PROGRAM and BOOK of ABSTRACTS

JOCLAD2020

22 - 24 OCTOBER

LISBOA, PORTUGAL

UNIVERSIDADE LUSÓFONA

CLAD
Associação Portuguesa de Classificação e Análise de Dados

XXVII MEETING OF THE PORTUGUESE ASSOCIATION FOR CLASSIFICATION AND DATA ANALYSIS
XXVII JORNADAS DE CLASSIFICAÇÃO E ANÁLISE DE DADOS

Program and Book of Abstracts

XXVII Meeting of the Portuguese Association for Classification and Data Analysis (CLAD)

22–24 October 2020

Lisboa, Portugal

www.joclad.ipt.pt/joclad2020/

Sponsors

Associação Portuguesa de Classificação e Análise de Dados (CLAD)
Universidade Lusófona de Humanidades e Tecnologias
Centro de Pesquisa e Estudos Sociais(CPES) - Universidade Lusófona
Instituto Nacional de Estatística
Banco de Portugal
SAS Portugal - Analytics Software & Solutions
Museu Bordalo Pinheiro
Garrocha Estate Wines

Program and Book of Abstracts

XXVI Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD2020)

Editors: Ana Sousa Ferreira, Ana Lorga da Silva, José Dias, Anabela Marques, Fernando Borges, Conceição Rocha, Paula Vicente

Publisher: CLAD

ISBN 978-989-98955-7-7

Preface

Welcome to JOCLAD2020! The JOCLAD2020 - Meeting of the Portuguese Association for Classification and Data Analysis aims to bring together researchers and practitioners. This is already the twenty-seventh meeting of the CLAD in the field of Data Science. After many meetings all over Portugal - 2017 in Porto, 2018 in Almada, 2019 in Viseu - JOCLAD2020 was planned to be held on 2-4 April, in Lisbon, at the Universidade Lusófona, which co-organizes it.

The JOCLAD2020 program had two mini-courses, on April 2, taught by the guest professors Mark de Rooj (Methodology and Statistics Unit, Leiden University) and Gilbert Saporta (Conservatoire National des Arts et Métiers, Paris), three plenary sessions also of their responsibility and a third invited speaker, José Luís Ferreira (Senior Consultant, Quidgest), three thematic sessions - CLAD 2020 Scholarship, INE and Banco de Portugal - and 25 oral communications, and 17 posters. Unfortunately, due to the COVID-19 pandemic, we were forced to postpone JOCLAD2020 to be held 22-24th October in a digital format. This postponing has caused changes to the JOCLAD2020 program due to the specific incompatibilities of guests and participants/authors agendas.

The program for this online meeting results from the dedicated effort of many people. We thank the invited speakers: Mark de Rooj (Methodology and Statistics Unit, Leiden University) "The MELODIC family for simultaneous binary logistic regression of multiple outcome variables in a reduced space" and José Luís Ferreira (Senior Consultant, Quidgest) "AI and ML: It's all about data. Datas paradox: as the value of a single datum tends to zero, the value of all data tends to infinite". Their talks present a representative cross-section of research in data science. We also thank Gilbert Saporta, who cannot attend our meeting, but was the first invited speaker confirming readiness to be in Lisbon, in April.

A Thematic Session is devoted to the students granted with a 2020 CLAD scholarship, whose members of the evaluation committee were Ana Sousa Ferreira (Chair), Manuela Neves, and Paula Vicente. We also thank the organizers of the other Thematic Sessions: Carlos Marcelo (INE - Instituto Nacional de Estatística), Luís Teles Dias (Banco de Portugal), and Adelaide Figueiredo (CLAD-SPE).

Additionally, this volume contains all the abstracts of talks and posters presented at regular oral and poster sessions. Each abstract published in this volume has been double-blind evaluated by at least one anonymous member of the scientific committee. We thank all authors who submitted an abstract to our meeting and the reviewers, who supported the editorial process with their fast and constructive reactions. These procedures contribute to reinforce the overall quality of the JOCLAD2020 program. Additionally, we thank all the chairs of these sessions.

Our deep gratitude extends to the members of the board of CLAD, in particular Carlos Marcelo and Conceição Rocha, who volunteered their time to support the JOCLAD

organization. Last but not least, it is a pleasure to thank the sponsors for helping the organization of this meeting. Our institutional sponsors deserve a special mention: Instituto Nacional de Estatística (INE) and Banco de Portugal.
Finally, a big thank you goes to all of you for your support, helping us with keeping our annual meeting on their feet; albeit in a way that is appropriate to the times of the COVID-19 pandemic and makes this meeting a success. With your high-quality work, CLAD will continue its tradition of excellence in advancing Data Science!
We hope to meet you again for the JOCLAD2021!

Lisboa, October 2020

Chair of the Scientific Program
Ana Sousa Ferreira

Conference Chair
Ana Lorga da Silva

President of CLAD
José Gonçalves Dias

Organization

President of the CLAD

José Gonçalves Dias

Chair of the JOCLAD2020

Ana Lorga da Silva (Universidade Lusófona de Humanidades e Tecnologias)

Local Organizing Committee

Ana Lorga da Silva (Universidade Lusófona de Humanidades e Tecnologias)
Paula Vicente (Universidade Lusófona de Humanidades e Tecnologias)
Fernando Borges (Universidade Lusófona de Humanidades e Tecnologias)
Conceição Rocha (INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência)

Chair of the Scientific Program Committee

Ana Sousa Ferreira (Universidade de Lisboa)

Scientific Program Committee

A. Manuela Gonçalves (Universidade do Minho)
Adelaide Figueiredo (Universidade do Porto)
Adelaide Freitas (Universidade de Aveiro)
Ana Lorga da Silva (Universidade Lusófona)
Ana Matos (Instituto Politécnico de Viseu)
Anabela Afonso (Universidade de Évora)
Anabela Marques (Instituto Politécnico de Setúbal)
Carla Henriques (Instituto Politécnico de Viseu)
Carlos Ferreira (Universidade de Aveiro)
Carlos Soares (Universidade do Porto)
Catarina Marques (ISCTE - Instituto Universitário de Lisboa)
Conceição Amado (Universidade de Lisboa)
Conceição Rocha (INESC-TEC e Universidade do Porto)
Fátima Salgueiro (ISCTE - Instituto Universitário de Lisboa)
Fernanda Otília Figueiredo (Universidade do Porto)

Fernanda Sousa (Universidade do Porto)
Helena Bacelar-Nicolau (Universidade de Lisboa)
Irene Oliveira (Universidade de Trás-os-Montes e Alto Douro)
Isabel Silva Magalhães (Universidade do Porto)
José Gonçalves Dias (ISCTE - Instituto Universitário de Lisboa)
Luís Miguel Grilo (Instituto Politécnico de Tomar)
Manuela Neves (Universidade de Lisboa)
Margarida Cardoso (ISCTE - Instituto Universitário de Lisboa)
Maria Filomena Teodoro (Escola Naval-Marinha Portuguesa)
Paula Brito (Universidade do Porto)
Paula Vicente (ISCTE - Instituto Universitário de Lisboa)
Paulo Infante (Universidade de Évora)
Pedro Campos (Universidade do Porto)
Pedro Duarte Silva (Universidade Católica Portuguesa)
Rosário Oliveira (Universidade de Lisboa)
Susana Faria (Universidade do Minho)
Victor Lobo (Universidade Nova de Lisboa)

Contents

Program Overview

Thursday, 22 October

16:00	**Opening Session of the Meeting**	Zoom Room 1
16:30	**Plenary Session I**	Zoom Room 1
17:30	**Thematic Session I - Scholarship CLAD 2020**	Zoom Room 1

Friday, 23 October

9:00	**Parallel Session I**	Zoom Room 1 & Zoom Room 2
10:20	Break	
10:40	**Parallel Session II**	Zoom Room 1 & Zoom Room 2
12:00	**Plenary Session II**	Zoom Room 1
13:00	Lunch Time	
14:00	**Thematic Session II - Statistics Portugal - Challenges in Official Statistics IX**	Zoom Room 1
15:20	Break	
15:40	**Thematic Session III - Banco de Portugal Statistics**	Zoom Room 1

Saturday, 24 October

9:30	**Parallel Session III**	Zoom Room 1 & Zoom Room 2
11:10	Break	
11:30	**Thematic Session IV - Extreme values theory and their applications** Zoom Room 1	
13:00	Lunch Time	
14:00	**Poster Session**	Zoom Room 1
15:20	**Closing Session of the Meeting**	Zoom Room 1

Program

Thursday, 22 October

16:00 **Opening Session of the Meeting** - Zoom Room 1

16:30 **Plenary Session I** - Zoom Room 1
AI and ML: It's all about data. Data's paradox: as the value of a single datum tends to zero, the value of all data tends to infinite
José Ferreira, p. 5
Chair: Ana Lorga da Silva

17:30 **Thematic Session I - Scholarship CLAD 2020** - Zoom Room 1
Chair: Ana Sousa Ferreira

17:30 **Normalization of gait features using Multiple Regression approach to classify Fabry's Disease**
Carlos Fernandes, Flora Ferreira, Miguel Gago, Olga Azevedo, Wolfram Erlhagen and Estela Bicho, p. 11

17:50 **Parametric Joint Modelling of Longitudinal Data with Informative Dropout**
Pedro Afonso and Inês Sousa, p. 13

18:10 **Symbolic outlier detection applied to the analysis of drinking water consumption**
Pedro Borralho Gonçalo, p. 15

Friday, 23 October

9:00 **Parallel Session I**

	Zoom Room 1 **Data science applications I** Chair: Fernanda Sousa	Zoom Room 2 **Latent variables models** Chair: Paula Cristina Vicente
9:00	**Main factors of motivation in an organizational context by multivariate data analysis methods: An empirical study**, Áurea S. T. de Sousa, M. da Graça C. Batista; Sara Cabral and Helena Bacelar-Nicolau, p. 49	**A multilevel factor analysis of the cybercrime risk perception in the European Union**, Ana Gomes and José G. Dias, p. 57
9:20	**Preliminary statistical results of arugula and lamb's lettuce growth in an aquaponic system**, Fernando Sebastião, p. 51	**Deviations from normality: Effects on the goodness-of-fit of latent growth curve models**, Catarina Marques, M. de Fátima Salgueiro and Paula C. R. Vicente, p. 59
9:40	**Preliminary screening of probabilistic models for water flow measurement**, Flora Ferreira, Marisa Almeida, Duarte Silva and Wolfram Erlhagen, p. 53	**PLS-SEM to assess burnout state in industry workers**, Luís M. Grilo, Miguel Lopes, Vanda Lima, Aldina Correia and Ana Martins, p. 61
10:00	**Multiple-valued symbolic data clustering using regression mixtures of Dirichlet distributions**, José G. Dias, p. 55	**An Issue About the Improvement of an Intelligent System Design for Disaster Situations**, M. F. Teodoro, M.J. Simões Marques, I. Nunes and G. Calhamonas, p. 63
10:20	**Break**	

10:40 **Parallel Session II**

	Zoom Room 1 **Classification methods** Chair: Pedro Duarte Silva	Zoom Room 2 **Data science applications II** Chair: Paula Brito
10:40	**How perfect is a composite reference standard? A biomedical challenge**, Ana Subtil, M. Rosário Oliveira and António Pacheco, p. 65	**Case study: Glycemic control in Type 2 diabetes**, Ana Cristina Matos, Carla Henriques, Sara Machado, Rui Marques and Edite Nascimento, p. 71
11:00	**How to detect the manipulation of financial statements in EU financial incentives in Portugal**, Susana Fernandes, Raul Laureano and Luís Laureano p. 67	**Statistics for communication students**, Cláudia Silvestre, and Ana Meireles, p. 73
11:20	**Hyperband for clustering**, Diogo Alves, Carlos Soares and Paula Brito, p. 69	**Student motivations in choosing the country for Erasmus**, Carla Henriques, Suzanne Amaro, Cristina Barroco and Joaquim Antunes, p. 75

12:00 **Plenary Session II** - Zoom Room 1
The MELODIC family for simultaneous binary logistic regression of multiple outcome variables in a reduced space
Mark de Rooij, p. 7
Chair: José G. Dias

13:00 **Lunch Time**

14:00 **Thematic Session II - Challenges in Official Statistics IX** - Zoom Room 1

Chair: Carlos Marcelo

14:00 **Road Traffic Statistics - Odometer readings Methodology**
João Barão, p. 19

14:20 **Using territorial data to define sampling of HFCS**
Catarina F. Valente, Francisco Vala and João S. Lopes, p. 21

14:40 **The National Data Infrastructure in Statistics Portugal and the data access for scientific research purposes - evolution and challenges**
José A. Pinto Martins, Francisco Lima and Maria João Zilhão, p. 23

15:00 **Data integration: Stats Business**
Cristina Neves, p. 25

15:20 **Break**

15:40 **Thematic Session Session III - Banco de Portugal Statistics** - Zoom Room 1

Chair: Luís Teles Dias

15:40 **The return-risk paradox after IFRS adoption by European listed groups**
Diogo Silva and Ana Bárbara Pinto p. 29

16:00 **Calibrating quarterly estimates with cluster analysis:The case of Portuguese firms**
Francisco Conceição, Francisco Fonseca, Mariana Oliveira and Miguel Fonseca, p. 31

16:20 **The role of investment funds' sector as a source of portfolio diversification for households: The Portuguese use case**
Pedro Miguel Alves and Sónia Mota, p. 33

16:40 **Banks' assets structure: Debt securities vs loans**
André Fernandes, Fábio Santos, Ricardo Correia and Sofia Camacho, p. 35

Saturday, 24 October

9:30 **Parallel Session III**

	Zoom Room 1 **Multivariate data analysis** Chair: Margarida Cardoso	Zoom Room 2 **Time Series Analysis** Chair: A. Manuela Gonçalves
9:30	**Parameter estimation for mixtures of linear mixed models with the EM algorithm: two different initialization strategies**, Luísa Novais and Susana Faria, p. 77	**Performance of time series forecasting models applied to economic data**, A. Manuela Gonçalves, Susana Lima and Marco Costa, p. 87
9:50	**Symbolic Sensometrics**, Paula Brito and Pedro Duarte Silva, p. 79	**Modelling censored time series of counts**, Isabel Silva, M. Eduarda Silva, Isabel Pereira and Brendan McCabe, p. 89
10:10	**Interpreting all-subsets MANOVA and Canonical Variate Analysis: The additional information biplop**, Pedro Duarte Silva, p. 81	**Study of the variation of loans granted to families between December 2009 and July 2019**, João Lamy Gil, Joana I. Ramalho, Vasco Miguel S. Barata and Ana Lorga da Silva, p. 91
10:30	**Community Detection in Interval-Weighted Networks**, Hélder Alves, Paula Brito and Pedro Campos, p. 83	**Understand time series uncertainty: A first approach**, M. Almeida Silva, Conceição Amado, Dália Loureiro and Álvaro Ribeiro, p. 63
10:50	**Reducing Dimensionality in Multi-Layer Networks through Factorial Techniques**, Pedro Campos and Patrícia Gonçalves, p. 85	
11:10	**Break**	

11:30	**Thematic Session Session IV - CLAD - SPE** - Zoom Room 1 **Extremes values theory and their applications** Chair: Adelaide Figueiredo
11:30	**Statistics of extremes and possible earthquakes' prediction** M. Ivette Gomes and Lígia Henriques-Rodrigues, p. 39
11:50	**Extreme value parameter estimation and the role of computational procedures** M. Manuela Neves, p. 41
12:10	**Extremal index estimation: An application** Dora Prata Gomes and M. Manuela Neves, p. 43
12:30	**Linear combinations of generalized Hill estimators** Fernanda O. Figueiredo and M. Ivette Gomes, p. 45
12:50	**Lunch Time**

14:00 **Poster Session** - Zoom Room 1

Chair: Pedro Campos

15:30 **Closing Session of the Meeting** - Sala Zoom 1

Abstracts

Plenary Sessions

22 October, 16:30 - 17:30, Zoom Room 1

AI and ML: It's all about data. Data's paradox: as the value of a single datum tends to zero, the value of all data tends to infinite.

José Luis Ferreira

Quidgest
jose.ferreira@quidgest.com

The growing presence of Artificial Intelligence in the way we live, by the use of technology, is an undeniable fact, even if we don't always realize it. Data is all around, growing faster than ever, feeding the intelligent algorithms, in the most diverse ways. As the amount of data tends to infinite, our attention to details tends to zero, and we rely on algorithms to make sense of data, to sustain our decisions. We freely give our data (give it no value), yet owning data means power. Precision is no longer a must, speed is, and AI is the weapon in this new way of living.

---

23 October, 12:00 - 13:00, Zoom Room 1

The MELODIC family for simultaneous binary logistic regression of multiple outcome variables in a reduced space

Mark de Rooij

Methodology and Statistics Unit, Leiden University
ROOIJM@fsw.leidenuniv.nl

Logistic regression is a commonly used method for binary classification. Oftentimes, researchers have more than a single binary response variable and simultaneous analysis is beneficial because it provides insight into the dependencies among response variables as well as between the predictor variables and the responses. In this paper we propose the MELODIC family for simultaneous binary logistic regression modeling. In this family the regression models are defined in an Euclidean space of reduced dimension, based on a distance rule. The model may be interpreted in terms of logistic regression coefficients or in terms of a biplot. We discuss a fast MM algorithm for parameter estimation. Two applications are shown in detail: one relating personality characteristics to drug consumption profiles and one relating personality characteristics to depressive and anxiety disorders.

Thematic Session: CLAD 2020 Scholarship

22 October, 17:30 - 17:50, Zoom Room 1

Normalization of gait features using Multiple Regression Approach to classify Fabry's Disease

Carlos Fernandes[1], **Flora Ferreira**[2], **Miguel Gago**[3], **Olga Azevedo**[3], **Wolfram Erlhagen**[2], **Estela Bicho**[1]

[1] Algoritmi Center, Dept. of Industrial Electronics, University of Minho, carlos.rafael.fernandes@hotmail.com, estela.bicho@dei.uminho.pt

[2] Center of Mathematics, University of Minho flora.ferreira@gmail.com, wolfram.erlhagen@math.uminho.pt

[3] Neurology and Cardiology Service, Hospital Senhora da Oliveira miguelfgago@yahoo.com, olgaazevedo@hospitaldeguimaraes.min-saude.pt

---

The aim of this study is to use a multiple regression normalization strategy that accounts for subject age, height, weight, sex, walking speed and stride length to identify differences in gait variables between Fabry disease (FD) patients and controls. The results show that multiple regression approach reduced the correlations between gait measures and physical properties, speed, and stride length, and subsequently increases the performance of learning strategies, in particular, Support Vector Machines (SVM) and Random Forest (RF). Gait normalization using MR revealed significant differences in the percentage of stance phase spent in foot flat and pushing ($\boldsymbol{p} < 0.05$), with FD presenting lower percentages in foot flat and higher in pushing. Support Vector Machines was the superior classifier achieving a classification accuracy of 77.33% after gait normalization, compared to 55.67% using raw gait data. Gait normalization significantly improved the performance of all classifiers.

