XXV Congresso
Sociedade Portuguesa
de Estatística
2021 Évora

BOOK OF ABSTRACTS *of* SPE 2021

XXV CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA—SPE 2021

SPE PRESS
EDIÇÕES SPE—SOCIEDADE PORTUGUESA DE ESTATÍSTICA

UNIVERSITY *of* ÉVORA
Évora, Portugal
www.spe2021.uevora.pt
✉: spe2021@uevora.pt

Ficha Técnica:

Book of Abstracts of SPE 2021

Russell Alpizar-Jara, Dulce Gomes, Lígia Henriques-Rodrigues, Patrícia A. Filipe

Editora: Sociedade Portuguesa de Estatística

ISBN: 978-972-8890-48-3

ORGANIZING COMMITTEE
(COMISSÃO ORGANIZADORA)

Russell Alpizar-Jara, UNIVERSITY *of* ÉVORA
(President)

Dulce Gomes, UNIVERSITY *of* ÉVORA

Lígia Henriques-Rodrigues, UNIVERSITY *of* ÉVORA

Patrícia A. Filipe, ISCTE, UNIVERSITY INSTITUTE *of* LISBON

SCIENTIFIC COMMITTEE
(COMISSÃO CIENTÍFICA)

Miguel de Carvalho, UNIVERSITY *of* EDINBURGH, UK
(President)

Fátima Ferreira, UNIVERSITY *of* TRÁS-OS-MONTE E ALTO DOURO

João Andrade e Silva, UNIVERSITY *of* LISBON

Luís Meira-Machado, UNIVERSITY *of* MINHO

Marco Costa, UNIVERSITY *of* AVEIRO

Maria Eduarda Silva, UNIVERSITY *of* OPORTO

Marília Antunes, UNIVERSITY *of* LISBON

Paula Brito, UNIVERSITY *of* OPORTO

Regina Bispo, NOVA UNIVERSITY *of* LISBON

Rosário Oliveira, UNIVERSITY *of* LISBON

Russell Alpizar-Jara, UNIVERSITY *of* ÉVORA

Table of Contents
(Índice)

Plenary Sessions
(Sessões Plenárias) 1

Invited Sessions
(Sessões Convidadas) 7

Biometry (SPE Section & SGAPEIO)
(Biometria (Secção da SPE & SGAPEIO)) 9

Caucus for Women in Statistics (CWS) 15

Official Statistics—Statistics Portugal
(Estatísticas Oficiais—Instituto Nacional de Estatística) 17

Portuguese Central Bank
(Banco de Portugal) 23

RBras (Brazilian Region International Biometric Society)
Região Brasileira da Sociedade Internacional de Biometria 29

SPE–CLAD
Statistics, a tool for sustainability
(Estatística, uma ferramenta para a sustentabilidade) 35

SPE–FenStats
Accreditation of statisticians: Why and how? 41

Oral Sessions
(Comunicações Orais) **43**

Posters **139**

Authors / Autores **163**

Plenary Sessions

—Sessões Plenárias—

XXV Congresso
Sociedade Portuguesa
de Estatística

Learning Probabilistic Graphical Models from Large and Complex Data

Genevera I. Allen[a,b]
gallen@rice.edu

[a] *Departments of Electrical and Computer Engineering, Statistics, and Computer Science, Rice University*
[b] *Jan and Dan Duncan Neurological Research Institute, Baylor College of Medicine*

Keywords: data integration, genomics, graphical models, latent variables, neuroscience

Abstract: Probabilistic graphical models are widely used to explore, model, and visualize relationships in large data sets from areas such as physics, systems biology, computer vision, and finance, among others. These models represent multivariate probability distributions as a graph with edges denoting conditional dependence relationships between random variables. In this talk, I will highlight recent advances from my research group including new types of graphical models as well as methods and theory for learning graphs from high-dimensional and complex data. Specifically, I will discuss graph learning for non-Gaussian data including data with extreme events, for mixed data and data integration, for non-simultaneous or non-aligned data via Graph Quilting, and for learning in the presence of latent variables. Additionally, I will present several applications of these approaches to learn regulatory networks in genomics and functional connectivity in neuroscience.

How long could a human live?

Anthony C. Davison[a]
anthony.davison@epfl.ch

[a] *Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

Abstract: There is sustained and widespread interest in understanding the limit, if any, to the human lifespan. Apart from its intrinsic interest, changes in survival in old age have implications for the sustainability of social security systems. Recent analyses of data on the oldest human lifespans have led to competing claims about survival and to some controversy, due in part to inappropriate use of statistical methods. One central question is whether the endpoint of the underlying lifetime distribution is finite. This talk will discuss the particularities associated with such data, outlines correct ways of handling them and presents suitable models and methods for their analysis. We illustrate the ideas through novel analysis of data on semi-supercentenarian lifetimes, which suggests that any upper limit to human lifetimes lies well beyond the highest lifetime yet reliably recorded, with lower limits to 95% confidence intervals around 127 years, and maximum likelihood estimates upwards of 130 years.
The work is joint with Léo Belzile, Jutta Gampe, Holger Rootzén and Dmitrii Zholud.

Level Crossing Ordering of Stochastic Processes

António Pacheco[a]
apacheco@math.tecnico.ulisboa.pt

[a] *CEMAT, Instituto Superior Técnico, Universidade de Lisboa*

Abstract: Stochastic Ordering is an important area of Applied Probability tailored for qualitative comparisons of random variables, random vectors, and stochastic processes. In particular, it may be used to investigate the impact of parameter changes in important performance measures of stochastic systems, avoiding the exact computation of such performance measures. In this respect, the great diversity of performance measures used in applied sciences to characterize stochastic systems has inspired the proposal of many types of stochastic orderings.

In this talk we address the level crossing ordering, proposed by A. Irle and J. Gani in 2001, which compares stochastic processes in terms of the times they take to reach high levels (states). After introducing some motivation for the use of the level crossing ordering, we present tailored sufficient conditions for the level crossing ordering of (univariate and multivariate) Markov and semi-Markov processes. These conditions are applied to the comparison of birth-and-death processes with catastrophes, queueing networks, and particle systems.

Our analysis highlights the benefits of properly using the sample path approach, which compares directly trajectories of the compared processes defined on a common probability space. This approach provides, as a by-product, the basis for the construction of algorithms for the simulation of stochastic processes ordered in the level crossing ordering sense. In the case of continuous Markov chains, we resort additionally to the powerful uniformization technique, which uniformizes the rates at which transitions take place in the processes being compared.

Joint work with Fátima Ferreira (CM-UTAD and Universidade de Trás-os-Montes e Alto Douro).

Everybody want (think?) to play stats, will statistics play a key role in an everything-is-AI world?

Maurizio Sanarico[a]
maurizio.sanarico@sdggroup.com

[a] *SDG Group*

Keywords: AI, AI-statisitics, real-world applications, subject-mixing

Abstract: The recent surge of machine learning and artificial intelligence as fashionable topics, with strong emphasis of technology, is prompting a question: what is, and what could be, the role and the place of statistics in this context? I found an analogy looking at the mathematics translated into engineering. When we go to engineering, the mathematics is forgotten, as the focus is on the object, be it a smartphone, a video-camera or any other among the huge number of systems that would be not-existing without the mathematics. In my role of Chief Data Scientist of SDG Group, I am constantly facing the tension between technology-oriented and scientific-oriented thinking trying to bring the two to a harmonic synthesis, with the additional constraint/goal to produce sustainable and profitable business. Statistics found its way as the backbone to strength the scientific method in disciplines it was already used and to extend it to other, most difficult disciplines. The statistical science on the practical ground develops method to quantifying uncertainty in mathematical models based on data and allows to make inferences out of these models. In AI and machine learning models are operationalized and become parts of processes that support a number of real-life applications. I will try to give a point of view of the space for statistics as a key subject in the present and envisioned future times. Another topic I will present is the idea of subject-mixing, i.e., the requirement to combine knowledge and skills from different, sometimes conflicting, disciplines to put models at work in production systems.

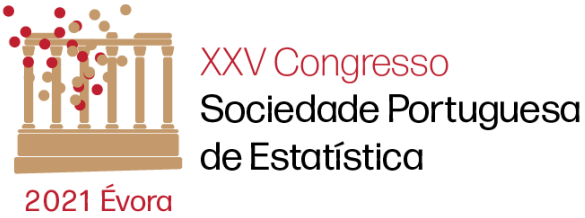
Invited Sessions

—Sessões Convidadas—

XXV Congresso
Sociedade Portuguesa
de Estatística
2021 Évora

Biometry
(SPE Section & SGAPEIO)

—**Biometria (Secção da SPE & SGAPEIO)**—



Inês Sousa
University of Minho, isousa@math.uminho.pt

Propensity scores approaches: challenges in observational studies

Rosemeire Fiaccone[a,b], Dandara Ramos[c,b]
fiaccone@ufba.br, dandara.ramos@ufba.br

[a] *Departamento de Estatistica, Universidade Federal da Bahia, Brasil*
[b] *Centro de Integracao de Dados e Conhecimentos para Saude Cidacs/Fiocruz, Brasil*
[c] *Instituto de Saude Coletiva, Universidade Federal da Bahia, Brasil*

Keywords: causality, electronic health records, observational study, propensity scores

Abstract: There is a major interest in evaluating the impact of the Bolsa Família Program (BFP) on health outcomes. The impact of policies and complex interventions is an extremely important area for methodological development. However, randomized trials cannot be used to evaluate the impact of policies already implemented. The alternative is to evaluate intervention in process by use of non-randomized methods (generically denominated quasi-experimental studies or observational studies). In recent years there has been an increase in the use of methodologies involving propensity scores to minimize selection bias in observational studies aimed at identifying causal relationships. Evaluation of BFP, in particular, brings additional methodological challenges due to: (i) nature of the cash transfers: the benefit is time-varying, given repeatedly over time (with varying durations) and distinct time delay to starting receiving the benefit; (ii) we might need to define multiple beneficiary groups according to duration, amount and dynamic of the benefit over time; (iii) there is great heterogeneity within the beneficiary group according to individual characteristics related to their poverty status and contextual factors deriving from higher levels such as municipalities, states and regions. In this work, we present some strategies for the use of propensity score methods taking into account the data structure in order to adequately evaluate the impact of BFP on health outcomes. In addition to this, novel methodologies should be presented.

Acknowledgements: FAPESB - Fundacao de Amparo a Pesquisa do Estado da Bahia

Identification of immune responses predictive of clinical protection against Malaria

André Fonseca[a], Clara Cordeiro[a,b], Nuno Sepúlveda[b,c]
a49406@ualg.pt, ccordei@ualg.pt, nunosep@gmail.com

[a] *Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*
[c] *Institute for Medical Immunology, Charité-Universitatsmedizin Berlin, Berlin, Germany*

Keywords: Box-Cox transformation, finite mixture model, malaria, random forest, skew-normal distribution

Abstract: Various statistical pipelines have been proposed to discover antibody responses associated with protection against clinical malaria. However, these have often produced inconsistent or even conflicting results among studies due to inadequate statistical assumptions such as the normality of the data.
In the present work, we have developed two new statistical pipelines to analyze data from IgG antibodies against 36 *Plasmodium falciparum* antigens from 121 Kenyan children [1]. The first pipeline relied on traditional statistical techniques for the normal distribution after Box-Cox transformation together with the use of flexible finite mixture models for seropositivity determination. The second pipeline was based on the identification of cutoff values in the antibody distributions that maximized the distinction between susceptible and protected individuals.
Our pipelines enabled to develop several classifiers based on antibodies against the msp2, msp4, msp7, msrp3 and pf110373 antigens, with an estimated area under the curve (AUC) of the Receiver Operating Characteristic reaching up to 89%, significantly outperforming the previous results based on random forest (AUC=68%).[2]
In summary, the good performance of our pipelines suggested their wide applicability in antibody data analysis aiming to identify antimalarial vaccine candidates.

Acknowledgements: André Fonseca has a PhD fellowship by FCT - Fundação para a Ciência e a Tecnologia (ref. SFRH/BD/147629/2019). Clara Cordeiro and Nuno Sepúlveda are partially funded by FCT (grant ref. UIDB/00006/2020).

References

[1] Osier FH, Mackinnon MJ, Crosnier C, et al. New antigens for a multicomponent blood-stage malaria vaccine.*Sci Transl Med*, 6(247):247ra102-247ra102. doi:doi:10.1126/scitranslmed.3008705

[2] Valletta JJ., Recker M. Identification of immune signatures predictive of clinical protection from malaria. *PLoS Comput Biol*, 13(10): e1005812. doi:https://doi.org/10.1371/journal.pcbi.1005812

Nonparametric inference for mixture cure model when the cure information is partially available

María Amalia Jácome[a], Wende Clarence Safari[b], Ignacio López-de-Ullibarri[c]
majacome@udc.es, wende.safari@udc.es, ignacio.lopezdeullibarri@udc.es

[a] *Department of Mathematics, Faculty of Science, University of A Coruña, CITIC, A Coruña, Spain*
[b] *Department of Mathematics, Faculty of Computer Science, University of A Coruña, CITIC, A Coruña, Spain*
[c] *Department of Mathematics, Escuela Universitaria Politécnica, University of A Coruña, Ferrol, Spain*

Keywords: bandwidth, kernel, latency, Nadaraya-Watson, survival

Abstract: When analyzing times to event in survival analysis, it is commonly assumed that all subjects in the population are susceptible to the event of interest when there is sufficient follow-up time. However, there are many instances where the event will not occur for all the individuals. Mixture cure model (MCM) assumes that the population is the mixture of two sub-groups: those whose event is certain not to occur are "cured" (or long-term survivors) and those who will experience the event are known to be "uncured" (or susceptible).
A difficulty encountered in the MCM is that the long-term survivors are never observed to be cured, rather they are censored at the end of the study. Hence, the cure status is unobserved (latent) in the right-censored subjects. Nonetheless, in several situations the cure status for some censored individuals is observed. For example, diagnostic procedures in medical studies might provide further information on whether a subject can be considered as cured or not. Also, for some types of cancer it is extremely unlikely to have any recurrence later than a given time after treatment, known as cure threshold, and consequently those patients with observed time surpassing the cure threshold can be considered as cured. In this talk, a completely nonparametric approach will be introduced to estimate the MCM when the cure status is partially known. Our example comes from a study of COVID-19 patients hospitalized in Galicia (Spain) during the outbreak of the pandemic.

A nonparametric estimation method based on the EM algorithm for the latency

A. López-Cheda[a], Y. Peng[b], M. A. Jácome[a]
ana.lopez.cheda@udc.es, yingwei.peng@queensu.ca,
maria.amalia.jacome@udc.es

[a] *Research group MODES, CITIC, Department of Mathematics, University of A Coruña*
[b] *Queen's Cancer Research Institute, Department of Public Health Sciences and Department of Mathematics and Statistics, Queen's University*

Keywords: bootstrap, censored data, mixture cure models, survival analysis

Abstract: Nonparametric methods have attracted much attention in the last few years for cure models. In the literature, to model the effects of covariates on the distribution of the failure time of the susceptible individuals (latency), it is assumed that the cure rate in the model either is a constant or depends on the same covariates as the latency distribution. We propose a new nonparametric estimator for the latency distribution that relaxes the assumption. The estimation, based on the EM algorithm, is readily available for mixture cure models. The finite sample performance of the proposed estimator is assessed in a simulation study. Finally, the proposed method is employed to model the effects of some covariates on the time to bankruptcy among commercial banks insured by the Federal Deposit Insurance Corporation.

Acknowledgements: The first author was sponsored by the BEATRIZ GALINDO JUNIOR Spanish Grant (reference BEAGAL18/00143) from MICINN (Ministerio de Ciencia, Innovación y Universidades) with code BGP18/00154.

Observational and comparative study between automatic and manual analysis of sleep studies

Ricardo São João[a,b], Andreia Cardoso[c], Tiago Domingues[a,b], Vânia Silva[c], Marta Fradinho[c], Laura Santos[c], Amélia Feliciano[d,e]
ricardo.sjoao@esg.ipsantarem.pt,andreia.neves.cardoso@hospitaldaluz.pt, tiago.domingues@esg.ipsantarem.pt,vania.carina.silva@hospitaldaluz.pt, marta.romao.fradinho@hospitaldaluz.pt,laura.simoes.santos@hospitaldaluz.pt, amelia.feliciano71@gmail.com

[a] *Escola Superior de Gestão e Tecnologia-IPSantarém*
[b] *Centro de Estatística e Aplicações Universidade de Lisboa-CEAUL*
[c] *Hospital da Luz Setúbal*
[d] *Lusíadas Cluster Clinics*
[e] *Trofa Saúde Loures/Amadora*

Keywords: automatic analysis, diagnosis, manual analysis, OSAS

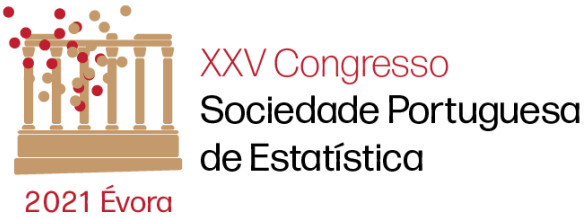
Abstract: Obstructive sleep apnea syndrome (OSAS) is a sleep-disordered breathing disorder. It consists of a set of symptoms and signs that result from recurrent episodes of intermittent upper airway obstruction. It is estimated that this syndrome is very common; however, its true prevalence is unknown, as it is underdiagnosed. OSAS is considered a systemic disease with several associated consequences, including cardiovascular and metabolic diseases, making it a public health problem. Polysomnography is the sleep study of choice for the diagnosis of OSAS, but the fact that it is a time-consuming test contributes to the underdiagnosis of this pathology. For this reason, one can opt for the Cardiorespiratory Sleep Study which, despite being a simpler test, allows for the diagnosis of this syndrome. The manual reading of these tests is one of the aspects that most contributes to their slowness, as well as the inter-observer variability in their reading. From a simplistic perspective, some studies have focused on the automatic analysis of these exams. However, there have been discrepancies between the two analyses, which can be significant in terms of final diagnosis and consequent therapeutic orientation. The Sleep Unit of the Hospital da Luz Setúbal, based on a sample of 3297 scans performed in 2011-2019, concluded that the automatic analysis may lead to an incorrect diagnosis of OSAS as well as its severity, being important that the diagnosis of this pathology be based on a manual analysis.

Acknowledgements: This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

References

[1] Pevernagie, D., Gnidovec-Strazisar, B., Grote, L., Heinzer, R., McNicholas, W., & Penzel, T. et al. On the rise and fall of the apnea-hypopnea index: A historical review and critical appraisal. *Journal Of Sleep Research, 29(4), 2020.* https://doi.org/10.1111/jsr.13066

Caucus for Women in Statistics



Vanda M. Lourenço
NOVA University of Lisbon, vmml@fct.unl.pt

Caucus for Women in Statistics: a perspective from the society and women in industry, government and academia

Tomi Mori[a], Nairanjana Dasgupta[b], Holly Shulman[c], Cynthia Bland[d], Wendy Lou[e], Vanda Lourenço[f]
Tomi.Mori@stjude.org, dasgupta@wsu.edu, hbs1@verizon.net, cbland@rti.org, wendy.lou@utoronto.ca, vmml@fct.unl.pt

[a] *Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, US*
[b] *Department of Mathematics and Statistics, Washington State University, Washington, Pullman, US*
[c] *Centers for Disease Control and Prevention (CDC), Atlanta, US*
[d] *RTI International, Research Triangle Park, North Carolina, US*
[e] *Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*
[f] *Department of Mathematics and CMA, NOVA School of Science and Technology, NOVA University of Lisbon, Lisbon, Portugal*

Abstract: The Caucus for Women in Statistics (CWS) is an international, professional statistical society formed in 1971, for the education, employment and advancement of women in statistics, whose membership is open to anyone who supports CWS's mission and vision, from academia, industry, government and elsewhere. The CWS vision is a world where women in the profession of statistics have equal opportunity and access to influence policies and decisions in workplaces, governments, and communities. Its mission is to advance the careers of women statisticians through advocacy, providing resources and learning opportunities, increasing their professional participation and visibility, and promoting and assessing research that impacts women statisticians. In this session, we will give the audience a more detailed insight of the aforementioned points in addition to three testimonies from women statisticians from academia, government and industry as to how they see and feel gender treatment differentiation, if any, in these three sectors.

Acknowledgements: The session organizer and speakers thank CWS for sponsoring this thematic session.

Official Statistics —Statistics Portugal

—Estatísticas Oficiais—Instituto Nacional de Estatística—

XXV Congresso
Sociedade Portuguesa
de Estatística

Pedro Campos
Statistics Portugal, pedro.campos@ine.pt

Infecundidade permanente e voluntária: as pessoas sem filhos e sem intenção de os vir a ter

Susana Clemente[a], **Joana Malta**[a], Rita Lages[a]
Susana.Clemente@ine.pt, Joana.Malta@ine.pt, Rita.Lages@ine.pt

[a] *Departamento de Estatísticas Sociais, Instituto Nacional de Estatística / Statistics Portugal*

Abstract: Este estudo foca-se na infecundidade (childlessness) permanente e voluntária, caraterizada pela decisão deliberada de uma pessoa sem filhos não querer ter filhos, com base nas intenções reprodutivas reportadas no momento da entrevista ao Inquérito à Fecundidade (IFEC) 2019. A análise permitiu traçar, a partir de uma Análise de Clusters, o perfil sociodemográfico destas pessoas, e, pela aplicação de uma Análise de Correspondências Múltiplas, identificar perfis de opiniões sobre a parentalidade e respetiva conciliação com a vida profissional, bem como conhecer o seu desejo em ter filhos. Portugal regista dos mais baixos níveis de fecundidade da Europa, resultado sobretudo da redução do número de filhos e do adiamento da parentalidade. Apesar de a baixa fecundidade observada no país não decorrer das pessoas não terem filhos, são raros os dados e estudos específicos sobre a infecundidade permanente e voluntária em Portugal. Dados do IFEC revelam que, em 2019, a grande maioria das pessoas tinha ou queria ter pelo menos um filho. No entanto, cerca de 9% não tinham nem pretendiam ter filhos (+1,5 p.p. face a 2013). O estudo conclui que, sendo uma população heterogénea, se evidenciam dois tipos de infecundidade voluntária: Resultante de uma escolha individual de não ter filhos, que pode variar ao longo da vida das pessoas ou ser uma decisão tomada desde cedo ("decisores precoces") - infecundidade convicta; Decorrente sobretudo das circunstâncias da vida, que podem incluir uma multiplicidade de fatores (inexistência de parceiro, idade, motivos de saúde ou ainda situações sociais e económicas) e que resultam em adiamentos persistentes ("decisores tardios"), que podem conduzir a uma situação em que a decisão de ter filhos pode ser mais difícil ou mesmo impossível para algumas pessoas - infecundidade condicionada. Pode ainda ser uma infecundidade voluntária potencial ou definitiva, já que algumas das pessoas que não tencionavam vir a ter filhos no momento do inquérito podem ainda vir a tê-los, e outras nunca o farão.

Consumption decisions and the COVID-19 pandemic: evidence from administrative data

Miguel Godinho de Matos[a], **Francisco Lima**[b]
miguel.godinhomatos@ucp.pt, francisco.lima@ine.pt

[a] *Católica Lisbon School of Business and Economics*
[b] *Statistics Portugal*

Keywords: age, consumption, COVID-19, income, risk aversion

Abstract: COVID-19 is an extreme event with large, negative consequences. One dimension of everyday life most affected by COVID-19 was individual consumption. Individuals adjusted their consumption patterns when facing different COVID-19 case-fatality rates—decisions on what, where, and when to consume changed dramatically since the COVID-19 outbreak. Using a sample of the population residing in Portugal, the study describes changes in individual consumption expenditures. The complete econometric analysis can be found in [1] where the implications of a model of risk-taking behavior are compared with the estimation results. The data originates from several administrative sources that have information on individual spending and sociodemographic characteristics. The exercise is part of Statistics Portugal's work on integrating administrative data sources and on the creation of the National Data Infrastructure, showing how it can expand the scientific and statistical capabilities for analyzing the Portuguese society.

References

[1] Eichenbaum, M.S., Godinho de Matos, M., Lima, F., Rebelo, S., Trabandt, M. How do People Respond to Small Probability Events with Large, Negative Consequences? *NBER Working Paper Series*, 27988, 2020. doi:10.3386/w27988

Contributos do INE no ensino e investigação em estatísticas oficiais

Francisco Lima[a], Pedro Campos[a], Carlos Marcelo[a]
francisco.lima@ine.pt, pedro.campos@ine.pt, carlos.marcelo@ine.pt

[a] *Instituto Nacional de Estatística / Statistics Portugal*

Abstract: Ao longo dos anos, o INE tem procurado acompanhar os desenvolvimentos científicos na área da Estatística e nas áreas que lhe estão associadas. Além disso, tem sido particularmente importante continuar a formar internamente os seus técnicos para que estes possam desempenhar as suas funções com conhecimentos mais atualizados. Essa formação tem ocorrido em várias vertentes, através de cursos internacionais organizados no âmbito do Eurostat, cursos à medida administrados por entidades externas e cursos de formação interna lecionados por outros colegas com conhecimentos e experiência nas áreas. Num futuro próximo, o INE planeia dar acesso a estas formações internas, de forma regular, a outras entidades que têm de lidar com estatísticas oficiais. As parcerias com a academia, com projetos de investigação, bolsas e estágios, também têm contribuído para o desenvolvimento de competências e avanços na produção estatística e científica. Recentemente, o INE passou a estar associado ao EMOS (European Master in Official Statistics). Este mestrado europeu em estatísticas oficiais é uma rede de programas de mestrado, criada para fortalecer a colaboração entre a academia e os produtores de estatísticas oficiais e ajudar a desenvolver profissionais capazes de trabalhar com dados oficiais europeus em diferentes níveis no sistema de produção em rápida mudança do século XXI. Nesta apresentação iremos desenvolver cada uma destas vertentes e apresentar algumas ideias sobre as competências em ciência dos dados em que o INE está a apostar.

Population mobility at the regional level during the COVID-19 pandemic: an analysis based on information from Facebook's "Data for Good" initiative

Francisco Vala[a], Cátia Nunes[a] and Miguel Godinho Matos[b]
francisco.vala@ine.pt, catia.nunes@ine.pt, miguel.godinhomatos@ucp.pt

[a] *Gabinete de Estudos Territoriais, Instituto Nacional de Estatística / Statistics Portugal*
[b] *Católica Lisbon School of Business & Economics and Carnegie Mellon University*

Keywords: experimental statistics, COVID-19, mobility, privately-held data

Abstract: The COVID-19 pandemic and the restrictive measures implemented since the first registered cases in the country – confinement, limitation of movement between municipalities, and mandatory teleworking – have had, over time, a heterogenous impact on the levels of mobility of the population. Context and epidemiological data are very important for understanding how COVID-19 and containment policies affect population behaviour, firms' performance and territorial interdependencies. With the increasing digitalization of societies and the massive use of new technologies in day-to-day activities and social interactions, several private entities produce large volumes of fine-grained individual behavioural data daily. These data support new digital business models and can serve as a relevant source of information of public interest ([3]), for example, to enable statistical authorities to capture and monitor new and unexpected social, economic, or environmental phenomena. Therefore, statistical offices have been looking into the potential of this type of data for official statistics ([1]; [2]; [4]). In this paper, we use an example of such data sources. We use Facebook's "Data for Good" initiative (e.g., [7]) and its indicator of the fraction of the population "staying put." to describe mobility patterns in Portugal during the different stages of the COVID-19 pandemic. These analyses were initially made public in Statistics Portugal's experimental statistics website (StatsLab) ([5], [6]).

References

[1] Berg, A. The Case for Regulated Access to Privately Held Data for Official Statistics. *Presentation at the High-Level Event "Data from and for Society" organised under the Portuguese Presidency of the Council of the EU, 3-4 June 2021*

[2] ESS. Position paper on access to privately held data which are of public interest. Opening up new data sources for a new generation of official statistics – in light of the growing European Digital Single Market and the revision of the Public Sector Information Directive, European Statistical System (ESS), 2017.

[3] EU. Towards a European strategy on business-to-government data sharing for the public interest. Final report prepared by the High-Level Expert Group on

Business-to-Government Data Sharing. Luxembourg: Publications Office of the European Union, 2020. doi: 10.2759/731415.

[4] Groves, R.M., Harris-Kojetin, B. A. (Eds.). Using Private-Sector Data for Federal Statistics. In Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy (pp. 55-72). Washington, DC: The National Academies Press, 2017. doi: 10.17226/24652

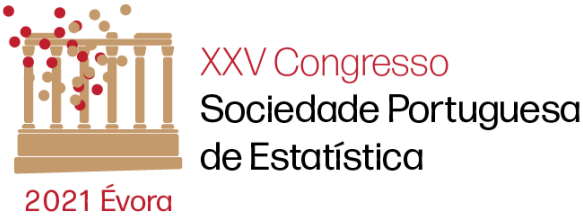
[5] INE. COVID-19: uma leitura do contexto demográfico e da expressão territorial da pandemia. Press Release, published on 12 May 2021.

[6] INE. Population mobility at regional level in the context of the COVID-19 pandemic. StatsLab Press Release, published on 20 November 2020.

[7] Maas, P. et al. Facebook Disaster Maps: Aggregate Insights for Crisis Response and Recovery. Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, Valencia, Spain, 2019.

Portuguese Central Bank

—Banco de Portugal—



Rita Sousa
Portuguese Central Bank, rcsousa@bportugal.pt

From BPstat to Stata

Nuno Azevedo[a,b]
ncazevedo@bportugal.pt

[a] *Banco de Portugal*
[b] *NIPE, Universidade do Minho*

Keywords: BPstat, bpstatuse, metadata, Stata, statistical data

Abstract: We will present a practical and introductory session to the *bpstat*. *bpstat* is a Stata package, developed by BPLIM - Banco de Portugal Microdata Research Laboratory, that allows users to import over 210000 statistical series from Banco de Portugal BPstat Database - an online statistical dissemination website that provides a large volume of statistical data and information. The package also contains some tools to help users search for series of interest and collect their metadata.

References

[1] https://github.com/BPLIM/Tools/tree/master/ados/General/bpstat
[2] https://bpstat.bportugal.pt/

Empirical applications on BdP data

Nuno R. Silva[a]
nrsilva@bportugal.pt

[a] *Banco de Portugal*

Keywords: extreme loss events, logit model, panel data, simulation-based multi-factor model

Abstract: In this talk, the use of Bank of Portugal databases is illustrated with two works. In the first work, titled "On the measurement of Portuguese firms' fixed operating costs", central balance sheet data is used to estimate fixed operating costs at the firm level using a high dimensional fixed effects regression model. This is an article that contributes to the corporate finance literature in operating leverage. Following the last two years lock-downs, this is a topic that became particularly relevant also for policy makers. The second work, titled "Sectoral concentration risk in Portuguese banks' loan exposures to non-financial corporations", implements a simulation-based multi-factor model to estimate the loss distribution associated with loans to non-financial firms of Portuguese banks. This was done using central credit responsibility information and a logit-based credit risk model estimated using mainly central balance sheet information. This work explores the role of sector concentration in the likelihood and severity of extreme loss events.

Using BPLIM to Access Data for Research Purpose

Rita Sousa[a,b]
rcsousa@bportugal.pt

[a] *Banco de Portugal*
[b] *Centro de Matemática e Aplicações, FCT-UNL*

Keywords: BPLIM, microdata, research

Abstract: The Banco de Portugal Microdata Research Laboratory (BPLIM) started its activity in 2016 and it is located in the Porto branch of Banco de Portugal. It is an autonomous unit within the Economics and Research Department, with the core mission of supporting the production of research projects and studies about the Portuguese economy.
Through BPLIM, both internal and external researchers gain access to well documented and anonymized micro data sets customized to their particular needs. Since BPLIM allows remote access to the data it hopes to attract the attention of both national and international researchers.
So far, nearly 200 projects have been opened, a large part of which are still active. Thus, BPLIM has actively contributed to scientific research.

References

[1] Banco de Portugal Microdata Research Laboratory (BPLIM). Webpage. https://bplim.bportugal.pt/

[2] Banco de Portugal Microdata Research Laboratory (BPLIM) (2021): Central Balance Sheet Annual Data. Extraction: June 2021. Version: V1. BANCO DE PORTUGAL. Dataset. https://doi.org/10.17900/CB.CBA.Jun2021.V1

[3] Banco de Portugal Microdata Research Laboratory (BPLIM) (2021): Central Balance Sheet Harmonized Panel. Extraction: June 2021. Version: V1. BANCO DE PORTUGAL. Dataset. https://doi.org/10.17900/CB.CBHP.Jun2021.V1

[4] Banco de Portugal Microdata Research Laboratory (BPLIM) (2019): Central Credit Responsibility Database - Firm Level Data. Extraction: June 2019. Version: V1. BANCO DE PORTUGAL. Dataset. https://doi.org/10.17900/CRC.FRM.Jun2019.V1

[5] Banco de Portugal Microdata Research Laboratory (BPLIM)(2021): Bank Balance Sheet Monthly Data. Extraction: June 2021. Version:V1. BANCO DE PORTUGAL. Dataset. https://doi.org/10.17900/BBS.Jun2021.V1

[6] Banco de Portugal Microdata Research Laboratory (BPLIM)(2020): Historical Series of the Portuguese Banking Sector Data. Extraction: October 2020. Version:V1. BANCO DE PORTUGAL. Dataset. https://doi.org/10.17900/SLB.Oct2020.V1

The Portuguese Central Credit Register

Marta Veloso[a]
mveloso@bportugal.pt

[a] *Banco de Portugal*

Keywords: central credit register, credit data, granular data

Abstract: The Portuguese Central Credit Register (CCR) is an information system managed by the Statistics Department of Banco de Portugal, which contains granular data on credit and credit risk, on a contract-by-contract basis. The main goal of the CCR is to provide information to the credit institutions to assist them in their credit risk assessment of their clients or potential clients. Based on CCR data, Banco de Portugal provides a service to the general population by making available the credit report of each individual or firm, which contains the full picture of the respective indebtedness vis-a-vis the resident financial system.
Additionally, according to CCR's legal framework, CCR data be used internally by Banco de Portugal to various purposes, namely, for prudential and conduct supervision of credit institutions, monetary policy, financial stability, compilation of statistics and economic research. Given these multi-purpose uses, credit attributes and other variables related to the classification of entities were defined in order to be meaningful for economic analysis and to be useful for all CCR purposes. Being a system with granular data, CCR allows moving beyond aggregates and compiling comprehensive and detailed statistics on credit without having to ask for more data to reporting institutions.
The richness of the CCR is leveraged by its integration with other granular datasets managed by Banco de Portugal like the Central Balance Sheet Database (CBSD) and the business register database for the accurate classification of entities. Moreover, the system implemented in 2018 introduced a new level of possibilities in terms of analysis and research given the new attributes related to each contract, protection or entity (and the relation between them) that started to be available in the CCR.

References

[1] Decreto-Lei n.º 208/2008, de 14 de outubro, que aprova o regime jurídico relativo à Central de Responsabilidades de Crédito.

[2] Instrução do Banco de Portugal n.º 17/2018, que regulamenta o funcionamento da Central de Responsabilidades de Crédito.

RBras
Brazilian Region International Biometric Society

—Região Brasileira da Sociedade Internacional de Biometria—

Paulo Canas Rodrigues
Federal University of Bahia, paulocanas@gmail.com

Detecting Spatial Clusters of Disease Infection Risk Using Sparsely Sampled Social Media Mobility Patterns

Renato Assunção[a,b], Roberto Souza[b], Wagner Meira[b], Daniel Neill[c]
rassuncao@esri.co, assuncao@dcc.ufmg.br

[a] *ESRI Inc.*
[b] *Departamento de Ciência da Computação, UFMG*
[c] *New York University*

Keywords: cluster detection, disease cluster, spatial cluster

Abstract: Standard spatial cluster detection methods used in public health surveillance assign each disease case a single location (e.g., patients' home address), aggregate locations to small areas, and monitor the number of cases in each area over time. However, this approach lacks the accuracy and specificity to deal with infectious disease outbreaks where human mobility plays a key role. Here, we use social media data to capture the individuals' mobility. We propose two new spatial scan methods (the unconditional and conditional spatial logistic models) which search for spatial clusters of increased infection risk in mobility patterns by maximizing a generalized log-likelihood ratio statistic over subsets of the data.

The methods correctly account for the multiple, varying number of spatial locations observed per individual, either by non-parametric estimation of the odds of being a case or by matching case and control individuals with similar numbers of observed locations. By applying our methods to synthetic and real-world scenarios, we demonstrate robust performance on detecting spatial clusters of infection risk from mobility data, outperforming competing baselines.

