Branco, João A., 52
Brandão, João, 71
Braumann, Carlos A., 19, 136
Brilhante, M. Fátima, 138, 174
Brites, Nuno M., 19, 136
Brito, Irene, 105, 180
Brito, Paula, 46, 59, 103, 111, 126, 127
Buescu, Jorge, 107
Bustorf, Manuela, 130

Cabral, Jorge, 84
Cadilhe, Raquel, 129

Caeiro, Frederico, 91–93, 145
Caiado, Jorge, 8
Campos, Pedro, 33, 111
Cardoso, Helena, 64
Cardoso, Margarida G. M. S., 128
Carinhas, Dora, 106
Carlos, Clara, 19, 136
Carmona, Ana, 34
Carrascal, Gloria, 146
Carrasquinha, Eunice, 71
Carrilho, João F., 183
Carvalho, Francisco, 162
Carvalho, Maria Lucília, 120
Carvalho, Mariana Reimão, 97
Carvalho, Pedro, 149
Carvalho, S., 175
Casaca, Pedro, 109
Casanova, José Luís, 185
Casquilho, Miguel, 107
Castro, Cecília, 107, 115, 117
Chambel, Filipa, 35
Chambel, Luís, 128
Cipriano, Fernanda, 83
Clancy, John P., 62
Coelho, Ricardo, 116
Coimbra, B., 175
Coletti, Roberta, 183
Comparada, Sofia, 53
Cordeiro, Daniel, 129
Correia, Elisete, 172
Correia, Iúri J. F. , 122
Costa, Beatriz, 147
Costa, Filipa, 73
Costa, Joana, 139
Costa, Mafalda T., 15
Costa, Marco, 99, 102
Costa e Silva, Eliana, 113
Crespo, Maria Teresa Barreto, 182
Cuyler, Christine, 122

da Silva, Bianca Rafaelle, 135
da Silva, Emanuel V. M., 86
de Carvalho, Miguel, 15, 90
de Miranda, M. Souto, 94
de Oliveira, Maria Helena, 96
de Sousa, Bruno, 112, 114
de Sousa, Jorge Alves, 186
de Sousa, Jossé Joedson Lima, 186
de Sá, Inês Rodrigues, 33
de Uña-Álvarez, Jacobo, 9, 76
del Puerto, Inés, 22
Diamantino, Maria Fernanda, 152
Dias, André, 170
Dias, Celeste, 56
Dias, Eduardo, 184
Dias, Inês, 121
Dias, Sónia, 126, 127
Dias, Sandra, 188, 191
Domingues, Nuno, 156
Dourado, Ricardo, 160
Duarte, Leandro, 139
Durão, Natércia, 130

Erlhagen, Wolfram, 177

Faria, Brígida Mónica, 64, 65, 108, 110
Faria, Susana, 39, 45, 69, 165, 181
Felgueiras, Miguel, 160
Fernandes, Eugénia, 109
Ferrandi, Giulia, 75
Ferreira, Fátima, 190
Ferreira, Flora, 146, 177, 179
Ferreira, Helena, 189
Ferreira, M. Conceição, 34
Ferreira, Marta, 188, 189
Ferreira, Susana, 173
Figueiredo, Adelaide, 157, 158
Figueiredo, Andreia, 156
Figueiredo, Fernanda Otília, 91, 158
Figueiredo, Ivone, 120
Fonseca, Marcelo B., 171
Fonseca, Válter R., 109
Forte, Anabel, 85
Fortuna, Inês, 110
Francisco, Carla, 162
Freitas, Adelaide, 60
Freitas, Ana, 117
Freitas, Elisabete, 165

Gago, Miguel F., 177
Gaio, Rita, 58, 70, 129
Garção, Eugénio, 168
García-Escudero, Luis Angel, 51
Garcez, Luís, 98
Garrido, Susana, 82