Keywords: walking, Machine learning, Multiple regression models, Fabry's disease

---

There has been growing evidence showing that gait assessment can be a powerful complementary tool in the diagnosis and management of patients with motor impairments [2]. However, gait characteristics of a subject are affected by his physical properties including age, gender, height, and weight, as well as by walking speed [3]. In [3] a multiple regression (MR) normalization method was employed on gait data to minimize the effect of inter-subject physical differences and self-selected speed thereby improving gait classification accuracy using machine learning methods [3]. It has been shown that the accuracy of Parkinson's disease diagnosis using Support Vector Machines (SVMs) and Random Forests (RFs) approaches improves from 81% to 89% and 75% to 93%, respectively, when gait data is normalized using the MR approach [3].

The aim of this study is to evaluate the effectiveness of machine learning strategies when distinguishing patients with FD from healthy controls based on normalized gait features.

As in [3] age, height, weight, gender, and self-selected walking speed were used as independent variables. Additionally, we also included the subjects' stride length as an independent variable, as it was shown to significantly affect foot clearance gait features [1]. Data from 36 FD patients and from 34 age-matched control subjects was acquired using two using foot-worn inertial sensors while the subjects walked a 60-meter continuous course at a self-selected walking speed. Using the control dataset, different multiple regression models were found for each gait variable considering different combinations of the independent variables. For each gait variable, the best regression model was selected based on adjusted R^2 and Akaike information criterion (AIC) values. Finally, each gait variable was normalized by dividing the raw value by the value estimated according the selected MR model (for more detail, see [3]).

Using raw data, no statistically significant differences were found. After normalization using MR approach, significant differences between controls and FD patients were observed in foot flat (mean difference: 0.11, 95%CI: [0.10;0.14], $p = .011$) and pushing (mean difference: 0.10, 95%CI: [0.08,0.12], $p = .019$), with FD presenting lower percentages in foot flat and higher in pushing. The MR normalized gait features with highest t-test scores were: foot flat, pushing, maximum toe clearance 2, minimum toe clearance, peak swing, and loading. Classification accuracy of FD gait using SVMs and RFs classifiers was highest at 77.33% and 73.67% when based on the MR normalized gait features, compared to 55.67% and 57.33% when based on raw gait features, respectively. The paired t-test revealed significant differences between the accuracy of the classifiers based on raw and MR normalized gait features, the significance values are 0.003 and 0.004 for the SVM and RF classifiers respectively.

Machine learning classifiers, in particular SVMs and RFs, based on gait variables normalized using a MR approach can reasonably support the diagnosis of FD with good accuracy.

Acknowledgements This work was partially supported by the projects NORTE-01-0145-FEDER- 000026 (DeM-Deus Ex Machina) financed by NORTE2020 and FEDER, and the Pluriannual Funding Programs of the research centres CMAT and Algoritmi.

References

[1] F. Ferreira, M. F. Gago, E. Bicho, C. Carvalho, N. Mollaei, L. Rodrigues, N. Sousa, P. P. Rodrigues, C. Ferreira, and J. Gama. Gait stride-to-stride variability and foot clearance pattern analysis in idiopathic parkinson's disease and vascular parkinsonism. *Journal of Biomechanics*, 92:98–104, 2019.

[2] K. J. Kubota, J. A. Chen, and M. A. Little. Machine learning for large-scale wearable sensor data in parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement disorders*, 31(9):1314–1326, 2016.

[3] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland. Classification of parkinson's disease gait using spatial-temporal gait features. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1794–1802, 2015.

22 October, 17:50 - 18:10, Zoom Room 1

Parametric Joint Modelling of Longitudinal Data with Informative Dropout

Pedro Afonso[1], **Inês Sousa**[2]
[1] Department of Mathematics & Centre of Molecular and Environmental Biology, University of Minho, Portugal, b8403@bio.uminho.pt
[2] Department of Mathematics & Centre of Molecular and Environmental Biology, University of Minho, Portugal, isousa@math.uminho.pt

---

A major challenge in the analysis of longitudinal data is missing data due to participants dropping out. In this work, we present a simulation tool for the R software to investigate how the characteristics of both the longitudinal dataset and missing observations influence inferences based solely on the observed data. Furthermore, we propose new correlation structures for the transformed Gaussian model [1]. This model describes the joint distribution of longitudinal and missing processes.

Keywords: longitudinal data, informative dropout, parametric joint model

---

The correct modeling of longitudinal data in the presence of missing data remains one of the greatest challenges in analyzing this type of data. If the missing mechanism is not random, the observed data may not resemble a random sample of the measurement process. This loss of information leads to reduced precision and, if not properly handled, to biased inferences and inaccurate conclusions.

In this work, a function is developed for the R software that allows the user to use a complete dataset to generate new datasets with missing observations while controlling for the missing mechanism and the overall subject dropout proportion. We conducted a simulation study using the developed function and simulated complete datasets to investigate how the characteristics of both the longitudinal dataset and missing observations influence inferences based solely on the observed data. Longitudinal data was generated from a linear mixed-effects model with a random intercept. The results show that the increase in the number of participants and the number of measurements per patient leads to a decrease in the mean percentage error (MPE) of the estimated parameters. These results suggest that, if the model considered is adequate and there is a high risk of informative dropout, the study design should consider an increase in the number of participants and/or the number of repeated measurements, to act as a counterweight to the possible dropout of some participants. However, it is important to note that our results suggest that the MPE rate of change seems to decrease as the number of participants in the study increases. Therefore, in studies involving a high number of individuals, the adjuvant of including

more participants may not justify the investment. An increased dropout rate translates to a worsening of the estimated parameters' MPE.

Diggle et al. (2008) [1] proposed a simple fully parametric model, the transformed Gaussian model (TGM), to describe the joint distribution of a longitudinal response $\mathbf{Y}_i$ and the log transformation of a single time-to-dropout $\log D_i$. The random variable D_i describes the time at which the subject i withdraws from the study for a reason linked to the longitudinal response $\mathbf{Y}_i$. The TGM assumes that the subject-specific response vector $(\mathbf{Y}_i, \log D_i)^\top$ is a realization of a multivariate Gaussian random variable. The model stands out from other approaches to joint modeling described in the literature due to its simplicity and fast computation. On the other hand, the purely empirical interpretation of the cross-correlation vector between the longitudinal measurements and time to dropout can be considered a weakness. In this work, we propose new correlation structures with a more intuitive interpretation, by introducing shared random effects between the two outcomes. The shared random effects is intended to explain unobservable characteristics that describe the association between the two processes. The inclusion of new terms allows for the factorization of the joint distribution as a random-effects model [2]. This new factorization makes it possible to apply the estimation-maximization algorithm, treating the random effects as missing data [3], to derive the maximum likelihood estimates of the model parameters, as an alternative to the differentiation of the log likelihood followed in the initial work [1]. This approach allowed us to achieve for the first time closed-form expressions for some of the model parameters when censored times are observed.

Acknowledgements This work was supported by project 028248/SAICT/2017 funded by COMPETE2020 (*Programa Operacional Competitividade e Internacionalização*) in its component FEDER (*Fundo Europeu de Desenvolvimento Regional*) and by FCT (*Fundação para a Ciência e a Tecnologia*, I.P.) in its component OE (*Orçamento de Estado*).

References

[1] Peter J. Diggle, Inês Sousa, and Amanda G. Chetwynd. Joint modelling of repeated measurements and time-to-event outcomes: The fourth armitage lecture. *Statistics in Medicine*, 2008.

[2] Inês Sousa. A review on joint modelling of longitudinal measurements and time-to-event. *Revstat Statistical Journal*, 9:57–81, 2011.

[3] Michael S. Wulfsohn and Anastasios A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

22 October, 18:10 - 18:30, Zoom Room 1

Symbolic Outlier Detection Applied to the Analysis of Drinking Water Consumption

Pedro Borralho[1], **M. Rosário Oliveira**[2], **Margarida Azeitona**[3]
[1] Baseform, Portugal and Instituto Superior Técnico, Universidade de Lisboa, pborralho.g@gmail.com
[2] CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, rosario.oliveira@tecnico.ulisboa.pt
[3] Baseform, Portugal, margarida.azeitona@baseform.com

An outlier detection method based on robust principal components for interval-valued data is developed and a simulation study is conducted to assess the performance of the method proposed. The potentialities of the developed methodologies are illustrated with their application to a real-life dataset of drinking water consumption of more than 90 000 clients, served by a large urban water supply system in a Portuguese utility.

Keywords: interval-valued data, symbolic principal component analysis, outlier detection, robust statistics, drinking water consumption

---

In 1987, Edwin Diday created a paradigm shift by introducing Symbolic Data Analysis (SDA), motivated by the need to reduce complex data without losing variability. This approach marks a transition from individual to higher-level observations and takes into consideration the inherently symbolic nature of the data.
A dataset may contain some observations that deviate considerable from the remaining data, usually designated outliers or anomalies, and we are interested in detecting them. Outliers can exist due to malicious activity or faulty data, and we may want to detect anomalies to prevent fraud detection for credit cards, for example. The motivation for this work was the prospect of creating new techniques for outlier detection in interval data, namely based on Symbolic Principal Component Analysis (SPCA).
Principal Component Analysis (PCA) is a popular statistical method for dimensionality reduction that forms new uncorrelated variables, called Principal Components (PC), by finding linear combinations of the original variables that maximise its variability. Successively, a new PC is defined as the linear combination of the original variables with the highest variance, uncorrelated with the previous PCs. The vector of weights defining the PCs are the eigenvectors of the covariance matrix of the original variables. PCA has been extended to interval-valued data by several authors. The four Symbolic PC (SPC) estimation methods we consider in this work are the centers (CPCA) and vertices (VPCA) methods, the first methods introduced, and, more recently, Complete Information PCA (CIPCA), and Symbolic Covariance PCA (SymCovPCA). These four methods follow a

common symbolic-conventional-symbolic strategy, applying PCA to a transformation of the symbolic data to conventional and then rebuilding the symbolic objects. By quantifying how the centers and the ranges contribute to the construction of the covariance matrix, [3] and [4] proposed a general formulation that unifies the four SPC estimation methods. The classical PC estimation methods can be extremely influenced by outlying observations in the conventional framework, which is naturally inherited by the symbolic counterparts, therefore the need for the proposal of robust SPC estimation methods. Since most SPC estimation methods follow a symbolic-conventional-symbolic strategy, we can apply the usual robust techniques on the conventional phase of the process, by obtaining robust estimates of location and scatter with the fast Minimum Covariance Determinant (MCD) estimator. This work explores the use of SPCA as a statistical outlier detection procedure. Similarly to the conventional approach, our methodology performs robust SPCA to estimate the SPCs and detect anomalies in the subspace spanned by the first SPCs, where Score and Orthogonal distances are used to detect outlying observations. A simulation study is conducted to assess the performance of the method proposed. The setup considered for the experiment is similar to that described in [1] for conventional data.
The potentialities of using symbolic methodologies in this context motivated the application of robust symbolic principal components to detect outliers on drinking water consumption data of more than 90 000 clients served by a large urban water supply system in a Portuguese utility. Knowledge of factors influencing water consumption is of vital importance in the planning, operation, and maintenance of water distribution systems [2]. It is essential, for example, when implementing and evaluating the impact of actions for efficient water use or when predicting consumption in new regions. The outlier detection procedure is used to identify groups of clients with anomalous consumption behaviours. Using the geographic identification of the clients provided by the utility, the results are also represented in maps through the use of the software Baseform.

References

[1] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[2] D. Loureiro. *Consumption analysis methodologies for the efficient management of water distribution systems*. PhD thesis, Instituto Superior Técnico, 2010.

[3] M. R. Oliveira, M. Vilela, A. Pacheco, R. Valadas, and P. Salvador. Extracting information from interval data using symbolic principal component analysis. *Austrian Journal of Statistics*, 46(3-4):79–87, Apr. 2017.

[4] M. Vilela. *Classical and Robust Symbolic Principal Component Analysis for Interval Data*. Master's thesis, Instituto Superior Técnico, 2015.

Thematic Session: Challenges in Official Statistics IX

23 October, 14:00 - 14:20, Zoom Room 1

Road Traffic Statistics – Odometer readings Methodology

João Barão[1]
[1] Statistics Portugal, joao.barao@ine.pt

To use centrally available data on VKm from odometer readings of technical control of vehicles, Eurostat in collaboration with Member States defined a set of indicators that can be constructed using aggregated information from odometer readings. Portugal has been complying with the roadworthiness package, and regular inspections apply to most road vehicles, with the exception of mopeds and motorcycles. So, there is information available to develop a methodology to produce statistics on vehicles-km performed by road vehicles using odometer readings from regular inspections. A specific methodology was developed by Statistics Portugal in order to be able to produce statistics on vehicles-km performed by road vehicles using odometer readings from regular inspections. The results produced were based on data provided by the Portuguese Road Agency, the Instituto da Mobilidade e dos Transportes (IMT).

Keywords: odometer, road, statistics, traffic

Currently there is no regulated and harmonised data collection of road traffic statistics and Eurostat is relying on the voluntary collection for the production of road traffic data.
The White Paper on Mobility highlighted the need to produce statistics on vehicle-kilometres (VKm) performed by road transport vehicles.
An important transport policy development in the European Union was the adoption of Commission's package on roadworthiness test of road vehicles. This package includes three new Directives to strengthen and harmonize technical inspections and road safety. One of this is Directive 45\2014, that introduces an obligation for Member states to centralize Vehicles Kilometers (Vkm) data obtained during technical control of vehicles.
To use centrally available data on VKm from odometer readings of technical control of vehicles, Eurostat in collaboration with Member States defined a set of indicators that can be constructed using aggregated information from odometer readings.
Portugal has been complying with the roadworthiness package and regular inspections apply to most road vehicles, with the exception of mopeds and motorcycles. Since 2011, and after a period of 8 years without available data, Statistics Portugal started to produce statistics on the stock of vehicles using the national vehicles register, with data provided by the Portuguese Road Agency. A similar approach was followed for vehicle-kilometres for road transport. In fact, Portugal has been complying with the roadworthiness package, and regular inspections apply to most road vehicles, with the exception of mopeds and

motorcycles. So, most of the information to be used in the production of this new type of statistics is already available.
Taking into account Eurostat recomendations and the definition of grants to support new developments under this area, Statistics Portugal developped a methodology to produce statistics on vehicles-km performed by road vehicles using odometer readings from regular inspections. The results produced were based on data provided by the Portuguese Road Agency, the Instituto da Mobilidade e dos Transportes (IMT).
The main objective was to identify problems with the odometer readings database and to create a methodology to produce results (estimations) concerning the vehicle-kilometres for road transport.
The most important milestones were related the conception of an estimation model for km performed by vehicles of the national road stock, taking into account the category of the vehicle, age and type of fuel. The proposed estimation model has different modules for different approaches, namely: - Vkm calculation from successive databases of subsequent years; - Vkm estimation for vehicles not yet subjected to inspections; - Vkm estimation for vehicles under a two-year break between inspections. It was possible to compute results for the period 2015-2018, corresponding to road traffic by the national stock had in consideration the vehicle category, age and type of fuel; those variables are already in the registers and can be used in the model.
In terms of data integration, the main purpose will be to complete the integration of all the data received from the IMT in the Data Warehouse, to be used as a repository for historical data and for the regular reception and updates of data.
Given the possibility to have access to updated data, a revisions process will need also to be further developed, in order to have access to all versions of data, to produce indicators concerning the degree of revisions, and to include, if needed, a coefficient in the estimation methodology, to better forecast the revisions.
The Data Warehouse will be the tool to be used (in a business objects environment) for the analysis of all variables and also for the longitudinal links between databases of consecutive time periods. This tool will be also used for the computation of the regular vehicle-km estimations, and for the production of the corresponded statistical indicators.
In conclusion, the results from the application of the defined methodology were very auspicious, and the production of regular results on the vehicles-km will be a reality for Statistics Portugal. Additional contacts with IMT will occur, in order to improve the administrative data available and to further analyse the possibility to provide it, in a regular and, if possible, infra-annual basis and with more territoriall breakdown or by type of road.

23 October, 14:20 - 14:40, Zoom Room 1

Using territorial data to define sampling of HFCS

Catarina F. Valente[1], **Francisco Vala**[2], **João S. Lopes**[3]
[1] Lisbon School of Economics & Management, catarinaifvalente@aln.iseg.ulisboa.pt
[2] Statistics Portugal, francisco.vala@ine.pt
[3] Statistics Portugal, joao.lopes@ine.pt

Household Finance and Consumption Survey was implemented to study economic inequality, while considering both income and net wealth. Its sampling procedure needs to account for asymmetries on these indicators. We propose a sampling scheme defining wealthiest classes at region-level (using national registries of Personal Income Tax and sale & purchase of real estates), coupled with previously information at household-level. We show that the new scheme greatly increases sampling efficiency.

Keywords: HFCS, wealth classes, territorial information, hierarchical clustering

---

Inequality studies have been mainly focused on wage income inequalities. However accounting for capital income is becoming increasingly important, especially regarding the wealthiest classes. Moreover, inheritances are returning to be a main factor in shaping economic inequality [1]. Household Finance and Consumption Survey (HFCS) has been implemented to gather information on income, assets and liabilities at household-level, thus allowing for in-depth studies on economic inequality. These studies are particularly important for the establishment of monetary and financial stability policies. Indeed, the characterization of income and wealth distributions is invaluable to understand macroeconomic shocks [2].

The distributions of income and wealth are typically skewed right, making them difficult to characterize correctly. Therefore the sampling scheme of HFCS should be carefully designed to take into account these asymmetries [3]. A poor characterization of the wealthiest classes, for example, can lead to underestimation of indexes of inequality and concentration. In order to overcome these issues, the European Central Bank suggested sampling a large proportion of wealthy households. In Portugal, the HFCS has been collected by over-sampling the wealthiest classes. From 2013 onwards, the over-sampling has been done using the Useful Area of the Household Main Residence (HMR) – a variable that has been shown to be highly correlated to both income and wealth of a household. Using information on Useful Area, we were able to define wealthiest classes at household-level.