An extended random-effects approach to modeling repeated, overdispersed count data

Clarice G.B. Demétrio[a], Geert Molemberghs[b], Geert Verbeke[c]
clarice.demetrio@usp.br, geert.molenberghs@uhasselt.be,
geert.verbeke@kuleuven.be

[a] *University of São Paulo, ESALQ, Piracicaba, SP, Brazil*
[b] *University of Hasselt, Dienpebeck, Belgium*
[c] *Katholieke Universiteit Leuven, Belgium*

Keywords: correlated data, Negative-binomial model, Poisson model, random effects

Abstract: Non-Gaussian outcomes are often modeled using members of the exponential family. The Poisson model for count data falls within this tradition. The family in general, and the Poisson model in particular, are at the same time convenient since mathematically elegant, but in need of extension. Two of the main rationales for existing extensions are (1) the occurrence of overdispersion, in the sense that the variability in the data is not adequately captured by the model's prescribed mean-variance link, and (2) the accommodation of data hierarchies owing to, for example, repeatedly measuring the outcome on the same subject, recording information from various members of the same family, etc. There is a variety of overdispersion models for count data, such as, for example, the negative-binomial model. Hierarchies are often accommodated through the inclusion of subject-specific, random effects. Though not always, one conventionally assumes such random effects to be normally distributed. While both of these issues may occur simultaneously, models accommodating them at once are less than common. Here we propose a model, accommodating overdispersion and clustering through two separate sets of random effects, of gamma and normal type, respectively. The model extends both classical overdispersion models for count data, as well as the generalized linear mixed model. Apart from model formulation, we briefly discuss several estimation options, and then settle for maximum likelihood estimation with both fully analytic integration as well as hybrid between analytic and numerical integration. The methodology is applied to data from a study in epileptic seizures.

References

[1] Molenberghs, G., Verbeke, G., Demétrio, C.G.B. An extended random-effects approach to modeling repeated, overdispersed count data. *LIDA*, 13, 513-531, 2007.

Regressão Quantílica aplicada ao estudo de associação genômica de incidência da ferrugem em *Coffea canephora*

Gabriela França Oliveira [a], Ana Carolina Campana Nascimento [a], Moysés Nascimento [a], Brunna de Figueredo Duarte[a], Maurício de Oliveira Celeri[a], Eveline Teixeira Caixeta Moura[a,b]
gabriela.franca@ufv.br, ana.campana@ufv.br, moysesnascim@ufv.br, brunna.duarte@ufv.br, mauricio.celeri@ufv.br, eveline.caixeta@embrapa.br,

[a] *Universidade Federal de Viçosa - UFV*
[b] *Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA, Embrapa Café*

Keywords: café, GWAS, marcadores moleculares, melhoramento genético, quantis condicionais

Abstract: A cafeicultura é uma atividade agrícola de grande importância, visto que o café é um dos produtos mais comercializados no mundo. Diante dessa importância, programas de melhoramento genético têm buscado novas estratégias para obtenção de genótipos mais produtivos, resistentes e adaptados. Uma técnica ainda pouco utilizada em estudos de Associação Genômica (GWAS) é a Regressão Quantílica (RQ). Essa abordagem permite identificar variações genéticas que podem estar associadas com características fenotípicas de interesse ao longo de toda sua distribuição de probabilidade, diferentemente dos métodos tradicionais baseados em média. Neste sentido, o objetivo deste estudo foi utilizar a RQ para identificar associações significativas ao longo de diferentes quantis para a característica fenológica incidência da ferrugem de Coffea Canephora. Foram utilizados 165 genótipos genotipados para 17.885 marcadores do tipo SNP. Esses dados são provenientes da Empresa Brasileira de Pesquisa Agropecuária e da Empresa de Pesquisa Agropecuária de Minas Gerais em colaboração com a Universidade Federal de Viçosa. Os efeitos dos marcadores foram estimados por meio da RQ nos quantis 0.2, 0.5 e 0.8. A significância dos efeitos foi avaliada utilizando o teste t, considerando um nível de significância de 5% corrigido para múltiplos testes pela taxa de falsas descobertas. O modelo ajustado para o quantil 0.2 apresentou um alto poder detecção, com 97 marcadores significativos e, para os quantis 0.5 e 0.8 foram encontrados 4 e 13 marcadores significativos, respectivamente. Esses resultados indicam que a RQ em quantis inferiores acarreta um aumento no número de marcadores significativos detectados.

Acknowledgements: O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

A complex trait with unstable QTLs can follow from component traits with stable QTLs: An illustration by a simulation study in pepper

Paulo Canas Rodrigues[a]
paulocanas@gmail.com

[a] *Federal University of Bahia, Salvador, Brazil*

Keywords: crop growth modelling, genotype-by-environment interaction, QTL detection, quantitative genetics

Abstract: Complex traits like yield are those in which phenotypic variation can be modelled as a linear function of a set of quantitative trait loci (QTLs) with environment dependency. This environment dependency can be observed at a phenotypic level as genotype-by-environment interaction (GEI) for yield itself and at an underlying genetic level as QTL-by-environment interaction (QEI). We show how GEI in yield may follow from pleiotropic QTLs for yield components that themselves are not environment dependent. We generated synthetic yield data via a crop growth model and analysed these data by common statistical models for GEI and QEI. QTLs for yield were pleiotropic with those for yield components. Such pleiotropy offers a path for improvement of yield under GEI. As model system, we used sweet pepper (Capsicum annuum L.) and developed an eco-physiological model for yield with seven genotype specific inputs or yield components. Synthetic yields were simulated for a back cross population of 500 lines across a factorial combination of four major environmental drivers. The yield components were given a simple QTL basis and produced credible amounts and patterns of GEI for yield. The QEI for yield could be interpreted from the expression of QTLs for yield components and the interaction of these components with the environmental drivers. We see the generation of synthetic yield data via crop growth models followed by an analysis with statistical models for GEI and QEI to quantify the contribution of yield components to GEI as a helpful step in the development of yield prediction models for complex traits across environments that can also serve as a basis for decisions on selection strategies of complex traits. Joint work with Ep Heuvelink (Wageningen University & Research, The Netherlands), Leo F.M. Marcelis (Wageningen University & Research, The Netherlands)„ Scott C. Chapman (CSIRO Agriculture and Food, Queensland Bioscience Precinct, Australia; The University of Queensland, Australia), and Fred A. Van Eeuwijk (Wageningen University & Research, The Netherlands).

SPE–CLAD
Statistics, a tool for sustainability

—Estatística, uma ferramenta para a sustentabilidade—

XXV Congresso
Sociedade Portuguesa
de Estatística

Clara Cordeiro
University of Algarve, ccordei@ualg.pt
Filomena Teodoro
Portuguese Naval Academy, mteodoro64@gmail.com

Climate change projections for heating and cooling degree-days for Portugal

Cristina Andrade[a,b], Sandra Mourato[c,d], João Ramos[c,e]
c.andrade@ipt.pt, sandra.mourato@ipleiria.pt, joao.ramos@ipleiria.pt

[a] *Instituto Politécnico de Tomar, Natural Hazards Research Center (NHRC.ipt), Estrada da Serra, Quinta do Contador, 2300-313 Tomar, Portugal*
[b] *CITAB, University of Trás-os-Montes and Alto Douro, PO Box 1013, 5001-801 Vila Real, Portugal*
[c] *Politécnico de Leiria, Apart. 4133, 2411-901 Leiria, Portugal*
[d] *MED, Universidade de Évora, Pólo da Mitra, Ap. 94, 7006-554 Évora, Portugal*
[e] *INESC Coimbra, DEEC, Rua Sílvio Lima, Polo II, 3030-290 Coimbra, Portugal*

Keywords: climate change, cooling degree-day (CDD), heating degree-day (HDD), energy demand, geostatistical methods

Abstract: Climate change is expected to influence cooling and heating energy demand of residential buildings and affect overall thermal comfort. Towards this end, the heating degree-day (HDD), the cooling degree-day (CDD) were computed from an ensemble of 7 high-resolution bias-corrected simulations attained from EURO-CORDEX under RCP4.5 and RCP8.5. These three indicators were analyzed for 1971-2000 (from E-OBS) and 2011-2040 and 2041-2070, under both RCPs. Evaluation of the Geostatistical methods using RMSE and ME showed that the estimation of HDD, CDD by OK, was the most accurate by comparison with OCK and IDW for all time periods and under both RCPs. The statistically significant (SS) anomalies were assessed by the Mann-Whitney-Wilcoxon test (MWW) and the SS trends by using the rank-based non-parametric Spearman's rho (SR) statistical test both at a 5% significance level. Results show that the overall spatial distribution of HDD trends for the 3 time-periods points out an increase of energy demand to heat internal environments in Portugal's northern-eastern regions, most significant under RCP8.5. It is projected an increase of CDD values for both scenarios; however, SS linear trends were only found for 2041-2070 under RCP4.5. The need for cooling is almost negligible for the remaining periods, though linear trend values are still considerably higher for 2041-2070 under RCP8.5. By the end of 2070, higher amplitudes for all indicators are depicted for southern Algarve and Alentejo regions, mainly under RCP8.5. For 2041-2070 the Centre and Alentejo (North and Centre) regions present major positive differences for HDD(CDD) under RCP4.5(RCP8.5), within the 5 NUTS II regions predicting higher heating(cooling) requirements for some locations. Therefore, to ensure thermal comfort, reduce energy consumption, and greenhouse gas emissions, new policies are needed.

Acknowledgements: This work was supported by National Funds by FCT - Portuguese Foundation for Science and Technology, under the project UIDB/04033/2020.

References

[1] Andrade, C., Mourato, S., Ramos, J. Heating and Cooling Degree-Days Climate Change Projections for Portugal. *Atmosphere*, 12, 715, 2021. https://doi.org/10.3390/atmos12060715

Brazilian Electric Power Distributors: Excellence Models

Andrade, M. A. P.[a,b], Carrasco, A.[a], Rosa, A.[a,c], Teodoro, M.F.[d,e]
marina.andrade@iscte-iul.pt

[a] *ISCTE – IUL, Lisbon University Institute*
[b] *ISTAR-IUL, Lisbon University Institute*
[c] *BRU-IUL, Lisbon University Institute*
[d] *CEMAT, Instituto Superior Técnico, Lisbon University*
[e] *CINAV, Portuguese Naval Academy, Military University Institute*

Keywords: energy, excellence, MEG

Abstract: The Brazilian Association of Electric Energy Distributors and the National Quality Foundation (FNQ), among others, supported Brazilian electric energy distributors that, as of 2006, use the Management Excellence Model (MEG) as an organizational diagnosis tool. MEG, according to FNQ, is a world-class business management system. In line with the main recognized international systems, the model is the result of the experience, knowledge and research of several national and international organizations and specialists. The latest version of the MEG was launched in October 2016 and is the FNQ's reference for the fulfillment of its mission: to encourage and support Brazilian organizations in the development of their management so that it becomes sustainable, cooperative and contributes to the generation of value for society. It is therefore important to assess the impact of the use of the MEG reference model by Brazilian electricity distributors and its effects on the operational indicators of continuity of supply DEC (Equivalent Outage Duration per Consumer Unit) and FEC (Equivalent Outage Frequency per Consumer Unit); compliance with DEC and FEC regulatory limits, in addition to customer satisfaction results, measured by the IASC.

References

[1] Boulter, L., Bendell, T., Dahlgaard, J.J. Total quality beyond North America: A comparative analysis of the performance of European excellence award winners. *Int. J. of Operations and Production Management*, 33, 197–215, 2013.

[2] Corredor, P., Goñi, S. Quality awards and performance: Is there a relationship? *The TQM Journal*, 22(5), 529–538.169, 2010.

[3] Buccelli, D.O., Costa Neto, P.L.O. Prêmio Nacional da Qualidade: Gestão da qualidade ou qualidade da gestão? *Proceedings of the XXXIII National Meeting of Production Engineering. A Gestão dos Processos de Produção e as Parcerias Globais para o Desenvolvimento Sustentável dos Sistemas Produtivos*, October, 8th-11st, Brasil, 2013.

[4] Escrig, A.B., Menezes, L.M. What characterizes leading companies within business excellence models? An analysis of "EFQM Recognized for Excellence" recipients in Spain. *Int. J. of Production Economics*, 169, 362–375, 2015. .

[5] Fundação Nacional da Qualidade Critérios de Excelência. *Melhores em Gestão. Instruções para candidatura 2018.* FNQ, São Paulo, 2018.

Uncovering Abnormal Water Consumption Patterns for Sustainability's Sake

Ana Borges[a], Clara Cordeiro[b], M. Rosário Ramos[c]
aib@estg.ipp.pt, ccordei@ualg.pt, marosram@uab.pt

[a] *CIICESI, Escola Superior de Tecnologia e Gestão, Politécnico do Porto*
[b] *FCT, Universidade do Algarve, and CEAUL, FCUL*
[c] *DCeT, Universidade Aberta, and CMAFcIO, FCUL*

Keywords: billed water consumption, change point, seasonal and trend decomposition using loess, time series

Abstract: Water has been recognized, in the last decades, as an essential resource for guaranteeing economic development and maintaining living standards [1]. Water companies' awareness for the responsible use of water has gained importance, with climate changes emphasizing this need. Controlling domestic water usage can help to reduce water consumption and protect the environment. In this context, water utilities feel the need to improve and develop mechanisms for predictive water planning based on data analysis. This study proposes a strategy that detects abnormal water consumption patterns, namely significant increases or decreases. An abrupt decrease can be related to apparent water losses, which could result in financial losses for the company. Detecting an anomalous increase will allow companies to take measures such as alert their consumers from a perspective of environmental sustainability. Thus, promoting sustainable consumption behaviors. The basis of the approach is a combination of methods to analyze billed water consumption time series. In the first step, the time series is decomposed using Seasonal-Trend decomposition based on Loess [2]. Next, breakpoint analysis is performed on the seasonally adjusted time series. The search for decreasing or increasing changes in the periods between breakpoints through the Mann-Kendall [3] test and Sen's [4] slope estimator, and an indicator for this change is presented. The strategy is applied to data on billed water consumption from the Algarve, Portugal.

Acknowledgements: Ana Borges, Clara Cordeiro and M.Rosário Ramos work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through projects UIDB/04728/2020, UIDB/00006/2020 and UIDB/04561/2020, respectively.

References

[1] Cámara, Á., Llop, M. Defining Sustainability in an Input–Output Model: An Application to Spanish Water Use. *Water*, 13(1), 1, 2021.
[2] Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I. STL: A seasonal-trend decomposition procedure based on loess. *Journal of official statistics*, 1990, 3–73.
[3] Kendall, M. Rank correlation methods *Charles Griffin*, 1975.
[4] Sen, P. K. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 1968 , 63(324), 1379–1389.

Análise de Dados SCADA de um Parque Eólico em Terreno Complexo

Alda Carvalho[a,b], Cláudia Casaca[b,c], Daniel Vaz[d], Tiago Silva[b,d]
alda.carvalho@isel.pt, claudia.casaca@isel.pt, dv@fct.unl.pt,
tan.silva@fct.unl.pt

[a] *CEMAPRE–REM, Universidade de Lisboa*
[b] *CIMOSM, ISEL–Instituto Politécnico de Lisboa*
[c] *ADEM, ISEL–Instituto Politécnico de Lisboa*
[d] *UNIDEMI–FCT, Universidade Nova de Lisboa*

Keywords: análise de dados, micro-sitting, parque eólico, sistema SCADA

Abstract: A produção de energia a partir de fontes renováveis tem um papel preponderante no desenvolvimento e implementação de políticas sustentáveis de energia e de ambiente. No entanto, a exploração de fontes renováveis, nomeadamente a eólica, apresenta alguns desafios; além do caracter de intermitência que este tipo de geração de energia apresenta, há ainda que ter em conta o retorno financeiro deste tipo de investimentos. A monitorização dos parques eólicos é feita através de sistemas de supervisão e aquisição de dados (SCADA – *supervisory control and data acquisition*). Existem sensores em vários componentes das turbinas eólicas, que recolhem diferentes variáveis de interesse, com o objetivo de monitorizar a produção elétrica, mas também de avaliar o estado de condição dos diversos componentes constituintes destas complexas estruturas.

A base para este trabalho foi a implementação de relatórios automáticos que integrem toda a informação recolhida pelo sistema SCADA. Estes relatórios permitem, não só uma leitura rápida do estado individual de cada turbina, mas também uma identificação de possíveis interações entre turbinas. O tratamento estatístico dos dados já permitiu perceber diferenças no vento que, num mesmo instante, chega a cada uma das turbinas[1]. No decorrer deste trabalho surgiram alguns desafios relacionados com a qualidade dos dados (valores não admissíveis, dados omissos, erros sistemáticos, intervenções humanas, etc.). Nesta comunicação apresenta-se um resumo do processo de tratamento e análise dos dados, focando a importância da multidisciplinaridade da equipa e das ferramentas de análise resultantes deste trabalho.

Acknowledgements: Este trabalho foi realizado no âmbito dos projectos *Modelos de diagnóstico para manutenção em parques eólicos* (MoDMaPE1 e MoDMaPE2, IDI&CA–IPL, 2019/2021).

References

[1] Casaca, C.S.S.L., Vaz, D.C., Silva, T.A.N. , Carvalho, A.C.J.V.N. An Analysis of Wind Farm Data to Evidence Local Wind Pattern Switches Near a Plateau. *4th International Conference on Numerical and Symbolic Computation: Developments and Applications*. April, 11-12, 2019, Portugal, ISBN: 978-989-99410-5-2, ©ECCOMAS (2019).

SPE–FENStatS
Accreditation of statisticians: Why and how?

XXV Congresso
Sociedade Portuguesa
de Estatística

Magnus Pettersson
FENStats, www.fenstats.eu/accreditation

FENStatS accreditation of statisticians: Why and how?

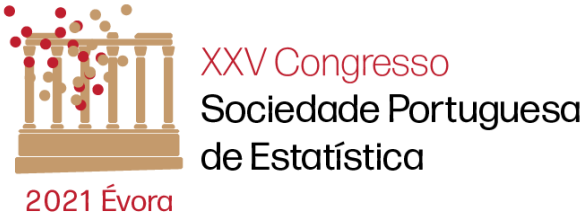
Magnus Pettersson[a]

[a] *FENStatS accreditation committee*

Abstract: In this talk we will present the motivation for the accreditation system for statisticians that FENStatS launched one year ago. This gives an opportunity, especially for applied statisticians, to summarize their credentials and will also promote quality in statistical work.

Oral Sessions

—Comunicações Orais—



A Joint Model for Multiple Longitudinal Outcomes and Terminal and Recurrent Events

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

[a] *Department of Biostatistics, Erasmus MC Rotterdam, the Netherlands*
[b] *Department of Epidemiology, Erasmus MC Rotterdam, the Netherlands*
[c] *Department of Pediatrics, Cincinnati Children's Hospital Medical Center, United States*
[d] *Department of Mathematical Sciences, University of Cincinnati, United States*
[e] *Cystic Fibrosis Foundation, United States*

Keywords: joint models, multivariate longitudinal data, recurrent events, registry analysis, survival analysis

Abstract: Cystic fibrosis (CF) is an inherited disease primarily affecting the lungs and gastrointestinal tract. It is of clinical interest to simultaneously investigate the association between the risk of recurrent acute respiratory events, lung function decline, the evolution of a patient's growth and nutritional status, and the risk of lung transplant or death using all the available U.S. CF Foundation (CFF) patient registry data. We intend to explore different forms of association between the longitudinal markers and the events of interest. We propose a joint modeling framework accommodating multiple longitudinal markers, a recurrent event process, and a terminal event. The terminal outcome considers informative censoring due to lung transplantation or death from respiratory failure. Novel elements of our approach, compared to previously proposed joint models for recurrent events, are: (i) allowance for multiple longitudinal markers with different distributions, (ii) specifying various functional forms to link these markers with the risk of a recurrent event and the risk of the terminating event, and (iii) accommodation of discontinuous intervals of risk, with the time being definable in terms of the gap or calendar timescale. Full MCMC algorithm implementation in C++ enables model fit in a timely fashion, despite its complexity. The developed model will be available in the R statistical package JMbayes2. The proposed multivariate joint model allows making more efficient use of all the available CFF registry data. It thereby brings new insights into CF disease progression and contributes to monitoring and treatment strategies.

References

[1] Rizopoulos, D., Papageorgiou, G., Afonso, P.M. JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. 2021. https://drizopoulos.github.io/JMbayes2/

Behaviour of Different Strategies for Cost-Sensitive Logistic Regression Applied to the Classification of Paediatric Familial Hypercholesterolemia

João Albuquerque[a,b,c], Ana Medeiros[c,d], Catarina Alves[c,d], Mafalda Bourbon[c,d], Marília Antunes[a,e]
joaodavid.alb@gmail.com

[a] *Centro de Estatística e Aplicações, FCUL, Portugal*
[b] *Departamento de Bioquímica, FMUP, Portugal*
[c] *Grupo de Investigação Cardiovascular, INSA, Portugal*
[d] *Instituto de Biossistemas e Ciências Integrativas, FCUL, Portugal*
[e] *Departamento de Estatística e Investigação Operacional, FCUL, Portugal*

Keywords: familial hypercholesterolemia, logistic regression, SMOTE, thresholding, weighting

Abstract: Familial hypercholesterolemia (FH) is an inherited disorder of lipid metabolism. When developing a classification model applied to FH diagnosis, misclassification costs should be accounted for. The main purpose of the current work was to set a cost-sensitive framework for the diagnosis of FH. Using a sample of 408 subjects at paediatric age (2-17y), a logistic regression (LR) model was trained using several biological and biochemical variables. In a first step, it was observed how the model behaved when increasing the importance of FH positive class, by a ratio of up to 5:1. Three different strategies were adopted for this purpose, resampling according to *Synthetic minority Oversampling Technique* (SMOTE), weighting and thresholding [1]. Performance measures included the area under the ROC curve, and several operating characteristics (OC), and were also compared with the values obtained with clinical diagnosis criteria. Secondly, an optimum model was defined for each of these methods, by calculating the minimum expected cost (MEC), adopting different scenarios for the cost matrix. All methods revealed similar behaviour for OC with the increase of FH positive class importance, with steeper changes in the first few proposed values. The required rate of increase to obtain the MEC was proportional to the ratio between costs attributed to False Positive (FP) and False Negative (FN) observations. The LR parameters could be adjusted to outperform clinical criteria in every cost matrix scenario. Results from this study suggest significant cost reduction can be achieved by the referred methods, in comparison with standard clinical diagnosis procedures.

Acknowledgements: Research supported by the programme Norte2020 (operação NORTE-08-5369-FSE-000018) and by FCT projects UID/MAT/00006/2020 and PTDC/SAU-SER/29180/2017.

References

[1] Branco et al. *ACM Computing Surveys (CSUR)*, 49(2), 1-50, 2016 doi:10.1145/2907070

Risk model with dependent frequency and severity, premium and ruin probability calculation

Renata G. Alcoforado[a,b], Alfredo D. Egídio dos Reis[a]
alcoforado.renata@ufpe.br, alfredo@iseg.ulisboa.pt

[a] *ISEG & CEMAPRE, Universidade de Lisboa*
[b] *Department of Accounting and Actuarial Science, Universidade Federal de Pernambuco*

Keywords: dependent risks, insurance risk model, premium calculation, real data application, ruin probability

Abstract: A common assumption made in classical risk theory with application in insurance modelling is that the "claim counts" and the "claim severity" are independent. It may not be the case in many situations. In recent times there have been authors working with models allowing some sort of dependence, these are good examples with application in motor insurance. They use different methods to capture dependence, eg. copulae, GLM and or even no particular hypothesis on the dependency structure.
In our case, by using real data from two different, but related, branches on housing and liability insurance from an anonymous insurer, we work numerical illustrations. We consider dependence between claim counts and severity of claims, as well as among individual claims in order to estimate subsequent premiums and corresponding ruin probabilities.
We work formulae and numerical results for both ruin probabilities and adjusted premiums in some sort of models. We start from studying two separate compound Poisson models and study existence of dependence, then its size using appropriate methods, either copulae, regression or mixing Poisson parametrization. Our database bring some challenges other than the automobile ones.

Acknowledgements: Authors were partially supported by the Project CEMAPRE/REM - UIDB/05069/2020 - financed by FCT/MCTES through national funds.

Distributional properties of the Lincoln-Petersen-like estimators under extreme lower recapture values

Russell Alpizar-Jara[a], Lígia Henriques-Rodrigues[a], Nilton Ávido[b]
alpizar@uevora.pt, ligiahr@uevora.pt, d44089@alunos.uevora.pt

[a] *CIMA-IIFA/DMAT-ECT, University of Évora, Portugal*
[b] *ISPH-CBDC, University of Mandume Ya Ndemufayo, Angola*

Keywords: abundance estimation, capture-recapture, extreme value theory, population estimation

Abstract: Capture-recapture studies have been widely used to monitor and quantify animal populations and species richness. The most basic and conceptual framework for capture-recapture models is based on a simple rule of three to intuitively derived the well-known Lincoln-Petersen estimator. The theoretical properties of this estimator date back to the work of Chapman, Bailey and others in 1950s, following various distributional assumptions for the number of recaptures. A common problem with its performance is related to low number of recaptures, leading to small sample biases and very low precision. In this work, we study its distributional properties based on results about the minimum of some discrete random variables. We will also highlight the importance of this estimator due to its wide range of applications and the use of mathematical statistics and extreme value theory to study the distributional properties.

Acknowledgements: This research has been partially supported by the Fundação Calouste Gulbenkian under the program Estágios Científicos Avançados em Matemática para docentes e investigadores dos PALOP, and for the Centro de Investigação em Matemática e Aplicações (CIMA), through the Project UIDB/04674/2020 of FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

[1] Nadarajah, S., Mitov, K. Asymptotics of Maxima of Discrete Random Variables, *Extremes*, 5, 287–294, 2002.

[2] Williams, B.K., Nichols, J.D., Conroy, M. *Analysis and Management of Animal Populations*. Academic Press, 2002.

Teaching and Learning Basic Statistics in Undergraduate Programs: an unexpected switch to online classes and assessment

Helena Alvelos[a], Ana Raquel Xambre[a]
helena.alvelos@ua.pt, raquelx@ua.pt

[a] *Department of Economics, Management, Industrial Engineering and Tourism & Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal*

Keywords: basic statistics, Moodle, online classes, online assessment, teamwork, Zoom

Abstract: Most Portuguese undergraduate programs in the Engineering, Sciences and Social Sciences include, at least, one course that addresses basic statistics techniques [1]. At the University of Aveiro one of such courses, named Statistical Techniques, is offered to first year Economics students and second year Industrial Engineering and Management students. This leads to an average total number of students attending the course of 165 (last five years). In the academic year of 2019/2020, when confinement was imposed due to the Covid-19 pandemic, it was necessary to change, in the period of a week, the teaching and learning system from a more traditional in person situation to a completely online version. It was necessary to resort to technology such as Zoom, for the online classes, and Moodle [2], for online assessment. The adaption to this new way of teaching and learning was abrupt and required the rethinking of the course and of its evaluation method. The teaching methodologies were adjusted and by changing the assessment so as to include teamwork it was possible to enhance students' engagement in the course, as well as promote autonomous learning. The purpose of this work is to show the main challenges faced when trying to offer students a successful learning experience in a very different context and within a very short timeframe. It will also be presented an analysis of the outcomes of the course, particularly students' grades, in order to better understand the impact of the changes.

Acknowledgements: This research was supported by the Portuguese National Funding Agency for Science, Research and Technology (FCT), within the Center for Research and Development in Mathematics and Applications (CIDMA), project UIDB/04106/2020.

References

[1] Guimarães, R.C., Cabral, J.S. *Estatística.* Verlag Dashöfer Lda., Lisboa, 2011. ISBN: 978-989-642-108-3

[2] Figueira, A., Figueira, C., Santos, H. *Moodle: criação e gestão de cursos online.* FCA – Editora de Informática, Lda., Lisboa, 2009. ISBN: 978-972-722-634-4

Robust alternatives of the hurdle model with Poisson regression

Conceição Amado[a], Manuela Souto de Miranda[b]
conceicao.amado@tecnico.ulisboa.pt, manuela.souto@ua.pt

[a] *CEMAT and IST, University of Lisbon*
[b] *CIDMA, University of Aveiro*

Keywords: hurdle model, Poisson regression, robustness

Abstract: Hurdle models are mixed models that can deal with excess of zeroes by considering two separate components, namely, a binary process and a truncated discrete distribution. They are particularly adequate for modelling counting processes when the occurrence of zero observations do not depend on the main generating process of the strictly positive counting. It is often the case when counting low-probability incidents, some Health Econometrics statistics or extreme values events *per* unity of time. The present study aims to compare robust estimators for the hurdle model when the positive component is modelled by a truncated Poisson regression. There are different proposals considered in the literature whose performance is investigated through a simulation study.

Acknowledgements: Research partially supported by National Funds through **FCT** – Fundação para a Ciência e a Tecnologia, projects UID/Multi/04621/2019 (CEMAT/IST-ID/ULisboa) and UIDB/04106/2020 (CIDMA/University of Aveiro).

References

[1] Cantoni, E., Zedini, A. A robust version of the hurdle model. *Journal of Statistical Planning and Inference*, 141, 1214–1223, 2011. doi:10.1016/j.jspi.2010.09.022

[2] Min, Y., Agresti, A. Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1 (1-2), 7–33, 2002.

[3] Hosseinian, S., Morgenthaler, S. Weighted maximum likelihood estimates in Poisson regression. *Book of Abstracts, International Conference on Robust Statistics*, Italy, 2011.

Abordagem Bayesiana para modelo de fração de cura inflacionado de zeros aplicado a dados de tempo de vida

Naiara Caroline Aparecido dos Santos[a,b], Talita Evelin N T de Moraes[c], Vera Tomazella[a,b], Danillo Magalhães Xavier Assunção[a,b]
naicaroline2@usp.br, talita_evel@usp.br, veratomazella@gmail.com, danilloxavier@usp.br

[a] *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP*
[b] *Departamento de Estatística, Universidade Federal de São Carlos - UFSCar*
[c] *Escola Superior de Agricultura Luiz Queiroz, Universidade de São Paulo - ESALQ/USP*

Keywords: bayesiana, diabete gestacional, fração de cura, inflação de zeros, stan

Abstract: Com os avanços nos tratamentos médicos, é esperado um aumento na expectativa de vida dos pacientes submetidos a novos tratamentos. Diante dessa nova realidade, o campo da estatística tem buscado apresentar modelos cada vez mais flexíveis para explicar melhor esses resultados. A metodologia proposta neste trabalho é fundamentada em dados de sobrevivência inflacionados de zero para lidar com situações nas quais existem uma fração de curados e uma grande proporção de zeros. Nossa abordagem nos permite acomodar três tipos de ocorrências: pacientes com tempo de sobrevida zero (ocorrência antes e/ou início do estudo); indivíduos que são suscetíveis e os não suscetíveis ao evento de interesse. Ilustramos a relevância prática do modelo através de um estudo com mulheres grávidas diagnosticadas com diabetes mellitus gestacional (DMG) antes das 24 semanas de gestação durante o pré-natal no Hospital das Clínicas da Universidade de São Paulo, Brasil. Consideramos a abordagem Bayesiana para estimação dos parâmetros do modelo proposto, por meio do algoritmo No-U-Turn Sampler (NUTS) usando a linguagem Stan em R. Os resultados mostram que o modelo proposto foi capaz de explicar ou prever o tempo de sobrevivência, bem como a fração de curados e a proporção de falhas zero.

Acknowledgements: À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro.

Nonparametric Bayesian Learning of GYMs: What Hides Behind the Shape of a Yield Curve?

Emmanuel Bernieri[a], Miguel De Carvalho[a]
s1771902@sms.ed.ac.uk, miguel.decarvalho@ed.ac.uk

[a] *School of Mathematics, University of Edinburgh*

Keywords: dependent dirichlet process, functional regression, measurement error, nonparametric Bayes, yield curves

Abstract: We propose a Bayesian nonparametric approach for modeling a scalar output conditioned on a functional input. We use this approach to model the production index growth of a country conditioned on its yield curve. To fully grasp the effect of a shift in the shape of a yield curve on the production index growth we develop a new functional parameter called GYM (growth–yield map) which maps economic growth reaction to different shapes of yield curves. The application of the proposed methods to real data reveals some interesting insights on how growth patterns of different European economies relate with different shapes of yield curves. Beyond uncovering the folk wisdom from Finance, that inverted yields are related with periods of lower growth, the proposed methods allows us to assess how the strength of that link changes over for different economies, as well as how much the intercept and slope of yields plays a role.

Using GAMs to understand a 30-year fish species Portuguese historical dataset from coastal transition ecosystems

Pedro Brandão[a], Luis da Costa[b], Susana França[a,b], Tiago A. Marques[a,c]
fc49345@alunos.fc.ul.pt, dacosta.luis@gmail.com, sofranca@fc.ul.pt
tiago.marques@st-andrews.ac.uk

[a] *Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboan, Portugal*
[b] *MARE - Center for Marine and Environmental Sciences Centre (MARE), Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*
[c] *Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland*

Keywords: climate change, CoastNet research infrastructure, estuarine fish communities, generalized additive models, historical data

Abstract: Coastal transition ecosystems like estuaries and lagoons are amongst the most productive aquatic ecosystems on the planet, recognized worldwide as a fundamental component of coastal areas in terms of biological relevance and anthropogenic use. Estuaries along the Portuguese coast differ in their geomorphological and hydrological characteristics. These systems play an important role in terms of nursery areas for economically important fish species. Although several authors have observed high specific variability in inter-estuarine fish communities along the Portuguese coast, few studies exploit the knowledge about factors that influence it. Most studies involving Portuguese estuarine fish communities focus on a single estuary and, when several are addressed, only a single factor is used to assess the specific variability of the communities. CoastNet is a research infrastructure whose main objective is to deepen knowledge about the functioning of national coastal ecosystems, through a remote monitoring system. In addition to autonomously collecting biological and environmental data, it includes historical datasets that allow to understand whether ecological functions performed by Portuguese estuaries have been altered in the last 30 years. The present work aims to demonstrate major challenges associated with scattered historical datasets using generalized additive models (GAMs) to analyse the ecological role played by the main Portuguese estuaries for fish species. We describe how fish species richness in Portuguese estuaries changes at different spatial and temporal scales, being influenced by abiotic factors like temperature and salinity. We discuss the difficulty in making inferences in long term datasets unbalanced over space and time.

Acknowledgements: TAM thanks partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020); LC and SF fully and partially supported, respectively, by COASTNET (PINFRA/22128/2016), in the framework of the National Roadmap of Research Infrastructures of strategic relevance.

Profit optimization for SDE cattle growth models using a general profit structure

Carlos A. Braumann[a,b], Gonçalo Jacinto[a,b], Patrícia A. Filipe[c,b]
braumann@uevora.pt, gjcj@uevora.pt, patricia.filipe@iscte-iul.pt

[a] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora*
[b] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora*
[c] *Iscte-Instituto Universitário de Lisboa, Iscte Business School, Quantitative Methods for Management and Economics Department*

Keywords: general profit structure, individual growth model, profit optimization, stochastic differential equations

Abstract: Regression on deterministic models is not appropriate to study individual growth of animals in randomly fluctuating environments. So, we use stochastic versions of the classical deterministic models, in the form of a general stochastic differential equation model. The aim is helping farmers to optimize the profit obtained by raising and selling an animal.
We have previously considered the case of a simple profit structure ([1]) with a constant price per kg paid to the farmers and raising costs that have a fixed component and a variable component (for costs like handling and feeding) proportional to the time the animal is being raised. We now generalize the results to the more challenging and realistic market situation where the price per kg paid to the farmers depends on the animal's age and weight category and to the case where the feeding cost component per unit time is proportional to the animal's weight. We obtain the profit probability distribution, its first two moments and other quantities of interest, which are required to determine the optimal selling age, i.e. the selling age that maximizes the expected profit. We also consider the particular case where one lacks information on the cost of the feeding component per unit time and unit weight of the animal, where we use an average feeding cost of an animal per unit time.

Acknowledgements: The Centro de Investigação em Matemática e Aplicações is supported by the Fundação para a Ciência e a Tecnologia, project UID/04674/2020. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais - Productivity improvement in the system of bovine raising for meat, funded by PDR 2020.

References

[1] Filipe, P.A., Braumann, C.A., Carlos, C. Profit optimization for cattle growing in a randomly fluctuating environment. *Optimization: A Journal of Mathematical Programming and Operations Research*, 64(6), 1393–1407, 2015. doi:10.1080/02331934.2014.974598

COVID-19 lockdown effect in COPD: a comparison of fixed-effects selection methods

Jorge Cabral[a,b], Vera Afreixo[a,c], Alda Marques[b,d]
jorgecabral@ua.pt, vera@ua.pt, amarques@ua.pt

[a] *Department of Mathematics, University of Aveiro, Aveiro, Portugal*
[b] *Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health Sciences (ESSUA), University of Aveiro, Aveiro, Portugal*
[c] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal*
[d] *Institute of Biomedicine (iBiMED), School of Health Sciences, University of Aveiro, Aveiro, Portugal*

Keywords: COPD, COVID-19, feature selection, linear mixed-effects models, longitudinal data

Abstract: Chronic obstructive pulmonary disease (COPD) is common and progressive. One of its major impacts on daily life is decreased functional status which can be assessed by the one-minute sit-to-stand test (1minSTS). The 2020 imposed lockdown due to the recent pandemic (COVID 19) is likely to have influenced the functional status of this population but this is still unknown. Few feature selection algorithms are available for longitudinal data. We aimed to compare different feature selection methods and describe the effect of the COVID-19 lockdown on the 1minSTS behaviour in people with COPD. Data from 59 people with COPD were collected at baseline (B), 34 of whom belonging to the no-lockdown group. 1minSTS was repeated after one (A1) and five months (A5), which corresponded to the assessments prior and after the lockdown in the lockdown group. Fixed-effects were included in different linear mixed-effects models (LMMs) according to the importance given by Random Forests, Boruta, Extreme Gradient Boosting, automatic backward elimination and L1-penalized estimation algorithms. The LMM with the lowest Akaike's information criterion (AIC) was chosen. The LMM obtained by automatic backward elimination achieved the lowest AIC (919.7) and was followed by the one using L1-penalized estimation algorithm (923.5) although this one produced a higher conditional R-squared. Boruta algorithm returned the highest AIC (964.2). Difference between B and A1 number of repetitions in 1minSTS was statistically significant in both COVID-19 groups. No difference was found between A1 and A5 in either group suggesting that the lockdown had no effect in the 1minSTS behaviour.