Gomes, Dora Prata, 67
Gomes, M. Ivette, 91–94, 158
Gomes, Manuel Carmo, 109
Gomes, Maria Isabel, 118
Gomes, Micael, 124
Gonçalves, A. Manuela, 102, 154, 164, 180
Gonçalves, Elsa, 81
Gonçalves, Francisco, 121
Gonçalves, Luzia, 54
González, Miguel, 22
Gouveia, Sónia, 123
Gouveia-Reis, Délia, 141
Graca, Luís, 109
Guerra, Sílvia, 83
Guimarães, Elsa, 129

Hennig, Christian, 51
Henriques, Lucas, 115
Henriques-Rodrigues, Lígia, 91–93
Hochstenbach, Michiel E., 75
Hoffbauer, Luísa, 150

Iglesias-Perez, Maria del Carmen, 76
Iglésias, Gustavo, 40
Infante, Paulo, 106, 159
Isorna, Francisco Caamaño, 28

Jacinto, Gonçalo, 159

Kanno, Gustavo de Oliveira, 169
Kateri, Maria, 7
Kort, Peter, 79
Kravchenko, Igor, 75

Laureano, Gonçalo, 156
Leder, Solange, 178
Leite, Pedro Pinto, 109
Leite, Rita, 101
Leão, Celina Pinto, 178
Lima, Felipe, 115
Lopes, Anselmo Ribeiro, 186
Lopes, João S, 35
Lopes, Luiz Guerreiro, 141
Lopes, Marta B., 57, 71, 183
Loura, Luísa, 152
Loureiro, Catarina P., 103
Lourenço, Vanda M., 171
Lousão, Maria Paula, 185
Lucas, Diana, 46

Macedo, Pedro, 84
Machado, Marcos, 152
Maciala, Faustino, 154
Madeira, Sara, 98
Magro, Felipe Antonio, 137
Maia, Marisa, 156
Malato, João, 109
Maltez, Marta, 60
Margalho, Luís, 148
Marinho, Ana, 69
Marques, Filipe J., 68
Marques, Tiago A., 122
Martinho, António, 106
Martins, Ana, 123
Martins, André, 72
Martins, André F. T., 48
Martins, João Paulo, 72, 142
Martins, Natália Costa, 166
Martins, Patrícia, 153
Martins, Rui, 87, 130
Martins, Susana Rafaela Guimarães, 76
Mateus, Ayana, 145
Matos, Cristina, 164
Matos, Gonçalo, 155
Mayo-Iscar, Agustín, 51
Medeiros, Ana Margarida, 131
Meira-Machado, Luís, 27, 86, 139, 161
Meireles, Ana, 112
Meireles, Paula, 139
Mendonça, Denisa, 163
Mendonça, Sandra, 138, 141, 151
Menezes, Raquel, 82, 104, 121, 170
Mexia, Tiago, 162
Milheiro-Oliveira, Paula, 125
Minuesa, Carmen, 22
Miranda, M. Cristina, 66, 94
Miranda, Rui, 58
Molina, Manuel, 20, 21
Monteiro, Andreia, 120
Monteiro, Magda, 60, 100
Morais, Jorge, 39
Moreira, Almiro, 34
Moreira, Ana, 181
Moreira, Carla, 139, 161

Moreira, Lucybell, 129
Morelli, Gianluca, 51
Morello, Judit, 73
Mota, Manuel, 20, 21
Mourão, Maria, 149
Mouriño, Helena, 98

Nakamura, Luiz R., 166
Nakyambadde, Betty, 89
Nascimento, Ana Paula, 108
Natário, Isabel, 83, 112, 116, 120
Neves, Cláudia, 78
Neves, M. Manuela, 67
Nieves, Paula Saavedra, 28
Noda, Lumy, 178
Norouzirad, Mina, 68
Novais, Luísa, 45
Nunes, Alcina, 142
Nunes, André, 54
Nunes, Cláudia, 79
Nunes, Rui, 127

Oliveira, Alexandra, 65, 124
Oliveira, Irene, 153
Oliveira, Lina, 103
Oliveira, M. Rosário, 75, 103
Oliveira, Mafalda, 70
Oliveira, Manuela, 162, 168
Oviedo-de la Fuente, Manuel, 104