In the present document, we studied the use of an alternative sampling scheme to over-sample the wealthiest classes. This sampling scheme should provide a more efficient sampling, allowing for a reduction of the sampling effort of the survey. Hence, we propose the use of wealthy regions defined using economic information at regional-level from two

data sources: Liquidation statements of the Personal Income Tax in 2016; and Registry of sale and purchase of real estates in 2017 [4]. These regions were created using a hierarchical clustering procedure, and were characterized by summary statistics so that they could be defined as wealthy or non-wealthy (Table 1). Prior this characterization, a merging procedure was implemented in order to enrich the information of the regional units.

Table 1: Characterization of regional-level clusters

	A	B	C	Total
Average income	32,077.31	28,742.79	19,394.30	21,726.39
Average HMR value	2,836.69	1,271.39	870.83	1,055.44
Regional units	54	153	700	907
Households	484,789	1,068,836	2,376,299	3,948,552

The use of the alternative territorial-based sampling scheme was evaluated using samples collected in the last wave of the HFCS (i.e. ISFF2017), and by considering the following comparison: (a) wealthy at household and region-levels vs. wealthy at household-level; (b) wealthy at region-levels vs. non-wealthy at region-level (within wealthy at household-level). The comparison was performed using summary statistics (central tendency and dispersion) and the two-sample Kolmogorov–Smirnov test, as well as classical economic indexes (e.g. Gini coefficient, Atkinson's index, Palma ratio). Each comparison was performed on the economic indicators gross income, net wealth, financial assets and real assets distributions. We found that for all the indicators the use of economic information at household-level and/or regional-level greatly increases the efficiency of sampling the wealthiest classes, however, information at household-level seems particularly important in the case of net wealth and financial and real assets. As for the definition of wealthy regions, choosing the two wealthiest clusters obtained by the hierarchical clustering seems to be a good compromise between sampling efficiency and number of available households to keep the sampling randomness.

References

[1] Atkinson A. B. Piketty T. Alvaredo, F. and E. Saez. The Top 1 Percent in International and Historical Perspective. *J Econom Persp*, 27(3):3–20, 2013.

[2] Household Finance and Consumption Network. The Household Finance and Consumption Survey: methodological report for the first wave. *ECB Statistical Paper Series No. 1*, 2013.

[3] A. B. Kennickell. Darkness Made Visible: Field Management and Nonresponse in the 2004 SCF. *FRB Working Paper*, 2005.

[4] C. F. Valente. Identificação e delimitação de territórios homogéneos de residentes com níveis de rendimento e património mais elevados, 2019.

23 October, 14:40 - 15:00, Zoom Room 1

The National Data Infrastructure in Statistics Portugal and the data access for scientific research purposes – evolution and challenges

José A. Pinto Martins[1], **Francisco Lima**[2], **Maria João Zilhão**[3]
[1] Statistics Portugal, pinto.martins@ine.pt
[2] Statistics Portugal, francisco.lima@ine.pt
[3] Statistics Portugal, mjoao.zilhao@ine.pt

The access to anonymized official microdata by researchers in Portugal is guaranteed by Statistics Portugal (SP).

The paper presents SP project for a National Data Infrastructure (NDI) already in implementation and that will allow data access maximization for the production of Official statistics and for research purposes.

Keywords: microdata databases, anonymized individual statistical data, free access, NDI

SP as the principal national statistical authority belonging to the European Statistical System has specific and unique competences that guarantee the independence and security of information. Within this context, the evolution towards a National Data Infrastructure is a logical and natural step for SP that is being gradually implemented.
The increasing digitization generates a large volume of data. Their transformation into knowledge is only possible with adequate processes of storage, treatment and analysis, in a safe, consistent and reliable way.
SP already uses considerable amounts of administrative data in its production and statistical analysis processes. As a national statistical authority, it has the proper structure that ensures the protection and integrity of the data, much like a data vault. The new information requirements led to the need to provide Public Administration with the capacity to manage and analyze large data sets and integrating this need into SP, evolving into a national data infrastructure, has obvious gains in scale.
Taking advantage of SP competencies, tasks and mission, the objective is to adopt a more intensive and integrated use of data in the production of statistical information and to take advantage of the entire SP production chain, from the development of platforms, applications and algorithms to the collection and validation of data, up to the analysis of statistical information. It is a cumulative process that will foster the development of new skills, gaining resources, space to intensify innovation throughout the entire organization, and providing higher returns to society.
The NDI will seek to respond to the need for SP scale up and gain critical mass in order to respond to an increasingly complex society that generates new expectations regarding

statistics. New services and statistical products are sought, with new approaches, with a guarantee of quality.
The development of the NDI will ensure SP the critical dimension to continue developing its skills and improving statistical production, benefiting the country by the increased processing and analysis capacity. Intensification of ownership and use of administrative data in SP production process anticipates a large increase in the volume of data and a substantial broadening of the covered areas.
As such, the NDI main purpose and associated benefits, among others, is to provide a set of related data and resources from a single point of entry, regardless of where the data is kept or how data can be accessed (opened, protected or safe) and assure its safety and quality by providing integrated data services and metadata.
The information available at the level of anonymized microdata with multiple intersections and scientific research outputs gains a new dimension and opens up numerous opportunities for partnerships and knowledge sharing, while assuring compliance with the statistical law, and in particular with respect to statistical confidentiality.
In summary, we can say that the NDI will increase the economic and social impact of statistical information as a public good.
Access to confidential data for scientific and research purposes is a key concern for SP.
A wider access to confidential data for scientific work and research without compromising the high level of secrecy protection will allow researchers in depth analysis.
SP commitment is to provide good information, and researchers' responsibility is to use data in full respect for the fundamental Principle of Statistical Confidentiality.
SP GOOD experience shows that COOPERATION for this specific purpose is an advantage: it allows, namely, split and share responsibilities' among parties with specific knowledge, and definitely providing a more efficient public service within the National Statistical System.

23 October, 15:00 - 15:20, Zoom Room 1

Data integration: Stats Business

Cristina Neves[1]
[1] Statistics Portugal, cristina.neves@ine.pt

Statistics Portugal has available a set of information about businesses which, through microdata linking, enables the data integration of different data sources, allowing a complete perspective on enterprises' economic performance. The most important set of information becomes from the IES – Simplified Business Information, corresponding to an administrative source containing all the annual accounting data for all the enterprises. Statistics Portugal has been developing a broader set of new statistical operations intended to disseminate information on diverse factors not covered by the European Statistical Program, but very relevant to capture different and important elements on business dynamics in an international context which poses new challenges and the need of constant attention to factors affecting their competitiveness: Perspectives of Exports of Goods; Management Practices; Framework Regulation Costs; Identification of Enterprises' Skills Requirements. All this information on the enterprises and additional available statistical data sources such as "Quadros de Pessoal", Community Innovation Survey, International Trade in Goods Statistics, Prodcom Survey, Survey on Information and Communication Technologies Usage in Enterprises are being used and linked at the microdata level, in order to allow the data integration and the construction of a Stats Business database. Thus, additional statistical indicators and analysis on business data is being produced, without increasing the statistical burden on enterprises.

Keywords: data integration, business, microdata linking, statistics

---

Since 2006 the compilation of Structural Business Statistics in Portugal is based on an administrative source available at Statistics Portugal (the Simplified Business Information - IES stands for "Informação Empresarial Simplificada"), created in 2006 within the framework of a government program for the simplification and modernisation of Public Administration named the SIMPLEX program, which aggregates the fulfilment of several legal obligations by the enterprises in a single act and only by electronic means that were previously dispersed and that implied the provision of information sometimes materially identical to different organisms of the Public Administration through different channels.
The annual accounting data for all the enterprises is only obtained by this way, without additional data collection via statistical surveys. Additionally, annual surveys on sectoral activities (Industry, Construction, Retail trade, Services, etc.) also benefit from IES data, both in the sample design (opportunity of having a complete knowledge of all the active

enterprises) and in the production of final results, with significant impacts in terms of quality while ensuring the consistency with global results for the total of the Portuguese business framework.

Statistics Portugal has been developing a broader set of new statistical operations intended to disseminate information on diverse factors not covered by the European Statistical Program, but very relevant to capture different and important elements on business dynamics in an international context which poses new challenges and the need of constant attention to factors affecting their competitiveness.

The implementation of new business surveys in Portugal does not impose a significant burden on respondents given the IES, which is used to reduce the sample size of surveys on enterprises, allowing a sound level of quality on data. Those new statistical operations that are being conducted by Statistics Portugal in order to get information on potential sources of economic growth correspond to:

i) Survey on Perspectives of Exports of Goods over exporting enterprises to obtain information about their activity expectations. This project took form in a context of increasing importance of Portuguese exports in the path towards recovery from the economic and financial crisis, and nowadays it keeps being important due to Portuguese exports are still playing a fundamental role to the correction of macroeconomic imbalances and to the stimulus of the Portuguese general businesses.

ii) Management Practices Survey, with the main purpose to obtain information on management practices, a survey mostly of a qualitative nature, addressed to top managers and dedicated to obtain information on characteristics of corporations' management that, although with no explicit monetary reflection on their financial statements, may constrain or foster their competitiveness in a context of growing integration within the overall economy.

iii) Survey on Framework Regulation Costs, a survey related to several types of costs affecting firms' performance besides wages and other inputs costs, conducted by SP with the main purpose of assessing these effects and it focused on nine potential areas of obstacle to businesses' activities: starting a business, licensing, costs of services produced by network industries, financing, judicial system, tax system, administrative burden, internationalisation and human resources.

iv) Survey on the Identification of Enterprises' Skills Requirements, a survey on the types of human resources qualifications needed by corporations in the near future in order to detect particular labour shortages.

All this information on the enterprises and additional available statistical data sources such as "Quadros de Pessoal", Community Innovation Survey, International Trade in Goods Statistics, Prodcom Survey (annual data on industrial production), Survey on Information and Communication Technologies Usage in Enterprises are being used and linked at the microdata level, in order to allow the data integration and the construction of a Stats Business database. Thus, additional statistical indicators and analysis on business data is being produced, without increasing the statistical burden on enterprises. This information is also available as anonymised statistical microdata, which may be supplied for scientific purposes to accredited researchers.

Thematic Session: Banco de Portugal Statistics

23 October, 15:40 - 16:00, Zoom Room 1

The return-risk paradox after IFRS adoption by European listed groups

Diogo Silva[1], **Ana Bárbara Pinto**[2],
[1] Banco de Portugal, dfsilva@bportugal.pt
[2] Banco de Portugal, apinto@bportugal.pt

Bowman (1980) found a negative association between corporate returns and corporate risk, which became known as the return-risk paradox because it is inconsistent with the positive association that is well documented in finance literature, namely the modern portfolio theory [5] and the CAPM [6]. This analysis tests the existence of the return-risk paradox in European listed non-finantial groups after the IFRS adoption and addresses its drivers by computing a hierarchical clustering analysis. The Ward's method is used for similarity parametrization, clusters are exclusive and the clustering process is complete. The analysis uses consolidated annual data available at ERICA database for the years of 2005 to 2018. The higher-risk cluster tends to perform worse, while the lower-risk cluster performs better, which is evidence of the return-risk paradox. Groups from the lower-risk and better performing cluster are more efficient and invest more.

Keywords: Return-risk paradox, hierarchical clustering

This analysis tests the existence of the return-risk paradox in European listed non-financial groups after the IFRS adoption and addresses its drivers. Bowman (1980) found a negative association between corporate returns and corporate risk [2]. This association became known as the return-risk paradox because it is inconsistent with the positive association between returns and risk that is well documented in finance literature [6]. Overall, a firm with lower risks and higher operational returns (negative association) may have its equity priced relatively higher in equity markets, lowering its return to the investors who buy it (positive association) [2]. Hence, the association between returns and risk may be different depending if one looks from a management perspective at groups' operational activity (negative association) or from a finance standpoint at equity markets (positive association). Another difference is related with the importance atributted to specific risk. The modern portfolio theory showed that one can diversify specific risk [5]. Thus, in line with the Capital Asset Pricing Model, returns depend positively on systematic risk, not on specific risk, because it is diversiafiable [6]. Still, strategic management literature actually suggests that managers are quite concerned about handling specific risk [1].
There are three theories that provide enligthment regarding the return-risk paradox. The prospect theory suggests that individuals measure outcomes relative to a reference point

and that individuals that are below their reference point tend to be risk seekers, while individuals above their reference point are risk avoiders [3]. The behavioral theory of the firm assumes that groups have an aspiration in terms of performance (similar to the reference point in the prospect theory). Groups with performance below their competitors will aspire to improve. Groups that desire to improve will take action. Taking action involves assuming risks. The agency theory accepts that managers may adopt strategies that are consistent with their individual preferences. For instance, managers' stock ownership is associated with more operational efficiency [4].

This analysis uses consolidated annual data available at ERICA database from 2005 until 2018. The database has data for non-financial listed groups of Austria, Belgium, France, Italy, Germany, Greece, Portugal, Spain and Turkey.

Two methods are applied to test the return-risk paradox. The standard approach consists in computing the correlation between a measure of returns and its volatility. This study also considers a completely different approach by computing a hierarchical clustering analysis. Clusters are complete and exclusive. The Ward's method is applied for similarity parametrization. Both methodologies point to the persistence of the return-risk paradox after IFRS adoption by European listed groups. The results suggest that the groups with less risk and higher returns tend to be more operationally efficient. The cash-flow from investing activities tends to be lower (usually more negative) for the groups from the lower-risk and better performing cluster, indicating that these groups invest more. Although this study documents that the return-risk paradox is associated with efficiency and investment intensity, one can only hypothesize about the casual relationship between these characteristics. Future research could look at these associations in greater detail.

Disclaimer The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

References

[1] R.A. Bettis. Modern financial theory, corporate strategy and public policy: Three conundrums. *Academy of Management Review*, 8(3):406–415, 1983.

[2] E.H. Bowman. A risk/return paradox for strategic management. *Sloan Management Review (pre-1986)*, 21(3):17, 1980.

[3] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[4] U.V. Lilienfeld-Toal and S. Ruenzi. Ceo ownership, stock market performance, and managerial discretion. *The journal of finance*, 69(3):1013–1050, 2014.

[5] H. Markowitz. Portfolio selectionn. *The journal of finance*, 1952.

[6] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

23 October, 16:00 - 16:20, Zoom Room 1

Calibrating quarterly estimates with cluster analysis: The case of Portuguese firms

Francisco Conceição[1], **Francisco Fonseca**[2], **Mariana Oliveira**[3], **Miguel Fonseca**[4]

[1] Banco de Portugal, fdconceicao@bportugal.pt
[2] Banco de Portugal, ffonseca@bportugal.pt
[3] Banco de Portugal, moliveira@bportugal.pt
[4] Banco de Portugal, masfonseca@bportugal.pt

This paper illustrates the potential benefits of using cluster analysis to improve the precision of sample estimates. The analysis is focused on the case of Portuguese firms, for which Banco de Portugal publishes quarterly statistics based on a sample, and annual statistics based on information with full coverage of non-financial corporations operating in Portugal. For the variables selected in this study, preliminary findings suggest that using a cluster analysis to calibrate the extrapolated values of a sample could bring valuable benefits, comparing to a calibration reliant on stratification.

Keywords: Corporations, Clusters, Extrapolation, Post-stratification, Sampling

---

Banco de Portugal publishes information of Portuguese non-financial h annual data is the IES (Simplified Corporate Information), which provides access to a wide range of information (mainly the financial statements) with full coverage of non-financial corporations operating in Portugal. On the other hand, quarterly statistics are compiled through the extrapolation of approximately four thousand responses submitted through the Quarterly Survey on Non-Financial Corporations (ITENF). The methodology involves selecting a representative sample from a sampling frame (stratum) provided by INE [3], and extrapolating the results in order to obtain population estimates for the variables of interest.
The corporations selected in the sample are then attributed an extrapolation factor (the inverse of its probability of selection) which is then applied to the reported values. After this stage, adjustments are made into the original stratification and a ratio estimator is applied to each *post-strata* in order to improve the precision of the final estimates. In other words, the extrapolated values are calibrated by the ratio estimator, which consists of re-weighting the observations in order to approximate the estimates to the population totals in each *post-strata* (for more details, please refer to [1]).
In this paper, we explore an alternative method to calibrate the extrapolated estimates by allowing the ratio estimator to be based on clusters instead of strata defined according to the sample design. With this procedure we expect to reduce the estimation error, since dividing the annual IES data into clusters and not into strata can lead to the identification of natural groups within the population.

We performed a preliminary analysis to test the proposed ratio estimator for two variables: trade payables and trade receivables (currently stratified by imports and exports, respectively, as ancillary variables). We calculated the extrapolated values for the last quarter of 2018 according to each calibration procedure (strata versus cluster based ratio estimator) and confronted those results with the population totals. We have also computed a performance indicator that assumes positive (or negative) values in the presence of a relative decrease (or increase) of the square errors, respectively [2].
For the cluster analysis, using the k-means algorithm and data from IES 2018, two financial ratios were chosen: imports as a percentage of purchases and exports as a percentage of turnover. Firms were divided into several clusters (from 2 to 100), in order to test different scenarios. Figure 1 displays the aggregated performance indicator of both trade payables and trade receivables plotted against the number of clusters (from 2 to 25).
In this case, for a partition between 2 and 8 groups, the cluster based ratio estimator outperforms the current one by almost 15%. When we consider more than 8 clusters, the behaviour becomes erratic, but in general we observe either marginal increases or gross decreases in performance. This conclusion is valid until 100 clusters (maximum tested).

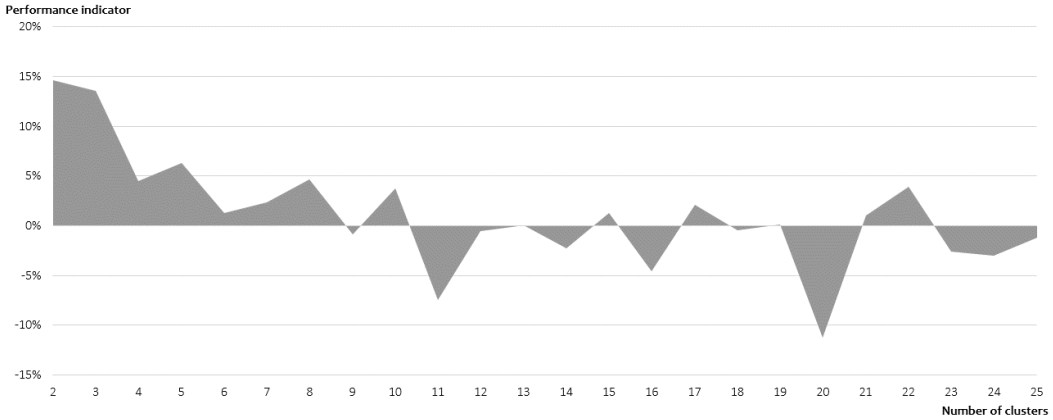

Figure 1: Performance indicator by partition of clusters

Future developments of this analysis involve broadening the cluster analysis to more clustering variables (to assess if there are marginal gains), applying this procedure to other variables of interest and extending the time span of the anaBlysis.