Acknowledgements: This work was partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA, University of Aveiro) through the Portuguese Foundation for Science and Technology (FCT – Fundação o para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020 and by FCT (PTDC/DTP-PIC/2284/2014, PTDC/SAU-SER/28806/2017), under projects UIDB/04501/2020 and UIDB/04106/2020.

Adaptive estimation of the Extreme Value Index with Probability Weighted Moments

Frederico Caeiro[a,b], M. Ivette Gomes[c,d]
fac@fct.unl.pt, ivette.gomes@fc.ul.pt

[a] *Universidade Nova de Lisboa, Portugal*
[b] *Centro de Matemática e Aplicações (CMA), Portugal*
[c] *Universidade de Lisboa (UL), Portugal*
[d] *Centro de Estatística e Aplicações (CEAUL), Portugal*

Keywords: extreme value index, heavy tails, probability weighted moment, semi-parametric estimation

Abstract: In statistics of extremes, the estimation of the extreme value index (EVI) is an important and central topic of research. In this work we consider the probability weighted moment estimator of the EVI, based on the largest observations of a Pareto type model. Due to the specificity of the properties of the estimator, a direct estimation of the threshold is not straightforward. In this work we provide and study an adaptive choice of the number of order statistics to be used in the estimation. We also apply the introduced methodology to a data set in the field of insurance.

Acknowledgements: Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/0006/2021 (CEA/UL) and UIDB/MAT/0297/2021 (CMA/UNL).

References

[1] Caeiro, F., Gomes, M.I. Semi-parametric tail inference through probability-weighted moments. *Journal of Statistical Planning and Inference*, 141, 937–950, 2011. https://doi.org/10.1016/j.jspi.2010.08.015

[2] Caeiro, F., Gomes, M.I., Vandewalle, B. Semi-Parametric Probability-Weighted Moments Estimation Revisited. *Methodology and Computing in Applied Probability*, 16(1), 1–29, 2014. https://doi.org/10.1007/s11009-012-9295-6

Mathematical modelling of the impact of non-pharmacological and vaccination strategies to control the COVID-19 epidemic in Portugal

C. Caetano[a,b], M. Morgado[b,c], P. Patrício[d], J. F. Pereira[a,c], A. Torres[e], A. Leite,[e,f], A. Machado[a,e], B. Nunes[a,e]
constantino.caetano@insa.min-saude.pt, luisam@utad.pt, pcpr@fct.unl.pt, pereira.jpf96@gmail.com, afm.torres@ensp.unl.pt, andreia.leite@ensp.unl.pt, ausenda.machado@insa.min-saude.pt, baltazar.nunes@insa.min-saude.pt

[a] *Departamento de Epidemiologia, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisbon, Portugal*
[b] *Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, University of Lisbon*
[c] *Department of Mathematics, University of Trás-os-Montes e Alto Douro, UTAD*
[d] *Centro de Matemática e Aplicações (CMA), FCT, UNL and Departamento de Matemática, FCT, UNL, Campus de Caparica, 2829-516, Caparica, Portugal*
[e] *NOVA National School of Public Health, Public Health Research Center, Universidade NOVA de Lisboa*
[f] *Comprehensive Health Research Center, Universidade NOVA de Lisboa*

Keywords: COVID-19, SEIR models, vaccination

Abstract: The COVID-19 public health emergency was declared a pandemic on March 11th 2020 by the World Health Organisation. It was nine days prior to this date that Portugal saw its first confirmed case. In order to curb the exponential growth of cases, the Portuguese government opted to implement non-pharmaceutical interventions (NPIs), such as the closure of schools on March 16th 2020 and general lockdown on March 22nd 2020. Vaccination strategies begun in late December 2020. The objective of this study is to develop an age-structured SEIR deterministic model to measure the impact of NPIs and vaccination strategies on disease spread. We used Portuguese data on the number of individuals in intensive care units (ICU) to calibrate the model. Other model parameters were obtained in COVID-19 literature and estimated from Portuguese case data. The vaccination simulations considered take into account seroprevalence data, current estimates of vaccine effectiveness and coverage by age-group. We estimate that changes in transmission coincided with either the implementation or lifting of NPIs and also predicted that the January 15th 2021 general lockdown needed to last at least two months in order to reduce the numbers of ICU cases to sustainable values. Vaccination scenarios present possible trajectories of the future number of hospitalisations, given certain assumptions. Early results show that NPIs should not be fully phased-out and instead be combined with vaccination strategies to keep transmission low. The results obtained help to inform public health policies to mitigate the burden of COVID-19 in Portugal.

References

[1] Caetano C, Morgado ML, Patrício P, Pereira JF, Nunes B. Mathematical Modelling of the Impact of Non-Pharmacological Strategies to Control the COVID-19 Epidemic in Portugal. *Mathematics*, 2021, 9(10):1084. https://doi.org/10.3390/math9101084

[2] Kislaya, I., Gonçalves, P., Barreto, M., de Sousa, R., Garcia, A.C., Matos, R., Guiomar, R., Rodrigues, A.P. Seroprevalence of SARS-CoV-2 Infection in Portugal in May-July 2020: Results of the First National Serological Survey (ISNCOVID-19). *Acta Médica*, Port. 2021, 34, 87–94.

The performance of a combined distance between time series

Margarida G. M. S. Cardoso[a], Ana A. Martins[b]
margarida.cardoso@iscte-iul.pt, ana.martins@isel.pt

[a] *Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL)*
[b] *Instituto Superior de Engenharia de Lisboa (ISEL)*

Keywords: clustering, distance measures, time series

Abstract: The use of dissimilarity measures between time series is critical in several data analysis tasks which range from simple querying to classification, clustering and anomaly detection. Recently [1], we proposed a new dissimilarity measure, a convex combination of four (normalized) distance measures which offer complementary perspectives on the differences between two time series: the Euclidean distance which captures differences in scale; a Pearson correlation based measure that takes into account linear increasing and decreasing trends over time; a Periodogram based measure that expresses the dissimilarities between frequencies or cyclical components of the series; and a distance between estimated autocorrelation structures, comparing the series in terms of their dependence on past observations. We conduct an experimental analysis, to evaluate the comparative performance of this combined distance measure, resorting to the UCR Time-Series Archive that includes time series data sets from a wide variety of application domains. We follow a methodology suggested in previous studies [2] that were conducted to compare several dissimilarity measures and their variants: we use one nearest neighbor (1NN) classifier on labelled data to evaluate the efficacy of the distance measures. In fact, since the distance measure used is critical to 1NN accuracy, this indicator directly reflects the effectiveness of the dissimilarity measure used. We conclude that the proposed combined measure is competitive in several settings. Finally, we suggest further research taking into account normalization methods.

Acknowledgements: This work was supported by Fundação para a Ciência e Tecnologia, grant UIDB /00315/2020.

References

[1] Cardoso, M.G.M.S., Martins, A., Lagarto, J.Combining various dissimilarity measures for clustering electricity market prices *Estatística: Desafios Transversais às Ciências dos Dados - Atas do XXIV Congresso da Sociedade Portuguesa de Estatística (Paula Milheiro et al. eds), Edições SPE*, 197 - 212 , 2021.

[2] Paparrizos, J., Liu, C., Elmore, A.J., Franklin, M.J. Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD '20). Association for Computing Machinery, New York, NY, USA*, 1887–1905, 2020. doi:10.1145/3318464.3389760

As marés ao serviço da variação do nível médio do mar

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

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

Keywords: costa portuguesa, marégrafos, modelos de regressão linear, nível médio do mar, redes neuronais

Abstract: Nos últimos anos temos ouvido falar de cheias, inundações, aluimentos de terras, assim como aumento dos níveis de erosão costeira. Isso é particularmente verdadeiro para Portugal; em 2008, a distribuição da população residente nas regiões costeiras em comparação com a população nacional era de 83% e 60%, respetivamente [2]. O aumento do nível do mar é causado principalmente por dois fatores associados ao aquecimento global: a água adicionada pelo derreter dos calotes polares e a expansão da água do mar à medida que esta aquece [1]. Este trabalho aborda a mudança do nível do mar a partir de dados de marégrafos ao longo da costa portuguesa. As longas séries de dados permitiram apurar que o nível médio do mar está a aumentar, por exemplo, em Sines, o incremento foi de 77 milímetros em 23 anos de análise, o que corresponde a uma tendência de 4.67 ± 0.71 mm/ano [4]. A tendência de subida do nível médio foi deduzida através da regressão linear e o resultado foi ainda comparado com o obtido através do modelo de redes neuronais autorregressivas. Foram ainda calculados os níveis médios mensais característicos, e verificada a existência de sazonalidade ao longo do ano. Observaram-se acentuadas variações anuais e semi-anuais do nível médio do mar devido a variações da pressão atmosférica, da densidade da água e da circulação oceânica. Em síntese, durante os meses de verão tendem a predominar as variações da densidade da água e nos meses de inverno as variações de origem meteorológica [3].

References

[1] Church, A., White, NJ. A 20th century acceleration in global sea-level rise. *Geophysical Research Letters*, 33, 2006. doi:10.1029/2005GL024826

[2] EUROSTAT. *Eurostat regional yearbook 2011*. Publications Office of the European Union, 2011.

[3] IOC. *Manual de Medição e Interpretação do Nível Médio do Mar*. Manuais e Guias No. 14, Reino Unido, 1985.

[4] Mendes, V., Barbosa, S., Carinhas, D. Vertical land motion in the Iberian Atlantic coast and its implications for sea level change evaluation. *Journal of Applied Geodesy*, 14, 361–378, 2014. doi:10.1515/jag-2020-0012

Estimation of differential entropy: a comparison study

Eunice Carrasquinha[a,b], M. Rosário Oliveira[c], Rui Valadas[b,d], António Pacheco[c]
eitrigueirao@fc.ul.pt, rosario.oliveira@tecnico.ulisboa.pt, rui.valadas@tecnico.ulisboa.pt, antonio.pacheco.pires@tecnico.ulisboa.pt

[a] *CEAUL and DEIO, Faculdade de Ciências, ULisboa*
[b] *Instituto de Telecomunicações, IST, ULisboa*
[c] *CEMAT and Departamento de Matemática, IST, ULisboa*
[d] *Departamento de Engenharia Electrotécnica e de Computadores, IST, ULisboa*

Keywords: differential entropy, mutual information, variable selection

Abstract: In an era of data abundance, of a complex nature, it is of utmost importance to extract, from the data, useful and valuable knowledge for solving real problems. In the literature, there are several proposals of variable selection methods, however, selecting relevant and non-redundant variables continues to be a challenging issue. Variable selection problems arise in a variety of applications, reflecting their importance. Here we considered classifier-independent, filter methods, which separate the classification and variable selection procedures. The methods use alternative ranking criteria balancing relevance, redundancy, and irrelevance of the candidate input variables to the classification problems [2]. In this work, we consider mutual information, which captures linear and non-linear association between variables as an indicator of variable relevance and redundancy for the classification problem. In the vast majority of real problems the distribution of the population from where the data was collected is unknown. Therefore, the estimation of the entropy, differential entropy and consequently mutual information cannot be addressed as a parametric problem. To overcome this, we have focused our study on a widely used family of non-parametric estimators based on the idea of the discretization of continuous random variables [1]. In this work, we used tested and public datasets to compare feature selection methods based on estimation of the mutual information by considering or not a correction factor in the discretization process.

Acknowledgements: This work was supported by FCT, Portugal, through projects: UIDB/00006/2020, UIDB/04621/2020, UIDB/50008/2020, and PTDC/EEI-TEL/32454/2017.

References

[1] Pascoal, C., Oliveira, M.R., Pacheco, A., Valadas, R. Theoretical Evaluation of Feature Selection Methods based on Mutual Information. *Neurocomputing*, 226, 168–181, 2017.

[2] Macedo, F., Oliveira, M.R., Pacheco, A., Valadas, R. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, 325, 67–89, 2019.

Fatores de risco para a demora no diagnóstico da tuberculose em Portugal

Ana Castanheira[a], Cristina Rocha[a,b], Patrícia Soares[c,d]
fc48142@alunos.fc.ul.pt, cmrocha@fc.ul.pt, patricia.soares@ensp.unl.pt

[a] *DEIO, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *NOVA National School of Public Health, Public Health Research Centre, Universidade NOVA de Lisboa, Portugal*
[d] *Comprehensive Health Research Centre - Universidade Nova de Lisboa, Portugal*

Keywords: análise de sobrevivência, dados omissos, imputação múltipla, modelo de Cox, tuberculose

Abstract: A demora no diagnóstico da tuberculose é um importante problema de saúde pública, dado que um tempo longo até ao diagnóstico resulta num aumento de transmissibilidade da infeção na comunidade e pode levar a um aumento da severidade da doença. Este trabalho tem como objetivo principal a identificação dos fatores que têm influência no tempo desde o início dos sintomas até ao diagnóstico. No entanto, a existência de dados omissos é uma situação frequente em estudos na área da saúde; a data de início de sintomas é muitas vezes desconhecida, sendo habitualmente excluídos da análise estatística os indivíduos para os quais tal acontece. Para evitar esta perda de informação, procedemos à imputação dos valores omissos do tempo até ao dignóstico através do método de imputação múltipla - *Predictive Mean Matching* ([1,2]), considerando que os dados são omissos ao acaso (MAR). Neste trabalho, os dados analisados são provenientes da base de dados do Sistema de Vigilância da Tuberculose (SVIG-TB) e correspondem aos pacientes diagnosticados entre 1 de janeiro de 2008 e 31 de dezembro de 2017. Além da data de início de sintomas (13,9% de valores omissos) e data de diagnóstico, os dados contêm informação sobre características clínicas e sociodemográficas dos indivíduos. Observou-se um aumento da mediana do tempo entre o início dos sintomas e o diagnóstico de 59 dias em 2008 para 71 dias em 2017. A identificação dos fatores de risco para a demora no diagnóstico é feita recorrendo ao modelo de regressão de Cox.

Acknowledgements: Este trabalho é financiado por Fundos Nacionais através da FCT —Fundação para a Ciência e a Tecnologia no âmbito dos projetos UIDB/00006/2020 (CEAUL) e PTDC/SAU-PUB/31346/2017 (URBAN-TB)

References

[1] Zhou, X., Eckert, G.J, Tierney, W.M. Multiple imputation in public health research. *Statistics in Medicine*, 20, 1541–1549, 2001. doi:10.1002/sim.689

[2] Morris, T.P., White, I.R., Royston, P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14:75, 2014. doi:10.1186/1471-2288-14-75

Modeling of Compositional Data: A Multilevel Approach to Benthic Cover Abrolhos Bank

Pamela M. Chiroque-Solano[a]
pamela@dme.ufrj.br

[a] *Instituto de Biologia and Núcleo Professor Rogério Vale de Produção Sustentável-SAGE/COPPE, Universidade Federal do Rio de Janeiro, 21941-900, Rio de Janeiro, RJ, Brazil*

Keywords: bayesian inference, ceterocedasticity, climate change, codel identifiability, skewness

Abstract: Coral reefs provide an important ecosystem for life underwater. Their conservation and protection of the coastal areas can be useful to develop treatments for many diseases. In this work, we developed an approach to modeling the benthic coral-reef dynamics for the data of the Abrolhos bank. The Abrolhos reefs, off the coast of Southern Bahia, Brazil, are the largest and the richest reef complexes of the Southwestern Atlantic. The reef structures are under severe stress from climate and anthropogenic stressors and the unequivocal contamination that resulted from the Fundão dam collapse in November 2015 (Minas Gerais, Brazil). Reef structures are built by categories such as corals, macroalgae, turf, bry-ozoans, sponge, fire coral, cyano-bacteries, and others. Each one of these categories decribe the benthic community and can be expressed as proportions of a whole. To understand this benthic cover composition over different reef locations, we extended the Dirichlet regression model including hierarchical effects by sites. The inference procedure was done under the Bayesian approach using Hamiltonian Monte Carlo (HMC) method to obtain the approximations to the posterior marginal distributions of interest. Decision theory supported choosing one component as a reference to avoid model identifiability issues. Based on the compositional data analysis proposed, the main results exhibit skewness and heteroscedasticity for specific components identifing the Abrolhos Coral reefs dynamics.

Acknowledgements: Field work and image processing were carried out by The Marine Biodiversity and Conservation Laboratory at Federal University of Rio de Janeiro. Research funded by The Fundação Espírito Santense de Tecnologia, FEST and The Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, FAPERJ - E-26/200.016/2021.

References

[1] Aitchison, J. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Springer Netherlands, 1986.

[2] Maier, Marco J. DirichletReg: Dirichlet Regression in R, 2020. R package version 0.7-0.

Estimating caribou abundance in West Greenland using distance sampling methods

Iúri J. F. Correia[a], Tiago. A. Marques[a,b], Christine Cuyler[c]
ijcorreia@fc.ul.pt, tam2@st-andrews.ac.uk, chris.cuyler@natur.gl

[a] *Faculty of Sciences of the University of Lisbon*
[b] *University of St. Andrews*
[c] *Greenland Institute for Natural Resources*

Keywords: caribou, density surface modelling/spatial density modelling, distance sampling, generalized additive model, Greenland

Abstract: The barren-ground caribou, *Rangifer tarandus groenlandicus*, is native to the West coast of Greenland, and has always been important for the human population. Its importance spans from cultural traditions and subsistence consumption to recreational and commercial harvesting. Hence, the importance of long-term monitoring to facilitate appropriate management strategies. To accomplish a robust monitoring method and to determine caribou density, Distance Sampling (DS) methods were used. These techniques are widely used for density and abundance estimation of a wide variety of taxa. In this project, the data from an aerial survey for caribou conducted by the Greenland Institute of Natural Resources (GINR) in the late winter of the year of 2018 was used to estimate abundance of caribou in the surveyed area. The survey data and covariates to fit the Density Surface Model (DSM) were provided by GINR. Starting from a Conventional Distance Sampling perspective, the data set was then used to create a DSM for caribou, *i.e.*, a model describing caribou density as a function of additional covariates collected during the survey. An introduction about the study species and the region of interest is provided as well as a brief description of the sampling design, DS methodology and some related methods. The results were consistent with previous studies in terms of distribution throughout the study region, but the spatial distribution map obtained a previously unavailable useful insight. The estimated confidence intervals for abundance overlap with estimates from previous studies. Even though the point estimates are smaller when compared to previous point estimates, these differences are not statistically significant.

References

[1] 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

[2] 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

[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

A calibration approach to update short-term forecasts of daily maximum temperature performed with distribution-free estimation

Marco Costa[a], F. Catarina Pereira[b], A. Manuela Gonçalves[b]
marco@ua.pt, up202010700@edu.fe.up.pt, mneves@math.uminho.pt

[a] *Águeda School of Technology and Management & Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal.*
[b] *Department of Mathematics & Center of Mathematics, University of Minho, Portugal.*

Keywords: forecasting calibration, GMM estimation, Kalman filter, maximum temperature, state space modeling

Abstract: Nowadays, there are several Application Programming Interfaces (API) to facilitate the access to real-time, historical and future weather information. Within the scope of the TO CHAIR project, the weatherstack API was adopted to obtain weather data, (https://weatherstack.com/).
The state space modeling is considered in order to improve the accuracy of the website's forecasts from a dataset of real observations. The proposed model establishes a stochastic linear relationship between the observed daily maximum temperature and the h-step-ahead forecast obtained from the website, [1]. This relation is modeled in a state space framework associated to the Kalman filter algorithm. Since the normality of disturbances was not a good assumption for this dataset based on previous work, alternative GMM estimators were considered in the estimation of parameters, [2]. The results show that this approach allows reducing the RMSE of the uncorrected forecasts in 16.90% considering the 6-step-ahead forecasts and in 60.45% considering the 1-step-ahead forecasts, compared with the initial RMSE. Additionally, empirical confidence intervals at the 95% level have a coverage rate similar to this confidence level. So, this approach has proven suitable to improving the accuracy of this type of short-term forecasts since it considers a stochastic calibration factor in order to model the time correlation of this type of variable.

Acknowledgements: This work has received funding from FEDER/COMPETE/ NORTE2020/POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/ IEEA/00147/ 006933-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247, and from the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and UIDB/04106/2020 and UIDP/04106/2020 of CIDMA-UA.

References

[1] Costa, M., Alpuim, T. Parameter estimation of state space models for univariate observations, *Journal of Statistical Planning and Inference*, 140, 1889–1902, 2010.

[2] Costa, M., Alpuim, T. Adjustment of state space models in view of area rainfall estimation. *Environmetrics*, 22, 530–540, 2011.

An extreme value Bayesian lasso for the conditional left and right tails

Miguel de Carvalho[a,b], Soraia Pereira[a], Paula Pereira[a,c], Patrícia de Zea Bermudez[a,d]
Miguel.deCarvalho@ed.ac.uk

[a] *CEAUL, Universidade de Lisboa, Portugal*
[b] *School of Mathematics, University of Edinburgh, UK*
[d] *ESTSetúbal, Instituto Politécnico de Setúbal, Portugal*
[d] *FCUL, Universidade de Lisboa, Portugal*

Keywords: conditional tail, extended generalized Pareto distribution, heavy-tailed response, lasso, L_1-penalization, nonstationary extremes, statistics of extremes, variable selection

Abstract: We introduce a novel regression model for the conditional left and right tail of a possibly heavy-tailed response. The proposed model can be used to learn the effect of covariates on an extreme value setting via a Lasso-type specification based on a Lagrangian restriction. Our model can be used to track if some covariates are significant for the lower values, but not for the (right) tail—and vice-versa; in addition to this, the proposed model bypasses the need for conditional threshold selection in an extreme value theory framework. We assess the finite-sample performance of the proposed methods through a simulation study that reveals that our method recovers the true conditional distribution over a variety of simulation scenarios, along with being accurate on variable selection. Rainfall data are used to showcase how the proposed method can learn to distinguish between key drivers of moderate rainfall, against those of extreme rainfall. The talk is based on [1].

Acknowledgements: We thank Vanda Inácio, Ioannis Papastathopoulos, Philippe Naveau, Antónia Turkman, and Feridun Turkman for helpful comments and fruitful discussions. This work 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] de Carvalho, M., Pereira, S., Pereira, P., de Zea Bermudez, P. An extreme value Bayesian lasso for the conditional left and right tails, *Journal of Agricultural, Biological and Environmental Statistics*, (2021, to appear). https://arxiv.org/abs/2010.07164

Spatial Temporal Analysis of COVID-19 in Elderly living in Residential Care Home in Portugal

Felipa de Mello-Sampayo[a,b]
fdmso@iscte-iul.pt

[a] *ISCTE -Instituto Universitário de Lisboa*
[b] *BRU-ISCTE Business Research Unit*

Keywords: COVID-19, elderly, geographic weighted regression, kernel density estimation, residential care home, spatial-temporal analysis

Abstract: Analysing COVID-19 spread pattern is fundamental to guide the next steps towards overcoming the damaging effects on the elderlies living in residential care homes. We aim to describe the evolution of COVID-19 in residential care home throughout the 278 municipalities of continental Portugal between March and December 2020. Spatial analysis used the Kernel density estimation (KDE), space-time statistic Scan, and geographic weighted regression (GWR) to detect and analyse clusters of infected elderly living in RHC. Between 3 March and 31 December 2021, the high-risk primary cluster was located in the regions of Bragança, Guarda, Vila Real, and Viseu, Northwest of Portugal (relative risk = 3.67), between 30 September and 13 December 2020. The priority geographic areas for attention and intervention for elderly living in care houses are the regions in the Northeast of Portugal, and around the big cities, Lisbon and Oporto, which had high risk clusters. The relative risk of infection was spatially not stationary and generally positively affected by both comorbidities and low-income level. The regions with a population with high comorbidities and low income are priority for healthcare, once there was an increased risk of outbreaks of COVID-19 in elderlies living in residential care homes.

Acknowledgements: Many thanks to: Bernardo Forbes Costa, Hugo Anjos, Pedro Casaca, André Peralta Santos, Ana Lúcia Figueiredo, Ana Sottomayor, José Martins, Pedro Pinto Leite, André Peralta Santos for making available the dataset from SINAVE.

Análise conjunta de contagens longitudinais de células CD4 e tempo de sobrevivência de longo prazo de pacientes com HIV/AIDS no Paraná, Brasil: uma abordagem bayesiana usando INLA

Talita Evelin N.T. de Moraes[a], Isolde T.S. Previdelli[b], Giovani L. Silva[c,d]
talita_evel@usp.br, itsprevidelli@uem.br, giovani.silva@tecnico.ulisboa.pt

[a] *Escola Superior de Agricultura Luiz Queiroz (ESALQ), Universidade de São Paulo, Brasil*
[b] *Programa de Pós graduação em Bioestatística, Universidade Estadual de Maringá, Brasil*
[c] *Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[d] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*

Keywords: fração de cura, HIV/SIDA, modelos mistos, análise de sobrevivência

Abstract: Recentemente, a análise de dados envolvendo pacientes com HIV/SIDA tem recebido atenção crescente com a modelação conjunta de medidas repetidas de biomarcadores e o tempo de sobrevivência. Essa análise conjunta tem mostrado ser mais adequada por levar em consideração a dependência entre os dois tipos de respostas, longitudinal e de sobrevivência, que têm sido comummente analisadas separadamente. Os modelos conjuntos de sobrevivência e longitudinal são aqui aplicados a um estudo de coorte de pacientes do estado do Paraná, Brasil, com HIV/SIDA durante os anos de 2002 a 2006. Neste caso, o tempo de sobrevivência de pacientes com HIV/SIDA é modelado conjuntamente com as medidas repetidas de contagens de linfócitos CD4, tendo em conta uma potencial fração de cura dos pacientes. Essas duas componentes do modelo conjunto partilham efeitos aleatórios que são mais facilmente implementados via modelos gaussianos latentes, sob uma perspectiva bayesiana. Todavia, apesar de os métodos Monte Carlo via cadeias de Markov (MCMC) serem amplamente usados na análise bayesiana, foi também mais conveniente utilizar aproximações de Laplace encaixadas e integradas (INLA) na obtenção de resultados inferenciais com base em efeitos aleatórios gaussianos. Os resultados da análise dos dados de pacientes com HIV/SIDA no Paraná indicam que em geral os modelos conjuntos apresentaram melhor desempenho comparativamente aos modelos analisados separadamente.

Acknowledgements: Os autores agradecem à Maria Goretti Fonseca, Cláudia Coeli, Valeska Andreozzi e Rui Martins por fornecerem os dados deste estudo. Talita de Moraes agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro.

Pessoas com deficiências, raça, análise de Correspondência, renda e COVID-19

Paulo Tadeu Meira e Silva de Oliveira[a]
poliver@usp.br

[a] *UNIVERSIDADE DE SÃO PAULO*

Keywords: análise de correspondência, análise exploratória de dados, pandemia COVID-19, pessoas com deficiências, situação econômica

Abstract: Pessoas com Deficiência constitui um grupo de excluídos que sempre despertou os mais variados sentimentos. Deficiência significa uma deficiência física, intelectual ou sensorial, de natureza permanente ou temporária, o que limita a capacidade de exercer uma ou mais atividades. Situações como estas proporcionam a pessoa com deficiência pior poder aquisitivo, menor participação social acarretando maior exclusão e desvantagens ao compará-la a pessoas sem deficiência. Nos séculos XIX e XX, a cultura brasileira tem promovido uma integração e miscigenação racial. No entanto, as relações raciais no Brasil não têm sido harmônicas, especialmente em relação ao papel de desvantagem dos negros brasileiros e indígenas que tendem a ocupar posições menos prestigiadas. Renda é a remuneração que inclui salários e ordenados, juros, aluguéis, lucros mais as transferências que uma pessoa recebe. Para especialistas as habilidades econômicas das pessoas são dependem fortemente do seu nível de instrução e contribui para a formação de capital humano, sendo determinante no bem-estar e riqueza pessoal. Por fim, completando este cenário a COVID-19 com alta taxa de contágio, internação em hospitais de alta complexidade e altos índices de mortalidade. A COVID 19 fez alterar condições de vida, bem como as rotinas de todas as pessoas, sejam elas infectadas ou não fazendo alterar suas condições de trabalho, moradia, sociais, econômicos, culturais, étnico-raciais, psicológicos e comportamentais que influenciam na saúde e seus fatores de risco na população. Neste trabalho consideramos dados dos respondentes do questionário completo do Censo Demográfico de 2010 com o objetivo de avaliar variáveis como renda, raça, sexo, moradia, deficiência, nível de instrução, trabalho, número de filhos e de situações decorrentes com o surgimento da pandemia da COVID-19 utilizando análises: exploratória de dados, correspondência simples e mútipla que ilustram impactos significativos na sua evolução nas variáveis moradores por dormitório e renda.

Acknowledgements: A Profa. Dra. Júlia Maria Pavan Soler pela indicação do tema e ao IBGE pela disponibilidade dos dados do Censo de 2010

References

[1] Figueira, E. *Caminhando em Silêncio*. Giz Editorial, São Paulo-SP, 2008

[2] Oliveira, P.T.M.S. Pessoas com deficiância: análise dos resultados do censo de 2010 e a sua evolução. *Sigmae*, 3(2), 1–23, 2013

Clustering of time series of COVID data

José G. Dias[a]
jose.dias@iscte-iul.pt

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

Keywords: COVID data, hidden Markov models, mixture models, model-based clustering, time series data

Abstract: In the digital age, data streams have been produced at an increasing pace from different sources for instance from biometric devices (sensors) and stock market (high frequency) data to digital platforms (feeds, audio, video). Time-dependent modeling has been applied in many contexts not only forecasting, but also outlier detection, matching, clustering, indexing, etc. We model COVID data time series using model-based clustering techniques that take time-dependency into account [1]. The application to time series of country-based data shows that country dynamics can be grouped into three distinct cross-sectional clusters. Moreover, the longitudinal dynamics can be summarized into three regimes.

Acknowledgements: This work was funded by the Portuguese Foundation for Science and Technology (Grant UIDB/00315/2020).

References

[1] Zucchini, W., MacDonald, I. L., Langrock, R. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC, Boca Raton, 2016.

Sobre o máximo de um processo max-INAR(1) bivariado

Sandra Dias[a], Maria da Graça Temido[b]
sdias@utad.pt, mgtm@mat.uc.pt

[a] *Universidade de Trás-os-Montes e Alto Douro, Dep. de Matemática, CEMAT*
[b] *Universidade de Coimbra, Dep. de Matemática, Fac. de Ciências e Tecnologia, CMUC*

Keywords: lei geométrica bivariada, max-semiestabilidade, processo max-INAR(1), teoria de valores extremos

Abstract: Neste trabalho estudamos o comportamento assintótico da sucessão de máximos de um modelo max-INAR(1) bivariado. Este processo é uma extensão do modelo univariado introduzido e estudado em [1]. Consideramos as inovações com lei geométrica bivariada e estabelecemos um comportamento assintótico quasi-max-estável para a sucessão bivariada de máximos. Em contexto de max-semiestabilidade, estabelecemos um comportamento assintótico max-semiestável para a sucessão bivariada de máximos, quando o número de variáveis de cada margem tem um crescimento geométrico.

References

[1] Scotto, M.G., Weiss, C.H, Möller, T.A., Gouveia, S. The max-INAR(1) model for count processes. *Test*, 27, 850–870, 2018. https://doi.org/10.1007/s11749-017-0573-z

A Classification model for Mertolenga cattle breed

Ana Paula Ferrari Januário[b], Patrícia A. Filipe[a,c], Gonçalo Jacinto[a,b]
ferrari.januario@gmail.com, patricia.filipe@iscte-iul.pt, gjcj@uevora.pt

[a] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Évora, Portugal*
[b] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, Évora, Portugal*
[c] *Iscte-Instituto Universitário de Lisboa, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia, Lisboa, Portugal*

Keywords: breeding costs, logistic regression, Mertolenga breed, PDO seal

Abstract: The Associação de Criadores de Bovinos Mertolengos (ACBM) performs the growing and finishing phases of young Mertolengo males, enabling breeders to breed and finish their cattle, allowing them to obtain a higher economic value than is normally achieved in weaning sales and, at the same time, helping to solve the problem of breeding and finishing when farms do not have technical and/or economic conditions for that purpose. When the animals achieve a certain weight at a given age, they are considered PDO (Protected Designation of Origin).

Using data of 716 male animals provided by ACBM, containing information about the general costs of the breeding process, zootechnical information, such as the animal's weight/age when entering and leaving ACBM and average daily gain, and information regarding the animal's genetic values, using a generalized linear model we obtain the factors that conduct the male bovines of the Mertolenga breed be considered as PDO. We concluded that the weight and age when entering in ACBM, and the genetic values of maternal capacity and calving interval were identified the most important variables for a given animal to be classified with the PDO seal.

Acknowledgements: The second and third authors belong to the research Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora), supported by Fundação para a Ciência e a Tecnologia, (project UID/04674/2020). We are grateful to GoBovMais Project (PDR2020-1.0.1-FEADER-031130), ACBM and José Pais (ACBM head engineer) for providing the data and for continuous support.

Sistemas $M^X/M/c/n$ com bloqueio e abandono

Fátima Ferreira[a,b], António Pacheco[a,c], **Helena Ribeiro**[a,d]
mmferrei@utad.pt, apacheco@math.tecnico.ulisboa.pt,
helena.ribeiro@ipleiria.pt

[a] *CEMAT, Instituto Superior Técnico, Universidade de Lisboa*
[b] *Universidade de Trás-os-Montes e Alto Douro, UTAD, Dep. Matemática*
[c] *Instituto Superior Técnico, Universidade de Lisboa, Dep. Matemática*
[d] *Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria*

Keywords: bloqueio e abandono de clientes, clientes servidos, filas de espera, períodos de ocupação contínua

Abstract: Neste trabalho estudam-se sistemas de filas de espera $M^X/M/c/n$ com bloqueio e abandono. O bloqueio é caracterizado por um mecanismo estocástico de entrada de clientes, modulado pelo estado do sistema nos instantes de chegada. Por seu lado, os abandonos ocorrem segundo uma taxa estocástica que é função do número de clientes em fila de espera.
Especificamente, caracteriza-se a distribuição de probabilidade conjunta do número de clientes servidos e do número de clientes perdidos, em períodos de ocupação contínua iniciados com múltiplos clientes no sistema. Para tal deriva-se a função geradora de probabilidades conjunta destas variáveis através de um método recursivo. Os resultados obtidos são ilustrados para diferentes políticas de bloqueio e abandono de clientes.
Este trabalho generaliza os resultados obtidos em *Ferreira et al.* [1] para filas $M^X/M/1/n$ com bloqueio.

Acknowledgements: Este trabalho foi elaborado com o apoio parcial da Fundação para a Ciência e a Tecnologia (FCT) pelo projeto UIDP/04621/2020 do CEMAT/IST-ID.

References

[1] Ferreira, F., Pacheco A., Ribeiro, H. Análise de filas $M^X/M/1/n$ com bloqueio. Em P. Milheiro, A. Pacheco, B. Sousa, I. F. Alves, I. Pereira, M. J. Polidoro & S. Ramos (eds.): *Atas do XXIV Congresso da SPE, Estatística: Desafios transversais às ciências com dados*, 73–87, Edições SPE, 2021.

Weighted maximum likelihood estimation for SDE individual growth models

Patrícia A. Filipe[a,c], Gonçalo Jacinto[a,b], Carlos A. Braumman[a,b]
patricia.filipe@iscte-iul.pt, gjcj@uevora.pt, braumann@uevora.pt

[a] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora*
[b] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora*
[c] *Iscte-Instituto Universitário de Lisboa, Iscte Business School, Quantitative Methods for Management and Economics Department*

Keywords: bootstrap estimation, cattle growth, stochastic differential equations, weighted maximum likelihood estimation

Abstract: We apply a class of stochastic differential equations to model individual growth in a randomly fluctuating environment using real weight data of males of the Mertolengo cattle breed. We have used maximum likelihood theory to estimate the parameters. However, for cattle data, it is often not feasible to obtain animal's observations at equally spaced ages nor even at the same ages for different animals and there is typically a small number of observations at older ages. For these reasons, maximum likelihood estimates can be quite inaccurate, being interesting to consider in the likelihood function a weight function associated to the elapsed times between two consecutive observations of each animal, which results in the weighted maximum likelihood method. We compare the results obtained from both methods in several data structures, simulated for scenarios corresponding to weights measured at equally spaced ages using different spacings and different maximum ages. The "real" ages scenario was also considered with simulation of weights at the effective observed ages of a group of animals randomly selected from the complete database. The weighted maximum likelihood revealed to improve the estimation when observations at older ages are scarce and the observation instants are unequally spaced, whereas the maximum likelihood estimates are recommended when animals are weighted at equally spaced ages. For unequally spaced observations, a bootstrap estimation method was applied in order to correct the bias of the maximum likelihood estimates, and revealed to be a more precise alternative, except for datasets with only young animals.