Paúl, Constança, 163
Pacheco, António, 190
Pacheco, Pedro, 179
Paiva, Francisco, 148
Palipana, Anushka, 62
Paulo, Rui, 85
Pedra, Ana Cristina, 180
Pedrosa, Ana, 179
Peleteiro, Bárbara, 129
Pereira, Cristiana Palmela, 143
Pereira, Diogo, 79
Pereira, F. Catarina, 102
Pereira, Isabel, 88, 89, 100
Pereira, Paula, 187
Pereira, Pedro, 149
Pereira, Rui, 15
Pereira, Soraia, 122
Pestana, Dinis D., 138, 151
Pestana, Pedro D., 138, 174
Pina, Helena, 144
Pinho, Ana, 130
Pinto, Helder, 56
Pires, Ana M., 52
Pires, Sara Ribeiro, 167
Plaza, Jairo, 146
Polidoro, Maria J., 15, 130
Pombal, José Maria, 48
Previdelli, Isolde, 96
Prudêncio, Cristina, 108

Quintino, Hugo, 159

Ramires, Thiago G., 135, 137, 166
Ramos, M. Rosário, 167
Ramos, Sandra, 80
Rasga, Célia, 147
Rebouças, Sílvia, 182, 187
Reis, Luís Paulo, 124
Riani, Marco, 51
Ribeiro, Conceição, 187
Ribeiro, Helena, 190
Ribeiro, Oscar, 163
Ribeiro, Pedro, 47
Ribeiro, Ruy M., 109
Ribeiro, Tiago, 118
Ribeiro, Vítor M., 176
Rizopoulos, Dimitris, 62
Roca-Pardiñas, Javier, 27
Rocha, Ana Paula, 56
Rocha, Anabela, 66
Rocha, Cristina, 95
Rocha, Filipa, 63
Rocha, J. Leonel, 175
Rocha, M.L., 174
Rodrigues, Catarina, 140
Rodrigues, Paulo C., 171
Rodrigues, Vítor, 112
Romão, Nuno, 35
Royé, Dominic, 28

Saldanha, Ricardo, 155
Santos, Ana, 126
Santos, Cláudia, 89
Santos, Daniel, 173
Santos, David, 34
Santos, Gabriela, 172

Santos, Jair, 165
Santos, Marlene, 72
Santos, Rui, 143, 173
Santos, Susana, 70
Scotto, Manuel G., 123
Seabra-Silva, Joana, 125
Semedo-Lemsaddek, Teresa, 182
Sepúlveda, Nuno, 61
Sequeira, Fernando, 174
Serra, Maria Conceição, 23
Sestelo, Marta, 27
Severino, Eduardo, 53
Silva, A. Pedro Duarte, 59
Silva, Alexandra A., 104, 121
Silva, Catarina Campos, 111
Silva, Daniela, 82, 121
Silva, Fernando, 47
Silva, Giovani, 54
Silva, Isabel, 88
Silva, João Falcão, 41
Silva, Marcelo F., 135, 137
Silva, Maria Eduarda, 10, 47, 88
Silva, Vanessa Freitas, 47
Silvestre, Cláudia, 144, 185
Simões, Paula, 120
Soares, Elsa, 74
Soares, Isabel, 176
Soares, José Maria, 179
Soares, Rui, 100
Sousa, Cristovão, 113
Sousa, Inês, 29, 74, 184
Sousa, Lisete, 147, 156, 182
Sousa, Rita, 39, 69, 184
Sousa-Ferreira, Ivo, 95
Soutinho, Gustavo, 86, 176
Szczesniak, Rhonda D., 62

Tavares, Ana Helena, 46
Teixeira, Laetitia, 163
Teles, Carlota, 182
Teles, Júlia, 55
Temido, M. G., 191
Tenreiro, Carlos, 77
Torres, Beatriz, 65
Torres, Margarida, 142

Vaz-Fernandes, Paula, 167
Velosa, Sílvio, 151
Vicente, Astrid, 147
Vieira, Cláudia, 110
Vieira, Francisca G., 57
Vieira, Mónica, 108
Vilela, Alice, 172
Villamayor, Mª José Ginzo, 28
Villanueva, Nora M., 27

Weiß, Christian H., 123