Disclaimer: The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

References

[1] Banco de Portugal. Statistics on non-financial corporations of the central balance sheet database – metodological notes. Supplement to the Statistical Bulletin 2, 2013.

[2] Banco de Portugal. Sector tables and enterprise and sector tables. Central Balance-Sheet Studies 36, February 2019.

[3] INE. Documento metodológico – inquérito trimestral às empresas não financeiras. Technical Report v.2.2, July 2014.

23 October, 16:20 - 16:40, Zoom Room 1

The role of investment funds' sector as a source of portfolio diversification for households: the Portuguese use case

Pedro Alves[1], **Sónia Mota**[2]
[1] Banco de Portugal, pmalves@bportugal.pt
[2] Banco de Portugal, scmota@bportugal.pt

By investing in investment funds (IF), households are indirectly investing in several financial assets, being able to access a widespread of financial markets. Their indirect holdings through IF are significantly different from their direct holdings, mainly explained by the greater weight of debt securities and equity on their indirect portfolio. Since indirect holdings via IF just represent circa 4% of total financial assets, the diversification effect is limited.

Keywords: investment funds, households' holdings

Households have at their disposal a set of alternative applications for their savings other than solely deposits, namely the investment in IF units. As financial intermediaries, IF channel funds between third parties with a surplus and those with a lack of funds. Looking at the composition of financial assets held by households (figure 1), it is possible to conclude that over time deposits are the most preferred asset, while debt securities and equity just represent 15% at the end of 2010 and 11% at the end of 2019. Regarding Portuguese IF, it represents only circa 4% over time, while non-Portuguese IF represent just 1%.

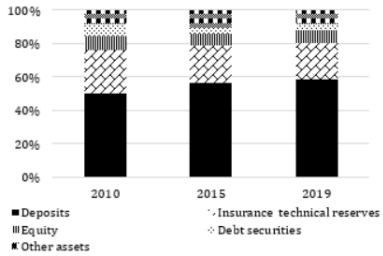

Figure 1: Composition of financial assets held by households; Source:Banco de Portugal

Despite of the weak representativeness of the IF units in the portfolio of households, they are the main holders of IF units issued by Portuguese IF. Analyzing the investment made by Portuguese households in Portuguese IF, at the end of 2010 and 2019, households held, respectively, 14,500 million euros and 12,700 million euros of IF units. Between 2010 and 2014, accumulated disinvestment reached about 7,600 million euros. In contrast, from 2015 to 2019, the accumulated investment amounts to about 3,900 million euros. One of the main purpose of investing in IF relates to the possibility of households to invest indirectly

in several financial assets, thus diversifying their portfolio. Looking at the decomposition of indirectly assets' holdings of households through the investment in Portuguese IF (figure 2), in December 2019, households are indirectly holding circa 4,900 million euros in equity shares and close to 4,500 million euros in debt securities. Furthermore, households also hold over 2,100 million euros in real estate assets and circa 1,700 million euros in deposits.

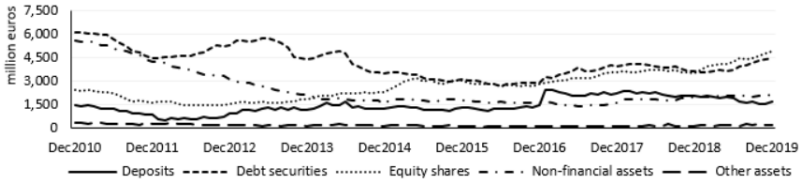

Figure 2: Development of indirectly assets' holdings of households through IFs; Source: Banco de Portugal

Comparing the direct and indirect investment of households, as realized by the European Fund and Asset Management Association [1], we can conclude that the composition of households' holdings is different comparing the two approaches (figure 3). Considering direct holdings, 60% of the portfolio is applied in deposits, 7% in equity and 3% in debt securities. On the other hand, indirectly investing in Portuguese IF, households are diversifying their financial assets, since 70% of those investment is channeled to debt securities and equity. Furthermore, through investment in real estate funds households have being able to invest in commercial and residential real estate.

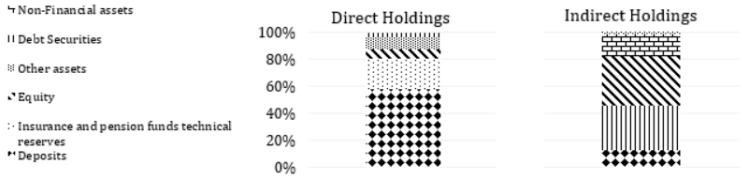

Figure 3: Direct and indirect households' holdings at the end of 2019;Source:Banco de Portugal

Taking in consideration data for households from financial accounts' statistics, it is possible to conclude that investment in Portuguese IF just represent circa 4% of total financial assets held by households in 2019. Thus, the combined portfolio of indirect plus direct holdings (excluding Portuguese IF) is not different from the composition of direct holdings, being the diversification effect limited. As a final point, by investing in Portuguese IF, households are also indirectly investing in non-Portuguese IF. Therefore, further research could be done in order to detail the direct and indirect investment in non-Portuguese IF.

Disclaimer: The analyses, opinions and findings of this paper represent the views of the author, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors. We are thankful to Filipa Lima and Luís Teles for comments and suggestions.

References

[1] European Fund and Asset Management Association. Ownership of investment funds in europe. Technical report, Brussels, 2019.

23 October, 16:40 - 17:00, Zoom Room 1

Banks' assets structure: debt securities vs loans

André Fernandes[1], **Fábio Santos**[2], **Ricardo Correia**[3], **Sofia Camacho**[4]
[1] Banco de Portugal, agfernandes@bportugal.pt
[2] Banco de Portugal, fpsantos@bportugal.pt
[3] Banco de Portugal, rncorreia@bportugal.pt
[4] Banco de Portugal, scamacho@bportugal.pt

After the 2007 financial and economic crisis, which was followed by a sovereign debt crisis in some countries of the euro area, two different behaviors emerged: while the banks from some euro area peripheral countries increased their investment in debt securities and tightened their loans activity, the banks from Germany presented the opposite trend. It can also be observed that after the recession of 2007, the total lending activity of banks decreased.

Keywords: deficit, public debt, interest rate spreads, banks' assets structure

In 2007, an economic and financial crisis emerged in the USA, triggering a global recession in the years that followed. Regarding Europe, this recession had a significant impact on the public finances of several EU countries and sparked a sovereign debt crisis in the GIIPS (Greece, Ireland, Italy, Portugal and Spain). During this period, part of the increase in the public debt stock is explained by the need in financing government deficits, which were at an all-time high.

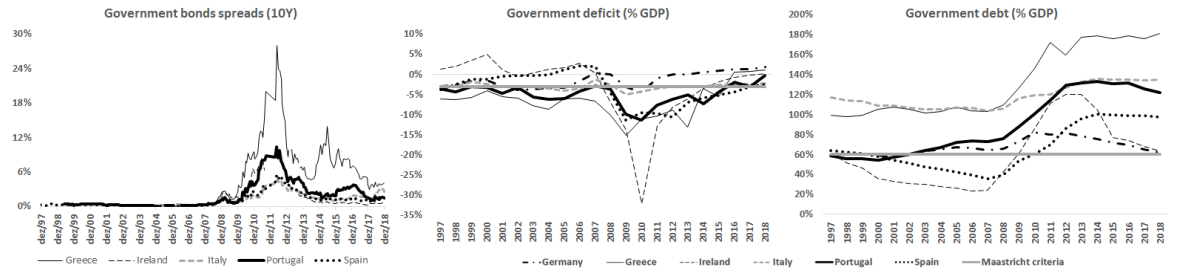

Figure 1: Government spreads, deficit and debt of GIIPS and Germany. Source: Banco de Portugal, Eurostat and Bloomberg.

This economic and financial crisis also had an impact on the banks' lending composition. As illustrated in Figure 2, until 2007, Portugal, Italy, Greece and Spain, presented a decrease in the weight of debt securities in the banks' total lending. In contrast, the loan activity has registered a significant increase regarding the banks' total assets, especially in the case of both Portugal and Greece, going from 31% to 57% and 47% to 62%, respectively.

However, from 2008 onwards, the trend was reversed with the weight of loans decreasing and a subsequent increase of the weight of debt securities in the banks' total lending. Conversely, Germany verified a rise in the weight of the debt securities up until the 2007 financial crisis, and a decline after this period.

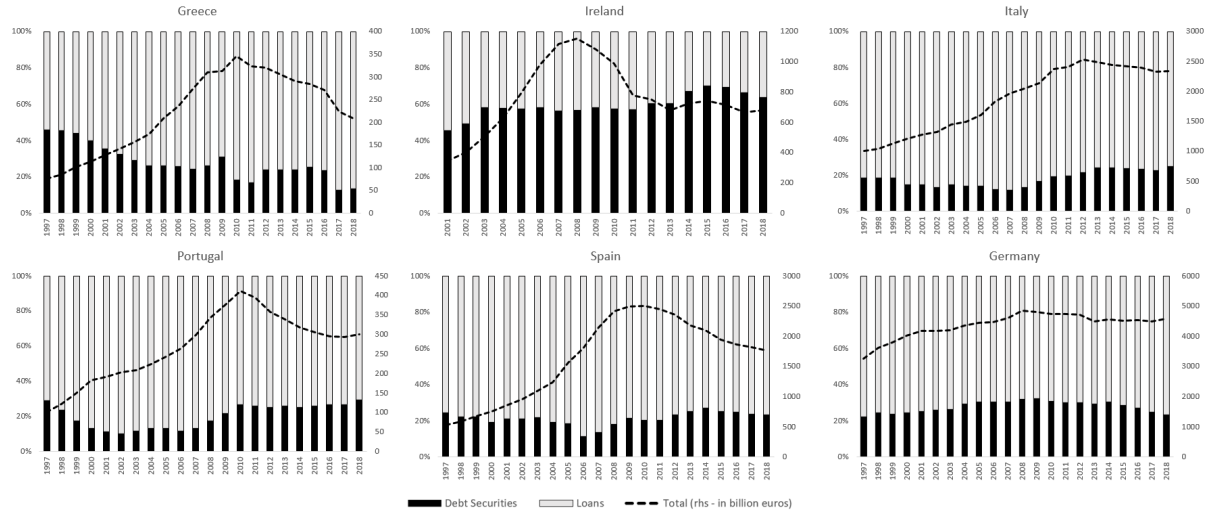

Figure 2: Banks' total lending composition. Notes: In the case of Ireland the data is only available after 2001. Source: Banco de Portugal and Eurostat.

It can also be observed that, after 2007, and despite its growing trend, the debt securities' weight has never represented more than 33% of the banks' total lending except for Ireland, which depicts a relatively larger investment in debt securities. This phenomena can be partially explained by the importance of the Irish securitization industry, as well as by the substantial activity of the money market funds (which are included in our analysis). Furthermore, for Portugal, Greece, Italy and Spain there has been a clear deleveraging of the the banks' total lending activity after the sovereign debt crisis. In the case of Ireland, the deleveraging of banks occurred earlier, namely after the financial crisis, going from a total value of 1116 billion euros in 2007, to 680 billion euros in 2018. Regarding Germany, even though there is no clear graphical evidence of deleveraging, there was also a decrease in the lending activity, with a reduction of 250 billion euros, from 2008 to 2018.
The results of our analysis suggest that after the financial crisis of 2007 there was a clear tightening of banks' total lending activity, especially for the GIIPS. After the crisis, there was also a switch in the composition of the banks' total lending, with a decrease of the loans' weight. These effects may have been induced by the crisis, the new regulatory requirements and the ECB monetary policy stimulus to credit.

Disclaimer The analyses, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

Thematic Session: Extreme values theory and their applications

24 October, 11:30 - 11:50, Zoom Room 1

Statistics of extremes and possible earthquakes' prediction

M. Ivette Gomes[1], **Lígia Henriques-Rodrigues**[2]
[1] CEAUL and DEIO, FCUL, Universidade de Lisboa, Portugal, ivette.gomes@fc.ul.pt
[2] CEAUL and Universidade de Évora, Portugal, ligiahr@uevora.pt

---

Statistics of extremes helps us to control potentially disastrous events, of a high relevance for society and a high social impact, like earthquakes. There are usually only a few observations in the tail of the distribution underlying the data. Estimates much below/above the observed minimum/maximum are required, and we thus need to consider reliable models for the tail.

Keywords: Extreme value theory, parametric and semi-parametric statistical inference, return periods

---

Statistics of extremes is a discipline that helps us to control potentially disastrous events, of a high relevance for society and a high social impact. The domains of application of statistics of extremes are quite diverse, as can be seen in [3], among others. Despite of being possible to find some historical articles related to extreme events, the field can be dated back to Gumbel, in articles published since 1935, which have been summarized in his 1958 book on statistics of extremes.

Generally speaking, the extremal types theorem provides the identification of the possible distributions of maxima with the general extreme value distributions. This type of distributions are also called *max-stable* (MS) laws, and can be defined through the functional equation $\mathrm{MS}^n(\alpha_n x + \beta_n) = \mathrm{MS}(x)$, $n \geq 1$, for $\alpha_n > 0$, $\beta_n \in R$. Under play, and for statistical purposes, we work with $\mathrm{MS}((x-\lambda)/\delta) \equiv \mathrm{MS}_\xi((x-\lambda)/\delta)$, where λ, δ, ξ are unknown location, scale and shape parameters, possibly dependent on adequate covariates, being ξ the so-called *extreme value index* (EVI). Indeed, the aforementioned models, and contrarily to the normal model, the most common model in classical statistics, provide a very reliable goodness-of-fit to data like the magnitudes of earthquakes, in the most diverse regions (see [4], [1], and [3], among others). More generally than the class of MS models, we can consider the class of *max-semi-stable* (MSS) models, introduced by Grienvich and Pancheva in 1992. The MSS distributions have revealed to be interesting to model some of the measurements associated with earthquakes, and their standard functional form is:

$$\mathrm{MSS}_{\xi,\nu}(x) = \begin{cases} \exp\left[-\nu\{\ln(1+\xi x)/\xi\}(1+\xi x)^{-1/\xi}\right], \ 1+\xi x > 0, & \text{if} \quad \xi \neq 0 \\ \exp\left\{-\nu(x)\exp(-x)\right\}, \ x \in R, & \text{if} \quad \xi = 0, \end{cases}$$

where $\nu(\cdot)$ is a positive, limited and periodic function, being $\mathrm{MS} \equiv \mathrm{MS}_\xi = \mathrm{MSS}_{\xi,1}$.

During the previous century, seismologists have observed and located millions of tremors all over the world. From these observations, what can we learn about the distribution of earthquakes in space, time, and magnitude? Can we find statistical models that accurately describe the distribution of seismicity? And, ultimately, can our models be extended into the future to make (probabilistic) forecasts or even predictions of the future seismicity? These are essentially the questions that need to be addressed by interdisciplinary researchers, including experts in *extreme value theory* (EVT).

Any map of seismicity shows that earthquakes cluster in space. However, there are occasional sequences of earthquakes that occur in places where no seismicity has been observed previously. This has largely to do with the fact that the instrumental record of seismicity stretches back barely 100 years, not enough to accurately assess the spatial distribution of earthquakes in low-seismicity areas. Nevertheless, the biggest challenge is not to identify the seismogenic regions of the earth but to estimate how frequent and how big events in a particular region might be. Data studies of this type, mainly related to earthquake's seismic moments, show that right-tails are often truncated and exhibit a high positive weight, i.e. a positive EVI ($\xi > 0$) (see [4], [1] and [2], among others). EVT can help us to predict the magnitude of future earthquakes, on the basis of the so-called return period of an extreme event, a topic to be discussed here.

We next refer one of the well-known sentences of Emil Gumbel: *'It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis'*. And to this sentence, we add the following one: '*Not only the rivers, but also the movements of the earth's crust get to know* EVT ...'

Acknowledgements Research partially supported by National Funds through **FCT** — Fundação para a Ciência e a Tecnologia, project UIDB/00006/2020 (CEA/UL).

References

[1] J. Beirlant, M.I. Fraga Alves, and Gomes M.I. Tail fitting for truncated and non-truncated pareto-type distributions. *Extremes*, 9:429–462, 2016.

[2] J. Beirlant, M.I. Fraga Alves, and T. Reykens. Fitting tails affected by truncation. *Electronic J. Statist.*, 11:2026–2065, 2017.

[3] M.I. Gomes, M.I. Fraga Alves, and Neves C. *Análise de Valores Extremos: uma Introdução*. SPE and INE, Lisbon, 2013.

[4] V.F. Pisarenko and Sornette D. Characterization of the frequency of extreme events by the generalized pareto distribution. *Pure and Applied Geophysics*, 160:2343–2364, 2003.

24 October, 11:50 - 12:10, Zoom Room 1

Extreme value parameter estimation and the role of computational procedures

M. Manuela Neves[1]
[1] Instituto Superior de Agronomia, and CEAUL, Universidade de Lisboa
manela@isa.ulisboa.pt

When modelling extreme events there are a few relevant parameters such as the *extreme value index* and the *extremal index* that need to be adequately estimated not only by themselves but also because they are the basis for the estimation of other relevant parameters. Motivated by real environmental problems we will present an introduction to extreme value analysis and will illustrate the application of the **R** statistical software. Recently computer intensive methodologies and adaptive algorithms have been proposed for an adequate estimation of the aforementioned parameters; they are here revisited and illustrated.

Keywords: extremal index, extreme value index, resampling procedures, R software

---

Extreme Value Theory (EVT) aims to study and to predict the occurrence of extreme or even rare events, outside of the range of available data. These events are part of the real world but environmental extreme or rare events may have a great impact on everyday life and may have catastrophic consequences for human activities. EVT has application in a number of different areas such as biostatistics, computer science, environment, finance, insurance, statistical quality control, structural engineering and telecommunications.
The classical assumption in EVT is that we have a set of independent and identically distributed (i.i.d.) random variables (r.v.'s), $X_1, \ldots, X_n$, from an unknown cumulative distribution function (c.d.f.) F and we are concerned with the limit behaviour of $M_n \equiv X_{n:n} = \max(X_1, \ldots, X_n)$ as $n \to \infty$. Whenever it is possible to linearly normalize M_n so that we get a non-degenerate limit, as $n \to \infty$, such a limit is of the type of the extreme value (EV) d.f.,

$$EV_\xi(x) := \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0 \quad \text{if} \quad \xi \neq 0 \\ \exp[-\exp(-x)], & x \in R \qquad\qquad\quad \text{if} \quad \xi = 0. \end{cases} \tag{1}$$

We then say that F is in the domain of attraction for maxima of EV_ξ, denoting this by $F \in D_{\mathcal{M}}(EV_\xi)$. The parameter ξ is the *extreme value index* (EVI) and it measures essentially the weight of the right tail function, $\overline{F} = 1 - F$.
In most fields of applications, the independence assumption is not valid. Stationary sequences are realistic for many real problems and dependence in stationary sequences can assume several forms.