Acknowledgements: The Centro de Investigação em Matemática e Aplicações is supported by the Fundação para a Ciência e a Tecnologia, project UID/04674/2020. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais - Productivity improvement in the system of bovine raising for meat, funded by PDR 2020.

Joint models for longitudinal and time-to-event data: comparing different computational approaches

Inês Fortes[a], Inês Sousa[b]
ines.fortes@gmail.com, isousa@math.uminho.pt

[a] *Centro Algoritmi, University of Minho, Braga, Portugal*
[b] *Department of Mathematics and Applications, University of Minho, Braga, Portugal*

Keywords: implementation, joint models, longitudinal, time-to-event, R

Abstract: Joint models for longitudinal and time-to-event data are common in data analysis when there is an association between (1) a longitudinally observed variable and (2) a variable measuring the time until the occurrence of a specific event. Given the usefulness of these models, there are currently at least two available package implementations in R [1-2]. However, these models are complex because they often include shared random effects between the longitudinal and time-to-event submodels. As a consequence, implementation of these joint models is computationally demanding. In this simulation study we explore a fully parametric model, where the joint distribution of longitudinal and transformed time-to-event data are assumed to follow a multivariate normal distribution [3]. The main advantage of this model is that it is computationally less demanding than the already implemented ones [1-2]. On the other hand, it depends on several assumptions, which makes it unsuitable for generalized application in complex problems. However, we believe it can be used for exploratory analysis, to select the more influencing explanatory variables. In this simulation study the different existing models and the details of their computational implementations will be described and compared, referring to their advantages and disadvantages.

References

[1] Rizopoulos, D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, 35(9), 1-33, 2010. doi:10.18637/jss.v035.i09

[2] Philipson, P., Diggle, P., Sousa, I., Kolamunnage-Dona, R., Williamson, P., Henderson, R. joineR: Joint modelling of repeated measurements and time-to-event data. R package, 2012

[3] Diggle, P. J., Sousa, I., Chetwynd, A. G. Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture *Statistics in Medicine*, 27(16), 2981-2998, 2008. doi:10.1002/sim.3131

Road effects on demographic traits of small mammal populations: a capture-recapture study

Ana Galantinho[a,b], Sofia Eufrázio[b], Carmo Silva[a,b],
Filipe Carvalho[d,e], Russell Alpizar-Jara[f], António Mira[a,c]
ana.galantinho@gmail.com, srle@uevora.pt, carmoms@uevora.pt,
filipescpcarvalho@gmail.com, alpizar@uevora.pt, amira@uevora.pt

[a] *UBC - Biology Department, University of Évora, Portugal*
[b] *MED–IIFA, University of Évora, Pólo da Mitra, Portugal*
[c] *MED–IIFA/DBIO-ECT, University of Évora, Pólo da Mitra, Portugal*
[d] *CIBIO-InBIO, University of Porto, Portugal*
[e] *Dept Zoo Entomology, SBES, University of Fort Hare, South Africa*
[f] *CIMA-IIFA/DMAT-ECT, University of Évora, Portugal*

Keywords: extended robust design models, population estimation, roadless area, wood mouse

Abstract: Recent studies have highlighted the positive effects of road verges on the abundance of small mammals. However, most of these studies occurred in intensively grazed or cultivated areas, where verges were the last remnants of suitable habitats, which could mask the true effects of roads on population traits. We analysed the effects of roads on small mammal populations living in a well-preserved Mediterranean forest. We used the wood mouse (Apodemus sylvaticus) as a model of forest-dwelling small mammals that probably are among the species most affected by road clearings. Our study compared populations in similar habitat areas with and without road influence. We assessed abundance, survival and temporary emigration using extended Pollock's robust design capture-recapture models. Moreover, we analysed population turnover, sex ratio, age structure and body condition. We found that wood mouse abundance and body condition were lower at the road bisected area, whereas the remaining population traits were similar. This suggests that the reduced habitat availability and quality due to the physical presence of the road and verge vegetation clearing are the main drivers of demographic differences in wood mouse populations between areas. Nevertheless, our results also suggest that in high quality habitats surrounding national roads, wood mouse populations present similar dynamics to others living in undisturbed areas, despite the decrease in abundance and body condition. Overall, the often-reported increased small mammal abundance in road surroundings should not be generalized independently of habitat quality or to other population traits.

References

[1] Galantinho, A., Eufrázio, S., Silva, C., Carvalho, F., Alpizar-Jara, R., Mira, A. Road effects on demographic traits of small mammal populations. *European Journal of Wildlife Research*, 63, 22, 2017. doi:10.1007/s10344-017-1076-7.

Symbolic Variance Maximisation for Interval Principal Component Analysis

Rodrigo Girão Serrão[a], **M. Rosário Oliveira**[a,b], Lina Oliveira[a,c]
rodrigogiraoserrao@tecnico.ulisboa.pt, rosario.oliveira@tecnico.ulisboa.pt, lina.oliveira@tecnico.ulisboa.pt

[a] *Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[b] *CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[c] *CAMGSD, Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Keywords: interval algebraic structure, symbolic data analysis, symbolic principal component analysis

Abstract: There have been a series of proposed adaptations of Principal Component Analysis to interval-valued symbolic data through the use of a symbolic - conventional - symbolic approach. This has the downside of having intermediate steps that deal with conventional data, so we put forward a fully symbolic adaptation of Principal Component Analysis for interval-valued data, the main contribution of this work. We develop a theoretical framework that allows for the definition of symbolic principal components which, among other advantages, provides the mathematical tools needed to use the symbolic principal components to transform the original data in a way that is mathematically coherent with the remainder of the framework, and defines the principal components as solutions of maximisation problems, similarly to what is done in conventional Principal Component Analysis. After the theoretical foundations are laid down, we explore real world data from the telecommunications sector, in an attempt to detect Internet redirection attacks in real-time. In particular, we use our symbolic method to improve and simplify an outlier-detection method that has been proposed in the literature.

Acknowledgements: This work was supported by Instituto de Telecomunicações grant BIL/N 63 - 14/10/2019 MaLPIS and by Fundação para a Ciência e Tecnologia, Portugal, through the projects UIDB/04621/2020, PTDC/EEI-TEL/32454/2017, and UID/MAT/04459/2020.

Estimation of the Weibull Tail Coefficient through the Power Mean-of-Order-p Generalized Mean

M. Ivette Gomes[a,b], Frederico Caeiro[c,d], Lígia Henriques-Rodrigues[e,f]
ivette.gomes@fc.ul.pt, fac@fct.unl.pt, ligiahr@uevora.pt

[a] *Universidade de Lisboa (UL), Portugal*
[b] *Centro de Estatística e Aplicações (CEAUL), Portugal*
[c] *Universidade Nova de Lisboa, Portugal*
[d] *Centro de Matemática e Aplicações (CMA), Portugal*
[e] *Universidade de Évora, Portugal*
[f] *Centro de Investigação em Matemática e Aplicações (CIMA), Portugal*

Keywords: power-mean-of-order-p, semi-parametric estimation, statistics of extremes, weibull tail-coefficient

Abstract: The *Weibull tail-coefficient* (WTC) is the parameter θ in a right-tail function of the type $\overline{F} := 1-F$, such that $H := -\ln \overline{F}$ is a *regularly varying* function at infinity with an index of regular variation equal to $\theta \in \Re^+$. In a context of extreme value theory for maxima, it is possible to prove that we have an *extreme value index* (EVI) $\xi = 0$, but usually a very slow rate of convergence. Most of the recent WTC-estimators (see Gardes and Girard, 2006, among others) are proportional to the class of Hill EVI-estimators (Hill, 1975), the average of the log-excesses associated with the k upper order statistics, $1 \leq k < n$. The interesting performance of EVI-estimators based on generalized means, lead us to base the WTC-estimation on the power *mean-of-order-p* (MO_p) EVI-estimators studied in Brilhante *et al.* (2013), among others (see also, Caeiro *et al.*, 2016, where MO_p EVI-estimators are dealt with under a third-order framework). Consistency and asymptotic normality of the estimators under study is put forward. The performance of the new estimators for finite samples is illustrated through a small-scale Monte-Carlo simulation study.

Acknowledgements: Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/0006/2021 (CEA/UL), UIDB/MAT/0297/2021 (CMA/UNL) and UIDB/MAT/04674/2021 (CIMA).

References

[1] Brilhante, M.F. Gomes, M.I. Pestana, D. A simple generalisation of the Hill estimator. *Comput. Statist. and Data Analysis*, 57:1, 518–535, 2013. doi:10.1016/j.csda.2012.07.019

[2] 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

[3] Gardes, L. Girard, S. Comparison of Weibull tail-coefficient estimators. *Revstat—Statist. J.*, 4, 163–188, 2006. https://www.ine.pt/revstat/pdf/rs060206.pdf

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

Daily temperature time series forecasting in the Optimal Challenges in Irrigation (TO CHAIR): a comparative study using TBATS and regression models

A. Manuela Gonçalves[a], Cláudia Costa[b], Marco Costa[c]
mneves@math.uminho.pt, claudiacostamf@hotmail.com, marco@ua.pt

[a] *Department of Mathematics & Center of Mathematics, University of Minho, Portugal*
[b] *Department of Mathematics, University of Minho, Portugal*
[c] *Águeda School of Technology and Management & Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal*

Keywords: forecasting, minimum temperature, regression with correlated errors, TBATS

Abstract: In a world where climate change and increasing social conflicts are a reality, a proper management of the existing scarce resources is vital. Predicting and forecasting weather time series has always been a difficult field of research analysis with a very slow progress rate over the years. The main challenge in this project - The Optimal Challenges in Irrigation (TO CHAIR) - is to study how to manage irrigation problems as an optimal control problem: the daily irrigation problem of minimizing water consumption. For that it is necessary to estimate and forecast weather variables in real time in each irrigation area, in order to determine, in particular, the evapotranspiration (related to the irrigation planning problem). This study presents a comparison of the forecasting performance of TBATS (Trigonometric Seasonal, Box-Cox Transformation, ARMA errors, Trend and Seasonal Components) [1], and regression with correlated errors models, [2]. These methods are chosen due to their ability to model trend and seasonal fluctuations present in weather data, particularly in dealing with time series with complex seasonal patterns (multiple seasonal patterns). The forecasting performance is demonstrated through a case study of weather time series: daily minimum air temperature.

Acknowledgements: This work has received funding from FEDER/COMPETE/NORTE2020/POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/IEEA/00 147/ 006933-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247, and from the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and UIDB/04106/2020 and UIDP/04106/2020 of CIDMA-UA.

References

[1] Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. *Forecasting with Exponential Smoothing: the State Space Approach.* Springer, Berlin, 2008.

[2] Alpuim, T., El-Shaarawi, A. Modeling monthly temperature data in Lisbon and Prague. *Environmetrics*, 20, 835–852, 2009.

Comparison of experimental designs for evaluation of intravarietal genetic variability and selection in ancient grapevine varieties

Elsa Gonçalves[a], Antero Martins[a]
elsagoncalves@isa.ulisboa.pt, anteromart@isa.ulisboa.pt

[a] *Linking Landscape, Environment, Agriculture and Food (LEAF), Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal*

Keywords: alpha designs, grapevine selection, partially replicated designs, resolvable row-column designs

Abstract: Plant breeding takes an important role in agriculture, providing plant materials with superior genetic quality to be used by farmers. In plant breeding the experimental design of field trials is crucial to provide a reliable evaluation of the genotypes under study. Grapevine is an outstanding example of a successful crop in Portugal, in which the results of selection have fostered important economic gains. Therefore, to increase the efficiency of the evaluation of intravarietal variability and the genetic gains of selection, studies related to the experimental designs implemented in the field are continuously conducted.

This work is focused on the experimental designs applied in the initial phases of grapevine selection, where a large number of genotypes is evaluated in field trials, concerning the most important economic traits. In this context, the comparison of the efficiency of several experimental designs to quantify intravarietal genetic variability and to perform selection within a variety was performed through simulation studies. Meanwhile, methodological studies with resolvable row-column designs, alpha designs, randomized complete block designs, and partially replicated designs were implemented in the field. In this work, these experimental designs are compared using real field data obtained in field trials constructed accordingly. The results point out the importance of the row-column designs to control the spatial variability present in large field trials. Additionally, the precision of the estimate of genotypic variance component, as well as, the precision of the prediction of genotypic effects of important traits, such as the yield, were higher in these latter experimental designs.

Modelação longitudinal binária da desnutrição aguda e crónica em crianças na província do Bengo em Angola

M. Helena Gonçalves[a,b], Carolina Gasparinho[c,d], Assucênio Chissaque[d,e], Giovani L. Silva[b,f], Filomeno Fortes[d], Luzia Gonçalves[b,d]
mhgoncal@ualg.pt, carolinagasparinho@gmail.com, assucenyoo@gmail.com, giovani.silva@tecnico.ulisboa.pt, filomenofortes@ihmt.unl.pt, LuziaG@ihmt.unl.pt

[a] *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade do Algarve*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[c] *Centro de Investigação em Saúde de Angola (CISA)*
[d] *Global Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa*
[e] *Instituto Nacional de Saúde de Moçambique*
[f] *Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa*

Keywords: dados longitudinais, desnutrição aguda, desnutrição crónica, desparasitação, parasitas intestinais

Abstract: Este trabalho visa modelar a desnutrição aguda e a desnutrição crónica, usando dados de crianças parasitadas, seguidas ao longo de dois anos num contexto comunitário, na província do Bengo em Angola. Para este estudo, com quatro intervenções (A1, A2, A3 e A4), foram ajustados diversos modelos para dados longitudinais binários, considerando duas definições: a recomendada pela Organização Mundial da Saúde (OMS) e outra definição baseada em Z-scores inferiores a -1 relativamente ao peso para a idade e à estatura para a idade. Esta última definição parece ter vantagens, quer do ponto de vista estatístico, quer dos cuidados de saúde a prestar a estas crianças. Para a análise dos dados foi utilizada a biblioteca *bild* do programa *R* que analisa dados longitudinais binários onde a dependência serial é regulada por um mecanismo de cadeias de Markov e o *odds-ratio* é usado para medir a dependência entre observações sucessivas no mesmo indivíduo. Os modelos selecionados indicam um efeito significativo das intervenções na desnutrição aguda das crianças, sendo o braço A3 (diagnosticar e tratar a criança individualmente) preferível ao tratamento habitual (A1). Para a desnutrição crónica, o modelo selecionado não evidencia um efeito das intervenções, mas nota-se uma diminuição da desnutrição crónica com o aumento da idade da criança (no momento inicial) e também um efeito temporal decrescente. Apesar destes desfechos não estarem previstos no estudo original, as análises estatísticas obtidas fornecem resultados com interesse para o acompanhamento das crianças em contexto comunitário, onde é difícil manter estudos por longos períodos de tempo.

Alguns determinantes para uma maior gravidade da sinistralidade rodoviária no distrito de Setúbal

Paulo Infante[a], Anabela Afonso[a], Gonçalo Jacinto[a], **Leonor Rego**[a], Rodrigo Cesar[a], Pedro Nogueira[a], Marcelo Silva[a], Vitor Nogueira[a], José Saias[a], Paulo Quaresma[a], Daniel Santos[a], Patrícia Gois[a], Paulo Rebelo Manuel[a]
pinfante@uevora.pt, aafonso@uevora.pt, gjcj@uevora.pt, lrego@uevora.pt, rcfs@uevora.pt, pmn@uevora.pt, marcelogs@uevora.pt, vbn@uevora.pt, jsaias@uevora.pt, pq@uevora.pt, dfsantos@uevora.pt, pafg@uevora.pt, pjsrm@uevora.pt

[a] *Universidade de Évora*

Keywords: Getis Ord-Gi*, Moran I local, regressão logística, sinistralidade

Abstract: A sinistralidade rodoviária é um problema com repercussões em várias dimensões: social, económica, saúde, justiça e segurança. Nos últimos anos tem-se assistindo a um agravamento deste problema, que se torna ainda mais preocupante num distrito em que existe o maior número de acidentes de viação com vítimas mortais (2017 e 2018), mas com menor número de acidentes que noutros distritos. Entre 2016 e 2019 neste distrito registaram-se 28103 acidentes na área sob jurisdição do Comando Territorial da GNR de Setúbal, dos quais resultaram 8260 vítimas, sendo 510 feridos graves e 167 feridos mortais. Neste trabalho analisam-se os dados do Boletim Estatístico de Acidentes de Viação, recolhidos e validados pelo Comando Territorial da GNR de Setúbal, com atualização da Autoridade Nacional de Segurança Rodoviária (ANSR) para vítimas a 30 dias, e complementados com informação meteorológica fornecida pelo IPMA. Inicialmente foi efetuada uma análise espacial dos acidentes, recorrendo-se à estatística Getis Ord-Gi* para identificação de hotspots e à estatística Moran I local para autocorrelação espacial, que permitiu identificar os concelhos com perfis idênticos para as vítimas mortais e feridos graves. Posteriormente, recorreu-se a um modelo de regressão logística para identificar alguns determinantes para a existência de vítimas mortais e/ou feridos graves nos acidentes de viação ocorridos no distrito de Setúbal. Os resultados obtidos foram confrontados com modelos de *machine learning*.

Acknowledgements: Este trabalho é uma contribuição para o projeto MOPREVIS "FCT DSAIPA/DS/0090/2018" financiado pela FCT-Fundação para a Ciência e a Tecnologia, no âmbito da Iniciativa Nacional em Competências Digitais e.2030, Portugal INCoDe.2030. Agradecemos também às entidades parceiras do projeto.

Association between weekly dengue cases and meteorological and geographical variables in the Dominican Republic

Adela Iutis[a], Helena Sofia Rodrigues[b,c], **Adelaide Freitas**[a,c], Natália Martins[a,c]
adelaiutis@ua.pt, sofiarodrigues@esce.ipvc.pt, adelaide@ua.pt, natalia@ua.pt

[a] *Department of Mathematics, University of Aveiro, Aveiro, Portugal*
[b] *School of Business, Instituto Politécnico de Viana do Castelo, Valença, Portugal*
[c] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal*

Keywords: Aedes mosquito, dengue disease, spatial analysis, weather predictors

Abstract: Dengue is an infectious viral disease transmitted by mosquitoes and a major public health problem in the Dominican Republic. There are no antiviral drugs available. Currently, vector control is key to controlling dengue transmission. This control must be guided by surveillance of the outbreak and implementation of measures to suppress the transmission of dengue and limit the risk of future spread. The main goal of this work is to analyse the relationship between weekly dengue cases in the Dominican Republic in 2019 and weather and geographical predictors for an early warning of dengue outbreak. The study was carried out for 9 provinces (out of 32 provinces), with weather data available. Two negative binomial regression models were built. One with the weekly dengue cases related to the weather variables of the corresponding week. Another one with a delay of two weeks, that is, the weekly cases of dengue related to the values of the weather variables from two weeks ago. While the results from a model without delay show that the average temperature is significantly related to dengue cases, a 2-week delay model shows that the accumulated precipitation and average temperature are the significant weather variables related to dengue cases. These two weather variables are known to be important to the development of the mosquito vector of dengue. Week, urbanization, GINI coefficient and the number of total health centres in each region are also statistically significant on the number of dengue cases in both considered models.

Acknowledgements: This work was partially supported by the BioMathematics thematic line of the Center for Research and Development in Mathematics and Applications (CIDMA, University of Aveiro) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. Adela Iutis was supported by an FCT research Grant Ref: BI-BIOMATH-1-2021.

Profit optimization with sensitivity analysis for cattle growth using Gompertz and Bertalanffy-Richards SDE models

Gonçalo Jacinto[a,b], Patrícia A. Filipe[c,b], Carlos A. Braumann[a,b]
gjcj@uevora.pt, patricia.filipe@iscte-iul.pt, braumann@uevora.pt

[a] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora*
[b] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora*
[c] *Iscte-Instituto Universitário de Lisboa, Iscte Business School, Quantitative Methods for Management and Economics Department*

Keywords: cattle growth, individual growth model, profit optimization, sensitivity analysis, stochastic differential equations

Abstract: Regression on deterministic models is not appropriate to study individual growth of animals in randomly fluctuating environments. So, we have used stochastic versions of the classical deterministic models, in the form of a general stochastic differential equation (SDE) model and obtained the expressions for the profit probability distribution, its first two moments and other quantities of interest for a general market realistic profit structure. The aim is helping farmers involved in the growing and finishing phases of bovine males to optimize the profit obtained by raising and selling an animal.

We now obtain explicit expressions for two particular models, the Gompertz and Bertalanffy-Richards SDE growth models, and apply the results to real weight date of Mertolengo cattle males. We then obtain the optimal selling age and optimal expected profit, assuming average feeding costs per unit time and using as model parameters their maximum likelihood (ML) estimates based on real data. We conclude that farmers are selling the animals a little earlier than the optimal selling age, resulting in a lower profit per animal.

Since ML estimates are not exact parameter values, we perform a sensitivity analysis on the estimates of the model parameters. It shows low sensitivity of the optimal profit and an almost negligible sensitivity of the optimal selling age.

Acknowledgements: The Centro de Investigação em Matemática e Aplicações is supported by the Fundação para a Ciência e a Tecnologia, project UID/04674/2020. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais - Productivity improvement in the system of bovine raising for meat, funded by PDR 2020.

Mixed models for individual growth in random environment through Laplace and delta method approximation methodologies

Nelson T. Jamba[a], Patrícia A. Filipe[a,c], Gonçalo Jacinto[a,b], Carlos A. Braumann[a,b]
d39830@alunos.uevora.pt, patricia.filipe@iscte-iul.pt, gjcj@uevora.pt, braumann@uevora.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] *Iscte - Instituto Universitário de Lisboa, ISCTE Business School, Quantitative Methods for Management and Economics Department*

Keywords: delta method, Laplace approximation, maximum likelihood estimation method, mixed models, stochastic differential equations

Abstract: We use a class of stochastic differential equations (SDE) to model the evolution of cattle weight, taking the form $dY_i(t) = \beta(\alpha - Y_i(t))dt + \sigma dW_i(t)$, $Y_i(t_0) = y_{i,0}$, $i = 1, \ldots, M$, where $Y_i(t) = h(X_i(t))$ is a transformed weight obtained by applying a strictly increasing C^1 function h to the actual weight $X_i(t)$ of the i^{th} animal at age t, α is the average transformed size at maturity, β is a growth parameter, σ measures the intensity of environmental fluctuations and $W_i(t)$ $(i = 1, \ldots, M)$ are independent standard Wiener processes. Depending on the function h chosen, we obtain stochastic versions of the most commonly used deterministic animal growth models. Since model parameters may vary from animal to animal, we have extended the study to SDE mixed models where the variation among animals of the parameters α, β or both is assumed to be random. The maximum likelihood estimation method was applied for these cases using Laplace and delta method approximation methodologies to solve the integrals involved in the maximum likelihood function. This type of methodologies has been addressed for the case where the time vector is the same for all trajectories. This is not our case since the animals are not weighted at the same ages and for this reason the existing methodology had to be adjusted. A comparison between methodologies is presented showing a very good approximation in estimating the parameters.

Acknowledgements: The Centro de Investigação em Matemática e Aplicações is supported by the Fundação para a Ciência e a Tecnologia, project UID/04674/2020. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais - Productivity improvement in the system of bovine raising for meat, funded by PDR 2020

Regression-type analysis for block maxima on block maxima

Alina Kumukova[a], Miguel de Carvalho[b], Gonçalo dos Reis[b,c]
s1874671@ed.ac.uk, Miguel.deCarvalho@ed.ac.uk, g.dosreis@ed.ac.uk

[a] *Maxwell Institute for Mathematical Sciences School of Mathematics, University of Edinburgh, Edinburgh UK*
[b] *School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3FD, UK*
[c] *Centro de Matemática e Aplicações (CMA), FCT, UNL, PT*

Keywords: Bernstein polynomials, block maxima, joint extremes, quantile regression, statistics of extremes

Abstract: This paper devises a regression-type model for the situation where both the response and covariates are extreme. The proposed approach is designed for the setting where both the response and covariates are themselves block maxima, and thus contrarily to standard regression methods it takes into account the key fact that the limiting distribution of suitably standardized componentwise maxima is an extreme value copula. An important target in the proposed framework is the regression manifold, which consists of a family of regression lines obeying the latter asymptotic result. To learn about the proposed model from data, we employ a Bernstein polynomial prior on the space of angular densities which leads to an induced prior on the space of regression manifolds. Numerical studies suggest a good performance of the proposed methods, and a finance real-data illustration reveals interesting aspects on the comovements of extreme losses between two leading stock markets.

References

[1] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. *Statistics of Extremes: Theory and Applications*. Wiley, Hoboken, NJ, 2004. doi:10.1002/0470012382

[2] Gudendorf, G., Segers, J. Extreme-Value Copulas. *Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*, 127–145, 2010. doi:10.1007/978-3-642-12465-5

[3] Koenker, R. *Quantile Regression*. Cambridge University Press, Cambridge, MA, 2005. doi:https://doi.org/10.1017/CBO9780511754098

[4] Petrone, S. Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26, 373-393, 1999. doi:https://doi.org/10.1016/j.spl.2017.03.030

[5] Hanson, T.E., de Carvalho, M., Chen, Y. Bernstein Polynomial Angular Densities of Multivariate Extreme Value Distributions. *Statistics and Probability Letters*, 128, 60–65, 2017. doi:https://doi.org/10.1016/j.spl.2017.03.030

Bayesian semi-parametric inferences for covariate-adjusted extreme-value copula with an application to cryptocurrency markets

Junho Lee[a,b], Miguel de Carvalho[a]
j.lee-63@sms.ed.ac.uk, Miguel.deCarvalho@ed.ac.uk

[a] *University of Edinburgh, U.K.*
[b] *Financial Supervisory Service, South Korea*

Keywords: angular measure, Bayesian semi-parametric methods, dependent Bernstein Dirichlet process, extreme-value copula, statistics of extremes

Abstract: Cryptocurrencies have been emerging as an alternative form of financial assets recently. Measuring extremal dependence of different financial assets is one of the most important subjects in financial risk modelling. To investigate the structure of the extremal dependence, we propose Bayesian inferences for covariate-adjusted angular densities and extreme-value copulae. Our methods resort to dependent Bernstein Dirichlet process prior with mean constraints to recover conditional angular densities and then induce covariate-adjusted extreme-value copulae. We apply the proposed methods to four major crypto-assets to reveal the structural change of extremal dependence using time-varying extreme-value copulae. Simulation studies are also provided to evaluate the performance of our methods.

Quantile-based fuzzy C-means clustering of multivariate time series: Robust techniques

Ángel López-Oriona[a], José A. Vilar[a,b]
oriona38@hotmail.com, jose.vilarf@udc.es

[a] *Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, 15071 A Coruña, Spain.*
[b] *Technological Institute for Industrial Mathematics (ITMATI), Spain.*

Keywords: exponential distance, multivariate time series, noise cluster, robust fuzzy C-means, trimming

Abstract: Three robust methods for clustering multivariate time series from the point of view of generating processes are proposed. The procedures are robust versions of a fuzzy C-means model based on: (i) estimates of the quantile cross-spectral density and (ii) the classical principal component analysis. Robustness to the presence of outliers is achieved by using the so-called metric, noise and trimmed approaches. The metric approach incorporates in the objective function a distance measure aimed at neutralizing the effect of the outliers, the noise approach builds an artificial cluster expected to contain the outlying series and the trimmed approach eliminates the most atypical series in the dataset. All the proposed techniques inherit the nice properties of the quantile cross-spectral density, as being able to uncover general types of dependence. Results from a broad simulation study including multivariate linear, nonlinear and GARCH processes indicate that the algorithms are substantially effective in coping with the presence of outlying series (i.e., series exhibiting a dependence structure different from that of the majority), clearly outperforming alternative procedures. The usefulness of the suggested methods is highlighted by means of two specific applications regarding financial and environmental series.

Acknowledgements: This research has been supported by the Ministerio de Economía y Competitividad (MINECO) grant MTM2017-87197-C3-1-P, the Xunta de Galicia through the ERDF (Grupos de Referencia Competitiva ED431C-2016-015), and the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01.

Improving estimation of the PCR duplicate rate in DNA sequencing experiments

Andy G. Lynch[a,b,c], Mike L. Smith[d], Matthew D. Eldridge[c], Simon Tavaré[e,c]
andy.lynch@st-andrews.ac.uk, mike.smith@embl.de,
matthew.eldridge@cruk.cam.ac.uk,st3193@columbia.edu

[a] *School of Mathematics and Statistics, University of St Andrews*
[b] *School of Medicine, University of St Andrews*
[c] *Cancer Research UK Cambridge Institute, University of Cambridge*
[d] *European Molecular Biology Laboratory. Heidelberg*
[e] *Herbert and Florence Irving Institute for Cancer Dynamics, Columbia University*

Keywords: whole-genome sequencing, maximum likelihood, duplicate rate, quality assurance, single nucleotide polymorphism (SNP)

Abstract: The volume of DNA in a sequencing experiment is often amplified by PCR, leading to the possibility that the same original DNA fragment will be sequenced twice - a 'PCR duplicate'. Generally indistinguishable from these are multiple sequences arising from identical but independent molecules, which can lead to an over-estimation of the PCR duplicate rate.

The PCR duplicate rate, and other measures derived from it, are important statistics for quality assurance, experimental design, and interpretation of sequencing experiments. Methods to improve the accuracy of duplicate rate estimates are therefore of value. Early approaches were based on predictions from the distribution of DNA fragment lengths [1], while we in our software [3] and others [2] have proposed that the locations of single nucleotide polymorphisms (SNPs) could provide a more direct approach to the problem.

Here we provide a full likelihood basis for an approach using SNP locations, including proof of a closed-form expression for key-coefficients in the model. We show the efficacy of the approach and how a naive SNP-based approach would fare in comparison. Finally we demonstrate the impact on two applications that make use of the duplicate rate.

Acknowledgements: We thank the Oesophageal Cancer Clinical And Molecular Stratification (OCCAMS) consortium for access to motivational data.

References

[1] Zhou W. *et al.* Bias from removing read duplication in ultra-deep sequencing experiments.. *Bioinformatics*, 30: 1073-1080, 2014.

[2] Bansal V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics*, 18(Suppl 3): 43, 2017.

[3] Lynch AG *et al.* LynchSmithEldridgeTavareFragDup (software) https://github.com/dralynch/duplicates 2016.

Modelos conjuntos para dados longitudinais no cancro da mama e tempo até recidiva

Sara Magalhães[a], Inês Sousa[b]
raquel_magalhaes00@hotmail.com, isousa@math.uminho.pt

[a] *Centro de Biologia Molecular e Ambiental, Escola de Ciências, Universidade do Minho, Guimarães, Portugal*
[b] *Departamento de Matemática e Aplicações, Universidade do Minho, Braga, Portugal*

Keywords: cancro da mama, longitudinal, modelos conjuntos, tempo até evento

Abstract: Modelos conjuntos para dados longitudinais e tempo até evento de interesse são usados em bases de dados onde o marcador longitudinal está associado com o evento de interesse. Para uma base de dados nacional de cancro da mama, consideram-se os dois marcadores tumorais mais utilizados no diagnóstico e acompanhamento deste tipo de cancro o CAE e o CA 15.3. O evento de interesse considerado é a recidiva de cancro após diagnóstico de cancro da mama. É expectável que o tempo até à recidiva esteja associado com a evolução da doença, traduzida na evolução do marcador tumoral.
Outros estudos foram já desenvolvidos usando como evento de interesse a morte do paciente, o que mostrou haver uma forte associação entre os dois processos. Agora pretende-se verificar se o mesmo tipo de associação está presente para a recidiva.
Desenvolveu-se uma extensa análise de uma base de dados com pacientes diagnosticados com cancro da mama acompanhados no Hospital de Braga. Consideram-se medidas repetidas dos dois marcadores tumorais, a cada 6 meses aproximadamente e o tempo desde o último diagnóstico até recidiva do cancro.

References

[1] Sousa, I. A review on joint modelling of longitudinal measurements and time-to-event. *REVSTAT*, 9, 57-81, 2011.

[2] Rodrigues, V. Manual de Ginecologia *Permanyer Portugal*, Vol. II, chap.34, 2011.

Impact of misclassification in association analyses of complex illnesses: the case of ME/CFS

João Malato[a], Luís Graça[a], Nuno Sepúlveda[b,c]
jmalato@medicina.ulisboa.pt, lgraca@medicina.ulisboa.pt, nunosep@gmail.com

[a] *iMM Lisboa, Universidade de Lisboa, Lisbon, Portugal*
[b] *CEAUL, Lisbon, Portugal*
[c] *Politechnika Warszawa, Warsaw, Poland*

Keywords: misclassification, monte carlo methods, power studies, stratification

Abstract: Misclassification can occur when diagnosing patients with complex illnesses without a biomarker. Despite great improvements in both clinical and statistical methods, and their subsequent interpretations, there are still areas where misclassification remains inevitable, with putative effects that can go uncontrolled or overlooked. Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is an example of a disease in which research is conducted under such uncertainty circumstances. For patients' diagnosis, there are currently more than 20 symptom-based case definitions for the disease. As a consequence, the respective diagnosis could vary from one case definition to another for the same patient. In this context, we investigated the impacts of misclassification and subgrouping of possible patients using hypothetical scenarios of association analyses from typical candidate gene and serology studies.

We estimated the power of each study through data simulation where a fraction of the sampled participants could be misdiagnosed as patients, depending on different rates for misclassification. Comparing these results to real-life scenarios showed the studies' consistency to discriminate cases from controls, weighting in for measurable variables such as the cohorts' sample sizes and differences in polymorphism/exposure risk for potential risk factors associated with ME/CFS. We concluded that patient's misclassification has indeed a deleterious impact in the respective findings and it should be taken into account when developing genetic and serological association studies of ME/CFS and other diseases in which the diagnosis is uncertain.

Acknowledgements: João Malato acknowledges a PhD fellowship by FCT – Fundação para a Ciência e a Tecnologia (grant ref. SFRH/BD/149758/2019). Nuno Sepúlveda is partially funded by FCT (grant ref. UIDB/00006/2020).

Using a Hierarchical GAM to model sea turtle sightings from 10 years of underwater surveys in Pearl Harbor, Hawaii

Ana Rita Marcelino[a], Tiago A. Marques[b,c], Sean F. Hanser[d], Stephen H. Smith[e], Robert K. Uyeyama[d]
fc47583@alunos.fc.ul.pt, tiago.marques@st-andrews.ac.uk, sean.hanser@navy.mil, stephen.h.smith@navy.mil

[a] *Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisbon, Portugal*
[b] *Centro de Estatistica e Aplicações, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisbon, Portugal*
[c] *Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland*
[d] *Naval Facilities Engineering Command, Pacific, 258 Makalapa Dr., Suite 100, Pearl Harbor, Hawaii*
[e] *Naval Facilities Engineering Services Center, 258 Makalapa Dr., Floor 3, Pearl Harbor, Hawaii*

Keywords: animal conservation, direct in-water observations, distribution modelling, hierarchical generalized additive models (HGAM), historical data

Abstract: Assessing presence and habitat use of marine turtles within foraging grounds provides valuable information for managing both populations and regions. Availability of food sources and several environmental factors such as water temperature and depth can influence sea turtle abundance and distribution in given regions. Pearl Harbor is a landlocked estuary controlled by the United States Navy (U.S. Navy) with restricted public access, situated on Oahu, Hawaii. To monitor the state of the natural resources present, the U.S. Navy has assembled observations from more than a decade of linear dive transects to search for sea turtles in Pearl Harbor and the coastal waters nearby. The present work aimed to assess temporal and spatial patterns in habitat use of green sea turtles (*Chelonia mydas*) in Pearl Harbor, by analyzing an historical dataset (2000-2011). We used a Hierarchical generalized additive model with a Zero-Inflated Poisson distribution to model the relationships between the number of turtles sighted per transect and temporal (year and month) and environmental (depth, sea surface temperature and underwater visibility) covariates. Our study allowed to reconstruct green turtles past use of an historic location such as Pearl Harbor, revealing their temporal and spatial distribution over ten years of sampling. Further, it was possible to identify the key areas used by the turtles, which allows the definition of priority zones, with a higher degree of protection, and contributes to the conservation of Hawaiian green turtles.