Provided that a stationary sequence $\{X_n\}_{n\geq 1}$ has limited long-range dependence at extreme levels, the maxima of this sequence follow the same distributional limit law as the associated independent sequence, $\{Y_n\}_{n\geq 1}$, but with other values for the parameters of EV d.f..
Let us assume to be working with a strictly stationary sequence of r.v.'s, $\{X_n\}_{n\geq 1}$, with marginal d.f. F, under the long range dependence conditions **D**, [3], and the local dependence condition **D"**, [2]. The stationary sequence $\{X_n\}_{n\geq 1}$ is said to have an extremal index (EI) θ, $0 < \theta \leq 1$, if for each $\tau > 0$, we can find a sequence of levels $u_n = u_n(\tau)$ such that, with $\{Y_n\}_{n\geq 1}$ the associated i.i.d. sequence (i.e. from the same F),

$$P\left(Y_{n:n} \leq u_n\right) = F^n(u_n) \underset{n\to\infty}{\longrightarrow} e^{-\tau} \quad \text{and} \quad P\left(X_{n:n} \leq u_n\right) \underset{n\to\infty}{\longrightarrow} e^{-\theta\tau}. \tag{2}$$

The estimation of ξ, in (1) and θ defined in (2) is of primordial importance not only by themselves but also because they are the basis for the estimation of other quantities of great importance, such as the probability of exceedance of a high level; the return period of a high level; the right endpoint of an underlying model and a high quantile of probability $1 - p$, with p small.
Accurately modelling extreme events has become more and more exigent and the analysis require tools that must be simple to use but also should allow to consider complex statistical models in order to produce valid inferences. It is our intention to show a comparative set of steps for performing a data analysis of extreme values in the **R** environment, see also [1] with different case-studies.
We will begin with a brief review on extreme value analysis, presenting recent estimators of the parameters and detailing the various tools for using the **R** software. Computational procedures that have revealed to improve the parameter estimators will be also illustrated.

Acknowledgements This work has been supported by FCT–Fundação para a Ciência e a Tecnologia through the projects UIDB/00006/2020(CEAUL).

References

[1] M.I. Gomes, M.I. Fraga Alves and C. Neves. *Análise de Valores Extremos: uma Introdução.* SPE, INE, Lisbon, 2013.

[2] M.R. Leadbetter and S. Nandagopalan. *On exceedance point process for stationary sequences under mild oscillation restrictions, In Extreme Value Theory. J. Husler and R. D. Reiss (edts),.* Springer Verlag, Berlim, 1989.

[3] R. Leadbetter, M. R. Lindgren and H. Rootzen. *Extremes and related properties of random sequences and series.* Springer Verlag, Berlim, 1989.

24 October, 12:10 - 12:30, Zoom Room 1

Extremal index estimation: an application

Dora Prata Gomes[1], **M. Manuela Neves**[2]
[1] Faculdade de Ciências e Tecnologia, and CMA, dsrp@fct.unl.pt
[2] Instituto Superior de Agronomia, and CEAUL, Universidade de Lisboa, manela@isa.ulisboa.pt

---

The objective of this work is to present an extremal index block estimation procedure that only depends on the block length, under some conditions on the underlying process. A procedure for verifying the validity of those conditions is given and illustrated. For finite samples is illustrated a stability criterion for choosing the block length and then obtaining the extremal index estimate. Simulated and real data will be considered.

Keywords: statistics of extremes, extremal index, block length

---

The main objective of Statistics of Extremes is the estimation of probabilities of rare events. When extending the analysis of the limiting behaviour of the extreme values from independent and identically distributed sequences to stationary sequences a key parameter appears, the *extremal index* θ, whose accurate estimation is not easy. Here we focus on the estimation of θ using blocks estimators, that can be constructed by using disjoint or sliding blocks. Both blocks estimators require the choice of a threshold and a block length. We will show how the threshold and the block size choices can affect the estimates. However the main objective of this work is to revisit another estimation procedure that only depends on the block length, although some conditions on the underlying process need to be verified. The proposed estimator, see [2], was defined in the following way: Denoting $N_i(r_n, v_{ni})$ as the number of up-crossing of v_{ni} in ith block, the estimator is defined by

$$\widetilde{\theta}_n^B(r_n) := \frac{k_n}{\sum_{i=1}^{k_n} N_i(r_n, v_{ni})}. \tag{1}$$

The associated estimator presents nice asymptotic properties, and for finite samples is here illustrated a stability criterion for choosing the block length and then obtaining the θ estimate, see [1], [3] and [4]. Simulated data and observations of daily mean flow discharge rate in the hydrometric station of Fragas da Torre in river Paiva, data collected from 1 October, 1946 to 30 April, 2012 will be considered for illustrating the proposed procedure. The plot of Figure 1 presents the application of estimator (1). The application of the stability algorithm led to a block length $r_n = 975$ and an EI-estimate equal to 0.9231.

Acknowledgements This work has been supported by National Funds through FCT Fundação para a Ciência e a Tecnologia, Portugal, through the projects UIDB/00297/2020 and UID/MAT/00006/2019 (CEA/UL) .

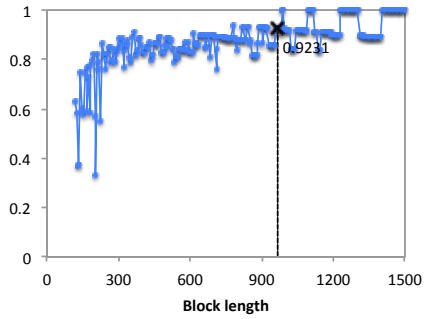

Figure 1: Estimates of $\widetilde{\theta}_n^B$ plotted against block length, with the choice of r_n for the daily mean flow discharge rate values.

References

[1] F. Caeiro and M. I. Gomes. Threshold selection in extreme value analysis. *Extreme value modeling and risk analysis: Methods and applications*, pages 69–87, 2015.

[2] L. Canto e Castro. Estudo de um método de estimação do indice extremal. *I Congresso Ibero-Americano de Estadística e Investigación Operativa, Salamandra*, 1992.

[3] M. M. Neves, M. I. Gomes, F. Figueiredo, and D. Prata Gomes. Modelling extreme events: Sample fraction adaptive choice in parameter estimation. *J. Stat. Theory Practice*, 9:184–199, 2015.

[4] D. Prata Gomes and M. M. Neves. Extremal index blocks estimator: the threshold and the block size choice. *Journal of Applied Statistics*, pages 1–16, 2020.

24 October, 12:30 - 12:50, Zoom Room 1

Linear combinations of generalized Hill estimators

Fernanda Otilia Figueiredo[1], **Maria Ivette Gomes**[2]
[1] Faculdade de Economia da Universidade do Porto and CEAUL, otilia@fep.up.pt
[2] FCUL, DEIO and CEAUL, Univerisdade de Lisboa, ivette.gomes@fc.ul.pt

The primary parameter in statistics of extremes is the tail index. Several estimators have been proposed in the literature for this parameter, but new refinements of such estimators can be considered to improve its efficiency. In this paper best linear unbiased estimators (BLUEs) of generalized Hill statistics are considered to estimate a positive tail index.

Keywords: BLUE, Hall-Welsh class of models, heavy-tails, Monte Carlo simulations, tail index

In statistics of extremes, one of the most important parameter related to extreme events is the tail index. For a heavy right-tail, several estimators have been proposed in the literature, being the most common estimators for a positive tail index, the class of Hill estimators.
Let $\underline{\mathbf{X}}_n = (X_1, \ldots, X_n)$ be the data sample, and $(X_{1:n} \leq \cdots \leq X_{n:n})$ the sample of the corresponding ascending order statistics. The Hill estimators (see [4]), $H(k)$, defined by

$$\mathrm{H}(k) = \mathrm{H}(k; \underline{X}_n) = \frac{1}{k}\sum_{i=1}^{k} \ln\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right) = \ln\left(\prod_{i=1}^{k}\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^{1/k} =: \ln\left(\prod_{i=1}^{k} U_{ik}\right)^{1/k} \tag{1}$$

are based on the $k+1$ upper observations of the sample. They have the disadvantage of having high variance for small values of k and high bias for large values of k. To improve the performance of such class of estimators, the class of generalized Hill estimators (see [1]), $H_p(k)$, among others, have been proposed in the literature. These estimators are defined by

$$\mathrm{H}_p(k) = \mathrm{H}_p(k; \underline{X}_n) := \begin{cases} \left(1 - \mathrm{M}_p^{-p}(k)\right)/p, & \text{if} \quad p < 1/\xi,\ p \neq 0, \\ \ln \mathrm{M}_0(k) = \mathrm{H}(k), & \text{if} \quad p = 0, \end{cases} \tag{2}$$

where

$$\mathrm{M}_p(k) = \begin{cases} \left(\frac{1}{k}\sum\limits_{i=1}^{k} U_{ik}^{p}\right)^{1/p}, & \text{if} \quad p \neq 0, \\ \left(\prod\limits_{i=1}^{k} U_{ik}\right)^{1/k}, & \text{if} \quad p = 0. \end{cases} \tag{3}$$

In this work, for the Hall-Welsh class of Pareto-type models (see [3]), we propose best linear unbiased estimators (BLUEs) of generalized Hill statistics, BL_{H_p}, to estimate the underlying positive tail index. If we consider m statistics H_p computed at different intermediate levels, i.e., the vector

$$\mathbf{H_p} \equiv (H_p(k-m+i), \quad i = 1, 2, \cdots, m), \quad 1 \leq m \leq k, \tag{4}$$

the BLUE estimator based on these H_p statistics is of the form

$$BL_{H_p} = \sum_{i=1}^{m} a_i \, H_p(k-m+i) = \mathbf{a}'\mathbf{H_p}, \tag{5}$$

for an adequate set of constants $\mathbf{a}' = (a_1, a_2, \cdots, a_m)$.

The work is conducted as follows. First, following [2] we derive such class of BLUEs, i.e., we determine the constants $(a_1, a_2, \cdots, a_m)$. Then, using Monte Carlo simulation techniques, we present for several models commonly used in applications, the finite sample behaviour of some of the estimators under discussion. Finally, we present some overall conclusions.

Acknowledgements Research partially supported by FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto UIDB/00006/2020.

References

[1] M. F. Brilhante, M. I. Gomes, and D. Pestana. A simple generalization of the hill estimator. *Computational Statistics and Data Analysis*, 57:518–535, 2013.

[2] M. I.. Gomes, F. Figueiredo, and S. Mendonça. Asymptotically best linear unbiased tail estimators under a second order regular variation. *Journal of Statistical Planning and Inference*, 134:409–433, 2005.

[3] P. Hall and A. W. Welsh. Adaptive estimates of parameters of regular variation. *Annals of Statistics*, 13:331–341, 1985.

[4] B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3:1163–1174, 1975.

Contributed Sessions

23 October, 9:00 - 9:20, Zoom Room 1

Main factors of motivation in an organizational context by multivariate data analysis methods: an empirical study

Áurea Sousa[1], **M. da Graça Batista**[2], **Sara Cabral**[3], **Helena Bacelar-Nicolau**[4]
[1] Universidade dos Açores, CEEAplA, aurea.st.sousa@uac.pt
[2] Universidade dos Açores, CEEAplA, maria.gc.batista@uac.pt
[3] Universidade dos Açores, sara_crc@hotmail.com
[4] Universidade de Lisboa, Faculdade de Psicologia e ISAMB-FML, hbacelar@psicologia.ulisboa.pt

Leadership competencies include the ability to motivate employees. Therefore, successful organizations use positive strategies to motivate their employees. This work aims to know the perspectives of bank employees on the main motivational factors in the work context and on the role of leaders in their motivation. The data collected by a questionnaire were analysed using several statistical and data analysis methods. Multivariate data analysis methods, defining main factors and a typology of motivation, brought a better and powerful insight about the subject in study.

Keywords: leadership, motivation, bank employees, categorical principal components analysis, ascendant hierarchical cluster analysis

As the globalization increase, the leadership role has been highlighted as an effective lead to organizations accomplish success, competitiveness and revitalization. In general, leadership can be defined as the process of influencing the behaviour or actions of an individual or group of individuals, to achieve common goals, based on a vision of the future that is constituted on a coherent set of ideas and principles ([1]). Therefore, it is the (effective) leader who must know how to deal with individual, group and organizational goals.
The data set consists of 202 bank employees working in the banks that operate in the Autonomous Region of the Azores, who answered to a tested and validated questionnaire in 2017. The first part of the questionnaire was composed of sociodemographic variables. The second part includes twenty-nine items, used to evaluate the variable "Leadership" through the identification of leader's characteristics. The third part includes a set of twenty-seven items, that was used to identify the main motivational factors of the respondents in an organizational context. There also were two other questions in this section: "Do you feel motivated with your work?" and "Does your leader have an important role in your motivation?". For each item and question, the respondents only could select one of six modalities of response (1 = Totally Disagree (TD); 6 = Totally Agree (TA)). Here we focus our attention in the second and third parts of the questionnaire.

The collected data were analysed using several statistical and data analysis methods, including the Categorical Principal Components Analysis (CatPCA) and the Ascendant Hierarchical Cluster Analysis (AHCA). CatPCA is the nonlinear equivalent of standard Principal Components Analysis (PCA). Here, we deal with ordinal variables, so we used CATPCA, applied over the sub-matrix containing the items that aim to measure the aspects most valued by bank employees in the work context.
AHCA methods used in the present work were based on the affinity coefficient and some probabilistic aggregation criteria issued from a parametric family under the probabilistic V (Validity) L (Link) approach, developed by Lerman (e.g., [2]), Bacelar-Nicolau, Nicolau and collaborators (e.g., [3]). We compare clustering results achieved using the ACHA probabilistic approach with those obtained by Cabral ([4]) using the Spearman's rank correlation coefficient and classical aggregation criteria, in order to find a consensus partition concerning a typology of the items describing the motivational factors of the employees.
This study help bank' institutions to understand the importance of effective leaders and their impact on employees' motivation. The CatPCA allowed to extract four components (dimensions), explaining almost 70% of the total variance of the data, and representing, respectively, "Psychological well-being / Interpersonal relationships"; "Job stability and incentive system"; "Career progression / Professional achievement"; and "Compliance with objectives and timings to achieve them".
In what concerns AHCA approach, hierarchical results display four main branches and a few singletons. Comparing these results with those obtained in Cabral 2018, the differences observed are mainly related to three singletons. Moreover, the clusters of items, named, respectively, "Career progression", "Psychological well-being/Interpersonal relationships", "Organizational environment and working conditions"; and "Compliance with objectives and timings to achieve them" obtained from AHCA methods, are associated to different main motivational factors, bringing new information for them.

References

[1] F. Lacombe and G. Heilborn. *Administração: Princípios e Tendências.* Saraiva, São Paulo, 2008.

[2] I.C. Lerman. *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series: Advanced Information and Knowledge Processing* Springer-Verlag, London, 2016. doi: 10.1007/978-1-4471-6793-8.

[3] H. Bacelar-Nicolau, F.C. Nicolau, Á. Sousa, and L. Bacelar-Nicolau. Clustering of Variables with a Three-way Approach for Health Sciences. Testing. *Psychometrics, Methodology in Applied Psychology (TPM)*, 21 (4 Special issue): 435–447, 2014. doi: 10.4473/TPM21.4.5.

[4] S. Cabral. *O Impacto da Liderança na Motivação dos Colaboradores do Setor Bancário na Região Autónoma dos Açores.* Dissertação de Mestrado em Gestão de Empresas/MBA, Universidade dos Açores, 2018.

23 October, 9:20 - 9:40, Zoom Room 1

Preliminary statistical results of arugula and lamb's lettuce growth in an aquaponic system

Fernando Sebastião[1], **Judite Vieira**[2], **Luís Cotrim**[3], **Nelson Oliveira**[4], **Ana Costa**[5], **Maria Carlos Rodrigues**[6]

[1] Laboratory of Separation and Reaction Engineering - Laboratory of Catalysis and Materials (LSRE-LCM), School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, fsebast@ipleiria.pt

[2] Laboratory of Separation and Reaction Engineering - Laboratory of Catalysis and Materials (LSRE-LCM), School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, judite.vieira@ipleiria.pt

[3] Laboratory of Separation and Reaction Engineering - Laboratory of Catalysis and Materials (LSRE-LCM), School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, luis.cotrim@ipleiria.pt

[4] Laboratory of Separation and Reaction Engineering - Laboratory of Catalysis and Materials (LSRE-LCM), School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, nelson.oliveira@ipleiria.pt

[5] Laboratory of Separation and Reaction Engineering - Laboratory of Catalysis and Materials (LSRE-LCM), School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, ana.s.costa@ipleiria.pt

[6] School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal, maria.l.rodrigues@ipleiria.pt

---

The main objective of a sustainable aquaponic system is to produce food in a sustaining way, achieved only when production systems with a minimum ecological impact are used. Recirculating aquaculture systems provide opportunities to reduce water usage and to improve waste management and nutrient recycling.

This work aims to analise the growth evolution of two cultivars of arugula and two cultivars of lamb's lettuce, in an aquaponics system. It is intended to evaluate the existence or not of significant differences through morphological characteristics and growth between cultivars of the same Genus (with different light exposure conditions), analised weekly, during 5-7 weeks.

Keywords: *eruca vesicaria*, *eruca sativa*, *valerianella locusta*, *clarias gariepinus*, hypothesis tests

---

Aquaponics is an integration system of plant production (Hydroponics) and fish production (Aquaculture). The nutrients required for plant growth are extracted from the waste

that results from feeding fish [1]. In this project there is a tank with catfish (*Clarias gariepinus*) and another tank with two cultivars of arugula (*Eruca vesicaria* var. *sativa* and *Eruca sativa*) and two cultivars of lamb's lettuce (*Valerianella locusta* var. *Favor* and *Valerianella locusta* var. *de Hollande*) [4]. Nowadays, the commercial scale interest of products obtained from aquaponics is small and the number of experimental studies is not enough, namely for some kind of plants such as lamb's lettuce. Thus, the main purpose of this study is to determine the productivity levels of such plants and arugula [2, 3].
Eight treatments were evaluated, which correspond to a 4 x 2 factorial design, with 34 plants of each cultivar of arugula and 34 plants of each cultivar of lamb's lettuce in two crop environments (a single greenhouse with and without shading).
The morphological characteristics in weekly evaluation were height, colour and health of plants; diameter, length and number of leaves; length of roots; freshness and dryness of the above-ground matter. In the daily evaluation some physical and chemical parameters in the water of both tanks such as temperature, pH, dissolved oxygen, electric conductivity, oxidation reduction potential and total dissolved solids were measured. The temperature and humidity of the air inside the greenhouse were also measured as well as the water loss in 5-7 weeks due to plants transpiration.
The water quality remained safe and stable and the fish did not die. Healthy growth was generally seen at moderate temperatures and sunlight. At higher average temperatures and under more intense sunlight leaf yellowing and withering was observed.
Some elementary statistical approaches were used, including hypothesis tests, to study the existence or not of significant differences in the morphological characteristics and in the physical and chemical parameters concerning the growth of the plants between both cultivars of the same Genus.

Acknowledgements This work was financially supported by: Associate Laboratory LSRE-LCM - UID/EQU/50020/2019 - funded by national funds through FCT/MCTES (PIDDAC).

References

[1] W. Lennard and J. Ward. A comparison of plant growth rates between an NFT hydroponic system and an NFT aquaponic system. *Horticulturae*, 5(2):27, 2019.

[2] C. Maucieri, C. Nicoletto, R. Junge, Z. Schmautz, P. Sambo, and M. Borin. Hydroponic systems and water management in aquaponics: A review. *Italian Journal of Agronomy*, 13:1012, 2018.

[3] D. Schmidt, V.J. Gabriel, B.O. Caron, V.Q. Souza, R. Boscaini, R.R. Pinheiro, and C. Cocco. Hydroponic rocket salad growth and production according to different color profiles. *Horticultura brasileira*, 35, 1:111–118, 2017.

[4] D.C. Sikawa and A. Yakupitiyage. The hydroponic production of lettuce (Lactuca sativa) by using hybrid catfish (Clarias macrocephalus x C.gariepinus) pond water: Potentials and constraints. *Agricultural Water Management*, 97:1317–1325, 2010.

23 October, 9:40 - 10:00, Zoom Room 1

Preliminary Screening of Probabilistic Models for Water Flow Measurement

Flora Ferreira [1], **Marisa Almeida**[2], **Duarte Silva**[3], **Wolfram Erlhagen**[1]
[1] Center of Mathematics, University of Minho, fjferreira@math.uminho.pt, wolfram.erlhagen@math.uminho.pt
[2] Department of Mathematics, University of Minho, pg36983@alunos.uminho.pt
[3] Águas do Norte, duarte.silva@adp.pt

The present study presents a preliminary evaluation of eight different probabilistic models in order to discriminate the most suitable to describe the water flow measurement. Based on the L-variation vs. L-skewness diagrams the Log-normal, Gamma and Generalized Logistic distribution reveal to be the most appropriate.

Keywords: probabilistic models, L-moment ratio diagrams, water flow

---

The monitoring and control of water losses require rigorous flow measurement and management [3]. Recently, the rising deployment of smart flow metering technologies, making available large amount of flow measurement data, provides new perspectives towards a better understanding and modeling of the water flow. In particular, the evaluation of the performance of different distributions and the comparison between them provide valuable prior knowledge to select the most suitable distribution for the water flow management.During the last decades, the L-moments ratio diagrams, introduced by Hosking [1], have gained great popularity and have been widely used in many hydrological applications, such as the evaluation of probabilistic models to describe residential water demand [2]. In the present study we examined which probabilistic models are considered more suitable to describe water flow measurements, based on L-moment ratio diagrams of L-variation vs L-skewness. The dataset is composed of water flow measurements recorded by two flow meters located in front of two different district monitoring area inputs. The data recordings were obtained every 5 minutes during a complete year (from 9-12-2018 to 8-12-2019). First, we studied the seasonal variation of the representative statistics mean value, L-variation coefficient and L-skewness coefficient of the two flow meter records on a monthly, a daily and an hourly basis. No significant variation is observed for those statistics from month-to-month and day-to-day. In the hour-to-hour analysis, the data shows some variation mainly in the statistic values between night and day hours. Then, the preliminary identification of most suitable probabilistic models was conducted on the basis of 48 (24 hours $\times$ 2 flow meters) L-points (L-variation, L-skewness). As candidate models, we analyzed the probabilistic models usually used to model hydrological variables with similar characteristics [2]: the

Weibull distribution, the Gamma distribution, the Log-normal distribution, the Generalized Logistic (GL) bounded at zero from below, the Generalized Pareto (GP) bounded at zero from below, the Generalized Extreme Value (GEV) distribution bounded at zero from below, the Normal and Exponential (EXP) distribution. Figure 1 shows L-variation vs. L-skewness diagrams where all understudy distributions are represented by a curve except the Exponential distribution that is a point (L-variation= 0.50 and L-skewness = 1/3). The normal distribution is represented by a horizontal line since it is symmetric about the mean (L-skewness = 0) and the L-variation can be adjusted to take any value. A large number of L-points, as well as the average L-points, are closer to Log-normal distribution. However, there are L-points closest to either Gamma or Generalized Logistic distribution. Therefore, these three distributions seem to be the most suitable probabilistic models to describe these records. Further analysis is needed to select from these three candidate models the one that better describes the probabilistic modeling of water flow.

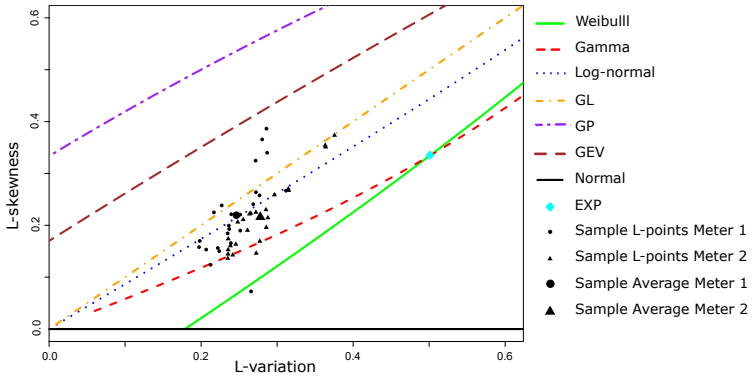

Figure 1: L-variation vs. L-skewness ratio diagrams that compare the observed L-points of the hourly water flow records of two flow meters against the theoretical L-points of the distributions under investigation.

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

References

[1] J. RM Hosking. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):105–124, 1990.

[2] P. Kossieris and C. Makropoulos. Exploring the statistical and distributional properties of residential water demand at fine time scales. *Water*, 10(10):1481, 2018.

[3] J. Sardinha, F. Serranito, A. Donnelly, V. Marmelo, P. Saraiva, N. Dias, R. Guimarães, D. Morais, and V. Rocha. Active water loss control. Technical report, EPAL, Empresa Portuguesa das Águas Livres S.A., 2017.

23 October, 10:00 - 10:20, Zoom Room 1

Multiple-valued symbolic data clustering using regression mixtures of Dirichlet distributions

José G. Dias[1]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, jose.dias@iscte-iul.pt

Symbolic data analysis (SDA) has been developed as an extension of the data analysis to handle more complex data structures. In this general framework the pair observation/variable is characterized by more than one value: from two (e.g., interval-value data defined by minimum and maximum values) to multiple-valued variables (e.g., frequencies or proportions). This research discusses the clustering of multiple-valued symbolic data using Dirichlet distributions. This new family of models explores the parameterization of compositional data in the regression setting, for instance regression/expert models. Results are illustrated with synthetic and demographic (population pyramids) data.

Keywords: multiple-valued symbolic data, clustering, mixture models, Dirichlet distribution, regression

The increase of storage capacity to handle big amounts of data has added pressure on the developing of data analytic tools. Quite often data sets need to be aggregated to allow the extraction of meaningful structures or classes for analysis. Symbolic data analysis [2] can be framed in this paradigm as it provides meaningful analysis of big data, otherwise difficult to be analyzed by traditional methodologies. This paper focuses on the multiple-value data that results from collapsing one or more dimensions in the data set. For instance, the aggregation over country produces proportions of the population based on sex, age groups, etc. This type of data shares similarities with compositional data [1].
Clustering is a specific goal of unsupervised statistical learning that assumes discrete hidden heterogeneity in the data set. The purpose of the analysis is to find the typology underlying the data that defines homogeneous groups. Different strategies of clustering have been proposed to handle multiple-value data from heuristic approaches (see e.g. [4]) to parametric settings [3]. This paper adds a predictive dimension to the analysis of histogram data by generalizing mixture of Dirichlet distributions to the regression setting. In particular, both mixture components and Dirichlet parameters are regressed on covariates, which enrich the understanding of data structures and move beyond descriptive analysis.
The proposed model is illustrated using synthetic data and an empirical data set. In particular, the application studies the country-level age structure of the population and uses the variables World Bank regions and Income group as covariates. The best solution

pertains three components in the mixture and both variables show predictive power to allocate observations into components.

Acknowledgements Funding from Fundação para a Ciência e Tecnologia (Portugal), UID/GES/00315/2019.

References

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B - Methodological*, 44(2):139–177, 1982.

[2] L. Billard and E. Diday. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487, 2003.

[3] N. Bouguila, D. Ziou, and J. Vaillancourt. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.

[4] A. Irpino, R. Verde, and F. D. T. de Carvalho. Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *Expert Systems with Applications*, 41(7):3351–3366, 2014.

23 October, 9:00 - 9:20, Zoom Room 2

A multilevel factor analysis of the cybercrime risk perception in the European Union

Ana Gomes[1,2], **José G. Dias**[2]
[1] Academia da Força Aérea, apgomes@academiafa.edu.pt
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, jose.dias@iscte-iul.pt

This study addresses the perception of cybercrime risk in the 28 European Union Member States. It aims to identify how the perception varies across the EU taking demographic characteristics of respondents and country effects into account. A Multilevel Factor Model with a Multiple Indicators and Multiple Causes structure is used to identify the typology of the EU citizens regarding the awareness and experience of cybercrime. Data are from the Eurobarometer 87.4/2017 and contain information on respondents from the 28 European Union Member States. The model shows a good fit and self-confidence, age, buying goods, gender, level of knowledge and previous experience of cybercrime are significant in explaining the Overall Cybercrime Risk Perception.

Keywords: cybercrime, european union, multilevel factor model, risk perception

Multilevel data structures are common in the social and behavioral sciences and are defined by nested levels of data, in which lower-level units are defined within macro units (e.g., countries) and share common characteristics. Multilevel analysis has been developed to account for specific statistical characteristics of this type of data. The Multilevel Factor Model (MFM) takes this structure into account by assuming the existence of two differewnt levels: the respondent level (Level 1), and an upper level e.g. country (Level 2) [3], [1]. Moreover, a MIMIC (Multiple Indicators and Multiple Causes) structure can be added to explain the latent variable in the multilevel structure.

The Multilevel Factor Model (MFM) used in this application takes into account two latent variables: a latent variable at the individual level (Level 1) that models the Overall cybercrime risk perception within each country and the between country latent variable (Level 2) that measures the cybercrime risk perception at country level to highlight the similarities (or differences) between European countries. The measurement model of Overall Cybercrime Risk Perception is based on ten items (e.g., Identity theft, Receiving fraudulent emails or phone calls asking for your personal details, Being asked for a payment in return for getting back control of your device). The model allows a MIMIC structure as the mean of the individual latent variable is regressed on a set of individual exogenous covariates (confidence in ones abilities to use the internet, previous experience of cybercrime and socio demographic variables) and the mean of the country-level latent variable is regressed on country-level indicator (Global Cybersecurity Index).

Data are obtained from the Eurobarometer 87.4/2017 [2] and contain information on respondents from the 28 countries of the European Union (sample of 27812 respondents reduces to 21657 for users of internet). The average age of the respondents is 48.49 years (s.d. = 18.75) and ranges from 15 to 99 years old. The European country weights were used to represent European Union population reproducing the real number of cases for each country, ponder the sample size with the universe size (derived from European Comission). Model estimation is performed by the maximum likelihood estimation using Matlab and shows a good fit.
The main results obtained shows: a significant and negative impact of Global Cybersecurity Index on country-level Overall Cybercrime Risk Perception; Less self-confidence and higher age increases Overall Cybercrime Risk Perception and Buying goods and being a male presents a significant and negative impact on the individual-level Overall Cybercrime Risk Perception.

Acknowledgements Funding from Fundação para a Ciência e Tecnologia (Portugal), UID/GES/00315/2019.

References

[1] E. S. Kim, R. F. Dedrick, C. Cao, and J. M Ferron. Multivariate behavioral research. *Structural Equation Modeling*, 51(6):881–898, 2016.

[2] Report TNS Opinion & Social. Cyber security report. Technical report, European Commission, September 2017.

[3] R. Varriale and J.K. Vermunt. Multilevel mixture factor models. *Multivariate Behavioral Research*, 47(2):247–275, 2012.

23 October, 9:40 - 10:00, Zoom Room 2

PLS-SEM to assess burnout state of industry workers

Luís M. Grilo[1,2,3,4], **Miguel Lopes**[4], **Vanda Lima**[4], **Aldina Correia**[4], **Ana Martins**[4]
[1] Instituto Politécnico de Tomar (IPT), Portugal, lgrilo@ipt.pt
[2] Centro de Matemática e Aplicações (CMA), FCT, UNL, Portugal
[3] Centro de Investigação em Cidades Inteligentes (Ci2), IPT, Portugal
[4] CIICESI, ESTG, Politécnico do Porto, Portugal, aic@estg.ipp.pt, aml@estg.ipp.pt, vlima@estg.ipp.pt, 816003@estg.ipp.pt

The COPSOQ questionnaire was used to assess the psychosocial risks to which the workers of a Portuguese industrial company are exposed and to find out which factors can cause them. A model was estimated using the partial least squares structural equation modeling approach. The latent construct 'quantitative demands' has a direct effect on 'stress' and this has a direct effect on 'burnout' state. A multi-group analysis was also conducted to compare the estimated model by gender (female and male) where some differences on the path coefficients are statistically significant.

Keywords: latent constructs, multi-group analysis, survey, health and wellness

Stress, anxiety or depressive states are usually considered to cause a generalized and persistent feeling of weakness, lack of vitality and energy, which is felt both physically and intellectually and affects the ability to work or perform simple tasks. According to a recent study, based on statistical evidence, "depressed or anxious people may be more likely to die from certain types of cancer" [1]. Moreover, excessive and prolonged stress can lead to burnout syndrome, which is characterized by World Health Organization (WHO) as "a feeling of exhaustion, cynicism or negativistic feelings linked to work and reduced professional effectiveness". Finally, in 2019, burnout has officially entered the WHO list of the International Classification of Diseases. According to a study developed during 2018, by DECO (the Portuguese Association for Consumer Protection), "one in three Portuguese workers is at risk of burnout" [4]. Plus, according to the report of the health sector (published in November 2019) of the Organization for Economic Cooperation and Development (OECD), after analyzing 29 countries, there is a general increase in the antidepressants consumption and Portugal is the fifth OECD country with the highest antidepressant consumption.
This study aims to identify which variables can cause stress and burnout state in workers of a Portuguese industrial company, in order to avoid negative situations not only for workers and the company, but also for the economy in general. With a better knowledge of the phenomenon, perhaps prevention mechanisms can be developed in order to reduce absences from work due to illness and, consequently, increase productivity and competitiveness.

The medium version of the Copenhagen Psychosocial Questionnaire (COPSOQ II) – a reliable and internationally validated instrument with 76 questions – was applied in April 2018 to assess psychosocial risks of workers in a Portuguese industrial company. This questionnaire was made available to workers in paper format and distributed individually throughout the four company sections. To characterize the stratified sample of 268 workers, some socio-demographic and labour issues were also considered, such as gender (female registered the highest percentage of 73.1%), age (55.2% of workers lie in]20, 39[years old), education level (38.4% of workers have at least the 9th grade) and work section (54.9% work in section two). The Structural Equations Modeling (SEM) has become used in the fields of social and health sciences, because it has the capacity to model constructs, to consider various forms of measurement error and to test entire theories and concepts. The consistent PLS (PLSc) version, which corrects for bias to consistently estimate SEM's with common factors, was applied, since it maximizes the explained variance of the endogenous constructs while it simultaneously relaxes the demands on data [2, 3]. The estimated SEM allows a better understanding of which variables have potential effect on company workers' stress and burnout. The total explained variance of the endogenous construct "burnout" is $R^2 = 73.5\%$. Based on multi-group analysis, by gender, for a significance level of 5%, the difference of the path coefficient between latent constructs 'justice' and 'job satisfaction' (higher for women) and between 'meaning of work' and 'job satisfaction' (higher for men) are statistically significant; the difference between constructs 'stress' and 'burnout' (higher for women) is almost significant. These results are important for companies that seek to improve working conditions, organizational and relational factors, in order to improve the well-being of workers and consequently increase their productive capacity.

Acknowledgements This work has received funding from FEDER funds through P2020 program and from FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) under the project UID/GES/04728/2020.

References

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

[2] L. M. Grilo, H. L. Grilo, and E. Martire. Sem using pls approach to assess workers burnout state. *AIP Conf. Proc.*, 2040:110008–1–110008–5, 2018.

[3] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2nd Ed., Thousand Oakes, CA: Sage, 2007.

[4] F. Ramos. Burnout: um terço dos inquiridos em risco. https://www.deco.proteste.pt/saude/doencas/noticias/burnout-um-terco-dos-portugueses-em-risco, 2018. [accessed in 03.Feb.2020].

23 October, 10:00 - 10:20, Zoom Room 2

An Issue About the Improvement of an Intelligent System Design for Disaster Situations

M.F. Teodoro[1,2], **M.J. Simões Marques**[2], **I. Nunes**[4], **G. Calhamonas**[4]
[1] CEMAT, IST, Lisbon University,Portugal
[2]CINAV, Naval Academy, Portuguese Navy, Portugal
[3]UNIDEMI,FCT, New Lisbon University, Portugal
[4]FCT, New Lisbon University, 2829-516 Caparica, Portugal

The goal of our work is to contribute to a Decision support system (DSS) with the ability to prioritize certain teams performance in a disaster occurrence situation, taking into account the importance of each team that acts, and the priority of the tasks to perform. In a recent work, were collected the tasks that shall be done in the case of an emergency intervention, which team can perform each task. The Delphi method, allowed to create an index that prioritize the teams that shall do each task and the order that each task shall be done. In the present work we propose a different way to compute the level of experience that weights the opinion of the experts in determination of such index.