Modelling COVID-19 positivity rate as function of wastewater viral load

Carolina S. Marques[a], Mónica V. Cunha[b], Manuel Carmo Gomes[c], Ricardo Santos[d], Sílvia Monteiro[d], Nuno Brôco[e], Marta Carvalho[e], Rita Lourinho[f], Pedro Álvaro[f], João Vilaça[g], Fátima Meireles[g], Norberta Coelho[h], Marco Silva[h], Tiago A. Marques[a,i]
carolinasegmarques@gmail.com

[a] *Centro de Estatística e Aplicações, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Portugal*
[b] *Centre for Ecology, Evolution and Environmental Changes (cE3c), Biosystems & Integrative Sciences Institute (BioISI), Faculdade de Ciências, Universidade de Lisboa, Portugal.*
[c] *Plant Biology Department, Universidade de Lisboa, Portugal.*
[d] *Laboratorio de Análises, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal.*
[e] *AdP VALOR, Serviços Ambientais, S.A., Portugal.*
[f] *Águas do Tejo Atlântico, Portugal.*
[g] *SIMDOURO, ETAR de Gaia Litoral, Portugal.*
[h] *Águas do Norte, Portugal.*
[i] *Centre for Research into Ecological and Environmental Modelling, University of St Andrews, Scotland*

Keywords: covid-19 cases, generalized additive model, surveillance, viral loads, wastewater

Abstract: Understanding the full extent of the SARS-Coronavirus-2 (SARS-CoV-2) impact on human lives is an ongoing challenge. Every single epidemiological indicator has its own limitations and bias. The insufficiency of the testing capacity and the fact that people with mild symptoms or asymptomatic disease are not easily tracked are widespread problems. Experiences from other viral diseases have shown that monitoring sewage for traces of a pathogen allows for effective surveillance of entire communities, giving a sensitive signal for the presence of the pathogen in the population and for whether transmission is increasing or declining [1]. It is known that in the current COVID-19 pandemic, a significant proportion of infected individuals shed SARS-CoV-2 with their faeces [2]. To identify and quantify SARS-CoV-2 in wastewaters during the emergence of COVID-19 in Portugal, wastewaters from five ETARs (Alcantara, Beirolas, Gaia, Guia, and Serzedelo) and three Hospitals (Lisboa, Gaia, and Guimarães) were tested using the envelope gene (E). Using smooth viral loads summed across sites as an overall indicator of the health of the country, we modelled the positivity rate on COVID tests at the national level using a Generalized Additive Model model. We compared a number of modelling approaches including (1) a beta regression model, a (2) binomial model and (3) modelling the counts of positive tests with total numbers of tests as an offset. We compare the

results of the modelling approaches, which all show clearly that as the viral load increases the positivity rate on performed tests also increases.

Acknowledgements: We thank funding from Programa Operacional de Competitividade e Internacionalização (POCI) (FEDER component), Programa Operacional Regional de Lisboa, and Programa Operacional Regional do Norte (Project COVIDETECT, ref. 048467). Strategic funding from Fundação para a Ciência e a Tecnologia (FCT), Portugal, to CEAUL, cE3c and BioISI Research Units (UIDB/00006/2020, UIDB/00329/2020 and UIDB/04046/2020) is also gratefully acknowledged.

References

[1] Larsen, D. A., Wigginton, K.R. Tracking COVID-19 with wastewater. *Nature Biotechnology*, 38, 1151-1153, 2020. https://doi.org/10.1038/s41587-020-0690-1

[2] Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environmental Science & Technology Letters*, 7, 511–516, 2020. https://doi.org/10.1021/acs.estlett.0c00357

Estimating sperm whales echolocation click production rates to inform passive acoustic density estimation

Tiago A. Marques[a,b], Kalliopi C. Gkikopoulou[a], Carolina Marques[a,b]
tiago.marques@st-andrews.ac.uk, kg366@st-andrews.ac.uk,
carolinasegmarques@gmail.com

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

Keywords: abundance, acoustic tags, cue rate, distance sampling, sperm whale

Abstract: Passive acoustic monitoring has become a standard way of estimating animal abundance. A possible way of doing so is estimating a density of sounds and convert it into to a density of animals by dividing it by a cue production rate. However, suitable cue production rates are lacking for many species. In particular, rather than constant cue production rates, sound production might be dependent on different factors, like season, behavioral state, sex, etc. The standard way to collect information about sound production rates for cetaceans is to deploy animal-borne tags with acoustic sensors. One of the cetacean species for which abundance and density are most often estimated using passive acoustics are sperm whales *Physeter macrocephalus*, known to produce intense echolocation clicks for foraging. Here we present a dataset that covers around 150 acoustic tags placed on sperm whales at a set of 8 locations in 14 different years (in the 2001-2019 period) to investigate how cue production rates change over time and space, and in particular how sound production depends on depth. We investigate two different approaches to estimate cue production rates: (1) a "traditional" approach as was presented by [1] for beaked whales, using conventional regression models, and (2) a point process framework that might be more suitable since it avoids the need to discretize time into arbitrary time units for modelling. Accounting for variability in differences across individuals and how to incorporate those in species specific estimates based on Tags with different duration will be discussed.

Acknowledgements: This research is part of the ACCURATE project, funded by the US Navy Living Marine Resources program. TAM and CM thank partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020). We thank the researchers that made the DTAG data available.

References

[1] Warren, V.E., Marques, T. A., Harris, D., Tyack, P.L., Thomas, L., de Soto, N.A., Hickmott, L., Johnson, M.P. Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation. *The Journal of the Acoustical Society of America*, 141, 1962–1974, 2017. doi:10.1121/1.4978439

Outlier detection in histogram-valued variables

Ana Martins[a], Paula Brito[b], Sónia Dias[c], Peter Filzmoser[d]
a.r.martins@ua.pt, mpbrito@fep.up.pt, sdias@estg.ipvc.pt,
peter.filzmoser@tuwien.ac.at

[a] *Institute of Electronics and Informatics Engineering of Aveiro, Aveiro, Portugal*
[b] *Fac. Economia, Univ. Porto & LIAAD-INESC TEC, Porto, Portugal*
[c] *Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal*
[d] *Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria*

Keywords: histogram-valued data, Mallows distance, outlier detection

Abstract: The symbolic data approach was developed to properly describe and analyse data with intrinsic variability. Many data analysis methods have been developed for histogram-valued data. However, outlier analysis has been mostly overlooked. In this work a method for multivariate outlier detection of histogram-valued variables based on the Mallows distance is presented. This method is inspired on the approach by Hubert et al. (2015) thus, an outlyingness measure based on a one-dimensional projection of the observed data is computed. To this purpose, we use the definition of linear combination proposed by Dias and Brito (2015), based on the representation of histograms by the associated quantile functions, under uniformity hypothesis, which solves the problem of the semi-linearity of the representation space. To classify observations as outliers, different cut-offs for the outlyingness Mallows measure are tested, namely, Tukey's $Q_3 + 3(Q_3 - Q_1)$ cut-off and the P_{95} and the $P_{97.5}$ of a Chi-Square distribution with p degrees of freedom (p = number of variables). Simulation studies considering $p = 3$ normally distributed histogram-valued variables contaminated with observations from log normal, uniform and normal distributions were conducted. Overall, outliers from the log normal and the uniform distributions are easily identified. Normally distributed outliers are more easily identified if the variance of the distribution is shifted, rather than the mean. The Tukey cut-off shows the highest specificity (i.e., ability to identify non-outlier observations correctly) and sensitivity (ability to identify true outliers), suggesting this is the best option.

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

[3] Hubert, M., Rousseeuw, P.J., Segaert, P. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), 177–202, 2015.

Modelling the population dynamics of the Blackspot Seabream with a bayesian state-space model

Rui Martins[a,b], Lisete Sousa[a,b], Iúri Correia[b], Inês Farias[c], Ivone Figueiredo[c]
rmmartins@fc.ul.pt, lmsousa@fc.ul.pt, iuri96@outlook.pt,
ifarias@ipma.pt, ifigueiredo@ipma.pt

[a] *Faculdade de Ciências da Universidade de Lisboa (FCUL)*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[c] *Instituto Português do Mar e da Atmosfera (IPMA)*

Keywords: bayesian, fisheries, nimble, state-space, stock-assessment

Abstract: Modeling the population dynamics of the blackspot seabream (Pagellus bogaraveo) on the Portuguese coast (International Council for the Exploration of the Sea – Subarea 9 – Atlantic Iberian waters) is accomplished through a Bayesian state-space model based on the length of the fish [1]. This has the advantage of structuring the population in terms of a quantity that is directly observable, requiring no indirect estimation of age distributions according to length, since monitoring the evolution of a population by age is notoriously difficult.
A state-space model (SSM) models two time series. The observations, Y_t, made at discrete points in time $t = 1, 2, \ldots, T$, are conditioned on the states, Z_t (not observable), defined at the same time points.
Generally the observation and states equations characterize conditional expectations, as is most often the case in fisheries models, with an initial state Z_0 to be estimated (or predicted).
Over a period of time, there are several processes that affect the population's transition from the state Z_{t-1} to the state Z_t. These processes that act on the population gradually and simultaneously, will be modelled as being instantaneous and partially sequential. The main constituents of the population dynamics model are growth, survival and mortality, which affect the composition and size of the existing population, and also reproduction and recruitment, which allow adding new individuals to the population. In the case of blackspot sea bream there is another important process to account for – sex change.
This is a first attempt, yet simplistic, to model population dynamics of blackspot seabream, intended as a prototype around which further developments can be built. Several assumptions were made because the model was fitted with very limited observed data. Adding the knowledge of stakeholders in the form of highly informed prior distributions improved its precision.

References

[1] Mäntyniemi, SP, Whitlock, RE, Perälä, TA, Blomstedt, P, Vanhatalo, JP, Rincón, MM, Kuparinen, AK, Pulkkinen, HP, Kuikka, OS. General state-space population dynamics model for Bayesian stock assessment. *ICES Journal of Marine Science*, 72(8), 2209–2222, 2015. doi:10.1093/icesjms/fsv117

A two-array system for localization of fish using passive acoustic data

André B. Matos[a,b], M. Clara P. Amorim[a,b], Manuel Vieira[b,c], Tiago Marques[d,e], Paulo Fonseca[c]
bmatos.andre@gmail.com, mcamorim@fc.ul.pt, manuel_1990_v@hotmail.com, tiago.marques@st-andrews.ac.uk, pjfonseca@fc.ul.pt

[a] *Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *MARE, ISPA – Instituto Universitário, 1149-041, Lisboa, Portugal*
[c] *Dpt. de Biologia Animal e cE3c, Faculdade de Ciências, Lisboa Portugal*
[d] *CREEM, University of St Andrews, Scotland, United Kingdom*
[e] *DBA, CEAUL, Faculdade de Ciências da Universidade de Lisboa, Portugal*

Keywords: *Argyrossomus regius*, *Halobatrachus didactylus*, localization, passive acoustics, Tagus estuary

Abstract: Passive acoustic localization methods use the time lag of a sound detected at several receivers (or at a single receiver via directed and reflected paths) to estimate the position, range or direction of a sound source. The use of this method applied to vocal fish is still scarce in scientific literature despite its important potential applications. These include the elucidation of reproductive behaviour of vocal fish and their reaction to noise, the estimation of the source level of vocalisations, the construction of reference libraries with the sounds produced by each species, and the understanding of the role of vocal communication in social interactions (e.g. those involved in reproductive competition).
In this communication we present a low-cost system that can both estimate the positions of vocalising fish and intuitively convey the uncertainty of such estimates to the user. Some features in which the present method differs from previously published work include (1) the usage of multiple arrays to overcome the limitation of point localization to distances comparable to the size of the array, (2) an output that informs on both the estimated location and the associated uncertainty, (3) the modelling of the dependency structures between observed lags of different pairs of hydrophones, and (4) a method for estimating uncertainty that takes advantage of the known behaviour of the animals.

Acknowledgements: We acknowledge the Portuguese Air Force for facilitating the collection of data, the Science and Technology Foundation, Portugal (FCT) for funding this research (FISHNOISE - PTDC/BIA-BMA/29662/2017), and both MARE-ISPA and cE3c-FCUL for the institutional and logistical support.

Modeling and forecasting wind energy production by Stochastic Differential Equations

Paula Milheiro–Oliveira[a], Paulo Cabral[b]
poliv@fe.up.pt, paulo.cabral.anjos@gmail.com

[a] *Faculdade de Engenharia and CMUP, Universidade do Porto*
[b] *Faculdade de Ciências and CMUP, Universidade do Porto*

Keywords: Ornstein-Uhlenbeck process, parameter estimation, predictive modelling, renewable energy, stochastic differential equations

Abstract: Renewable energies have increased their relevance in the composition of the world's energy matrix. In particular, in Portugal, wind energy corresponds to 27.5 % of total energy production, being the second largest energy source in the country. The diversification of the energy matrix requires adequate tools for forecasting production from different sources, so that the management of energy resources is automatic and efficient. Wind energy presents a special challenge in this context, as it is a highly complex phenomenon with intensely non-linear behavior and with high variability. This work addresses this issue by a first tentative modeling the energy production of the wind turbines placed in continental Portugal using Stochastic Differential Equations (SDEs), based on available hourly observations. With this goal, we resort to parametric models of SDEs proposed in the literature of wind energy research (the Ornstein-Uhlenbeck model and a transformed Ornstein-Uhlenbeck model), we estimate the model parameters, we do the residual analysis and the short term forecasting. We found that SDEs have produced useful results for the management of wind energy production. However there would be an interest in evolving towards SDEs models that better explain the data in short periods of time, in order to obtain more reliable forecasts.

Acknowledgements: This research was partially supported by CMUP (UID/MAT/-00144/2013), funded by FCT (Pt) with National and European structural funds through FEDER, under partnership agreement PT2020, and by project STRIDE (NORTE-01-0145-FEDER-000033), supported by NORTE2020, through ERDF. The authors are very grateful to Cláudio Monteiro, from the Faculty of Engineering of the University of Porto, for motivating this research and for providing the data used.

References

[1] Nielsen, J.N., Madsen, H., Young, P.C. Parameter estimation in stochastic differential equations: an overview. *Annual Reviews in Control*, 24, 83–94, 2000.

[2] Verdejo, H., Awerkin, A., Kliemann, W., Becker, C. Modelling uncertainties in electrical power sysare more resilient to large variations in short periods of time, in order to obtain mortems with stochastic differential equations. *International J. of Electrical Power & Energy Systems*, 113, 322–332, 2019.

Moodle tests: not so much of a fuss when you have R

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

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

Keywords: evaluation tests, markdown, moodle, questions bank, R, randomized questions

Abstract: Evaluation is one of the components of the teaching-learning process. It is a periodic task that all teachers desire to continuously improve. In higher education institutions it not only supports the learning process but also allows for accountability and certification. In that process teachers aim to achieve equity, suitability, reliability and efficiency. Its preparation is one of the most time-consuming activities and so it is highly desirable to reduce the time used in that process. Moodle statistics show that this is one of the most adopted platform to support the process of learning-teaching in educational institutions of all degrees. It provides some tools to perform formative as well as summative evaluation. Pandemic disease brought the need for learning how to use those tools and incremented its application. Recent studies show new problems raised with home evaluation tests; one of which is the easier possibility of fraud. One way to fight it is to increase the number of different questions presented to different students. With particularly advantages to statistics teachers, the R package *exams* is a powerful tool that gives some answers to those problems: it produces questions that can be exported directly to Moodle quiz format and it allows for random generation of parametrized questions. In doing so, it allows to rapidly multiply the number of questions with the same level of difficulty and same topics included in the bank of questions. This paper aims to show how these tools combined with *latex* and *markdown* environments may contribute to help Statistics teachers activity.

Acknowledgements: Research partially supported by National Funds through **FCT**, —Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/0006/2021 (CEA/UL) and UIDB/MAT/04106/2021 (CIDMA).

References

[1] Gamage, S.H.P.W., Ayres, J.R., Behrend, M.B. et al. Optimising Moodle quizzes for online assessments. *IJ STEM*, Ed 6, 27, 2019. doi:https://doi.org/10.1186/s40594-019-0181-4

[2] Zeileis A, Umlauf N, Leisch F. Flexible generation of E-learning exams in R: Moodle quizzes, OLAT assessments, and beyond.*J Stat Softw*,58(1),1-36,2014.

[3] Guangul, F.M., Suhail, A.H., Khalit, M.I. et al. Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educ Asse Eval Acc* 32, 519-535 (2020). https://doi.org/10.1007/s11092-020-09340-w

Intensity-Dependent Point Processes

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

[a] *CMA, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa, Portugal*
[b] *CINAMIL, Instituto Universitário Militar, Portugal*
[c] *CEAUL, Faculdade de Ciências da Universidade de Lisboa, Portugal*
[d] *DM, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Portugal*
[e] *IPMA, Instituto Português do Mar e da Atmosfera*
[f] *CIDMA, Centro Investigação e Desenvolvimento em Matemática e Aplicações*

Keywords: log Gaussian Cox process, marked point process, preferential sampling

Abstract: A practical and theoretically interesting problem in the contex of Point Processes are Marked Point Patterns where the statistical properties of marks depend locally on point intensity. Such dependence can be observed e.g. in fishery data, where catches (marks) are certainly associated with the locations where the fisheries take place (points), in order to optimize capture effort. In intensity-marked point processes for the stationary log Gaussian Cox process the marks are allowed to be marginally correlated and the mark size depends locally on the point density. In this work we analyse the relationship between these models and the geostatistical model under preferential sampling. Detecting dependence between marks and locations of marked point processes is an important issue because predictions of the process can be severely biased when standard statistical methodologies are applied to data where the distribution of a mark varies along the point density. Exploring the abovementioned relation, we evaluate the existence of this dependency in real data provided by the Instituto Português do Mar e da Atmosfera (IPMA) which correspond to the black scabbardfich catches in the fishing grounds of the South zone of Portugal, from 2009 to 2013.

Acknowledgements: This work was partially supported by Portuguese Foundation for Science and Technology through the project PREFERENTIAL, PTDC/MAT-STA/28243/2017, PREFERENTIAL and UIDB/00297/2020.

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] Myllymäki, M., Penttinen, A. Conditionally heteroscedastic intensity-dependent marking of log Gaussian Cox processes *Statistica Neerlandica*, 63(4), 450–473, 2009. doi.org/10.1111/j.1467-9574.2009.00433.x

Non-parametric change point detection applied to climate change in Europe

Magda Monteiro[a,b], Marco Costa[a,b]
msvm@ua.pt, marco@ua.pt

[a] *Águeda School of Technology and Managment, University of Aveiro, Portugal*
[b] *CIDMA – Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal*

Keywords: air temperature, change point detection, climate change, distribution-free estimation, state space models

Abstract: The study of long temperature time series is of particular interest in understanding climate dynamics, allowing efficient monitoring of environmental processes.
This work presents a statistical modeling of time series of monthly average temperatures in fifty European locations using a state space approach, where it is considered a model comprising a deterministic seasonal component and a stochastic trend component. The aim of this analysis was to identify, through a non-parametric detection method, the change points from which the pattern growth has changed and use them to accurately estimate growth rates for the forthcoming decades.
Maximum Mann-Whitney type tests were applied to the one-step-ahead innovations obtained from the application of a classical decomposition approach to estimate the model's parameters. A combination of least squares estimation of the seasonal parameters with a distribution-free estimators developed to state space models in order to estimate the remaining parameters.
In Northern Europe the change points were, almost all, identified in the late 1980s while in Central and Southeastern Europe was, for the majority of cities, in the 1990s and later.

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 - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

Longitudinal analysis of viral shedding in astronauts before, during, and after a mission to the International Space Station

Frederico Moreira[a], Marília Antunes[a,b], Nuno Sepúlveda[b]
fc48046@alunos.fc.ul.pt, mcreis@fc.ul.pt, nunosep@gmail.com

[a] *Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*

Keywords: longitudinal models, missing data, multiple imputation based on chained equations, space flight

Abstract: In space, astronauts experience loss of gravity, lack of exposure to natural light and confinement. These stressful conditions could lead to the reactivation of dormant infections by common herpesviruses. These viral reactivations have been linked to different pathologies, such as chronic fatigue syndrome, cancer and arteriosclerosis. To understand these viral reactivations, a recent study analysed data on viral shedding in 23 astronauts participating in a long-duration mission to the International Space Station [1]. Viral shedding occurs when a virus is able to replicate productively within the host cell. The data of this study refers to viral counts measured before, during and after the flight and there were some missing data during the flight. The analysis of this data started with the imputation of the missing values using the method of multiple imputation by chained equations (MICE) [2]. Initially, MICE was used with the transformed binary data where 1 represented a detected reactivation in that given moment and 0 otherwise. In general MICE creates m imputation chains which are iteratively updated until a convergence is achieved. For each imputation a logistic mixed model was fitted describing the probability of reactivation in each timepoint. The pooling method proposed by Rubin estimated that the probabilities of reactivation for the Epstein-Barr virus were 0.126, 0.239 and 0.453 for the inflight timepoints "early", "mid" and "late", respectively. After testing the hypothesis, the only inflight timepoint with a model parameter significantly different from the baseline (referring to measurements taken 180 days before launch) was the "late" timepoint.

References

[1] Mehta S. K., Ladenslager M.L., Stowe R.P., Crucian B.E., Feiveson A.H., Sams C.F., Pierson D.L., Latent virus reactivation in astronauts on the International space station, *Npj Microgravity* 3:11, 2017.

[2] Van Buuren, S., Groothuis-Oudshoorn, K., mice: Multivariate Imputation by Chained Equations, *R. Journal of Statistical Software*, 45(3), 1-67, 2011.

[3] Rubin, D. B., Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91:434, 473-489, 1996.

Fitting flexible parametric detectability functions in distance sampling

Jaime Mosquera[a], Russell Alpizar-Jara[b]
jmosquerag@unal.edu.co, alpizar@uevora.pt

[a] *Universidad Nacional de Colombia*
[b] *CIMA-IIFA/DMAT-ECT, University of Évora, Portugal*

Keywords: abundance, density estimation, detectability function, model selection, survival analysis

Abstract: Population density estimation in distance sampling requires fitting a probability density function denoted by $f(y|\theta)$, where y represents the perpendicular (or radial) distance from a detected animal (or object) to a transect line (or point), and θ represents the vector parameter indexing this family of probability density functions. The most popular approach to estimate $f(\cdot)$, is based on a semi–parametric methodology proposed by [1, 2] using typical forms such as half-normal and hazard-rate functions. The main idea is to find the maximum likelihood estimator for θ using a parametric functional form combined with a series expansion. We present an R package with several possible shapes of detectability functions based on already implemented survival functions from novel distributions currently used in parametric survival analysis [3].

Acknowledgements: This research has been partially supported by the Centro de Investigação em Matemática e Aplicações (CIMA), through the Project UIDB/04674 /2020 of FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

[1] Buckland, S.T. Maximum likelihood fitting of the Hermite and simple polynomials densities. Applied Statistics, 41, 241–266, 1992.

[2] Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L., *Introduction to Distance Sampling: Estimating Abundance of Biological Populations.* Oxford University Press, Oxford, UK, 2001.

[3] Almalki, S. J., Nadarajah, S. Modifications of the Weibull distribution: A review. Reliability Engineering and System Safety, 124, 32–55, 2014. https://doi.org/10.1016/j.ress.2013.11.010

Estimation of the Effective Reproduction Number for the COVID-19 Epidemic in Switzerland

Marcelo Henrique de Oliveira Mrtvi[a,b], Isolde Previdelli[a], Anthony C. Davison[b]
m.mrtvi@gmail.com, isoldeprevidelli@gmail.com, anthony.davison@epfl.ch

[a] *Universidade Estadual de Maringá (UEM)*
[b] *École polytechnique fédérale de Lausanne (EPFL)*

Keywords: Bayesian estimation, COVID-19, Markov Chain Monte Carlo, Metropolis-Hastings algorithm, reproductive number

Abstract: The COVID-19 pandemic created a harsh dilemma for our society, in which the application of public health measures and restrictions has to be balanced with their economic consequences. To guide decisions during this crisis, one of the main indicators used by governments is the effective reproductive number (R_t). It represents the expected number of secondary cases that an infected individual will cause during his infectious period. This project uses a Bayesian approach with a Metropolis-Hastings algorithm to estimate the reproduction number in Switzerland and other countries. While most approaches reconstruct the incidence curve in a separate step from the estimation of R_t, our approach estimates the weekly pattern, the incidence curve and R_t in the same algorithm. Also, the use of splines as a prior for R_t allow us to create a dependence between days avoiding sudden jumps in the estimate. The developed method is applied to simulated data, real data for the COVID-19 pandemic and compered to the official estimates from the Swiss government generated by the Swiss National COVID-19 Science Task Force.

Acknowledgements: We thank EPFL for funding the project.

References

[1] Cori, A., Ferguson,N., Fraser, C., Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *AmericanJournal of Epidemiology*, 178(9):1505–1512, 2013. doi:10.1093/aje/kwt133

[2] Gostic, K. et al., Practical considerations for measuring the effective reproductive number, R_t. *PLOS computational biology*, 16(12):e1008409, 2020.doi:10.1371/journal.pcbi.1008409

[3] Sciré, J. et al., Reproductive number of the Covid-19 epidemic in Switzerland with a focus on the cantons of Basel-Stadtand Basel-Landschaft. *Swiss Medical Weekly*, 150(19-20):w20271, 2020. doi:10.4414/smw.2020.20271

Tail Index Estimation in the Presence of Covariates

João Nicolau[a], **Paulo M. M. Rodrigues**[b], Marian Stoykov[c]
nicolau@iseg.ulisboa.pt, prodrig@novasbe.pt, marian.stoykov@novasbe.pt

[a] *ISEG, Universidade de Lisboa and CEMAPRE*
[b] *Banco de Portugal and NOVA School of Business and Economics*
[c] *NOVA School of Business and Economics*

Keywords: covariates, extreme value theory, pareto-type distributions, tail index

Abstract: This paper provides novel theoretical results for estimation of the conditional tail index of Pareto and Pareto-type distributions in a time series context. We show that both the estimators and relevant test statistics are normally distributed in the limit, both when independent and identically distributed data is considered as well as when the data is dependent. Simulation results provide support for the theoretical findings and highlight the good finite sample performance of the approach in a time series context. The methodology developed in this paper is then used to compute a systematic risk measure which exploits the relationship between the market volatility from the Chicago Board of Exchange Volatility Index and the probability of observing extreme values in the tails.

Acknowledgements: The authors gratefully acknowledges financial support from the Portuguese Science Foundation (FCT) through project PTDC/EGE-ECO/28924/2017, and (UID/ECO/00124/2013 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

O desempenho do algoritmo EM na estimação de misturas de modelos lineares mistos para diferentes procedimentos de inicialização

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

[a] *Universidade do Minho, CBMA*

Keywords: algoritmo EM, estudo de simulação, método de máxima verosimilhança, modelos de mistura

Abstract: Os modelos de mistura têm sido muito utilizados na modelação de conjuntos de dados que provêm de populações heterogéneas. Em particular, em análise de regressão é recorrente o uso de modelos de mistura de regressões para modelar a heterogeneidade não observada da população. Dentro do contexto dos modelos de mistura de regressões, as misturas de modelos lineares mistos aplicam-se a diversas áreas, dado que permitem explicar as correlações entre observações do mesmo indivíduo, através da incorporação de efeitos aleatórios e, simultaneamente, modelar a heterogeneidade entre diferentes indivíduos.

Apesar do algoritmo EM ser o algoritmo mais utilizado na estimação dos parâmetros em modelos de mistura, uma das principais dificuldades consiste na seleção dos valores iniciais dos parâmetros, pois o algoritmo apenas garante a seleção de um máximo local. Deste modo, a seleção de valores iniciais para os parâmetros é crucial para o bom funcionamento do algoritmo. Para solucionar o problema é conveniente escolher diversos valores iniciais para os parâmetros, quer seja aleatoriamente ou recorrendo a algum critério, para garantir a escolha de um máximo adequado, isto é, por forma a selecionar a solução que proporcione o maior valor da função de log-verosimilhança.

O objetivo deste trabalho consiste em analisar o problema da seleção de valores iniciais para os parâmetros de misturas de modelos lineares mistos aquando da inicialização do algoritmo EM. Para isso comparam-se distintas medidas de desempenho recorrendo aos verdadeiros valores dos parâmetros ou a valores aleatórios, através de um estudo de simulação.

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] 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), 1–38, 1977. doi:10.1111/j.2517-6161.1977.tb01600.x

[2] McLachlan, G., Peel, D. *Finite Mixture Models*. John Wiley & Sons, 2000. doi:10.1002/0471721182

[3] Scharl, T., Grün, B., Leisch, F. Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics*, 26(3), 370–377, 2009. doi:10.1093/bioinformatics/btp686

Em que se parecem e em que diferem entre si os municípios Portugueses

Inês Oliveira[a], Luísa Canto e Castro[a,b]
nesaaoliveiraa@gmail.com, ldloura@fc.ul.pt

[a] *Faculdade de Ciências da Universidade de Lisboa*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa*

Keywords: análise de *clusters*, componentes principais, estatísticas oficiais, estatística multivariada

Abstract: Este trabalho teve por objectivo traçar um perfil dos 308 municípios portugueses, perfil esse que possa servir de base à elaboração de cenários prospetivos e que permita identificar as áreas que poderão vir a configurar uma situação de risco e a necessitar de intervenção. Nesse sentido, de entre o vasto leque de dados estatísticos oficiais disponíveis ao nível do município, olhou-se de forma multivariada para aqueles que incidem sobre: envelhecimento, escolarização, habitação, mercado de trabalho, finanças autárquicas, população estrangeira e actividade turística. Para as variáveis escolhidas foram considerados, tanto os valores mais recentes como as respectivas taxas de variação nos últimos 5 anos. Quanto à metodologia estatística de análise dos dados, esta passou, num primeiro passo, pela identificação das componentes principais e, num segundo passo, pela constituição dos grupos de municípios com valores similares nessas componentes.

Enfoque decisional num quadro bayesiano paramétrico para verificar conformidade de água de lastro de navios com normas prescritas

Carlos Daniel Paulino[a], Eliardo G. Costa[b], Julio M. Singer[c]
daniel.paulino@tecnico.ulisboa.pt, eliardocosta@ccet.ufrn.br,
jmsinger@ime.usp.br

[a] *Universidade de Lisboa, CEAUL*
[b] *Universidade Federal do Rio Grande do Norte*
[c] *Universidade de São Paulo*

Keywords: água de lastro, intervalo de credibilidade, modelo bayesiano Binomial Negativa/Gama, risco de Bayes

Abstract: Normas estabelecidas na comunidade marítima internacional impõem limites superiores para a concentração média de micro-organismos com dadas dimensões nos tanques de água de lastro antes de se efetivar a descarga desta à medida que se processa o carregamento dos navios. O cumprimento de tais normas pode ser verificado através de estimativas intervalares da concentração média a obter de coleta de uma amostra de alíquotas de água cujo tamanho (n) deve ser previamente determinado por algum método. Considera-se para o efeito o modelo bayesiano Binomial Negativa/Gama para as hipotéticas contagens de organismos e associada concentração média.
Encarando o problema de determinação de n e de intervalos de credibilidade (IC) à luz da teoria de decisão, com especificação de uma função perda baseada na amplitude e viés do IC, o critério de minimização de uma função do risco de Bayes e do custo operacional da amostragem permite determinar o número ótimo de alíquotas de água para vários valores dos hiperparâmetros distribucionais fixados. Concretizando então a amostragem desse número ótimo de alíquotas e a contagem dos correspondentes números de organismos, calcula-se o IC de Bayes recorrendo a métodos de simulação Monte Carlo, com base no qual se pode tomar uma decisão sobre a conformidade ou não da água do tanque com a norma em questão.

Variable selection for black-box functions

Rui Paulo[a], Gonzalo García-Donato[b]
rui@iseg.ulisboa.pt, gonzalo.garciadonato@uclm.es

[a] *ISEG, Departament of Mathematics and CEMAPRE/REM, Universidade de Lisboa*
[b] *Department of Economics and Finance, Universidad de Castilla-La Mancha*

Keywords: black-box, computer model, Gaussian process, screening, variable selection

Abstract: The black-box functions that we have in mind are what are generally referred to as computer models, i.e., computer implementations of (deterministic) mathematical models built to replicate specific real life processes. Inputs come in, in the form of a p-dimensional vector describing the system, and the computer program outputs the model's prediction of the real process for that configuration. Generally speaking, computer models are computationally too intensive to allow traditional techniques of numerical analysis to be used, and typically one only has access to runs of the model at carefully designed set of inputs.

In this context, we approach the screening problem — i.e. detecting which inputs of a computer model significantly impact the output — from a formal Bayesian model selection point of view. That is, we treat the function as unknown, place a Gaussian process prior on it, and consider the 2^p Bayesian statistical models that result from assuming that each of the subsets of the p inputs affect the response. The goal is to obtain the posterior probabilities of each of these models and use this object to answer the screening problem. Hence, the methodology is very much reminiscent of Bayesian variable selection from a model selection perspective in the context of linear regression.

We focus on the specification of objective priors for the parameters of the Gaussian process of each model and on the use of the Laplace approximation to compute the associated marginal likelihood. This results in methodology that is computationally quite fast and fully automatic. We illustrate ideas using synthetic examples taken from the screening literature.

Calibration of daily maximum temperature: a comparative study using regression and state space models

F. Catarina Pereira[a], Marco Costa[b], A. Manuela Gonçalves[a]
up202010700@edu.fe.up.pt, marco@ua.pt, mneves@math.uminho.pt

[a] *Department of Mathematics & Center of Mathematics, University of Minho, Portugal.*
[b] *Águeda School of Technology and Management & Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal.*

Keywords: calibration, Kalman filter, maximum temperature, state space model

Abstract: At present, it is essential to manage water in a sustainable way. Economic development, population growth, and climate change have all contributed to the dwindling of this limited natural resource. Water is present in practically all human activities, with agriculture being one of the activities that consumes the highest share of water. Thus, it is critical to find the best technical solutions to improve water use efficiency, particularly in irrigation systems. This work is carried out in the context of project "TO CHAIR - Optimum Challenges in Irrigation", which aims to estimate and predict water losses by evapotranspiration. In this study, we present a comparison of forecast models, namely the state space models associated with the Kalman filter, particularly the calibration model, and linear regression models, which constitute a particular class of the calibration model when considering the deterministic state. The objective is to improve short-term forecasts of the initial h-steps-ahead ($h = 1, \ldots, 6$ days) of the maximum temperature in real time, obtained from the `weatherstack.com` website. These models allow dealing with time series with unstable behavior, which is a predominant characteristic in meteorological data, due to their ability to incorporate periodic structure, trend, seasonality, temporal correlation, as well as important covariates for the explanation of the process, [1]. To estimate the unknown parameters of the state space model, the classical approach was used, through the maximum likelihood estimation, assuming the normality of the disturbances, [2].

Acknowledgements: Funding from FEDER/COMPETE/ NORTE2020/POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/ IEEA/00147/ 006933-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247, and from the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and UIDB/04106/2020 and UIDP/04106/2020 of CIDMA-UA. This work was also financed by national funds from FCT through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM.

References

[1] Costa, M., Alpuim, T. Adjustment of state space models in view of area rainfall estimation. *Environmetrics*, 22, 530–540, 2011.

[2] Shumway, R. H., Stoffer, D. S. Time Series Analysis and Its Applications: with R examples. Springer, 4th edition, 2017.

An auction where prices only decrease: a probability distribution for the number of bids

Julio Pereira[a,b], Heung Yeung Lam[a]
J.Pereira@massey.ac.nz, H.Y.Lam@massey.ac.nz

[a] *Massey University, New Zealand*
[b] *Federal University of São Carlos, Brazil*

Keywords: auction, average profit, geometric distribution, non-constant probability

Abstract: An auction website, created by a young student and entrepreneur, auctioned items in an unusual way. An item was entered for sale with an initial price equal to the market value, but the person interested in buying it could only see it but not its price when accessing the auction. To find this information they had to pay the website $1. After this payment, the price of the item was automatically reduced by $0.50 and that current price was shown to the bidder. The bidder had two minutes to decide whether to buy the item or not. After two minutes the price was hidden again and the item was relisted. Therefore, the more viewings, the more the item's price dropped, until someone made the purchase. We proposed a function to calculate probabilities associated to the random variable 'number of bids required for the item to be sold', in which the probability of success at each bid is not constant. We established the necessary conditions and proved that under such conditions the proposed function is a probability mass function. This function was known as 'Ale's distribution', after the student who first considered this problem. Finally we presented an example, where we fitted 'Ale's distribution' to a dataset resulting from 60 independent auctions of the identical item and we estimated the expected profit for that item. We compared the results to those obtained assuming the geometric distribution as a probability model for the data. We concluded that the average profit estimate based on Ale's distribution is more reliable, since this distribution fits the data better.

Bayesian Lasso Tail Index Regression

Soraia Pereira[a], Miguel de Carvalho[a,b], Carlos da Câmara[c], Ricardo Trigo[c]
sapereira@fc.ul.pt

[a] *CEAUL, Universidade de Lisboa, Portugal*
[b] *School of Mathematics, University of Edinburgh, UK*
[c] *IDL, Universidade de Lisboa, Portugal*

Keywords: extreme value index, Lasso, statistics of extremes, variable selection

Abstract: In extreme value statistics, the extreme value index is an important measure to understand the heavy-tailed behavior of a distribution. Some estimators for this index have been proposed in the extremes literature, among them the tail index regression estimator proposed by [2]. The proposed approach was motivated by the importance of the relationship between the extreme response and covariates in real data analysis. Under Pareto-type distributions, the authors employed the logarithmic function to link the tail index to the linear predictor induced by covariates. Here we introduce a novel regression model for the extreme value index based on a Bayesian Lasso specification. The Bayesian Lasso was introduced by [1] as a Bayesian version of Tibshirani's Lasso ([3]), which is a regularization method that shrinks some regression coefficients and sets others to zero, being naturally tailored for variable selection. To fully examine the finite-sample performance of the proposed methodology, we report the main findings of a Monte Carlo simulation study. Forest fires data are used to illustrate the application of our method in a real data problem.