Keywords: decision support system, expert, Delphi method, questionnaire, catastrophe, hierarchical classification, multidimensional scaling

---

Taken into consideration that a DSS is an information system that supports business or organizational decision-making activities. DSSs serve the management, operations and planning levels of an organization and help people make decisions about problems that may be rapidly changing and not easily specified in advance, a DSS in an catastrophe occurrence (e.g. floods, storms, tsunamis, fires, wars) is an excellent tool the help decision makers to prioritize certain teams for specific issues, taking into account the importance of each team that acts in each task. The reconnaissance, search and rescue, medical or logistics issues are important if we think about hurt people, dead animals, falling buildings, no water, no food, others. What shall be done in first place? Which task needs more urgently to be performed? Maybe to attend hurt people....or not, maybe to save people closed in a building to fall apart. This kind of decision can be taken in a shorter time and be optimized using a DSS. Noticing that the DSS must be feed with adequate information, it is important in advance to know, for each possible situation, the opinion of experts.
Noticing that the Delphi method is useful where the opinions of individuals are needed to dilute the lack of agreement or incomplete state of knowledge, is adequate to prioritize the teams that shall do each task and the order that each task shall be done. This method is important due its ability to structure and organize group communication. The idea is to get a maximum number of consensus between the experts with several rounds of questions. In [3, 4], we have performed the Delphi method based on [1, 2].

Detailing the process, the topic of interest was distributed (in a series of rounds) between the participating experts who comment on it and modify the opinion(s) until a certain degree of mutual consensus is reached. In our case, the collection and summary of knowledge of a group of experts from a given area was done through various phases of questionnaires, accompanied by an organized feedback. It consisted in 3 rounds of completed questionnaires responded by experts. The measure of consensus between the experts used a rule based on the inter quartile range (IQR).
After the consensus on all issues, could be computed and evaluated an index allowing to associate a number in scale which evaluate the order of teams that must be called to perform a certain task. It were identified which of tasks were the most indicated for each team, and the team that shall be called to respond under a certain order of priority. It can be observed that majority of the tasks are carried out by a SAR brigade team, followed by the Reconnaissance team, the Technical brigade teams, the Logistics brigade and the Medical team respectively. In [4] the authors have detailed the tasks associated with priority to each team.
In this manuscript we propose an alternative way of weighting the experts experience that contributes to the index that evaluates the order of teams that are going to perform the necessary task in a maritime environment. Firstly this proposal happens due the application of hierarchical classification, where in the same group of similarity of experts opinion, some individuals have very different experience level coefficient. We can classify the experience of each expert evaluating the similarity/distance between the individuals in the group of proposed experts using the multidimensional scaling technique and compare with the results presented in [4].
Acknowledgements This work was supported by Portuguese funds FCT, through the CEMAT, University of Lisbon, Portugal, project UID/Multi/04621/2019, and CINAV, Portuguese Naval Academy.

References

[1] M. Adler and E. Ziglio. *Gazing into the Oracle: The Delphi method and its application to social policy and public health*. Kingsley Publishers, London, 1996.

[2] H.M. Gunaydin. *Impact of Information Technologies on Project Management Functions*. PhD thesis, Chicago University, USA, 1999.

[3] I.L. Nunes, G. Calhamonas, M. Simões Marques, and M.F. Teodoro. Building a decision support system to handle teams in disaster situations - a preliminary approach. *Advances in Intelligent Systems and Computing*, 923:551–559, 2018. In: A. Madureira, A. Abraham, N. Gandhi, M. Varela, M. (eds.) Hybrid Intelligent Systems. HIS 2018.

[4] M. Simões-Marques, M. F. Teodoro, G. Calhamonas, and I.L. Nunes. Applying a variation of delphi method for knowledge elicitation in the context of an intelligent system design. *Advances in Intelligent Systems and Computing*, 959:386–398, 2020. In: Nunes, I. L. (Eds) Proceedings of Advances in Human Factors and System Interactions, AHFE 2019 Conference on Human Factors and System Interactions.

23 October, 10:40 - 11:00, Zoom Room 1

How perfect is a composite reference standard? A biomedical challenge

Ana Subtil[1], **M. Rosário Oliveira**[2], **António Pacheco**[3]

CEMAT and Instituto Superior Técnico, Universidade de Lisboa
[1] anasubtil@tecnico.ulisboa.pt
[2] rosario.oliveira@tecnico.ulisboa.pt
[3] apacheco@math.tecnico.ulisboa.pt

In the absence of a gold standard, the sensitivity and specificity of a new dichotomous diagnostic test may be estimated by comparison with a composite reference standard defined by combining multiple imperfect diagnostic tests according to a fixed rule. In order to get a better understanding of these methods and the way they are affected by various factors, we adopt a theoretical approach, based on analytical expressions derived for the new test's sensitivity and specificity biases arising from the use of composite reference standards based on different rules.

Keywords: composite reference standard, diagnostic test, sensitivity, specificity, imperfect gold standard

---

Before a new dichotomous diagnostic test is put into practice, its ability to correctly discriminate between the presence or absence of the target condition (disease, injury, parasite,...) must be evaluated. Sensitivity (Se) and specificity (Sp) are statistical measures commonly used in assessing the performance of diagnostic tests. Se is the probability of a positive test outcome given that the subject has the target condition, and Sp is the probability of a negative test outcome given that the subject does not have the condition. In the best-case scenario, the new test's Se and Sp are estimated by comparison with a perfect reference test or gold standard. When budgetary, ethical or technical restrictions prevent the use of a gold standard, the new test's Se and Sp may be estimated based on imperfect diagnostic tests. A straightforward approach is to adopt as reference an imperfect test which is perceived as the best one available for the target condition. Since this test, designated imperfect gold standard (IGS), is not error-free, it will potentially misclassify some subjects regarding the target condition, and thus bias the estimates of the new test's performance measures [3].

We focus on the estimation of the new test's Se and Sp by comparison, not with a single imperfect test, but with a composite reference standard (CRS) defined by combining multiple imperfect tests according to a fixed rule [1]. We address two rules that can be used to combine diagnostic tests: the "and" rule (CRS_A), in which a positive outcome emerges

if all component tests indicate a positive result, and the "or" rule (CRS_O), in which a positive result occurs if any of the component tests is positive. We also investigate an approach called dual composite reference standards (dCRS) [2], that resorts to the CRS_A to estimate the Se of the new test, and to the CRS_O to estimate its Sp.
To get further insight into these methods, we adopt a theoretical approach, deriving analytical expressions for the biases of the new test's Se and Sp determined by applying a CRS_A or a CRS_O. These biases are functions of various population (true) values: Se and Sp of the tests, prevalence of the condition, and conditional covariances between pairs of tests. Besides the basic scenario of conditional independence between the tests, we also consider disjoint pairwise conditional dependencies among the tests.
Both CRS_A and CRS_O can be understood as comparisons with an imperfect reference built using multiple imperfect diagnostic tests. Based on derived analytical expressions, we show how CRS_A, CRS_O, and IGS relate to each other, revealing a unified framework to deal with their performance measures and associated biases.
The adopted theoretical approach underlies our discussion on key issues regarding CRS_A, CRS_O, and dCRS, such as: the trade-off in Se and Sp between alternative CRS and their component tests, and how it affects the new test's Se and Sp biases; the impact on the various CRS of changing the number of component tests; the effect on the CRS of conditional dependence between the tests under consideration. We explore real diagnostic accuracy studies to illustrate our findings about the CRS behaviour, and the way it is affected by various factors.
Adopting the "and" rule or the "or" rule to combine imperfect tests has opposite effects on the Se and Sp of the resulting imperfect reference, improving one of them and impairing the other, thus exposing the great difficulty in obtaining good estimates of both Se and Sp of the new test either with the CRS_A or CRS_O. The analytical expressions of the Se and Sp biases shed a light on the dCRS performance, revealing that the dCRS can yield accurate estimates of both Se and Sp of the index test, particularly under conditional independence. In spite of the dCRS accurate estimates in various scenarios, the conditional dependence between the tests has an important effect on the Se and Sp biases and can undermine the dCRS results.

Acknowledgements Research supported by CEMAT and FCT, through projects PTDC/EEI-TEL/32454/2017 and UID/Multi/04621/2019.

References

[1] A. L. Baughman, K. M. Bisgard, M. M. Cortese, W. W. Thompson, G. N. Sanden, and P. M. Strebel. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clinical and Vaccine Immunology*, 15(1):106–114, 2008.

[2] S. Tang, P. Hemyari, J. A. Canchola, and J. Duncan. Dual composite reference standards (dCRS) in molecular diagnostic research: A new approach to reduce bias in the presence of imperfect reference. *Journal of Biopharmaceutical Statistics*, 28(5):951–965, 2018.

23 October, 11:00 - 11:20, Zoom Room 1

How to detect the manipulation of financial statements in EU financial incentives in Portugal

Susana Fernandes[1], **Raul Laureano**[2], **Luis Laureano**[3]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal , susana.fernandes@iscte-iul.pt
[2] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, BRU-IUL, Lisboa, Portugal, raul.laureano@iscte-iul.pt
[3] Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Lisboa, Portugal, luis.laureano@iscte-iul.pt

---

Portugal has been receiving funds from the EU for 33 years to help develop its economy. However, there are still those who question the success of the EU funds received. Indeed, firms try by all means to have their investment projects approved. This work predicts the propensity of firms to manipulate their financial statements, using the Beneish's M-Score and classification techniques. The results help public organizations identify which projects are more prone of being unsuccessful and thus alert for the need of a better scrutinization of these projects and the firms behind them.

Keywords: M-Score, financial statements, EU funds, manipulation, data analytics.

Incentives for business investment have become a fundamental instrument of public policies for economic dynamism and to achieve European convergence through economic growth [1]. There exist doubts about the effectiveness of the European Union (EU) funds, because Portugal has a medium-low rate of implementation of Europe 2020 strategies [1]. Many firms, either to apply to the EU funds or to get reimbursed of the approved incentives, may be tempted, among others, to manipulate their financial statements [2]. It is in this context of mistrust regarding the behavior of firms, that public organizations (e.g., AICEP, IAPMEI) analyze the projects and try to identify anomalies in the financial statements that can make the project approved or some investment costs eligible, while otherwise would not be. In view of the problem which the Member States are facing, on the one hand, of improving the efficiency and effectiveness of European funds and, on the other, to ensure that only compliant firms have access to the financial incentives, that the following research question arises: How can data analytics techniques help to improve the efficiency and effectiveness of the financial aid process to the Portuguese SMEs? To answer this question this study aims to: i) identify firms that manipulate their financial statements, ii) estimate the propensity to manipulate; and iii) identify the financial ratios related with manipulation.
To meet these objectives, we adopt the cross industry standard process for data mining

(CRISP-DM), a well-proven methodology to solve business problems. Keeping in mind the problem, the objectives and the data collected for 518 SMEs that obtained financial support, the data preparation, the modelling and the evaluation phases followed. First,many financial ratios (e.g., profitability, solvability, liquidity and operational ratios) were created in order to be used as inputs in the modelling techniques. Second, the Beneish's M-Score [3], a mathematical model that uses financial ratios to identify whether a firm has or has not manipulated its financial statements, was used to define the target attribute.

Third, different classification techniques were used, namely, logistic regression and decision trees with different algorithms, to classify firms in one of the two categories of the target attribute and to obtain the relative importance of the predictors. Finally, to statistically evaluate the models the most used quality metrics were selected and computed for both train and test samples. The results identified 43 percent of the firms committing manipulation. These firms tend to be from the Centre region, operate in the services or tourism industry and apply to the SME qualification - individual incentive program. Among the different classification models created and having in mind not only the quality metrics but also the parsimony and interpretability criteria, a CART balanced model was chosen. In the test sample the model shows a good performance (accuracy: $76,8$ percent, sensibility: $81,2$ percent, specificity: $71,4$ percent) and identifies equity growth and receivables growth as the two most important predictors of manipulation.
As firms are able to turn a bad project into a falsely good one, the results suggest that the public organizations need to perform a deeper analysis of the financial statements when evaluating the projects submitted by firms more prone to manipulate their accounts. Therefore, this work contributes to help the country to use EU funds more efficiently and effectively and to not compromise the EU objectives for Portugal.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia,grants UIDB/00315/2020, FCT UIDB/04466/2020 e UIDP/04466/2020.

References

[1] M. Beneish, C. Lee, and D. Nichols. Earnings manipulation and expected returns. *Financial Analysts Journal*, 69(2):57–82, 2013.

[2] OLAF. *The role of Member States' Auditors in Fraud Prevention and Detection.* European Union. Directorate D — Policy, European Anti-Fraud Office, Brussels, 2015.

[3] M. Stec and M. Grzebyk. The implementation of the Strategy Europe 2020 objectives in European Union countries: The concept analysis and statistical evaluation. *Quality & Quantity - International Journal of Methodology*, 52(1):119–133, 2018.

23 October, 11:20 - 11:40, Zoom Room 1

Hyperband for Clustering

Diogo Alves[1], Carlos Soares[2], Paula Brito[3],
[1] Fac. Economics, Univ. Porto, Portugal, 090402077@fep.up.pt
[2] Fraunhofer AICOS, Lab. for Artificial Intelligence and Computer Science, Fac. Engineering, Univ. Porto, Portugal, csoares@fe.up.pt
[3] Fac. Economics, Univ. Porto & LIAAD INESC TEC, Portugal, mpbrito@fep.up.pt

---

Choosing the best algorithm for a given task is difficult, because the number of alternatives is large. This choice becomes even harder because an additional decision must be made concerning the values of the hyperparameters of the selected algorithm(s). The most common approach is by trial and error, which is computationally expensive and may easily lead to sub-optimal choices, as it is driven by the knowledge of the data scientist, which is naturally limited and focused. Recently, several AutoML (i.e. automated machine learning) approaches have been developed for that purpose. One approach which has been particularly successful is *Hyperband*. *Hyperband* was originally developed for supervised learning tasks and for tuning parameters of a single algorithm. In this project, we adapt *Hyperband* for clustering, simultaneously selecting the algorithm and the corresponding hyperparameter values. We conduct an extensive empirical study, involving two algorithms, k-means and DBSCAN. Our results provide some interesting information about the behavior of the proposed approach as well as the clustering algorithms tested.

Keywords: Hyperband, K-means, DBSCAN, Calinski-Harabasz index, Silhouettes

---

The main goal of a cluster analysis is to find a finite and discrete structure of groups in a finite dataset where objects within the same cluster are similar and objects in different clusters are distinct. When conducting a cluster analysis, the process of selecting a clustering algorithm, the parameter configuration and the cluster validation technique to address a given problem is usually a hard task and contains a high degree of uncertainty since there is an extensive variety of alternatives. Each clustering algorithm has its strengths and weaknesses and no single one outperforms the remaining in all situations; moreover, clustering outcomes are very sensitive to parameter setting.
Recently, a bandit-based strategy known as *Hyperband* [3] was proposed to perform hyperparameter value optimization in other data analysis tasks. In this work we adapt *Hyperband* for clustering with the goal of selecting the clustering algorithm and the parameter configuration that best suits a given problem considering a given quality criterion.
Hyperband is a heuristic based on a principled early-stopping strategy that adaptively allocates a predefined budget of resources (size of the training set, number of variables, the

number of iterations,...) to randomly sampled configurations. This approach focuses on accelerating the configuration step, allocating more resources to promising hyperparameter configurations while eliminating the ones with worst results [2].
The *Hyperband* algorithm was originally developed to select the parameter values of a single algorithm. The focus was on supervised learning tasks, where quantitative evaluation is more natural. In this project, we have adapted it not only for clustering but also to select the algorithm as well as its parameters. For that purpose, we made several adaptations to the algorithm, most importantly, the performance function and the definition of bounds.
We tested *Hyperband* using two clustering algorithms, namely K-means and DBSCAN. Experiments were executed to optimize the parameters of each of the algorithms separately and to select both the algorithm and the hyperparameters. The partitions obtained are evaluated using two internal validation measures, the Silhouette coefficient [4] and the Calinski-Harabasz index [1]. Datasets from the UCI repository and simulated data with pre-fixed configurations were used to run experiments and validate the proposed approach. The conclusions can be summarized as:

- The results obtained with simulated data validate the proposed method.
- The configurations obtained when simultaneously selecting algorithms and optimizing hyperparametes sometimes lead to similar configurations as when the method was applied on the corresponding algorithm alone.
- The method proposed values for the parameters MinPts and Eps of DBSCAN that lead to reasonable results.
- The method sometimes proposed hyperparameter values that lead to a significantly different number of classes, depending on the evaluation criterion used.

Acknowledgements This work was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER, where applicable.

References

[1] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, 3(1):1–27, 1974.

[2] A. Klein, S. Falkner, S. Bartels, P. Hennig, F. Hutter, et al. Fast bayesian hyperparameter optimization on large datasets. *Electronic Journal of Statistics*, 11(2):4945–4968, 2017.

[3] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

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

23 October, 10:40 - 11:00, Zoom Room 2

Case study: Glycemic control in Type 2 diabetes

Ana Matos[1], Carla Henriques[2], Sara Brandão[3], Rui Marques[3], Edite Nascimento[3]

[1] School of Technology and Management, Polytechnic Institute of Viseu, Research Centre in Digital Services (CISeD), amatos@estgv.ipv.pt

[2] School of Technology and Management, Polytechnic Institute of Viseu, Centre for Mathematics, University of Coimbra, (CMUC), carlahenriq@estv.ipv.pt

[3] Department of Internal Medicine, Tondela-Viseu Hospital Center, sarabranmac@gmail.com

---

Capillary glycemic self-monitoring in diabetic patients, combined with individualized knowledge of therapy, is an essential vehicle for the control of blood glucose levels. In this work, we use regression modelling to assess the influence of glycemic self-monitoring on glycemic control (measured through glycated hemoglobin A) in three groups of patients with type 2 diabetes. Patients are divided in three groups according to the medication used in the treatment: with insulin, with oral antidiabetics and with combination therapy.

Keywords: regression modelling, type 2 diabetes, glycemic control

---

Glycemic control in diabetic patients is very important and should be an objective to be achieved in order to reduce the risk of developing late diabetes complications. Capillary glycemic self-monitoring, combined with individualized knowledge of therapy (understanding the results and knowing what to do in the face of a very high or very low result) is a valuable instrument because it allows the definition of individualized controlled objectives [1].