Acknowledgements: This work 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] Park, T., Casella, G. The Bayesian Lasso. *Journal of the American Statistical Association*, 103, 681–686, 2008. doi:10.1198/016214508000000337

[2] Wang, H., Tsai, C. L. Tail Index Regression. *Journal of the American Statistical Association*, 104, 1233-1240, 2009. doi:10.1198/jasa.2009.tm08458

[3] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267-288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x

Bivariate generalized Gompertz distribution in presence of mixture and non-mixture cure fraction models

Marcos Vinicius de Oliveira Peres[a], Jorge Alberto Achcar[b], Edson Zangiacomi Martinez[b]
mvperes1991@alumni.usp.br,achcar@fmrp.usp.br, edson@fmrp.usp.br

[a]*State University of Maringá (UEM), Cidade Gaúcha, PR, Brazil.*
[b]*University of São Paulo (USP), Ribeirão Preto, SP, Brazil*

Keywords: Bayesian estimates, copula function, survival analysis

Abstract: In survival analysis is common when for each patient we observe two lifetimes. In this cases, the time to the events can be modeled by a bivariate distribution function. A method to generate bivariate distributions and to model dependence between the lifetimes is the use of copula functions. Copulas are basically functions that link univariate distributions to generate multivariate distributions. The copula functions enable to define different distributions for the marginals, with a dependence structure, creating a multivariate distribution with the selected margins. It is important to observe that different copula functions have different structures of dependence among the lifetimes.
A other common situation in survival analysis is the presence of a fraction of individuals not expecting to experience the event of interest, thus individuals are immune to the event or cured for the disease during the study and known as long-term survivors. In this situation, commonly the mixture model is considered, it assumed that the population is a mixture of susceptible individuals who experience the event of interest and non-susceptible individuals that supposedly will never experience it. On other hand, we can also consider the non-mixture model defines an asymptote for the survival function, that is associated to the fraction of cured.
In this paper, the main goal is to explore the use of the Generalized Farlie-Gumbel-Morgenstern, Ali-Mikhail-Haq, Clayton, Gumbel-Hougaard, and Frank copula in the analysis of different bivariate lifetime data assuming a generalized Gompertz distribution introduced by [1] in the presence of mixture and non-mixture models to estimate the cure rate.

References

[1] El-Gohary, A., Alshamrani, A.,Al-Otaibi, A. The generalized Gompertz distribution. *Applied mathematical modelling*, 37, 13-24, 2013. doi:10.1016/j.apm.2011.05.017

Explorando o poder da memória das redes neuronais LSTM na modelação e previsão do PSI 20

F.R. Ramos[a], A.R. Costa[b], D.A. Mendes[a], D.R. Lopes[c]
frjrs@iscte-iul.pt, anabela.costa@iscte-iul.pt, diana.mendes@iscte-iul.pt, dro.lopes@campus.fct.unl.pt

[a] *ISCTE-IUL e BRU-IUL*
[b] *ISCTE-IUL e CMAF-CIO*
[c] *UNL e CEDOC-UNL*

Keywords: deep neural networks, LSTM, previsão, séries temporais

Abstract: A articulação de técnicas estatísticas, matemáticas e computacionais, na modelação e previsão de séries temporais, manifesta-se num claro suporte de apoio à tomada de decisão. Especificamente para séries temporais económico-financeiras, como as relativas aos mercados financeiros, a aplicação de metodologias de *Machine Learning*, em particular de *Deep Learning*, tem sido apontada como uma opção promissora.

Trabalhos desenvolvidos anteriormente [3, 1] apresentam resultados concordantes com a literatura científica, apontando para algumas limitações das metodologias clássicas lineares no processo de modelação e previsão de séries financeiras. Os modelos não lineares mostram-se mais adequados do que os lineares para este propósito, sendo as Redes Neuronais Artificiais um exemplo bastante bem-sucedido. Deste modo, analisamos neste trabalho o poder da 'memória de longo prazo' presente em algumas arquiteturas de *Deep Neural Networks*, em particular nas redes *Long Short-Term Memory* (LSTM). Posteriormente, mediante a construção de rotinas computacionais completas e automatizadas (disponibilizadas em [1]), avaliam-se as potencialidades destas arquiteturas (LSTM) face às arquiteturas *Multilayer Perceptron* na modelação e previsão do PSI20. Para efeitos de analise crítica e comparativa, são objeto de discussão a qualidade preditiva dos modelos e o custo computacional implícito.

Apesar de reconhecidas vantagens nas redes LSTM, pela análise ao erro de previsão (*Mean Absolute Percentage Error*), apontam-se limitações em termos do tempo de execução computacional abrindo perspetivas a desenvolvimentos futuros.

References

[1] Costa, A., Ramos, F., Mendes, D., Mendes, V. Forecasting financial time series using deep learning techniques. *In IO 2019 - XX Congresso da APDIO 2019*. Instituto Politécnico de Tomar - Tomar, 2019.

[2] Lopes, D., Ramos, F. Univariate Time Series Forecast. Retrieved from https://github.com/DidierRLopes/UnivariateTimeSeriesForecast

[3] Ramos, F., Costa, A., Mendes, D., Mendes, V. Forecasting financial time series: a comparative study. *In JOCLAD 2018, XXIV Jornadas de Classificação e Análise de Dados*. Escola Naval – Alfeite, 2018. doi:10.13140/RG.2.2.11548.41606

Estimating probability of detection of Blaineville's beaked whales (*Mesoplodon densirostris*) groups using regression models

João Ribeiro[a], Tiago A. Marques[b,c]
fc47713@alunos.fc.ul.pt, tiago.marques@st-andrews.ac.uk

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

Keywords: blainville's beaked whales, distance estimation, echolocation, passive acoustic, probability of detection

Abstract: Blainville's beaked whale (*Mesoplodon densirostris*, Md) is known for associating in small groups and tends to vocalize only during deep foraging dives. These sounds are ultrasonic echolocation clicks, and are routinely detected year-round on the Atlantic Undersea Test and Evaluation Center (AUTEC) range. The range consists of a 93 bottom-mounted hydrophone array. We used a data set described by [1], with automated Md group dive detections, for which group size per dive was estimated also by [1]. Highly likely false positives were removed through manual inspection, removing dives with biologically infeasible characteristics. We estimated the average location for each of 8271 deep dives detected at the AUTEC using three different approaches: (1) the centroid of hydrophone locations a group was detected at, and the weighted average of these coordinates, weighted by (2) the count of detected clicks by hydrophone and (3) the log of these counts. We modelled the deep dive detection probability as a function of the distance to the hydrophones, group size and hydrophone type. The most important variable to explain detection probability was the distance to the hydrophone, using the log of the counts as weights. Detectability varied depending on hydrophone type and group size. As expected, the probability of detection is higher for larger groups and for recent hydrophones. These results support findings from literature, contribute to the *Md* species information repository and shed further light regarding passive acoustic detection of beaked whale groups.

Acknowledgements: TAM thanks partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020)

References

[1] Marques, T. A., Jorge, P. A., Mouriño, H., Thomas, L., Moretti, D. J., Dolan, K., Claridge, D., Dunn, C. Estimating group size from acoustic footprint to improve Blainville's beaked whale abundance estimation *Applied Acoustics*, 156, 434–439. doi:10.1016/j.apacoust.2019.07.042

Análise multivariada para identificação da influência dos fatores antropométricos no dispêndio metabólico na marcha durante a utilização de um exosqueleto para atuação do tornozelo

Nuno Ribeiro[a], Luís Quinto[a,b], Sérgio Gonçalves[b], Ivo Roupa[b], Miguel Silva[b], Paula Simões[a,c]
ribeiro.ngmfa@exercito.pt, luis.quinto@academiamilitar.pt,
sergio.goncalves@tecnico.ulisboa.pt, iroupa@gmail.com,
miguelsilva@tecnico.ulisboa.pt, paula.simoes@academiamilitar.pt

[a] *CINAMIL, Academia Militar-Instituto Universitário Militar, Portugal*
[b] *IDMEC, Instituto Superior Técnico - Universidade de Lisboa, Portugal*
[c] *CMA, Faculdade de Ciências e Tecnologia - Universidade NOVA de Lisboa, Portugal*

Keywords: dados antropométricos, estatística multivariada, exosqueleto, militar, regressão linear múltipla

Abstract: Atualmente, os militares são incumbidos de funções a que estão associados esforços físicos intensos, aumentando a sua probabilidade de lesão e a consequente redução do nível de operacionalidade da força, situação que preocupa decisores políticos e responsáveis militares. Os exosqueletos assumem-se como uma possível solução para a atenuação do risco de lesão. No âmbito do projeto ELITE - Enhancement LITe Exoskeleton, que pretende desenvolver um exosqueleto passivo, para redução do risco de lesão dos militares e o aumento do seu nível de operacionalidade, reduzindo o custo metabólico despendido, pretende-se inferir a existência de uma relação entre os parâmetros antropométricos e a rigidez do elemento elástico durante a marcha, para definição de um procedimento de seleção deste elemento tendo em conta as características físicas do utilizador. Uma análise estatística com recurso a técnicas multivariadas foi aplicada com o objetivo de analisar a relevância dos dados antropométricos nos melhores resultados metabólicos obtidos nos ensaios laboratoriais, possibilitando a seleção da mola mais adequada, evitando morosos ensaios em laboratório. O estudo permitiu estabelecer um método quantitativo para seleção do elemento de força mais adequado para cada militar, sendo este um ponto crucial para garantir a boa performance e conforto do exosqueleto.

Acknowledgements: Ao Estado Maior do Exército(ELITE2/2021/CINAMIL) e à Fundação para a Ciência e Tecnologia - projeto LAETA (UIDB/50022/2021). Ao Laboratório de Biomecânica de Lisboa.

References

[1] Collins, S., et al. Reducing the energy cost of human walking using an unpowered exoskeleton. *Nature*, 522(7555), 212–215, 2015.

[2] Pinheiro, P. et al. Analysis of the Performance of a Passive Ankle Exoskeleton for Reduction of the Metabolic Costs in Gait. *Congresso Nacional de Biomecânica, Covilhã*, 2019.

A robust version of the FGLS estimator for panel data

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

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

Keywords: FGLS, panel data, robust

Abstract: Panel or longitudinal data sets are frequent in financial and economic studies. This type of data aggregates cross-sectional with time series data, providing extra information and allowing to evaluate and measure statistical effects that would otherwise keep unknown. Different degree of restrictions upon the structure of the data leads to different approaches with least squares methodology. This results in estimators that can be highly affected by violation of those assumptions. The Feasible Generalized Least Squares estimator (FGLS) is an estimator that preserves good properties without requiring strong distribution requisites. In spite of this it is highly affected with the presence of observations to much different from all the rest. These are known as atypical observations or outliers. Economical and financial real data often present this type of data and the FGLS estimator may be seriously affected by those observations. This might be avoided if a robust option is chosen. Robustness is not yet spread among econometricians and panel data analysis follows that tendency. Recent studies in those fields point to the advantage of using robust estimators. With this work we want to contribute for the use of robust methodologies in the estimation of panel data models and present a robust version of FGLS, the RFGLS (Robust Feasible Generalized Least Squares). We compare the performance of the proposed estimator with the FGLS using a real data previously analysed by some authors.

Acknowledgements: Research partially supported by National Funds through **FCT**, —Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/0006/2021 (CEA/UL) and UIDB/MAT/04106/2021 (CIDMA).

References

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

[2] Bramati, M. C. Robust Estimators for the Fixed Effects Panel Data Model. *Econometrics Journal*, 10(3), 521-540, 2007.

[3] Baltagi, B. H. *Econometric Analysis of Panel Data*. New York: John Wiley, 2016.

Application of dimensionality reduction methods to high dimensional data as a tool for predicting the geographic origin of the saltwater clam *Rudipates philippinarum*

Clara Yokochi Sampaio[a], Fernando Ricardo[b], Ricardo Calado[b], Regina Bispo[c]
c.sampaio@campus.fct.unl.pt, fafr@ua.pt, rjcalado@ua.pt, r.bispo@fct.unl.pt

[a]*MSc in Mathematics and Applications, Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*
[b]*ECOMARE, CESAM - Centre for Environmental and Marine Studies, Department of Biology, University of Aveiro, Santiago University Campus, 3810-193 Aveiro, Portugal*
[c]*Center for Mathematics and Applications and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*

Keywords: elastic net, high-dimensional data, lasso, multinomial regression, ridge

Abstract: Food industry market globalization encouraged the importance in answering questions regarding food safety issues, triggering the growth of consumer awareness of food traceability. Hence, determining the geographic origin of seafood is crucial for controlling product quality and safeguarding consumer interest. Unfortunately, we are faced with an increase of fraud connected to forgery of product origin.

Rudipates philippinarum is a species of saltwater clam that is commercially harvested for consumption, being the second most important bivalve grown in aquaculture worldwide. This species' location of origin can be predicted by modeling features like their organic and chemical composition. The exploited dataset constitutes 30 clam samples, detailing information on 44 composition features, with the purpose of identifying which features distinguish between three geographic origins: Ria de Vigo, Ria de Aveiro, Estuário do Tejo, i.e, a classical Multinomial Logistic Regression problem. However, given the high-dimensionality of the dataset (number of variables higher than number of observations), the estimation of the model coefficients is compromised as Fisher's Information Matrix is no longer invertible. To overcome this problem, three dimensionality reduction methods were applied to model the origin of the clams: *Ridge*, LASSO and *Elastic Net*. Additionally, since datasets of only 30 samples challenge the process of model validation, the re-sampling method of *Monte Carlo Cross-Validation* was also implemented. We finalize comparing the results between the three methods, identifying which has the best predictive performance and comparing the estimation errors in each category of the response.

Acknowledgements: RB work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/00297/2020 (Center for Mathematics and Applications).

A Supervised Clustering Algorithm for Attributed Networks

Santos, B.[a,b], Campos ,P.[a,b,c]
up201903440@fep.up.pt, pcampos@fep.up.pt

[a] *Statistics Portugal*
[b] *Faculty of Economics, University of Porto*
[c] *Laboratory of Artificial Intelligence and Decision Support, LIAAD INESC TEC*

Keywords: attributed networks, subgroup discovery, supervised clustering

Abstract: A new method of supervised clustering with attributed networks is proposed, based on Single Representative Insertion/Deletion Hill Climbing with Restart (SRIDHCR) algorithm [1]. The goal is to obtain class-uniform clusters, while minimizing the number of clusters. This method deals with representative-based supervised clustering, where a set of initial representatives is randomly chosen. By assigning each observation to the closest representative, clusters are obtained. With the new methodology, the way nodes are associated to clusters does not only depend on their network distance, but also on the distances between their attributes. This can be accomplished through a combination of weights between the matrix of distances between nodes and their attributes, when defining the clusters. Hence, the method considers both structural and compositional characteristics of the network. As a benchmark, we use the Subgroup Discovery [2] on attributed network data. Subgroup Discovery focuses on detecting subgroups described by specific patterns that are interesting with respect to some target concept and a set of explaining features. Therefore, interesting patterns among subgroups can be revealed, for example, by inductive and exploratory data analysis tasks that find relations between a dependent and independent variables [2], considering the compositional aspect of the networks. For this work, SD-Map, a fast algorithm for exhaustive Subgroup Discovery [3], will be used to perform Subgroup Discovery on attributed networks. The proposed methodologies are applied to an inter-organizational network, denominated by EuroGroups Register, a central register that contains statistical information on companies from European countries, provided by Statistics Portugal.

References

[1] Atzmueller, M. Subgroup Discovery. *WIREs Data Mining Knowledge Discovery*, 5,3, 35–49, 2015. doi:49.doi:10.1002/widm.1144

[2] Atzmueller M., Puppe, F. SD Map A Fast Algorithm for Exhaustive Subgroup. *Proceedings of Fürnkranz J., Scheffer T., Spiliopoulou M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science*, Vol 4213, 2006. doi:10.1007/11871637_6

[3] Zeidat, N., Eick, C. K-medoid-style Clustering Algorithms for Supervised Summary Generation. *Proceedings of the International Conference on Artificial Intelligence*, 2004.

Approximations to the Binomial distribution, its bounds, relative and absolute precision

Jorge Santos[a], Marilia Pires[a], Russell Alpizar-Jara[a]
jmas@uevora.pt, marilia@uevora.pt, alpizar@uevora.pt

[a] *CIMA-IIFA/DMAT-ECT, University of Évora, Portugal*

Keywords: binomial, continuity correction, normal approximation

Abstract: Approximations to the Binomial by a Gaussian distribution are not always consensus and can be tedious, especially when we are dealing with the probability of a range of values of extreme values of the random variate. In the last century, some research led to the use of cumulative normal distribution tables, as they are easily available. This approximation is best for large samples and evenly symmetric situations. We discuss the 3 usual criteria for the applicability of this method. We show that not only sample size and symmetry are important, but also that the error rates are crucial. Small values of the success probability p become unacceptable when we try to calculate tail probabilities that sometimes have the same order of magnitude as the error. Besides, we show that the most restrictive criterion demands more than 50 trials with a probability of success belonging to the range [0.1;0.9], the criteria based on a variance greater than 5 lead to a parabolic shape criterion and the most liberal ones lead to the intersection of two hyperbolic regions. With the increase of computer capabilities this is not a practical problem, but the main idea is helping to provide some guidelines to this subject that is pervasive for recommendations to usual introductory probability and statistics courses.

Acknowledgements: This research has been partially supported by the Department of Mathematics, School of Sciences and Technology and by the Centro de Investigação em Matemática e Aplicações (CIMA), through the Project UIDB/04674/2020 of FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

[1] Afonso, A., Nunes, C. *Estatística e Probabilidades: Aplicações e soluções em SPSS*, Escolar Editora, 2011.

[2] Agresti, A. *An Introduction to Categorical Data Analysis*, John Wiley and Sons, 2002.

[3] Rosner, B.A. *Fundamentals of Biostatistics*, Duxbury Press,1995.

How to increase the visibility of a statistician in the modern world of collaborative research?

Nuno Sepúlveda[a,b]
nunosep@gmail.com

[a] *CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Lisbon, Portugal*
[b] *Institute for Medical Immunology, Charité-Universitatsmedizin Berlin, Berlin, Germany*

Keywords: collaboration management, communication, networking, statistical leadership

Abstract: Modern science is highly collaborative and its execution often requires the assembly of multidisciplinary research teams. In this context, the role of a statistician is invariantly perceived as essential but it usually comes short in terms of the appreciation recognized by the public in general and, in some extreme cases, by his or her own team members. A statistician needs to fight against the prevailing stigma as the "p-value" or the service provider with little or no influence on the decisions to be made along a given project. This paradox between essential contribution but reduced external and internal visibility raises the question on how to increase the value of a statistician in a multidisciplinary and collaborative environment. This question is particular important in nowadays world given the emergence of competing jobs with "sexy" titles, such as the infamous data scientists or bioinformaticians, whose technical and non-technical competences are more appealing than the ones of a traditional statistician in health and biomedical sciences. In this talk, I will leverage from my own experience as a past member of the European Network on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (EUROMENE) to discuss different aspects of statistical leadership. In particular, I will raise awareness of the need to develop soft skills, such as active listening, networking, and communication, in the curricula of modern statisticians.

Acknowledgements: Nuno Sepúlveda was partially funded by FCT (grant ref. UIDB/00006/2020) and he was a member of the managing committee of the European Network on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (EUROMENE) funded by the European Union via the COST Action number 15111.

Contribution to the diagnosis of skewness and kurtosis

Pedro Serrasqueiro[a,b], José Dias Curto[a,b]
pnsoa@iscte-iul.pt, dias.curto@iscte-iul.pt

[a] *ISCTE-IUL Instituto Universitário de Lisboa*
[b] *ISCTE-IUL Instituto Universitário de Lisboa*

Keywords: autocorrelation, generalized method of moments, heteroskedasticity, kurtosis, skewness

Abstract: This paper is motivated by the widespread interest in higher-order moments for financial risk management, namely skewness and kurtosis. We derive the asymptotic sampling distributions of skewness and kurtosis coefficients in the case of non-i.i.d. random variables applying the Generalized Method of Moments. We add to the existing literature by simulating a conditionally heteroskedastic process and compute heteroskedasticity-autocorrelation consistent (HAC) standard errors for hypothesis testing. The proposed skewness test significantly outperforms the traditional alternatives, while the kurtosis test corroborates known difficulties of accurately estimating kurtosis even for very large samples. In an illustrative example we analyze the daily returns of four major currency pairs in the period 2010-2020.

Time series periodicity detection using area biplots

Alberto Silva[a,b], Adelaide Freitas[a,b]
albertos@ua.pt, adelaide@ua.pt

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

Keywords: area biplots, nipals, periodicity detection, singular spectrum analysis

Abstract: Based on multivariate visualization techniques, an exploratory analysis method to estimate the dominant periodicities of a time series (*ts*) is proposed. The application of the NIPALS algorithm to the trajectory matrix of the Singular Spectral Analysis (SSA) is presented, resulting in i) a diagonal matrix containing the norms of the score vectors (singular values); ii) a matrix formed by the normalized score vectors (left single vectors); and iii) another one formed by the loadings vectors (right singular vectors). Pairs of singular values close to each other suggest the respective principal components (PCs) are associated with the periodicity of the *ts*. The proposed method consists of the construction of the biplot of these PCs, pinning a biplot vector of interest (i.e., some loading vector associated with a lagged vector), and the 90° rotation of the others. Depending on the percentage of variability explained by the PCs involved, the areas of the triangles formed by the origin of the factorial axes and the endpoints of the pinned vector and each of the rotated vectors will provide visual information regarding the magnitude of the autocorrelation between the corresponding lagged vectors. In addition, the periodicity will emerge from the appearance of groups of similar triangles, because of the strong autocorrelation between groups of lagged vectors. If the data are not well represented in the biplot, the periodogram should be used to confirm the analysis. In addition to the method, the R package *areabiplot* was developed, already published and available for use in CRAN.

Acknowledgements: This work was supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), reference UIDB/04106/2020.

References

[1] Gower, J.C., Groenen, P.J.F. Area Biplots. *J. Comput. Graphical Stat.*, 19, 46–61, 2010. doi:10.2307/25651299

[2] Golyandina, N., Nekrutkin, V., Zhigljavsky, A. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC, Boca Raton, 2001.

Variação espaço-temporal da distribuição e abundância da sardinha ao largo da costa continental portuguesa: efeitos ambientais

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

[a] *Centro de Matemática, Universidade do Minho*
[b] *Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera*

Keywords: efeitos ambientais, geoestatística, modelo com barreira, modelo de distribuição de espécies, *Sardina pilchardus*

Abstract: Nas últimas décadas, o desafio da sustentabilidade tem representado uma preocupação devido à depleção dos recursos naturais, à degradação ambiental e, à consequente perda da biodiversidade. Desta forma, melhorar o conhecimento da dinâmica das espécies representa um passo crucial para a sua conservação. No ambiente marinho, as espécies, principalmente as exploradas comercialmente, estão sujeitas a pressões ambientais e antropogénicas. O estudo da relação entre a distribuição espacial de espécies marinhas e as alterações ambientais permite um conhecimento dos processos de mudança nos indicadores de abundância, identificação de habitats potenciais e, na melhoria da capacidade de prever tendências nas dinâmicas dessas espécies. O presente estudo visa a estimação da distribuição espacial da sardinha (*Sardina pilchardus*), relacionando a variabilidade espaço-temporal da biomassa com as condições ambientais. A modelação em duas partes (modelo *hurdle*), adotada neste trabalho, permite incorporar as especificidades dos dados: dinâmicas espaço-temporais complexas, excesso de zeros e diferença entre processos de ocorrência e abundância sob ocorrência. Para além de incorporar variáveis ambientais e a estrutura espaço-temporal no modelo, avalia-se o impacto das condições ambientais com desfasamentos temporais em relação aos dados do indicador de biomassa. Opta-se por modelar separadamente a costa ocidental da costa sul da Península Ibérica, devido à geometria da costa e às diferentes condições oceanográficas. A utilização desta abordagem deverá permitir a compreensão da dinâmica espaço-temporal da sardinha na costa continental portuguesa e no Golfo de Cádiz, como também a relação do indicador com as condições ambientais, podendo contribuir para uma melhor gestão desta espécie.

Acknowledgements: Os autores agradecem à Fundação FCT (Fundação para a Ciência e Tecnologia) pelo financiamento através da Bolsa de Investigação Individual PD/BD/ 150535/2019, do Projeto de I&D PTDC/MAT-STA/28243/2017, dos Projetos UIDB/ 00013/2020 e UIDP/00013/2020; ao MAR2020 pelo financiamento através do projeto SARDINHA2020 (MAR-01.04.02-FEAMP-0009) e a todos os colegas envolvidos no presente trabalho.

The impact of wind speed on athletics triple-jump – an approach with the r largest order statistics

Domingos Silva[a,b], Frederico Caeiro[c,d], Manuela Oliveira[a,e]
domingosjlsilva@gmail.com, fac@unl.fc.pt, mmo@uevora.pt

[a] *Centro de Investigação em Matemática e Aplicações (Universidade de Évora)*
[b] *Escola Secundária de Barcelinhos*
[c] *Centro de Matemática e Aplicações*
[d] *Universidade Nova de Lisboa*
[e] *Universidade de Évora*

Keywords: athletics triple-jump, probability of exceedance, r largest order statistics method, return levels, right endpoint

Abstract: The r largest order statistics method is an important extension of the block maxima approach since it allows us to use more information from the data. However, the choice of r is not easy. If r is too large, bias can occur, and if r is too small, the variance of the estimator can be very high. So, we must deal with a bias-variance trade-off. In the present study, we use the largest order statistics of athletics triple-jump, men and women, between 1980-2020 and 1993-2020, respectively. In addition to the tail inference, we will analyse the wind effect on the athlete's performance.

Acknowledgements: This research was partially supported by the Fundação para a Ciência e a Tecnologia, Portugal (Portuguese Foundation for Science and Technology), through the projects UIDB/MAT/04674/2020 (CIMA, Centro de lnvestigação em Matemática e Aplicações, Universidade de Évora) (Research Centre for Mathematics and Applications, University of Évora) and UID/MAT/00297/2020 (Centro de Matemática e Aplicações, Universidade Nova de Lisboa) (Centre for Mathematics and Applications, Nova University of Lisbon)

References

[1] Silva, D.; Caeiro, F., Oliveira, M. Men's Performance in Triple Jump: an approach. *5th Stochastic Modelling Techniques and Data Analysis International Conference, 12-15 June 2018, Chania, Crete, Greece*, 127-135, 2018.

[2] Silva, D., Caeiro, F., Oliveira, M. Método das r-maiores observações anuais na estimação de quantis extremais no triplo-salto masculino. *Atas do XXIII Congresso da Sociedade Portuguesa de Estatística, 18-21 de outubro de 2017, Instituto Universitário de Lisboa (ISCTE-IUL)*, 59-73, 2020. http://www.spestatistica.pt/images/spe/Livro_de_Atas_Congresso_SPE_2017.pdf.

[3] Silva, D., Caeiro, F. Modelling the athletics long jump performance – an approach with r largest order statistics. Extreme Value Analysis 2021, The University of Edinburgh, 2021.

Mixed moment estimator for inference on space-time extremes

Jessica Silva Lomba[a], Maria Isabel Fraga Alves[a], Cláudia Neves[b]
jslomba@fc.ul.pt, mialves@fc.ul.pt, c.neves@reading.ac.uk

[a] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *Department of Mathematics and Statistics, University of Reading, UK*

Keywords: extreme rainfall, extreme value statistics, non-identical distributions, semi-parametric inference

Abstract: Extreme Value Theory provides the ideal framework for forecasting the frequency of extreme and hazardous events that are unlikely to occur and hard to predict. Within the aim of extreme value statistics lies the estimation of probabilities of extreme events rarely observed in the past, to which end the estimation of the extreme value index is key. Due to accelerating climate change, extreme meteorological phenomena such as heavy precipitation seem to be growing more severe and frequent, but estimation of this evolution remains subject to large uncertainty. Thus, inferential methods for the underlying non-stationary spatio-temporal processes are currently object of widespread interest.

A recent development is the concept of scedasis, through which a trend in extremes of space-time indexed observations can be captured and tactfully modeled within a semi-parametric framework. In this talk, we will look at how one can use series of data collected at several isolated locations to model extremes of the whole space-time process, enabling the mixed moment estimator of the extreme value index ([2]) to seamlessly incorporate space-time non-stationarity and dependence.

As a stepping-stone for the development of the estimator's asymptotic properties, and building on the foundational works by [1], we investigate a second order approximation to both tail empirical and tail quantile processes, making explicit a crucial term for assessing the dominant component of bias in estimation. Finally, application of the extended mixed moment estimator is illustrated with daily rainfall data from a homogeneous region in the UK.

Acknowledgements: J. Silva Lomba and M.I. Fraga Alves' research is supported by Fundação para a Ciência e a Tecnologia, I.P. through PhD grant SFRH/BD/130764/2017 and project UIDB/00006/2020. C. Neves gratefully acknowledges support from EPSRC-UKRI Innovation Fellowship grant EP/S001263/1.

References

[1] Einmahl, J. H. J., Ferreira, A., de Haan, L., Neves, C., Zhou, C. Spatial dependence and space-time trend in extreme events. *The Annals of Statistics*, to appear, 2021.

[2] Fraga Alves, M. I., Gomes, M. I., de Haan, L., Neves, C. Mixed moment estimator and location invariant alternatives. *Extremes*, 12, 149–185, 2009. doi: 10.1007/s10687-008-0073-3

Geostatistical sampling designs under preferential sampling for Black Scabbardfish

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

[a] *CMA, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa, Portugal*
[b] *CINAMIL, Instituto Universitário Militar, Portugal;*
[c] *CEAUL, Faculdade de Ciências da Universidade de Lisboa, Portugal;*
[d] *DM, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Portugal;*
[e] *Instituto Português do Mar e da Atmosfera (IPMA)*

Keywords: geostatistics, inla, preferential sampling, sampling design

Abstract: In Portugal, the spatial distribution and abundance of the black scabbardfish (BSF) is mostly unknown, the existing information relying on data from commercial fisheries. Available data refers to areas where fisherman expect to have higher catches of the species, which results that fishing locations are not selected randomly but preferentially. The BSF captures in Portuguese waters were previously modelled, taking the sampling preferentiability into account, using a Bayesian approach and INLA methodology, considering stochastic partial differential equations (SPDE) for geostatistical data jointly with a with a Log-Cox point process model. Based on this work, the aim of this study is to construct a new survey design to improve the BSF capture estimates and, to analyse the effect of preferential sampling on the choice of new sampling locations and its influence in the sampling design choise. Within this approach, different design classes are investigated, namely random, simple inhibitory and adaptive geostatistical sampling designs, regarding the problem of spatial prediction, in order to achieve the optimal BSF design towards the objective of the analysis.

Acknowledgements: This work was partially supported by Portuguese Foundation for Science and Technology through the project PREFERENTIAL, PTDC/MAT-STA/28243/2017 and UIDB/00297/2020 (Centro de Matemática e Aplicações).

References

[1] Chipeta, M. et al. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, 28.1,e2425, 2017. doi:10.1002/env.2425

[2] 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

[3] Krainski, E. T. et al. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press, 2019. doi:10.1201/9780429031892

Zero-distorted generalized geometric distribution with application to time series of counts

D. Sousa[a], E. Gonçalves[b]
diogo.sousa.1997@hotmail.com, esmerald@mat.uc.pt

[a] *Department of Mathematics, University of Coimbra, 3001-501 Coimbra Portugal*
[b] *CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra Portugal*

Keywords: asymptotic behaviour of estimators, generalized geometric distribution, ingarch time series, zero-distorted law

Abstract: The probability laws commonly used to describe counting data generally include zero as an element of its support. Nevertheless, situations often arise, in areas of statistical application, in which the number of zeroes expected according to a given counting model differs significantly from the number of zeros actually observed. In such cases, it becomes necessary to adapt the standard counting models, inflating or reducing the probability initially assigned to zero. In this sense, we consider a distribution that generalizes the geometric law ([1]) and that, through an additional parameter, allows changing the probability attributed to zero observation. After reviewing some probabilistic characteristics of this distribution, the asymptotic behaviour of the estimators obtained by the method of proportions, moments and maximum likelihood is established, and their performances are compared by means of numerical studies in medium and high sample sizes. The suitability of this law for real data sets is assessed, proving to be more effective than that obtained with other classical ones. The study continues with the introduction of a model for counting time series with conditional law to the past belonging to the family of these generalized geometric laws, and the stationarity at order one is established. The analysis of the number of new cases of Hantavirus infection per week in a Germany state, between 2005 and 2018, using zero-distorted integer-valued time series models concludes the study.

Acknowledgements: CMUC, UIDB/00324/2020, funded by the Portuguese Government through the FCT/MCTES.

References

[1] Sastry, D.V.S., Bhati,D., Rattihalli, R.N., Gómez-Déniz, E. On zero-distorted generalized geometric distribution. *Communications in Statistics – Theory and Methods*, 45,18, 5427–5442, 2016. doi:https://doi.org/10.1080/03610926.2014.942437

Análise de Dados Longitudinais Multivariados

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

[a] *Departamento de Matemática da Universidade do Minho*
[b] *CMAT-UM*
[c] *CBMA-UM*
[c] *CEAUL*

Keywords: longitudinal, modelação, multivariado

Abstract: Em estudos observacionais longitudinais é usual fazerem-se medições repetidas de mais do que um processo que evolui ao longo do tempo, sendo o objetivo principal entender a associação dos vários processos e se existe algum que seja mais relevante. Por exemplo, num processo de psicoterapia semanal várias avaliações são feitas aos clientes, como a sintomatologia, empatia do cliente com psicoterapeuta ou quantidade de momentos de inovação, entre outras. Nestes estudos é de particular interesse entender de que forma a evolução da sintomatologia e dos momentos de inovação se associam e qual deles é o maior impulsionador para a mudança. Nesta apresentação iremos propor um modelo multivariado para duas variáveis longitudinais no contexto de modelos lineares mistos, explorando diferentes estruturas de correlação possíveis, para responder a estas questões. Serão discutidas também diferentes distribuições para as variáveis longitudinais.

References

[1] Wang, Z., Cheng, Y., Seaberg, E.C., Rubin, L.H, Levine, A.J., Becker J.T. Longitudinal multivariate normative comparisons. *Statistics in Medicine*, 40(6), 1440–1452, 2020. doi:10.1002/sim.8850

Estimação de modelos de regressão linear multivariada para dados censurados

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

[a] *Universidade de Aveiro*
[b] *CIDMA*
[c] *Universidade do Porto*

Keywords: algoritmo EM, algoritmo Gibbs, ampliação de dados, dados censurados, regressão linear multivariada

Abstract: Muitas vezes, em problemas reais envolvendo modelos de regressão linear (RL), pode-se obter melhor descrição da realidade considerando-se várias variáveis respostas em simultâneo, onde se analisam $m \geq 2$ caraterísticas do mesmo fenómeno. Nestes casos, deve-se considerar o modelo de RL multivariada, em que cada indivíduo $\mathbf{Y}_i^*, \;\; i = 1, \ldots, n$ é descrito pela equação

$$\mathbf{Y}_i^* = \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

em que $\mathbf{Y}_i^* = (y_{i1}, \ldots, y_{im})$ é um vetor $1 \times m$ correspondente à $i-$ésima observação, $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})$ é um vetor $1 \times (k+1)$ dos preditores, $\boldsymbol{\beta}$ é uma matriz $(k+1) \times m$ dos coefficientes e o termo dos erros $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})$ é independente e idnticamente distribuído e tem as mesmas dimensões de $\mathbf{Y}_i^*$.

Devido a eventuais limitações dos aparelhos de medição ou da forma como se planeiam as experiências, os dados observados podem estar censurados, pelo que a variável realmente observada é $\mathbf{Y}_i$. Assim, $Y_{ij} = max\{Y_{ij}^*, L_j\}, \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, m$ (no caso da censura à esquerda). Geralmente, a análise estatística deste tipo de dados baseia-se no algoritmo EM (*Expectation-Maximization*) ou na ampliação de dados, visto que resultam numa reconstrução de dados permitindo aplicação de técnicas desenvolvidas para dados não censurados.

Atendendo à pertinência do problema e à escassez de litertura no tema, neste trabalho analisamos, através de um estudo de simulação, o desempenho de três métodos de estimação de modelos de RL multivariada para dados censurados: o algoritmo EM, a ampliação de dados e o algoritmo Gibbs com ampliação de dados. No estudo, foram consideradas diferentes percentagens de censura e dimensões de amostras, e os resultados sugerem que os três métodos resultam em estimativas consistentes.

Acknowledgements: Este trabalho foi realizado com o suporte financeiro da Fundação Calouste Gulbenkian e da Fundação para a Ciência e a Tecnologia (FCT) através do projecto UIDB/04106/2020

The extended Chen-Poisson marginal rate model for recurrent gap time data

Ivo Sousa-Ferreira[a], Cristina Rocha[a], Ana Maria Abreu[b]
ivo.ferreira@staff.uma.pt, cmrocha@fc.ul.pt, abreu@staff.uma.pt

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

Keywords: extended Chen-Poisson distribution, gap times, non-homogeneous Poisson process, parametric model, recurrent events

Abstract: In recent years, there has been a renewed interest in developing new gap time models to better describe real-life phenomena. A large part of the statistical methods for the analysis of recurrent events is based on Poisson processes. In this context, Zhao and Zhou [2] proposed a semiparametric model based on a marginal rate function that is derived from a non-homogeneous Poisson process (NHPP). Nonetheless, when the research interest focus on how the recurrence rate evolves over time, a suitable parametric model could be more advantageous.
Therefore, in this work we present a detailed study on the properties of a fully-parametric model derived from a NHPP, named extended Chen-Poisson (ECP) marginal rate model. The model is characterized by a marginal rate function based on the ECP distribution that was recently proposed by Sousa-Ferreira *et al.* [1]. Under this approach, the conditional distribution of the gap times is deduced. Moreover, it is shown that this model has the Chen marginal rate model as a limiting case. The maximum likelihood (ML) method is applied for parameters estimation in the presence of right-censoring. A simulation study is conducted to evaluate the properties of the ML estimators in several scenarios with different sample sizes, number of recurrences, percentages of censoring and shapes of the marginal rate function. An application to the bowel motility data is considered to illustrate the potential of the ECP marginal rate model in comparison with other competing models.

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). I. Sousa-Ferreira also acknowledges FCT for the PhD grant DFA/BD/6459/2020.

References

[1] Sousa-Ferreira, I., Abreu, A.M., Rocha, C. *A new lifetime distribution and its applications in recurrent events analysis*. Notas e Comunicações CEAUL 02/20, 2020. ISBN: 978-989-733-064-3.

[2] Zhao, X., Zhou, X. Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data Analysis*, 56(2), 370–383, 2012. doi:10.1016/j.csda.2011.07.015

Parametric landmark estimation of the transition probabilities in survival data with multiple events

Gustavo Soutinho[a], Luís Meira-Machado[b], Pedro Oliveira[a]
gdsoutinho@gmail.com, lmachado@math.uminho.pt, pnoliveira@icbas.up.pt

[a] *EPIUnit, Instituto de Saúde Pública da Universidade do Porto (ISPUP)*
[b] *Departamento de Matemática da Universidade do Minho*

Keywords: generalized gamma distribution, landmark approach, multi-state models, transition probabilities

Abstract: The estimation of transition probabilities is of major importance in the analysis of survival data with multiple events. These quantities play an important role in the inference in multi-state modeling providing in a simple and summarized manner long-term predictions of the process. Recently, de Uña-Álvarez and Meira-Machado (2015) proposed nonparametric estimators based on subsampling which have already proved to be more efficient than other nonparametric estimators in case of strong violation of the Markov condition. However, since this approach uses a specific portions of data when the subsample sizes are reduced or in the presence of heavily censored data this may lead to higher variability of the estimates. To avoid this problem, we introduce parametric estimators for the transition probabilities that are also based on subsampling. We have considered several flexible distributions to handle this issue appropriately. One of the proposed approaches, which provides good results, with high flexibility, is based on the generalized gamma distribution. Results of simulation studies confirm the good behavior of the proposed methods. We also illustrate and compare the new methods to the nonparametric landmark estimator through a real data set on colon cancer.

Acknowledgements: This research was financed by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the research grants PTDC/MAT-STA/28248/2017 and PD/BD/142887/2018.

References

[1] de Uña-Álvarez, J., Meira-Machado, L. Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study. *Biometrics*, 71, 364–375, 2015.

[2] Meira-Machado, L., Sestelo, M. Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal*, 61, 245–263, 2019. doi: 10.1002/bimj.201500038

A robust hurdle Poisson model in the estimation of the extremal index

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

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

Keywords: blocks estimator, extremal index, hurdle model, robustness

Abstract: In statistical extreme value theory, the occurence of clusters of exceedances of high thresholds is related to the extremal index (*EI*), when that parameter exists. In such cases, the *EI* represents the reciprocal of the mean cluster dimension in the limit distribution. The most known *EI* estimator is the blocks estimator, which is not robust. But previous studies show its good performance in m-dependent structures with low values of *EI*, which are very interesting since the *EI* takes values between zero and the unity, the last case corresponding to independence. We consider the estimation of the mean cluster size, modelling the clusters dimension with a hurdle Poisson regression model. A simulation study explores and compares different robust proposals. We propose the robust hurdle estimation based on the best estimates results.

Acknowledgements: Research partially supported by National Funds through **FCT**, —Fundação para a Ciência e a Tecnologia, projects UIDB/MAT/0006/2021 (CEA/UL) and UIDB/MAT/04106/2021 (CIDMA).

References

[1] Cantoni, E., Zedini, A. A robust version of the hurdle model. *J. Statist. Plann. and Infer.*, 141, 121–1223, 2011. doi:10.1016/j.jspi.2010.09.022

[2] Gomes, I., Miranda, C., Souto de Miranda, M. A Note on Robust Estimation of the Extremal Index. In: La Rocca M., Liseo B., Salmaso L. (eds.) *Nonparametric Statistics:* ISNPS 2018. Springer Proceedings in Mathematics and Statistics, vol. 339, 213–225, 2020. doi:10.1007/978-3-030-57306-5_20

[3] Robert, C. Y. Inference for the limiting cluster size distribution of extreme values. *The Ann. of Statist.*, 37, 271–310, 2009. doi:10.1214/07-AOS551

Dimension Reduction with Histogram Principal Component Analysis for the Detection of Internet Attacks

Ana Subtil[a], M. Rosário Oliveira[a], Lina Oliveira[b]
anasubtil@tecnico.ulisboa.pt, rosario.oliveira@tecnico.ulisboa.pt,
lina.oliveira@tecnico.ulisboa.pt

[a] *CEMAT e Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[b] *CAMGSD e Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Keywords: anomaly detection, histogram principal component analysis, internet data, symbolic data analysis

Abstract: Internet communications are vulnerable to attackers that unlawfully intercept traffic between two parties, to observe, modify, or disrupt the communication. We use unsupervised anomaly detection methods, based on histogram principal components, to detect this type of attacks. The methods draw on round-trip-times (RTT) measured between a set of monitoring devices (probes) and a target computer, the potential victim of attacks.
The data was gathered on an infrastructure comprising 12 worldwide distributed probes, periodically measuring the RTT to each monitored target. Whenever an attack was being perpetrated, the traffic was routed from each of the probes to the attacker and then towards the target, presumably leading to anomalous RTT which the anomaly detection methods should signal.
At 120-second intervals, each probe sent 10 packets to the monitored target, obtaining 10 RTT measurements, from which minimum, median, and maximum were computed. Thus, for each target, we consider a symbolic data set with 12 histogram-valued variables and propose anomaly detection methods based on histogram principal component analysis. From the histogram-valued input data set, we compute symbolic covariance matrices and determine their eigenvalues and eigenvectors. Then, we use a Moore's based histogram algebraic structure to obtain linear combinations of the initial histogram-valued data, with loadings determined by the calculated eigenvectors. The resulting weighted sums are the histogram-valued PC scores. We use the projected data on the first principal component as the basis for anomaly detection methods, applying conventional methods that draw on the symbolic means of the histogram-scores.

Acknowledgements: Work supported by Fundação para a Ciência e Tecnologia, Portugal, through the projects UIDB/04621/2020, PTDC/EEI-TEL/32454/2017, and UID/MAT/04459/2020.

RM-SMOTE: A new robust balancing method

Rasool Taban[a], Cláudia Nunes[a], M. Rosário Oliveira[a]
rasooltaban@tecnico.ulisboa.pt, claudia.nunes@math.tecnico.ulisboa.pt,
rosario.oliveira@tecnico.ulisboa.pt

[a] *CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa*

Keywords: binary classification, imbalanced learning, outlier detection, robust Mahalanobis distance, SMOTE

Abstract: Imbalanced data is a common characteristic in vast real-world applications, such as health cancer diagnose, fraud detection, etc. In problems involving classification, regression, or anomaly detection learning patterns from classes where a small number of observations are available may bias the results, and hence the learned model become useless, in practice.
Balancing techniques are a common strategy to address these problems. As the name suggests, these techniques try to overcome the problem by balancing the number of observations in each class.
In addition to imbalance problems, datasets coming from real problems may have atypical observations or outliers. As a result, learning from these datasets become even more challenging, since atypical data may gain unpleasant importance after balancing. This effect can be more severe when one uses classical balancing techniques, leading to biased and poor results.
In order to overcome these issues, we propose a robust approach to imbalanced learning - which we call RM-SMOTE - that combines the idea of SMOTE with robust Mahalanobis distance. We propose to automatically down weight atypical Minority class observations in such a way that potentially outliers from this class have a low chance to be selected in the resampling step.
The mean performance of this method is evaluated using simulated and real imbalanced datasets with different levels of contamination. The results indicate the superiority of RM-SMOTE in terms of performance metrics when handling divergent proportion of outliers while balancing the dataset.

Acknowledgements: This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement Spin- 812912, and by FCT, Portugal, through projects PTDC/EGE-ECO/30535/2017 and UIDB/04621/2020.

Statistical challenges in mutational signature analyses

Víctor Velasco-Pardo[a], Michail Papathomas[a], Andy G. Lynch[a,b]
vvp1@st-andrews.ac.uk,
m.papathomas@st-andrews.ac.uk, andy.lynch@st-andrews.ac.uk

[a] *School of Mathematics and Statistics, University of St Andrews, U.K.*
[b] *School of Medicine, University of St Andrews, U.K.*

Keywords: biostatistics, cancer, genomics

Abstract: Cancer is a disease driven and characterised by mutations in the DNA. Different categorisations of DNA mutations have allowed the identification of patterns that can act as signatures for the processes that have governed the life of the cancer. Over the last decade, research groups have identified more than 100 such signatures [1, 2].

These analyses are improving our understanding of cancer aetiology and have the potential to play a role in diagnosis, prognosis and treatment choice. Consisting of the estimation of probability mass functions or weights determining non-negative weighted combinations, mutational signature analyses are perhaps unique amongst comparable analyses in the medical literature, in that no confidence intervals or other representations of uncertainty are demanded when reporting the results.

In this talk we review the key statistical challenges for the field, assess the potential of existing approaches to adapt to those challenges, and comment on what we think are promising directions. We evaluate how to perform simultaneous inference on model dimension and parameters. As data arrives over time, we assess whether it is possible to update parameters sequentially in a statistically sound manner. As practitioners can trade off a large number of participants in a study for measurement accuracy in individual tumour samples, a fully probabilistic approach to modelling will help to determine sample sizes. Lastly, as we deal with data that is highly heterogeneous, we will argue that all sources of heterogeneity should be modelled explicitly.

Acknowledgements: We thank The Melville Trust for the Care and Cure of Cancer for financial support.

References

[1] Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149, 979–993, 2012. doi:10.1016/j.cell.2012.04.024

[2] Alexandrov, L.B., Kim, J., Haradhvala, N.J. *et al.* The repertoire of mutational signatures in human cancer. *Nature*, 578, 94–101, 2020. doi:10.1038/s41586-020-1943-3

Testing for additivity in nonparametric heteroscedastic regression models

Adriano Zanin Zambom[a], Jongwook Kim[b]
adriano.zambom@csun.edu, jki5@iu.edu

[a] *California State University Northridge*
[b] *Indiana University Bloomington*

Keywords: analysis of variance, heteroscedasticity, local alternatives, nearest neighbor, principal components

Abstract: This talk introduces a novel hypothesis test for additivity in nonparametric regression models. Inspired by recent advances in the asymptotic theory of analysis of variance when the number of factor levels is large, we develop a test statistic that checks for possible nonlinear relations between the available predictors and the residuals from fitting the additive model. The asymptotic distribution of the test statistic is established under the null and local alternative hypotheses, demonstrating that it can detect alternatives at the rate of $n^{1/4}$. An advantage over some methods in the literature is that the proposed method maintains its level close to nominal under heteroscedasticity and can be applied to both fixed and random designs. Extensive simulations suggest that the proposed test outperforms competitors for small sample sizes, especially for fixed designs, and performs competitively for larger sample sizes. The proposed method is illustrated with a real dataset.

Posters

XXV Congresso
Sociedade Portuguesa
de Estatística
2021 Évora

Can researchers ignore a complex sampling design on the estimation of Mokken scales?

Marcia Andrade[a]
armsandrade@gmail.com

[a] *Universidade Cândido Mendes*

Keywords: complex non-parametric Mokken scale analysis, complex sample design, complex survey, Mokken's scalability coefficients, nonparametric item response theory

Abstract: Up till now, the construction of Mokken scales [1], [4] has still assumed that complex samples are treated as SRS samples, but there are very few exceptions [2], [3]. This means that the Popularities (Pi), the Scalability Coefficients Hij, Hi and H, and their standard errors are estimated, for instance, ignoring a stratified multi-stage cluster sampling design. Consequently, this decision can lead to erroneous conclusions. From this purpose, the current study briefly illustrates how to use complex probabilistic sampling designs in the Mokken scaling analysis, including the statistical software Complex Mokken, the 2001 SAEB complex survey data (Brazilian Basic Education Assessment System), and an economic capital measure from the fifth grade students enrolled in urban primary schools. The results suggest that the combined effects of stratification, clustering and sampling weights via SAEB design should not ignored in the estimation. In summary, this study strongly recommends that researchers should be aware when using complex sample data for the construction of Mokken scales [5].

Acknowledgements: I am grateful for the National Institute of Educational Studies and Research Anísio Teixeira (INEP), CAPES, and CNPq.

References

[1] Mokken, RJ. *A theory and procedure of scale analysis*. Mouton, The Hague, 1971.

[2] Andrade, MS. *Uma nova abordagem para a estimação dos coeficientes de escalonabilidade associados à teoria de resposta ao item não paramétrica*. Tese (Doutorado), Pontifícia Universidade Católica do Rio de Janeiro, 2012.

[3] Andrade, M. *Mokken scale & complex sampling designs: insights*. Autografia Editora, Rio de Janeiro, 2016.

[4] Sá, SPC., Santos, DM., Andrade, MS et al. A capacidade de autocuidado dos idosos usuários do Programa de Enfermagem Gerontogeriátrica da Universidade Federal Fluminense. XVIII Congress of the Portuguese Statistical Society.

[5] Myszkowski, N. A Mokken scale analysis of the Last Series of the Standard Progressive Matrices (SPM-LS). *Journal of Intelligence*, 8, 22, 2020.

vsd – Visualizing survival data

Pedro Daniel Camacho[a], Ana Maria Abreu[a,b]
2016711@student.uma.pt, abreu@staff.uma.pt

[a] *Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal*
[b] *Centro de Investigação em Matemática e Aplicações, Portugal*

Keywords: ggplot2, R packages, survival analysis

Abstract: We present a newly developed R package, vsd [1], meant to visualize data in survival analysis.
This package, taking use of other R packages, such as the packages ggplot2, survminer and flexsurv, allows the user to visualize several plots for right-censored data, as per example: the Kaplan-Meier estimate of the survival function, the smoothed estimate of the hazard function, the residuals and forest plots associated to the coefficients of a Cox model, as well as the survival function estimates for parametric models.

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

References

[1] Camacho, D., Abreu, A. *vsd: Graphical Shim for Visual Survival Data Analysis.* R package version 0.1.0., 2021. https://CRAN.R-project.org/package=vsd

Análise de sobrevivência de pacientes com insuficiência cardíaca

Inês Carvalho[a], Margarida Oliveira[a], Luís Machado[a,b]
pg42544@alunos.uminho.pt, pg42546@alunos.uminho.pt,
lmachado@math.uminho.pt

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

Keywords: análise de sobrevivência, estimador de Kaplan-Meier, insuficiência cardíaca, modelo de regressão de Cox

Abstract: A doença cardiovascular é uma das principais causas de morte em ambos os géneros no mundo. Para compreender melhor os efeitos de diferentes fatores de prognóstico realizou-se um estudo de sobrevivência que retrata o tempo desde o diagnóstico de insuficiência cardíaca até ao término do estudo ou à sua morte [1]. Numa fase inicial, obtivemos através do estimador de Kaplan-Meier, as estimativas de sobrevivência. De forma a comparar a sobrevivência em determinados grupos recorreu-se a alguns métodos não paramétricos, nomeadamente, o teste de log-rank. Para estudar o efeito de um conjunto de covariáveis no tempo de sobrevivência foram ajustados vários modelos de regressão de Cox [2]. Como algumas das variáveis quantitativas em estudo revelaram um efeito não linear, procedeu-se ao ajustamento das mesmas através de splines penalizadas [3]. Utilizou-se o método de seleção regressiva para selecionar as variáveis a reter no modelo final. As variáveis que se mostraram significativas foram a idade, os níveis da enzima CPK, os níveis de creatinina sérica, a anemia, a hipertensão e a fração de ejeção, indo de acordo aos resultados apresentados em [1]. Com a finalidade de verificar o pressuposto dos riscos proporcionais recorreu-se aos resíduos de Schoenfeld e a um teste de hipótese [4], a partir dos quais o pressuposto foi validado. Por último, através da análise dos resíduos de Cox-Snell e dos desvios residuais constatou-se o bom ajustamento do modelo.

References

[1] Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. Survival analysis of heart failure patients: A case study. *PLoS ONE*, 12 (7), 2017.

[2] Cox, D.R. *Regression models and life tables (with discussion).* Journal of the Royal Statistical Society, Series B, 34, 187-200, 1972.

[3] Eilers, P.H.C., Marx, B.D. *Flexible smoothing with B-splines and penalties.* Statistical Science, 11, 89-121, 1996.

[4] Grambsch, P., Therneau, T. *Proportional hazards tests and diagnostics based on weighted residuals.* Biometrika, 81, 515-526, 1994.

A model to predict the compressive strength of binders' constructed using fly ash and glass waste

Elisete Correia[a], Nuno Cristelo[b], Adelaide Cerveira[c]
ecorreia@utad.pt, ncristel@utad.pt, cerveira@utad.pt

[a] *CEMAT, Department of Mathematics, University of Trás-os-Montes e Alto Douro, Vila Real, Portugal (Orcid:0000-0002-1121-2792)*
[b] *Centro de Química - Vila Real, Department of Engineering, University of Trás-os-Montes e Alto Douro, Vila Real, Portugal (Orcid:0000-0002-3600-1094)*
[c] *INESC-TEC, UTAD's Pole, Department of Mathematics, University of Trás-os-Montes e Alto Douro, Vila Real, Portugal (Orcid:0000-0002-7494-6566)*

Keywords: activator-precursor ratio, dummy variables, glass waste, multiple linear regression, uniaxial compression strength

Abstract: The use of industrial waste in the production of new types of cement replacement binders, especially through the alkaline activation technique, has been gathering increasing attention, [1, 2]. In the present work, different mixtures were prepared with glass waste and fly ash considering different precursor activator solutions. The aim is to obtain a model to predict the compressive strength as a function of different ratios between glass waste and fly ash. Five different types of the mixture were prepared with different glass waste/fly ash ratios within three different levels of precursor activator solutions. The five mixtures correspond to the following glass waste/fly ash ratios: 0/100, 25/75, 50/50, 75/25 and 100/0. The levels of precursor activator solutions are denoted by L, M and H, where level L is the one with the lowest quantity of cleaning solution while level H is the one with the highest quantity of cleaning solution. All the specimens were submitted to a uniaxial compression strength (UCS) test.

For each level of precursor activator, it is performed a multiple linear regression with dummy variables to determine the influence of the ratio between glass waste and fly ash on the compressive strength [3].

References

[1] Cerveira, A., Correia, E., Cristelo, N., Miranda, T., Castro, F., Fernandez-Jimenez, A. Statistical Analysis of the Influence of Several Factors on Compressive Strength of Alkali Activated Fly Ash. *Procedia Structural Integrity*, Vol 5, 1116–1122, 2017.doi:10.1016/j.prostr.2017.07.099.

[2] Correia, E., Cerveira, A., Fernandez-Jimenez, A., Coelho, J., Miranda, T., Castro, F., Cristelo, N. Statistical study of curing conditions in alkali activation of Portuguese mine tailings. *Environmental Geotechnics*, 1–12, 2019. doi:10.1680/jenge.18.00013.

[3] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. *Multivariate Data Analysis*. Seventh. ed. Pearson, 2014.

Regressão Quantílica Lomax Exponencializada

Guilherme Da Silva Machado[1,a], Cleber Bisognin[2,a], Vanessa Siqueira Peres Da Silva[3,a]
[1]guilhermesv2015@gmail.com, [2]cleber.bisognin@ufsm.br, [3]vanessa@ufsm.br

[a]*Departamento de Estatística - UFSM*

Keywords: EMV, regressão quantílica, simulações de Monte Carlo

Abstract: Na literatura tem surgido cada vez mais modelos de regressão, com o objetivo de melhorar o ajuste de dados reais, utilizando distribuições de probabilidades não gaussianas. Em algumas situações é necessário relacionar uma variável com distribuição Lomax Exponencializada com séries de covariáveis. Devido a essa necessidade o objetivo deste trabalho é propor um novo modelo de regressão quantílica utilizando a distribuição Lomax Exponencializada, proposta por [1]. Seja $y \sim \text{LE}(\alpha, \delta, \lambda)$. Inicialmente, reparametrizamos a distribuição em termos dos quantis (μ) utilizando a seguinte relação $\lambda = [(1-\tau^{\frac{1}{\alpha}})^{-\frac{1}{\delta}} - 1]/\mu$. Assim, cada y_t possui distribuição LE reparametrizada em termos do quantil pode ser escrita por meio de uma estrutura de regressão, dada por $g(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$, para $t = 1, \ldots, n$, $\boldsymbol{x}_t^\top = (1, x_{t1}, \ldots, x_{tp})$, $p \in \mathbb{N}$, um vetor com as variáveis explicativas, com vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \delta, \boldsymbol{\beta}^\top)^\top$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)^\top \in \mathbb{R}^k$ e $k \in \mathbb{N}$ e $g(\cdot)$ é uma função de ligação monótona e duas vezes diferenciável. A estimação dos parâmetros foi realizada via Estimador da Máxima Verossimilhança (EMV). Foram realizadas simulações de Monte Carlo para avaliar o viés, o desvio padrão, o Erro Quadrático Médio (EQM), o coeficiente de curtose e o coeficiente de assimetria. Durante as simulações, para cada cenário, foram analisadas 10.000 replicações com tamanho amostral $n \in \{50, 100, 500\}$, $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ e valores dos parâmetros $\alpha = 0.5$, $\delta = 5.0$ e $\boldsymbol{\beta} = (0.2, 0.5, 0.3)^\top$. As simulações de Monte Carlo foram analisadas e verificou-se que o viés dos parâmetros $\boldsymbol{\beta}$ são menores em comparação ao viés dos parâmetros α e δ. Dentre todos os parâmetros o δ foi o que apresentou maior viés. O EMV de α apresentou menor EQM, seguido do vetor $\boldsymbol{\beta}$ e do parâmetro δ, respectivamente, em ambos os cenários. Analisando o EMV de todos os parâmetros pode-se observar que quando o tamanho amostral aumenta, o víes e o EQM diminuem. O estimador EMV de $\boldsymbol{\beta}$ apresentou menor desvio padrão, seguido pelo desvio padrão de α e de δ, respectivamente. O desvio padrão diminui conforme o tamanho amostral aumenta. Esta análise indica que o EMV é um estimador consistente para os parâmetros do modelo proposto. Analisando os valores dos coeficientes de curtose e assimetria, para o vetor de parâmetro $\boldsymbol{\beta}$, estes estão próximos de zero e três, respectivamente, indicando a normalidade assintótica do EMV.

References

[1] Abdul-Moniem, I.B., Abdel-Mameed, H.F. On expoenentiated lomax distribution. *International Journal of Mathematical Archive*, 3, 1 – 7, 2012.

Assessing comorbid chronic diseases as predictors of COVID-19 severity: a meta-analysis

Marco Encarnação[a], Isabel Natário[b], **Regina Bispo**[b]
ma.encarnacao@campus.fct.unl.pt, icn@fct.unl.pt, r.bispo@fct.unl.pt

[a]*MSc in Mathematics and Applications, Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*
[b]*Center for Mathematics and Applications and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*

Keywords: comorbidities, COVID-19, disease severity, meta-analysis

Abstract: COVID-19 is a disease caused by the virus SARS-COV-2. As a new disease, its poor understanding can result in delayed identification and treatment. Hence, is of crucial importance to understand the factors that may influence the disease severity. In this study, we run a meta-analysis on published articles (www.covidanalytics.io/dataset), to evaluate comorbid chronic diseases as predictors for COVID-19 severity. Specifically, pre-existing chronic conditions, such as hypertension, diabetes and cardiovascular diseases, were analyzed as risk factors for developing severe COVID-19.
Meta analysis with fixed-effects or random effects models, used as appropriate, were considered with estimation of odds ratio (OR) of severe COVID-19 for the aforementioned comorbidities and corresponding 95% confidence intervals (95% CI). The number of studies considered ranged from 14 to 18, depending on the comorbidity considered, with numbers of COVID-19 patients with severe disease ranging from 8 (15%) to 181 (39%), and resulting in a overall sample with a total of patients with severe COVID-19 ranging from 1071 (27%) to 1182 (27%).
The presence of comorbidities such as hypertension (*random effects model*: OR 2.72; 95% CI $[2.06, 3.60]$), diabetes (*fixed effects model*: OR 2.51; 95% CI $[2.03, 3.10]$) and cardiovascular diseases (*fixed effects model*: OR 2.73; 95% CI $[2.00, 3.72]$) were associated with significantly higher risk of developing severe COVID-19 ($p < 0.0001$).
The results also show that those who had a severe disease manifestation, compared with those who had not, differed significantly on symptoms such as the presence of fever (*random effects model*: OR 2.00; 95% CI $[1.38, 2.91]$; $p = 0.0002$) and cough (*random effects model*: OR 1.11; 95% CI $[1.03, 1.20]$; $p = 0.0081$).
This analysis reinforces that patients with such today life's comorbidities, in case of catching COVID-19, should be early and carefully monitored in order to prevent severe disease.

Acknowledgements: IN and RB work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/00297/2020 (Center for Mathematics and Applications).

Modeling residential adoption of solar PV systems

Carolina Goldstein[a], José Miguel Espinosa[b], Regina Bispo[c]
c.goldstein@campus.fct.unl.pt, miguel.espinosa@galp.com, r.bispo@fct.unl.pt

[a] *MSc in Analytics and Big Data Engineering, Department of Computer Science and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*

[b] *Tech Garage, IT & Digital Department. Galp Energia, SGPS, S.A., 1600-209 Lisboa, Portugal*

[c] *Center for Mathematics and Applications and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*

Keywords: PV system adoption, social effects, spatial modeling, technology diffusion

Abstract: The world is on the verge of a necessary energy transition. Solar energy, obtained through PhotoVoltaic (PV) systems, is the most viable green energy source to be produced at the domestic level. This allows every individual to take matters into their own hands and make a valuable contribution towards this common goal.
Understanding the factors that influence the adoption of domestic solar energy, how it changes throughout the different regions of Portugal and how spatial dependent factors contribute to the promotion of this technology is of the utmost importance to stimulate adoption. As to this day, to the best of my knowledge, these are not yet known. This study will be a key contribution to enable adherence efforts to be channelled to where adoption is more likely, ultimately accelerating Portugal's energy transition.
Hence, the goal of this study is to build a spatial model that estimates for each region in Portugal the probability of individuals adopting domestic PV systems. To better visualize and understand the underlying spatial distribution, a map will be produced where the variation of this probability is reflected.
To fulfil this goal, the study uses a sample from around 440 domestic PV systems, installed between June and November 2020, spread over 278 municipalities in continental Portugal as well as environmental, socioeconomic and demographic data from public sources (*Instituto Nacional de Estatística, IP - Portugal*).
To this end, different spatial regression models will be considered and their results compared.

Hierarchical clustering algorithms in classification the spread of COVID-19 cases and deaths in European countries

Dulce Gomes[a], Thaís Zamboni Berra[b], Antônio Ramos[b], Ricardo Alexandre Arcêncio[b]
dmog@evora.pt, thaiszamboni@live.com, antonio.vieiraramos@outlook.com, ricardo@eerp.usp.br

[a] *Center for Research in Mathematics and Applications (CIMA), IIFA, University of Évora, Portugal*
[b] *Department of Maternal-Infant and Public Health Nursing, University of São Paulo at Ribeirão Preto College of Nursing, Ribeirão Preto, São Paulo, Brazil*

Keywords: ARIMA, COVID-19, dynamic time warping, hierarchical clustering, time series

Abstract: The coronavirus disease outbreak, which is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), seems to have started in the Chinese city of Wuhan in December 2019 and, in March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic. Since then, the outbreak of the COVID-19 undoubtedly poses the biggest public health challenge since the Spanish flu in 1918 and 1919[2].
Time series hierarchical clustering algorithms will be applied in order to classify European countries where the 14-day notification rates of newly reported COVID-19 cases and deaths per 100 000 population by week have similar patterns of behaviour. For measuring similarity between time series, an accurate and computationally efficient distance measures, such as a *shape Dynamic Time Warping*, will be performed. After an appropriate characterization of time series data, where countries with similar/different trends, seasonality and hidden patterns were identified, forecasting ARIMA models will be applied to forecast a 14-day notification rates of cases and deaths.
The presented methodologies are applied to confirmed cases of Coronavirus from week 13 of 2020 to the latest most updated date. The data were gathered from the European Centre for Disease Prevention and Control (ECDC). Weekly updates on the number of cases and deaths reported worldwide and aggregated by week are published every Wednesday[1].

Acknowledgements: This communication was partially financed by the Center for Research in Mathematics and Applications (CIMA), through the Project UIDB/04674/2020 of the FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

[1] European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/en/covid-19/data

[2] Stubinger, J., Schneider, L. Epidemiology of Coronavirus COVID-19: Forecasting the Future Incidence in Different Countries, *Healthcare*, 8(2), 99, 1–15, 2020. doi:10.3390/healthcare8020099

Irradiação solar e áreas das regiões faculares

E. Gonçalves[a], N. Mendes Lopes[a], J.M. Fernandes[b]
esmerald@mat.uc.pt, nazare@mat.uc.pt, jmfernan@mat.uc.pt

[a] *CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra Portugal*
[b] *Univ. Coimbra, CITEUC, OGAUC, Departamento de Matemática, 3001-501 Coimbra Portugal*

Keywords: atividade magnética solar, dependência linear, séries temporais

Abstract: Dispomos de dados diários, de 1975 a 2006, sobre as áreas das regiões faculares nos hemisférios norte e sul do Sol, medidas no OAUC. Estes dados são indicadores da atividade magnética do Sol e foram tratados anteriormente ([1]), tendo sido constatada a existência de uma dinâmica assimétrica nos dois hemisférios. Dispomos também de dados diários sobre a irradiação solar (intensidade luminosa do Sol recebida pela Terra) entre 1976 e 2015, registados no National Solar Observatory Kitt Peak (NSOKP, EUA).
Na sequência da breve análise da dependência existente entre estas variáveis através de um modelo de regressão linear simples feita em [2], exploramos, neste trabalho, a adequação de tal modelo linear àquelas séries temporais recorrendo aos dados recolhidos no OAUC e no NSOKP.
A série temporal bivariada que consideramos é relativa ao maior período de observação comum à série das regiões faculares e à série da irradiação solar. Além disso, para ultrapassar o problema de dados omissos, especialmente presente na segunda série, os elementos da série bivariada são os pares de médias mensais (mês solar com 27 dias, segundo a rotação de Bartel) das variáveis em estudo.
O estudo desenvolvido mostra que o modelo de regressão linear simples obtido, com coeficientes semelhantes aos de [2], é insuficiente para exprimir a média mensal das áreas das regiões faculares em termos da média mensal da irradiação solar. Observa-se, no entanto, a correta adequação a um modelo linear mais complexo que permite concluir que a média mensal das áreas das regiões faculares é bem descrita pelos seus dois valores mais recentes e pelos valores actual e mais recente da irradiação solar média mensal.

Acknowledgements: Ao CMUC, UIDB/00324/2020, com fundos do Governo Português via FCT/MCTES, e aos projectos UID/00611/2020 e UIDP/00611/2020.

References

[1] Gonçalves, E., Mendes-Lopes, N.M., Dorotovi, I., Fernandes, J.M., Garcia, A.. North and South Hemispheric Solar Activity for Cycles 21-23: Asymmetry and Conditional Volatility of Plage Region Areas. *Solar Physics*, 289, 6, 2283–2296, 2014. DOI10.1007/s11207-013-0448-8

[2] Shapiro, A.I., Solanki, S.K., Krivova, N.A., Schmutz, W.K., Ball, W.T., Knaack, R., Rozanov, E.V., Unruh,Y.C.. Variability of Sun-like stars: reproducing observed photometric trends. *Astronomy and Astrophysics*, 569, A38, 2014. doi: https://doi.org/10.1051/0004-6361/201323086

Using SRSM for prediction and risk analysis: an application to censored/uncensored medical data

C. Leal[a,b], T.A. Oliveira[a,b], A. Oliveira[a,b]

[a] *Universidade Aberta*
[b] *CEAUL*

Keywords: polynomial chaos, risk analysis, stochastic response surface methodology, uncertainty

Abstract: Risk Analysis has assumed a crucial relevance over the past few years, particularly in dynamical systems with increasing complexity. Thanks to the recent technological advances, the use of simulation techniques became current to estimate models allowing predict systems' behaviors, with respect to the probability of occurrence of a specific event and the consequences of this occurrence. Uncertainty associated with the simulation, either in model parameters or in experimental data, reveals its quantification as a prerequisite in probabilistic risk assessment.
The computational costs of numerical simulation are often very high, thus the use of metamodels arises as a pressing necessity. Stochastic Response Surface Methodology (SRSM) is known to be a suitable tool, both for the estimation of metamodels for the behaviors of systems and risk assessment, as for the quantification of uncertainty.
Although engineering applications of the methodology predominate in the literature, it has also been possible to extend and explore this methodology in many areas, namely in studies connected with Health problems, more specifically, in Medicine. We present a study based on Wisconsin Breast Cancer Prognostic database (http://pages.cs.wisc.edu/$\sim$olvi/uwmp/cancer.html#prog) with uncensored data and right censored data, in order to compare both censored/uncensored results and to attain more realistic results. The aim is to provide more appropriate models to prediction and to risk analysis, since the models obtained with uncensored data suffer from a right bias. Computational results and graphics were implemented using R software.

References

[1] Isukapalli, S. An uncertainty analysis of transport transformation models. *Ph.D. Thesis, New Brunswick, New Jersey:The State University of New Jersey*, (1999).

[2] Leal, C, Oliveira, T., Oliveira, A. Stochastic Response Surface Methodology–a study on polynomial chaos expansion. *Stochastic Modeling Data Analysis & Statistical Applications, 283* (2015).

[3] Oliveira, T., Leal, C., Oliveira, A. Stochastic response surface methodology: A study in the human health area. *AIP Conference Proceedings*(Vol. 1648, No. 1, p. 840012) (2015).

[4] Oliveira, T., Leal, C., Oliveira, A. Response Surface Methodology: A Review of Applications to Risk Assessment. *Theory and Practice of Risk Assessment*, Springer Verlag, 136, 385–397 (2015).

Forecasting models for time-series: a comparative study between classical methodologies and Deep Learning

D.R. Lopes[a], F.R. Ramos[b], D.A. Mendes[b], A.R. Costa[c]
dro.lopes@campus.fct.unl.pt, frjrs@iscte-iul.pt, deam@iscte-iul.pt
anabela.costa@iscte-iul.pt

[a] *UNL and CEDOC-UNL*
[b] *ISCTE-IUL and BRU-IUL*
[c] *ISCTE-IUL and CMAF-CIO*

Keywords: ARMA, DNN, ETS, forecasting, time-series

Abstract: In a year where the word "forecast" has been extensively used, it's more important than ever to have accurate forecasting models. In particular, in economics, finance and business areas; forecasting techniques are used to support enterprises to decide future directions, which determine the success of the same enterprises. However, in order for the forecasting techniques to be efficient, these must be truly understood and tested in real data-driven context, by taking in account existing models and new approaches. Based on the scientific literature, the classical methodologies are the most utilised by professionals, the autoregressive moving average (e.g. ARMA) and the exponential smoothing models (e.g. ETS), are the classical methodologies which are the most utilised by professionals [3]. Nonetheless, due to promising results, the literature has been keen on Deep Learning methodologies, in particular Deep Neural Networks (DNN) [2].

In fact, investigating what type of models should be used for each time-series based on their characteristics is the goal of this work. Three distinct models – ARMA, ETS and DNN – are assessed in the forecast of time-series with distinct patterns (see [1]). The discussion of the results will take into account not only the forecasting ability, but also its interpretability and computational cost.

This study shows that the additional computational power required in more complex models may not justify the improved accuracy. Although in time-series with strong perturbations, advantages are recognised in DNN models (lower prediction error), in series with a clear trend and/or seasonality, classical methodologies (e.g. ETS) outweighs the former.

References

[1] Lopes, D., Ramos, F. Univariate Time Series Forecast. Retrieved from https://github.com/DidierRLopes/UnivariateTimeSeriesForecast

[2] Tkáč, M., Verner, R. Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, 38, 788–804, 2016. doi:10.1016/J.ASOC.2015.09.040

[3] Wilson, J. H., Spralls III, S. A. What do business professionals say about forecasting in the marketing curriculum?. *International Journal of Business, Marketing, & Decision Science*. 11(1), 1–20, 2018.

Análise conjunta do tempo de sobrevivência global e do tempo livre de doença baseada em cópulas: uma aplicação ao cancro da mama

Beatriz Lourenço[a], **Giovani L. Silva**[a,b], António E. Pinto[c]
bds.lourenco96@gmail.com, giovani.silva@tecnico.ulisboa.pt,
aepinto@ipolisboa.min-saude.pt

[a] *Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa*
[b] *Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*
[c] *Departamento de Patologia, Instituto Português de Oncologia de Lisboa*

Keywords: análise de sobrevivência, cancro da mama, cópulas, modelos mistos

Abstract: Este trabalho foi motivado por um estudo envolvendo pacientes do sexo feminino com cancro da mama, realizado no Instituto Português de Oncologia de Lisboa. Os modelos de sobrevivência conjuntos baseados em cópulas foram ajustados a esses dados, contemplando diversos fatores de risco, tais como, a idade do paciente, a histologia e grau de diferenciação do tumor, a ploidia do ADN e o estadiamento clinicopatológico do paciente. Modelos de sobrevivência univariados foram também construídos, permitindo a comparação de resultados com os modelos conjuntos. Para a análise conjunta dos dados de cancro da mama, os resultados computacionais foram obtidos através do pacote do R, GJRM, desenvolvido para implementar Modelos de Regressão Conjuntos Generalizados.

O objetivo deste trabalho é avaliar a relação entre o tempo de sobrevivência global e o tempo livre de doença através de funções cópula, considerando distribuições marginais de sobrevivência Weibull. O estadiamento clinicopatológico dos pacientes e a ploidia do ADN são considerados estatisticamente relevantes para ambos os tempos de sobrevivência. Em relação à idade dos pacientes, o seu efeito apenas é relevante para o tempo de sobrevivência global de pacientes mais velhos. O risco de ocorrer uma reincidência do cancro ou morte, dadas as características clínicas do paciente, é diferente para os modelos univariado e bivariado. Para além disso, a estimativa do parâmetro de cópula demonstra uma elevada associação entre os tempos em estudo. Esta associação é também influenciada pela idade dos pacientes, i.e., pacientes mais velhos apresentam uma maior correlação entre os tempos de sobrevivência.

Modelo Hierárquico Poisson-binomial na subnotificação de casos e óbitos por COVID-19 no Brasil

Janaíne Geralda Mesquita Martins[a], Guaraci Requena[a]
janaine.martins@ufv.br, requena@ufv.br

[a] *Universidade Federal de Viçosa, Minas Gerais, Brasil – UFV*

Keywords: COVID-19, modelo hierárquico Poisson-binomial, subnotificação

Abstract: Um dos grandes problemas na análise e compreensão da COVID-19 está no processo de registro dos dados de casos e óbitos. Diversas pesquisas têm apontado que, no Brasil, há subnotificação na contagem oficial devido a diversos fatores, como a baixa testagem e a dificuldade de diagnóstico em diversas regiões. No entanto, muitos estudos levam em consideração os números oficiais e influenciam diretamente em tomadas de decisões. A Estatística nos fornece modelos que auxiliam na correção de subnotificação e, neste trabalho, estudamos e apresentamos as ideias por trás do modelo hierárquico Poisson-binomial. No modelo de Poisson, supomos que o número verdadeiro de casos, digamos y, que não é diretamente observado, segue uma distribuição de Poisson(λ) e o número notificado, digamos z, segue uma distribuição Binomial(y, p), sendo p a probabilidade de notificação. Desse modo, estabelecemos uma relação hierárquica entre y e z tal que $0 \leq z \leq y$. Tanto no caso Poisson quanto no Binomial, encaramos o problema adicionando covariáveis, como número de testes disponíveis, IDH, dentre outros, referentes aos estados brasileiros. Com isso, espera-se que os resultados obtidos tragam uma visão mais crítica e estatística dos dados divulgados no Brasil. Este trabalho é fruto de uma Iniciação Científica que se encontra em desenvolvimento na UFV e é financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Acknowledgements: Agradeço primeiramente a Deus pelo dom da vida. Agradeço imensamente ao meu orientador Guaraci por tanto me ensinar, pelo incentivo e por ser um grande exemplo, e à toda minha família pelo grande apoio e por estarem sempre ao meu lado.

The chi-squared goodness of fit test revisited

Sandra Mendonça[a,b], Délia Gouveia-Reis[a,b]
smfm@uma.pt, delia.reis@staff.uma.pt

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

Keywords: chi-squared goodness of fit test, degrees of freedom, estimators, type I error, type II error

Abstract: Pearson's influential work of 1900 [1] introduced the chi-squared goodness-of-fit test, that can be found today in almost every textbook on elementary statistics. This test is in fact a test of goodness of fit for a multinomial distribution. Many results concerning the test were discovered since Pearson's work. For example, it is known that the way distribution parameters are estimated and the way data is partitioned affects the quality of the test. Nonetheless, most of the times, important details like these are omitted in the referred textbooks, for the sake of simplicity or space saving, or any other reason. Trying to avoid the effect of the result of "a lie repeated a thousand times", this work collects some important milestones in the history of the chi-squared goodness-of-fit test, hoping to help the practitioners to design their goodness of fit tests.

Acknowledgements: This work is partially financed by national funds through FCT - Fundação para a Ciência e a Tecnologia under the project UIDB/0006/2020.

References

[1] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*. Series 5. 50 (302): 157–175, 1900. doi:10.1080/14786440009463897.

Modelação de nutrientes e elementos de qualidade ecológica para a classificação do estado das massas de água superficiais (Albufeiras do Norte)

Irene Oliveira[a,b], Joaquim Jesus[a], Rui Cortes[a]
ioliveir@utad.pt, jjesus@utad.pt, rcortes@utad.pt

[a] *Centro Investigação e de Tecnologias Agroambientais e Biológicas (CITAB), Universidade de Trás-os-Montes e Alto Douro, Portugal*
[b] *Centro de Matemática Computacional e Estocástica (CEMAT-IST-UL)*

Keywords: ambiente, ecologia, modelação, regressão

Abstract: Apresentam-se alguns dos resultados de modelação para o estabelecimento do Bom Potencial Ecológico em massas de água fortemente modificadas - Albufeiras do Norte, no âmbito do projeto *Serviços para Melhorar e Complementar os Critérios de Classificação do Estado das Massas de Água Superficiais Interiores*, promovido pela Agência Portuguesa do Ambiente. Destaca-se o trabalho estatístico realizado para uma das tipologias (Albufeiras do Norte), relativo ao uso de parâmetros físico-químicos, de suporte aos elementos biológicos que melhor discriminam o Estado Ecológico determinados a partir dos EQR (*Ecological Quality Ratio*), ou seja o desvio em relação à situação de referência (massas de água não perturbadas). Consideraram-se os dados disponibilizados relativos às campanhas de monitorização de âmbito nacional, realizadas nos anos de 2006, 2010, 2017 e 2019. Na análise exploratória usaram-se transformações de variáveis para selecionar as relações e modelos mais relevantes e estabelecer os limiares entre classes de qualidade do Estado Ecológico. Foram avaliados modelos de regressão procurando definir, se possível, as classes de qualidade para Potencial Ecológico, a partir dos inversos dos limiares definidos para o EQR. Teve-se em consideração o modelo de regressão clássico, modelos aditivos generalizados e outros alternativos (regressão segmentada e regressão quantílica). Os procedimentos utilizados seguiram a publicação de referência [1].

Acknowledgements: Este trabalho foi financiado por Fundos Nacionais através dos Projetos UID/AGR/04033/2019 e UID/MULTI/04621/2019 da Fundação para a Ciência e a Tecnologia, Agência Portuguesa de Ambiente e o Projeto *Serviços para Melhorar e Complementar os Critérios de Classificação das Massas de Água Superficiais Interiores*.

References

[1] Phillips, G., Birk, S., Bohmer, J., Kelly, M., Willby, N., Poikane, S. *The use of pressure response relationships between nutrients and biological quality ele-ments as a method for establishing nutrient supporting element boundary values for the Water Framework Directive, Publications Office of the European Union, Luxembourg*, 2018.

Unravel shopper behaviour: analysis of consumption associations in fuel stations and their convenience stores

André Pedrinho[a], José Miguel Espinosa[b], Regina Bispo[c]
a.pedrinho@campus.fct.unl.pt, miguel.espinosa@galp.com, r.bispo@fct.unl.pt

[a]*MSc in Analytics and Big Data Engineering, Department of Computer Science and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*
[c]*Tech Garage, IT & Digital Department. Galp Energia, SGPS, S.A., 1600-209 Lisboa, Portugal*
[b]*Center for Mathematics and Applications and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal*

Keywords: association rules, big data, clustering, market basket analysis, shopping mission

Abstract: The population growth as well as technology development have caused the volume of data to increase exponentially. In case of commerce and retail, the available amount of data increases almost by the second, due to high number of transactions that are processed in a short space of time. This high amount of data is very important to a company success. However, to take insights of high amounts of data is a hard and complex task.
This study addresses the problem of analyzing transactions based on a dataset characterized by high dimensionality and large volume. In particular, it aims (1) to establish relationships between items which are purchased together (baskets), using Market Basket Analysis, through the analysis of Association Rules and (2) to group different types of baskets using (mixed data) clustering algorithms, defining and discovering different segments of baskets.
The available data was from 2 years and included 120 760 325 shopping baskets of 297 fuel stations and respective convenience stores describing the purchased items and the location and date of purchase. Prior to the analysis, a pre-processing step was taken to evaluate the levels of the products hierarchy. To avoid granularity problems, this analysis focuses on the second level of the products hierarchy. The Market Basket Analysis was implemented using the FP-Growth algorithm while the Clusters Analysis was performed with K-Medoids, K-Means and DBSCAN algorithms. A comparison between the performance of these algorithms is also presented.

Acknowledgements: RB work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/00297 /2020 (Center for Mathematics and Applications).

Comparação das abordagens semi-paramétricas à ANOVA com dois fatores na presença de heterocedastidade

Dulce G. Pereira[a], Anabela Afonso[a]
dgsp@uevora.pt, aafonso@uevora.pt

[a] *Departamento de Matemática/ECT, Centro de Investigação em Matemática e Aplicações/IIFA, Universidade de Évora*

Keywords: estatística de Wald, interação, robustez, testes de permutação

Abstract: A ANOVA fatorial é usada para comparar a média de vários grupos de dados e assenta nos pressupostos; i) normalidade da distribuição dos erros, ii) homocedasticidade, e iii) independência dos resíduos. Na prática, estes pressupostos são facilmente violados. No final da década de 90, foram propostos dois testes alternativos que permitem relaxar estes pressupostos: o Wald Type Statistics (WTS)[1] e Anova Type Statistics (ATS)[3], mas que requerem amostras de grande dimensão. Nos últimos anos têm ganho muita atenção os testes baseados na permutação de observações que são muito robustos à violação do pressuposto de normalidade, mas existem poucos estudos sobre o seu desempenho na presença de heterogeneidade [5, 2, 4]. Este trabalho pretende dar um contributo nesta última análise. É realizado um estudo por simulação, em que se consideram delineamentos equilibrados com igual e desigual número de níveis dos fatores, vários tipos de distribuições com diferentes graus de dispersão.

Acknowledgements: Este trabalho é parcialmente suportado pelo Centro de Investigação em Matemática e Aplicações (CIMA), através do Projeto "UIDB/04674/2020" da FCT-Fundação para a Ciência e a Tecnologia, Portugal

References

[1] Akritas, M. G., Arnold, S. F., Brunner, E. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437), 258–265, 1997.

[2] Basso, D., Chiarandini, M., Salmaso, L. Synchronized permutation tests in replicated $I \times J$ designs. *Journal of Statistical Planning and Inference*, 137(8), 2564-2578, 2007.

[3] Brunner, E., Dette, H., Munk, A. Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92(440), 1494–1502, 1997.

[4] Pauly, M., Brunner, E., Konietschke, F. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 77(2), 461–473, 2015.

[5] Salmaso, L. Synchronized permutation tests in 2^k factorial designs. *Communications in Statistics-Theory and Methods,* 32(7), 1419–1437, 2003.

COVID-19 Hospitalization in Portugal: Results from hospital discharge data

João F. Pereira[a,b], Constantino Caetano[a,c], Liliana Antunes[a,f], Maria Luísa Morgado[a,c], Paula Patrício[d], Baltazar Nunes[a,e],
pereira.jpf96@gmail.com, constantino.caetano@insa.min-saude.pt,
liliana.antunes@insa.min-saude.pt, luisam@utad.pt, pcpr@fct.unl.pt,
baltazar.nunes@insa.min-saude.pt

[a] *Dep. de Epidemiologia, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisbon, Portugal*
[b] *Dep. of Mathematics, University of Trás-os-Montes e Alto Douro, UTAD*
[c] *Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, University of Lisbon*
[d] *Centro de Matemática e Aplicações (CMA), FCT, UNL and Dep. de Matemática, FCT, UNL, Portugal*
[e] *Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Lisbon, Portugal*
[f] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Keywords: COVID-19, data analysis, hospitalization, parameter estimation

Abstract: In December 2019, a novel coronavírus emerged in the Wuhan province in China. As it spread across the world rapidly, the World World Health Organization officially declared a pandemic on March 11, 2020.
As the number of COVID-19 cases increased all over the world the capabilities of the different health systems of each country were put to the test, and there was a necessity to better understand the hospital dynamics of this decease.
From the beginning of the pandemic, mathematical models were constructed worldwide to study the spread of COVID-19, but the lack of information about the novel coronavirus and its dynamics of transmission made parametrization of these models extremely difficult and heavily reliable on curve fitting, what lead to not so realistic models constructions and very deviated predictions when compared with the reality.
The goal of this study is to describe the COVID-19 Hospital discharge data in Portugal. We analysed the evolution of the number and duration of hospital stays in infirmary and intensive care units (ICU), risk of being hospitalized and admitted to ICU, and explored the best probabilistic distributions to describe the duration of hospital stays in infirmary and ICU. From this description, parameters were calculated and successfully used in the COVID-19 in-CTRL model[1] and can be used in other modelling studies.

References

[1] Caetano C, Morgado ML, Patrício P, Pereira JF, Nunes B. Mathematical Modelling of the Impact of Non-Pharmacological Strategies to Control the COVID-19 Epidemic in Portugal. *Mathematics*, 2021, 9(10):1084. doi:10.3390/math9101084

COVID-19 em Portugal: uma análise de riscos competitivos

Margarida Ribeiro[a], Cristina Rocha[a,b], André Peralta-Santos[c]
fc49811@alunos.fc.ul.pt, cmrocha@fc.ul.pt, aperaltasantos@dgs.min-saude.pt

[a] *DEIO, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[b] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[c] *Direção de Serviços de Informação e Análise, DGS, Portugal*

Keywords: análise de sobrevivência, COVID-19, fatores de risco, riscos competitivos

Abstract: A COVID-19 é uma doença respiratória infecciosa grave causada pelo coronavírus SARS-CoV-2 que originou uma pandemia que se tem prolongado pelos anos de 2020 e 2021. Identificar os fatores de risco determinantes é fundamental para minimizar o risco de eventos adversos, especialmente a morte.
Este estudo tem como objetivo analisar o impacto da pandemia na população portuguesa. Propõe-se um estudo no âmbito da análise de sobrevivência, recorrendo a modelos de riscos competitivos, considerando a recuperação, morte por COVID-19 e morte por outra causa como acontecimentos de interesse, para uma análise mais rigorosa. Pretende-se identificar quais os fatores que têm influência significativa no tempo até à morte por COVID-19. Este estudo inclui todos os casos positivos desde 2 de março a 31 de dezembro de 2020.
Neste período foram infectadas 360 914 pessoas, das quais 5 197 tiveram como causa básica de morte COVID-19, 411 morreram por outra causa e 313 377 recuperaram. O tempo médio desde a infecção até a morte por COVID-19 foi 11 dias, até à morte por outra causa foi 14 dias e até à recuperação foi 16 dias. Para todos os grupos, definidos por género, grupo etário e Administração Regional de Saúde (ARS), a recuperação destaca-se por ser o acontecimento com maior probabilidade de ocorrência. Indivíduos do sexo masculino e com uma idade mais avançada têm um maior risco de morte por COVID-19 do que por outra causa, após infetados pelo vírus.

Acknowledgements: À Direção-Geral da Saúde, em especial, à Direção de Serviços de Informação e Análise pela disponibilização dos dados. Este trabalho é parcialmente financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] Haller, B., Schmidt, G., Ulm, K. Applying competing risks regression models: an overview. *Lifetime Data Analysis*, 19, 33-58, 2013. doi:10.1007/s10985-012-9230-8

[2] Rocha, C., Papoila, A.L. *Análise de Sobrevivência*. Edições SPE, Lisboa, 2009 ISBN:978-972-8890-22-3

[3] Zhang, Z. Survival analysis in the presence of competing risks. *Annals of Translational Medicine*, 5, 3–47, 2017. doi:10.21037/atm.2016.08.62

Identificação de requisitos para o desenvolvimento de Exoesqueletos no âmbito da Infantaria

Joni Santos[a], Luís Quinto[a,b], Sérgio Gonçalves[b], Ivo Roupa[b], Miguel Silva[b], Paula Simões[a,c]
santos.jg@exercito.pt, luis.quinto@academiamilitar.pt,
sergio.goncalves@tecnico.ulisboa.pt, iroupa@gmail.com,
miguelsilva@tecnico.ulisboa.pt, paula.simoes@academiamilitar.pt

[a] *CINAMIL, Academia Militar-Instituto Universitário Militar, Portugal*
[b] *IDMEC, Instituto Superior Técnico - Universidade de Lisboa, Portugal*
[c] *CMA, Faculdade de Ciências e Tecnologia - Universidade NOVA de Lisboa, Portugal*

Keywords: exoesqueleto, militar, requisitos operacionais

Abstract: Os militares são atualmente submetidos a esforços físicos significativos, potenciando a ocorrência de lesões e, consequentemente, a redução do seu nível de operacionalidade, sendo os exoesqueletos apontados como uma possível solução para mitigar tais efeitos. O objetivo desta investigação foi a identificação dos requisitos operacionais para implementação de um exosequeleto para utilização militar, enquadrada no projeto ELITE2 - Enhanced LITe Exoskeleton, que pretende desenvolver exoesqueletos para apoio ao movimento humano. Neste âmbito, foram realizadas entrevistas e inquéritos por questionário aos militares pertencentes a várias Forças Nacionais Destacadas projetadas para a República Centro-Africana, entre Março e Abril de 2021. A análise de dados foi efetuada recorrendo a diversos métodos estatísticos, implementados com recurso ao software IBM SPSS® Statistics, nomeadamente, técnicas de estatística descritiva, essencialmente na fase de resumo de informação, bem como ao nível da inferência estatística, como forma de avaliar diversas hipóteses tendo em conta a problemática em causa. Conclui-se que um exoesqueleto, para ser aplicado no âmbito militar, deve apoiar diversas posições de atirador de pé, bem como a realização de movimentos como saltar obstáculos, arrombamento de portas e corrida. Deve também auxiliar o transporte de carga, de modo a mitigar lesões associadas ao peso do equipamento. Adicionalmente, deverá ser capaz de resistir a elevadas temperaturas, humidade e poeira, elementos caraterísticos do Ambiente Físico deste Teatro de Operações.

Acknowledgements: Ao Estado-Maior do Exército (ELITE2/2021/CINAMIL) e à Fundação para a Ciência e Tecnologia - projeto LAETA (UIDB/50022/2021). Ao Laboratório de Biomecânica de Lisboa.

References

[1] Collins, S., et al. Reducing the energy cost of human walking using an unpowered exoskeleton. *Nature*, 522(7555),212–215, 2015.

[2] Pinheiro, P. et al. Analysis of the Performance of a Passive Ankle Exoskeleton for Reduction of the Metabolic Costs in Gait. *Congresso Nacional de Biomecânica, Covilhã*, 2019.

Desempenho do algoritmo EM na estimação dos parâmetros de misturas pseudo-convexas

Rui Santos[a,b], Miguel Felgueiras[a,b], João Paulo Martins[a,b]
rui.santos@ipleiria.pt, mfelg@ipleiria.pt, jpmartins@ipleiria.pt

[a] *Escola Superior de Tecnologia e Gestão do Politécnico de Leiria*
[b] *CEAUL – Centro de Estatística e Aplicações*

Keywords: algoritmo EM, distribuições estáveis para extremos, misturas generalizadas, misturas pseudo-convexas, simulação

Abstract: As misturas pseudo-convexas geradas por distribuições estáveis para extremos formam uma família de distribuições útil para modelar dados em questões de fiabilidade, como a mistura pseudo-convexa gerada pela distribuição exponencial ou pela distribuição em função potência. No entanto, a estimação dos seus parâmetros não é precisa em alguns casos, nomeadamente quando o peso ω é negativo. Neste trabalho, o desempenho do algoritmo EM (*Expectation-Maximization*) aplicado na estimação dos parâmetros em misturas pseudo-convexas geradas por distribuições estáveis para extremos é avaliado via simulação.

Acknowledgements: Este trabalho foi financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito dos projetos UIDB/00006/2020 e UIDP/00006/2020.

References

[1] 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 (Methodological)* **39**(1), 1–38, 1977. doi:10.1111/j.2517-6161.1977.tb01600.x

[2] Felgueiras, M., Martins, J.P., Santos, R. Pseudo-convex Mixtures, Numerical Analysis and Applied Mathematics ICNAAM 2012, American Institute of Physics, *AIP Conf. Proc.* **1479**, 1125–1128, 2012. doi:10.1063/1.4756346

[3] Santos, R., Felgueiras, M., Martins, J.P. Pseudo-convex Mixtures Generated by Shape-extended Stable Distributions for Extremes, *Journal of Statistical Theory and Practice* **10**(2), 357–374, 2016. doi:10.1080/15598608.2016.1146929

[4] Santos, R., Felgueiras, M., Martins, J.P. Estimação em misturas pseudo-convexas, *Atas do XXII Congresso Anual da Sociedade Portuguesa de Estatística*, 211–222, 2016.

Uso de *machine learning* na classificação de notificações de dengue a partir de atributos clínicos de triagem

João Carlos Zayatz[a], Gislaine Camila Lapassini Leal[a], Paulo César Ossani[b], Greicy Cezar do Amaral[c]
joaoczayatz@gmail.com, gcllleal@uem.br, ossanipc@hotmail.com, amaralgreicy722@gmail.com

[a] *Programa de Pós-graduação em Engenharia de Produção/Universidade Estadual de Maringá - PGP/UEM*
[b] *Departamento de Estatística/Universidade Estadual de Maringá - DES/UEM*
[c] *15ª Regional de Saúde, Secretaria de Saúde do Paraná - SESA/PR.*

Keywords: aedes, dados epidemiológicos, *knowledge discovery in databases*, modelo de classificação, *naive bayes*

Abstract: A crescente disponibilidade de dados epidemiológicos, da área de saúde, desperta o interesse para investigação de padrões que contribuam para melhorar o atendimento clínico. Modelos preditivos baseados em informações clínicas podem ser relevantes em âmbitos epidemiológicos em que não estejam presentes as possibilidades de investigação laboratorial. Neste contexto, é apresentado um modelo em *machine learning* para a classificação de notificações de dengue, doença viral transmitida por mosquitos do gênero Aedes, responsável por 400 milhões de casos por ano no mundo [1]. As amostras correspondem à registros do Brasil, durante a epidemia dos anos de 2019 e 2020. Dados epidemiológicos registrados por unidades de saúde foram pré-processados seguindo etapas do processo *Knowledge Discovery in Databases - KDD* [2]. O conjunto de dados desta análise contou com 48.088 instâncias, que correspondem a casos notificados como suspeitos e classificados como confirmados (57,25%) ou descartados (42,75%), após exame laboratorial. Foram considerados 25 atributos categóricos como variáveis preditivas, constituídos por informações socioeconômicas e sinais clínicos aparentes, registrados durante triagem clínica. Para gerar o modelo, foi utilizado o *software* Weka. Um processo de seleção de atributos *CfsSubsetEval*, com método de busca *BestFirst*, permitiu reduzir o número de variáveis preditivas para 6. Como técnica de classificação, foi utilizado o algoritmo *Naive Bayes*. Os dados foram divididos em 70% para teste e 30% para treinamento. O modelo gerado possibilitou classificação correta de 63,51% das instâncias, com sensibilidade de 72% e especificidade de 55,5%. A utilização de modelos em *machine leaning* como ferramenta auxiliar na busca de aprimorar os diagnósticos da dengue mostra-se viável.

Referências

[1] Wilder-Smith, A. et al. Dengue. *The Lancet*, v. 393, n. 10169, p. 350-363, 2019.
[2] Fayyad, U. et al. The KDD process for extracting useful knowledge from volumes of data *Communications of the ACM*, v. 39, n. 11, p. 27-34, 1996.

Index

Álvaro, Pedro, 93

Abreu, Ana Maria, 132, 142
Achcar, Jorge Alberto, 114
Afonso, Anabela, 82, 157
Afonso, Pedro M., 45
Afreixo, Vera, 55
Albuquerque, João, 46
Alcoforado, Renata, 47
Allen, Genevera I., 3
Alpizar-Jara, Russell, 48, 76, 104, 121
Alvelos, Helena, 49
Alves, Catarina, 46
Alves, Maria Isabel Fraga, 127
Amado, Conceição, 50
Amaral, Greicy Cezar do, 162
Amorim, Clara, 98
Andrade, Cristina, 36
Andrade, Marcia, 141
Andrade, Marina A. P., 38
Andrinopoulou, Eleni-Rosalina, 45
Antunes, Liliana, 158
Antunes, Marília, 46, 103
Arcêncio, Ricardo Alexandre, 148
Assunção, Danillo, 51
Assunção, Renato, 30
Azevedo, Nuno, 24

Bernieri, Emmanuel, 52
Bisognin, Cleber, 145
Bispo, Regina, 119, 146, 147, 156
Bland, Cynthia, 16
Borges, Ana, 39
Bourbon, Mafalda, 46
Brôco, Nuno, 93
Brandão, Pedro, 53
Braumann, Carlos A., 54, 74, 84, 85
Brito, Paula, 96

Cabral, Jorge, 55
Cabral, Paulo, 99
Caeiro, Frederico, 56, 78, 126
Caetano, Constantino, 57, 158
Calado, Ricardo, 119
Camacho, Pedro Daniel, 142
Campos, Pedro, 20, 120
Canto e Castro, Luísa, 108
Cardoso, Andreia, 14
Cardoso, Margarida, 59
Carinhas, Dora, 60
Carrasco, Alexandre, 38
Carrasquinha, Eunice, 61
Carvalho, Alda, 40
Carvalho, Filipe, 76
Carvalho, Inês, 143
Carvalho, M. Lucília, 101, 128
Carvalho, Marta, 93
Casaca, Cláudia, 40
Castanheira, Ana, 62
Celeri, Maurício, 32
Cerveira, Adelaide, 144
Cesar, Rodrigo, 82
Chiroque-Solano, Pamela M., 63
Chissaque, Assucênio, 81
Clancy, John P., 45
Clemente, Susana, 18

Coelho, Norberta, 93
Cordeiro, Clara, 11, 39
Correia, Elisete, 144
Correia, Iúri, 64, 97
Cortes, Rui, 155
Costa, Anabela R., 115, 151
Costa, Cláudia, 79
Costa, Eliardo G., 109
Costa, Marco, 65, 79, 102, 111
Cristelo, Nuno, 144
Cunha, Mónica V., 93
Curto, José Dias, 123
Cuyler, Christine, 64

da Câmara, Carlos, 113
da Costa, Luis, 53
da Silva Machado, Guilherme, 145
Dasgupta, Nairanjana, 16
Davison, Anthony C., 4, 105
de Carvalho, Miguel, 52, 66, 86, 87, 113
de Mello-Sampayo, Felipa, 67
de Moraes, Talita, 68
de Oliveira, Paulo, 69
de Zea Bermudez, Patrícia, 66
Demétrio, Clarice G. B., 31
Dias, José Gomes, 70
Dias, Sandra, 71
Dias, Sónia, 96
Domingues, Tiago, 14
dos Reis, Gonçalo, 86
dos Santos, Naiara, 51
Duarte, Brunna, 32

Egídio dos Reis, Alfredo, 47
Eldridge, Matthew D., 89
Encarnação, Marco, 146
Espinosa, José Miguel, 147, 156
Eufrázio, Sofia, 76

Faria, Susana, 107
Farias, Inês, 97
Felgueiras, Miguel, 161
Feliciano, Amélia, 14
Fernandes, João M., 149
Ferreira, Fátima, 73
Fiaccone, Rosemeire, 10
Figueiredo, Ivone, 97, 101, 128
Filipe, Patrícia A., 54, 72, 74, 84, 85
Filzmoser, Peter, 96
Fonseca, André, 11
Fonseca, Paulo, 98
Fortes, Filomeno, 81
Fortes, Inês, 75
Fradinho, Marta, 14
França, Susana, 53
Freitas, Adelaide, 83, 124

Galantinho, Ana, 76
García-Donato, Gonzalo, 110
Garrido, Susana, 125
Gasparinho, Carolina, 81
Girão Serrão, Rodrigo, 77
Gkikopoulou, Kalliopi C., 95
Godinho de Matos, Miguel, 19, 21
Gois, Patrícia, 82
Goldstein, Carolina, 147
Gomes, Dulce, 148
Gomes, M. Ivette, 56, 78, 134
Gomes, Manuel Carmo, 93
Gonçalves, A. Manuela, 65, 79, 111
Gonçalves, Elsa, 80
Gonçalves, Esmeralda, 129, 149
Gonçalves, Luzia, 81
Gonçalves, M. Helena, 81
Gonçalves, Sérgio, 117, 160
Gouveia-Reis, Délia, 154
Graça, Luís, 91

Hanser, Sean F., 92
Henriques-Rodrigues, Lígia, 48, 78

Infante, Paulo, 60, 82
Iutis, Adela, 83

Jácome, María Amalia, 12, 13
Jacinto, Gonçalo, 54, 72, 74, 82, 84, 85
Jamba, Nelson T., 85
Januário, Ana Paula, 72
Jesus, Joaquim, 155

Kim, Jongwook, 138
Kumukova, Alina, 86

López-Cheda, A., 13

López-de-Ullibarri, Ignacio, 12
López-Oriona, Ángel, 88
Lages, Rita, 18
Lam, Heung Yeung, 112
Lapassini Leal, Gislaine Camila, 162
Leal, Conceição, 150
Lee, Junho, 87
Leite, Andreia, 57
Lima, Francisco, 19, 20
Lomba, Jessica, 127
Lopes, Didier R., 115, 151
Lou, Wendy, 16
Lourenço, Beatriz, 152
Lourenço, Vanda M., 16
Lourinho, Rita, 93
Lynch, Andy G., 89, 137

Machado, Ausenda, 57
Magalhães, Sara, 90
Malato, João, 91
Malta, Joana, 18
Manuel, Paulo Rebelo, 82
Marcelino, Ana Rita, 92
Marcelo, Carlos, 20
Marques, Alda, 55
Marques, Carolina S., 93, 95
Marques, Tiago A., 53, 64, 92, 93, 95, 98, 116
Martinho, António, 60
Martins, Ana, 96
Martins, Ana A., 59
Martins, Antero, 80
Martins, Janaíne Geralda Mesquita, 153
Martins, João Paulo, 161
Martins, Natália, 83
Martins, Rui, 97
Matos, André B., 98
Medeiros, Ana, 46
Meira, Wagner, 30
Meira-Machado, Luís, 133, 143
Meireles, Fátima, 93
Mendes Lopes, Nazaré, 149
Mendes, Diana A., 115, 151
Mendonça, Sandra, 154
Menezes, Raquel, 125
Milheiro-Oliveira, Paula, 99
Mira, António, 76
Miranda, M. Cristina, 100, 118, 134
Molemberghs, Geert, 31
Monteiro, Andreia, 101, 128
Monteiro, Magda, 102
Monteiro, Sílvia, 93
Moreira, Frederico, 103
Moreno, Ana, 125
Morgado, Leonel, 150
Morgado, Maria Luísa, 57, 158
Mori, Tomi, 16
Mosquera, Jaime, 104
Moura, Eveline, 32
Mourato, Sandra, 36
Mrtvi, Marcelo, 105

Nascimento, Ana Carolina, 32
Nascimento, Moysés, 32
Natário, Isabel, 101, 128, 146
Neil, Daniel, 30
Neves, Cláudia, 127
Nicolau, João, 106
Nilton, Ávido, 48
Nogueira, Pedro, 82
Nogueira, Vitor, 82
Novais, Luísa, 107
Nunes, Baltazar, 57, 158
Nunes, Cátia, 21
Nunes, Cláudia, 136

Oliveira, Gabriela, 32
Oliveira, Inês, 108
Oliveira, Irene, 155
Oliveira, Lina, 77, 135
Oliveira, M. Rosário, 61, 77, 135, 136
Oliveira, Manuela, 126
Oliveira, Margarida, 143
Oliveira, Pedro, 133
Oliveira, Teresa, 150
Ossani, Paulo César, 162

Pacheco, António, 5, 61, 73
Palipana, Anushka, 45
Papathomas, Michail, 137
Patrício, Paula, 57, 158
Paulino, Carlos Daniel, 109
Paulo, Rui, 110
Pedrinho, André, 156
Peng, Y., 13

Peralta-Santos, André, 159
Pereira, Dulce G., 157
Pereira, F. Catarina, 65, 111
Pereira, Isabel, 131
Pereira, João F., 57, 158
Pereira, Julio, 112
Pereira, Paula, 66
Pereira, Soraia, 66, 113
Peres, Marcos, 114
Pettersson, Magnus, 42
Pinto, António E., 152
Pires, Marília, 121
Previdelli, Isolde, 68, 105

Quaresma, Paulo, 82
Quinto, Luís, 117, 160

Ramos, Antônio, 148
Ramos, Dandara, 10
Ramos, Filipe R., 115, 151
Ramos, João, 36
Ramos, M. Rosário, 39
Rego, Leonor, 82
Requena, Guaraci, 153
Ribeiro, Helena, 73
Ribeiro, João, 116
Ribeiro, Margarida, 159
Ribeiro, Nuno, 117
Ricardo, Fernando, 119
Rizopoulos, Dimitris, 45
Rocha, Anabela, 100, 118
Rocha, Cristina, 62, 132, 159
Rodrigues, Helena Sofia, 83
Rodrigues, Paulo C., 33
Rodrigues, Paulo M. M., 106
Rosa, Álvaro, 38
Roupa, Ivo, 117, 160

São João, Ricardo, 14
Safari, Wende Clarence, 12
Saias, José, 82
Sampaio, Clara, 119
Sanarico, Maurizio, 6
Santos, Bárbara, 120
Santos, Daniel, 82
Santos, Joni, 160
Santos, Jorge, 121
Santos, Laura, 14
Santos, Ricardo, 93
Santos, Rui, 161
Sepúlveda, Nuno, 11, 91, 103, 122
Serrasqueiro, Pedro, 123
Shulman, Holly, 16
Silva, Alberto, 124
Silva, Carmo, 76
Silva, Daniela , 125
Silva, Domingos, 126
Silva, Giovani L., 68, 81, 152
Silva, Marcelo, 82
Silva, Marco, 93
Silva, Maria Eduarda, 131
Silva, Miguel, 117, 160
Silva, Nuno R., 25
Silva, Tiago, 40
Silva, Vânia, 14
Simões, Paula, 101, 117, 128
Singer, Júlio M., 109
Siqueira Peres da Silva, Vanessa, 145
Smith, Mike L., 89
Smith, Stephen H., 92
Soares, Patrícia, 62
Sousa, Diogo, 129
Sousa, Inês, 75, 90, 130
Sousa, Lisete, 97
Sousa, Rita, 26
Sousa, Rodney, 131
Sousa-Ferreira, Ivo, 132
Soutinho, Gustavo, 133
Souto de Miranda, Manuela, 50, 134
Souza, Roberto, 30
Stoykov, Marian, 106
Subtil, Ana, 135
Szczesniak, Rhonda D., 45

Taban, Rasool, 136
Tavaré, Simon, 89
Temido, Maria da Graça, 71
Teodoro, M. Filomena, 38
Tomazella, Vera, 51
Torres, André, 57
Trigo, Ricardo, 113
Tristão de Moraes, Talita, 51

Uyeyama, Robert K., 92

Vala, Francisco, 21

Valadas, Rui, 61
Vasquez, Fernando, 60
Vaz, Daniel, 40
Velasco-Pardo, Víctor, 137
Veloso, Marta, 27
Verbeke, Geert, 31
Vieira, Manuel, 98
Vilar, Jose, 88
Vilaça, João, 93

Xambre, Ana Raquel, 49

Zambom, Adriano Zanin, 138
Zamboni Berra, Thaís, 148
Zangiacomi, Edson, 114
Zayatz, João Carlos, 162

www.spestatistica.pt