A total of 117 type 2 diabetic patients from the Diabetes Unit of the Tondela-Viseu Hospital Center were object of our study. The patients were divided in three groups according to the medication used in the treatment of diabetes: group O, with 28 patients, treated exclusively with oral antidiabetics; group I, with 28 cases, medicated with insulin only and group I+O of 68 patients with combined therapy.

Blood glucose control was measured using glycated hemoglobin A (HbA1c). The results showed that there was no statistically significant relationship between HbA1c and the number of capillary glycemia per month (Kendall's tau = 0.043, p = 0.751 for group I; Kendall's tau = 0.255, p = 0.062 for O group; Kendall's tau = -0.122, p = 0.17 for I + O group). Additionally, no statistically significant association was found between glycemic control and schooling, glycemic control and age and glycemic control and the type of area of residence (rural or urban), in any of the groups.

Multivariable logistic regression was used to estimate the odds of a patient being uncontrolled (HbA1c $\geq 7\%$) adjusting for gender, schooling and age. The odds of belonging to the

uncontrolled group are lower for the patients taking O medication, compared to the I group patients (OR=0.193, 95% confidence interval: 0.065-0.575). Furthermore, for the group of patients taking I + O medication, the odds of being an uncontrolled patient decrease with the increase of the number of capillary glycemia per month (interaction term I+O*number of capillary glycemia/month, OR= 0.986, 95% confidence interval: 0.971-1.001).
Statistical regression reveals that capillary glycemic self-monitoring (measured by the number of capillary glycemics/month) has a positive impact on glycemic control when comparing the group with combined therapy with the group medicated with insulin only, but it was expected to observe this positive impact in general.This raises a reflection on the patient's true ability to interpret the results of blood glucose analysis and the knowledge of the therapy to be performed, highlighting the need to implement therapeutic education programs so that these patients can intervene adequately in the therapeutic adjustment according to the information obtained by capillary blood glucose. So, maybe the results are different.

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

References

[1] American Diabetes Association. Glycemic Targets: Standards of Medical Care in Diabetes—2018. *Diabetes Care*, 41(Suppl. 1):55–64, 2018.

[2] S. Machado, R. Marques, E. Nascimento, A. Matos, and C. Henriques. Relationship Between HbA1c and Capillary Blood Glucose Self-Monitoring in Type 2 Diabetics. *Romanian Journal of Internal Medicine*, 57(2):125–132, 2019.

23 October, 11:00 - 11:20, Zoom Room 2

Statistics for communication students

Cláudia Silvestre[1], **Ana Meireles**[2]
[1] Escola Superior de Comunicação Social-IPL, csilmestre@escs.ipl.pt
[2] Escola Superior de Comunicação Social-IPL, ameireles@escs.ipl.pt

Communication professionals are not only consumers, but also producers of statistical information. As a consequence, statistics plays an important role in their education. So we need to make statistics personally meaningful to our students and show them the importance of statistics to their professional career. To achieve this goal we analyze piece of news that have statistical information, promote the debate in class and sometimes we try other way to communicate that information. Besides engage students, we develop their ability to interpret, critique, and communicate about statistical information.

Keywords: education, statistics, communication professionals, data visualization

In 2007, UNESCO proposed a complex concept - Media and Information Literacy (MIL) – which is a composite set of knowledge, skills, attitudes, competencies and practices. MIL covers a lot of areas including Statistics. This is specially relevant when we deal with communication student. Besides being consumer, they will produce and communicate statistical information. So, they need to develop skills in order to analyze, critically evaluate, interpret, use and create statistical information. Further more, they need to communicate effectively ([1]).

There is no doubt that statistical literacy is essential for communication student as a tool in their professional lives ([2]). In order to develop their statistical skills, they have to (i) understand the data-generation process (e.g. sampling procedure, the influence of the sample process and sample size); (ii) understand the meaning of sampling error ou margin error, that are frequently used in poll results; (iii) know how to use statistical terms correctly, like random, representative or reliable since may have different everyday meaning; (iv) deal with uncertainty; (v) choose the best way to represent the data accurately; among others.

Teaching statistics is a challenge since communication students do not feel at easy with mathematics or statistics. One core practice that engage students is to use real examples([3]). In our classes we pick examples from media (e.g. TV, newspapers, advertisements) and discuss in detail the available statistical information and its representation ([4]). Visual information is one of the discussion topics. For example, in figure 1 a) and b) the data are not represented accurately. In theses cases, we improve the representations or try other ways of visualizing data. These goals help to develop students' ability to discuss, interpret, evaluate and communicate statistical information.

This approach stimulate debate and action in statistical field. In addition, it emphasizes the power of critical thinking which is one top skills that professionals need for success. Students also learn more about the relevance of statistical methods including data collection.

Figure 1: (a) Poll results and (b) Distribution of members 2015 vs 2019

References

[1] I. Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 10(1):1–25, 2002.

[2] S. Gordon and J. Nicholas. Teaching with examples and statistical literacy: Views from teachers in statistics service courses. *International Journal of Innovation in Science and Mathematics Education*, 18(1):14–25, 2010.

[3] I. Ograjenšek and I. Gal. Enhancing statistics education by including qualitative research. *International Statistical Review*, 84(2):165–178, 2016.

[4] C. Silvestre and A. Meireles. Towards a statistically literate communication profissionals. *Proceedings of the Satellite conference of the International Association for Statistical Education (IASE) July 2017, Rabat, Morocco*, pages 1–5, 2017.

23 October, 11:20 - 11:40, Zoom Room 2

Student Motivations in choosing the country for Erasmus.

Suzanne Amaro[1], **Carla Henriques**[2], **Cristina Barroco**[3], **Joaquim Antunes**[4]
[1] Polytechnic Institute of Viseu, CISED samaro@estv.ipv.pt
[2] Polytechnic Institute of Viseu, CMUC, carlahenriq@estv.ipv.ptt
[3] Polytechnic Institute of Viseu, CISED cbarroco@estv.ipv.pt
[4] Polytechnic Institute of Viseu, CISED jantunes@estv.ipv.pt

---

Student mobility has been growing in recent years and it is pertinent to examine the underlying motivations that influence their choice of the host country for Erasmus. This study explores these motivations, identifying and describing the different segments. Four distinct segments were found: Language Attracted (motivated essentially by the language of the country), Enthusiasts (highly value all motivations), Rationalists (highly value the economic factor) and Academic Focused (oriented by the education characteristics).

Keywords: factorial analysis, cluster analysis, Erasmus, motivating factors

---

Student mobility across countries has a positive impact on student training but also on the economy of the countries that host these students [2]. This study investigates the factors that condition students in choosing the Erasmus country and seeks to find different segments of students according to these factors. The data were collected between May and June of 2018 through an online survey questionnaire and 5510 complete answers were obtained. The sample includes students from around 80 different countries, 70% female, who have done Erasmus in about 50 distinct countries. Eighteen questions about motivating factors for choosing the country of Erasmus were considered. Each item was measured on a 5-point Likert-type scale (1=Not important to 5=Very important). An exploratory factor analysis was applied to these eighteen motivation items revealing six factors, which were labed: Community Recommendation (related to third party recommendations), Novelty Seeking (looking for a country that is distant or unknown), Education Features (related to the quality, reputation and other education characteristics), Local Attractiveness (such as climate and gastronomy), Language interest (motivation centered on the language spoken) and Cost (related to travel expenses and living costs). Exploratory factor analysis was applied based on Pearson and Spearman correlations, and the same factors were obtained. To represent these six factors, additive scales were constructed, considering the average of the items associated to each factor. Cluster analysis was then applied to find motivational segments among Erasmus students. Cluster analysis involved two stages: first, two hierarchical techniques were applied, Ward's method and average linkage; the solutions of these methods were then considered as initial solutions for the k-means method. Guidance

on the number of clusters was initially obtained by examining the agglomeration coefficients of the hierarchical algorithms. Additionally, the k-means method was applied to 50 bootstrap samples and the similarity of cluster solutions for different numbers of clusters was examined using the rand index [1]. The four cluster solution emerged as an optimal solution, with a meaningful practical interpretation (Figure 1):
Cluster 1 - Language attracted - students who are very much guided by the language of the chosen country but do not follow recommendations, are not guided by "novelty seeking" and are not concerned about expenses or with education features.
Cluster 2 - Enthusiasts - group of students who value all factors more than others, except for the language factor in which they have lower values than those in Cluster 1.
Cluster 3 - Rationalists – These are the ones that most value the economic factor; this group has a profile very similar to those in Cluster 4 except that they highly value the possibility of spending less, while those in Cluster 4 do not value this at all. This cluster and cluster 4 distinguished from the other two because they value the attractive aspects and the language of the country less. The recommendations are not very important for students in these two groups.
Cluster 4 - Academic focused - They have a profile very similar to students in Cluster 3, but, unlike these, they have no concern for costs and are also less motivated by the search for a country that represents "novelty" and has attractive aspects. The aspects that most motivate these students are the education characteristics.

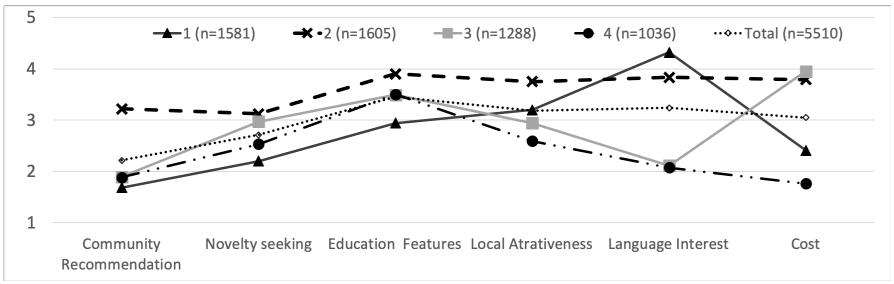

Figure 1: Cluster profiles

Other variables were used to further profile and distinguish these clusters and the results provide some insights regarding what Erasmus students look for in the country they choose.

Acknowledgements This research was supported by a grant from CI&DETS/IPV/CGD. The authors would also like to thank the CMUC for financial support.

References

[1] S. Dolnicar and F. Leisch. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, 21(1):83–101, 2010.

[2] OECD. Education at a glance 2017: OECD indicators paris: OECD publishing. URL: http://dx.doi.org/10.1787/eag-2017-en, 2017.

24 October, 9:50 - 10:10, Zoom Room 1

Symbolic Sensometrics

Paula Brito[1], **A. Pedro Duarte Silva**[2]
[1] Fac. Economics, Univ. Porto & LIAAD INESC TEC, Portugal, mpbrito@fep.up.pt
[2] Católica Porto Business School & CEGE, Univ. Católica Portuguesa, Portugal, psilva@porto.ucp.pt

We adress a sensometric study in a Symbolic Data Analysis framework. This is done by first aggregating individual repeated measurements as interval-valued observations, and then analysing the resulting interval data array, for which we rely on the parametric model proposed by Brito & Duarte Silva [2]. The proposed approach is applied to a data set of frozen pea samples.

Keywords: interval data, interval outlier, MANOVA, sensometics

Sensometric studies are usally based on (a set of) repeated measures on a collection of units/items, provided by a panel of evaluators, on different attributes. Frequently, the objective is to investigate the possible effect of some factor(s), thus leading to (M)ANOVA studies, or to put in evidence atypical variants of the product under study.
We address such problems in the context of Symbolic Data Analysis, by aggregating the repeated measurements as interval-valued data, thereby capturing the variability observed across observations. The resulting interval data array is then analysed considering the parametric model proposed by Brito & Duarte Silva [2].
The proposed approach is applied to a data set concerning a study on frozen peas [4]. In this study 16 frozen pea samples from the Danish market were profiled by 12 trained assessors in 3 replicates using 13 attributes; the data set contains some imputed values (data available at the Sensometric Society Data Set Repository `http://www.sensometric.org/datasets`).

Parametric models for interval-valued variables
Brito & Duarte Silva [2] proposed parametric models for interval data, relying on multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval-valued variables.
Given p interval-valued variables $Y_j, j = 1, \ldots, p$, the Gaussian model consists in assuming a joint multivariate Normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the MidPoints C_j and the logs of the Ranges R_j^*, with $\boldsymbol{\mu} = [\boldsymbol{\mu}_{\boldsymbol{C}}^t \ \boldsymbol{\mu}_{\boldsymbol{R^*}}^t]^t$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{CC}} & \boldsymbol{\Sigma_{CR^*}} \\ \boldsymbol{\Sigma_{R^*C}} & \boldsymbol{\Sigma_{R^*R^*}} \end{pmatrix}$ where $\boldsymbol{\mu_C}$ and $\boldsymbol{\mu_{R^*}}$ are p-dimensional column vectors of the mean values of, respectively, the MidPoints and Log-Ranges, and $\boldsymbol{\Sigma_{CC}}, \boldsymbol{\Sigma_{CR^*}}, \boldsymbol{\Sigma_{R^*C}}$ and $\boldsymbol{\Sigma_{R^*R^*}}$ are $p \times p$ matrices with their variances and covariances. This model has the advantage of allowing for a straightforward application of classical multivariate methods. The link that might or not exist between MidPoints C_j and

Log-Ranges R_j^* is modelled by specific configurations of the variance-covariance matrix $\mathbf{\Sigma}$. The most general formulation allows for non-zero correlations among all MidPoints and Log-Ranges, but other restricted cases may be considered - see e.g. [2].
A more general model may be obtained by considering the Skew-Normal distribution (see, for instance, [1]), which generalizes the Gaussian distribution by introducing an additional p-dimensional shape parameter $\boldsymbol{\alpha}$.
Using this framework, (M)ANOVA has been developped for interval data (see [2]). Interval outlier detection has been adressed in [3] for the Gaussian model, using Mahalanobis distances relying on robust estimates obtained following a trimmed-likelihood approach.

Analysis of peas sensometric data
We started by centering the values of each variable by evaluator, to cope with their possible different scaling. Then two different analysis are developed.
In a first approach, the $12 \times 3 = 36$ values of each variable have been aggregated, resulting in a data array of 16 pea varities (product) $\times 13$ interval-valued variables. Univariate interval outlier detection was performed - given the small sample size, multivariate outlier detection is not possible in this case. This resulted in the identification of several outlying products, some of which repeatly for different variables, and with varying degrees of outlyingness.
In a second approach, we aggregated the 3 values recorded for each combination product $\times$ evaluator, resulting in a matrix of 192 rows $\times$ 13 interval-valued variables. Eleven of those rows had to be eliminated due to the presence of degenerate intervals, leading to a final interval data array of 181 rows $\times$ 13 variables. Multivariate outlier detection flags 20 product $\times$ evaluator combinations as outliers, concerning 3 different products. A global MANOVA indicates that the 16 peas varieties are significantly different for the given variables; individual ANOVA's shows that those products are different for each of the 13 interval-valued variables.

Acknowledgements This work was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, including through project UID/GES/00731/2019, and co-funded by the FEDER, where applicable.

References

[1] A. Azzalini. A class of distributions which includes the Normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

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

[3] A. Pedro Duarte Silva, P. Filzmoser, and P. Brito. Outlier detection in interval data. *Advances in Data Analysis and Classification*, 12(3):785–822, 2018.

[4] T. Naes, P.B. Brockhoff, and O. Tomic. Detecting and studying sensory differences and similarities between products. In *Statistics for Sensory and ConsumerScience*, pages 47–66. John Wiley & Sons, 2011.

24 October, 10:10 - 10:30, Zoom Room 1

Interpreting all-subsets MANOVA and Canonical Variate Analysis: The additional information biplot

A. Pedro Duarte Silva[1],
[1] Católica Porto Business School & CEGE, Univ. Católica Portuguesa, Portugal, psilva@porto.ucp.pt

---

A common problem in Data Analysis is the identification of the relevant variable subsets to include in a given analysis. This problem may be addressed by all-subset comparison methods. However, all-subset methods often generate many almost equally performing alternative subsets, difficulting the interpretation of their results. These issues will be reviewed in the context of Canonical Variate Analysis (CVA), or more general Multivariate Analysis of Variance (MANOVA) effects, and a new graphical tool will be proposed: the MANOVA additional information biplot. This new biplot allows for the visualization of the specific additional contribution given by particular variable subsets. The usefullnes of this biplot will be illustrated in several applications.

Keywords: MANOVA, Canonical Variate Analysis, all-subset comparisons, biplots

---

Consider the MANOVA model

$$X = A\Psi + U \tag{1}$$

where we are interested in the description of a multivariate "effect", characterized by the violation of a linear hypothesis with general expression given by

$$H_0 : C\Psi = 0 \tag{2}$$

When the U rows are multivariate normally distributed with common covariance matrix Σ, hypothesis (2) can be tested by the well known statistics: $\Lambda = \det(E^{-1}T)$, $\lambda_1 = egval_1(E^{-1}H)$, $U = tr(T^{-1}H)$, or $V = tr(E^{-1}H)$, where E, H and $T = E + H$ are Error, Hypothesis and Total, Sum of Squares and Cross Product (SSCP) matrices (see *e.g.* [3])
A particularly important case is the one-way setup, where the different levels of the unique MANOVA factor can be interpreted as different groups in which the data is partitioned, hypothesis (2) states the equality of means across groups, and the matrices H and E reduce to the traditional Between ($B = H$) and Within ($W = E$) groups SSCP matrices.
Here, we will assume that hypothesis (2) has already been rejected, and that we are interested in describing deviations from it, from now on referred to as a MANOVA effect.
In the one-way setup the effect can be understood as group separation, and an usefull graphical aid is the Canonical Variate Analysis (CVA) biplot, where group means are

represented as points in a small r-dimensional space, at coordinates $Y = \bar{X}V_{r,int}$, with $\bar{X}$ representing group centroids, and $V_{r,int}$ interpolation axes. Matrix $V_{r,int}$ is formed by the first r columns of $V = W^{-1/2}V_{temp}$, with V_{temp} equal to the normalized eigenvectors of $W^{-1/2}BW^{-1/2}$ in non-increasing order of the corresponding eigenvalues. Biplot prediction axes can also be displayed, based on the firts r columns of $V_{pr} = (V^{-1})^T$ (see *e.g.* [4]).
Whatever the MANOVA setup, important research questions in this context include the identification of the number of dimensions responsible for the effect, a substantive interpretation of these dimensions, and the identification of the original variable subsets that contribute to the effect (see [3]).
Multivariate approaches to the latter question are based on the partition $X = [X_{.1}|X_{.2}]$ of variables already included ($X_{.1}$) or excluded ($X_{.2}$) from the model, and the hypothesis

$$H_0(2|1) : C\left(\Psi_{.2} - \Psi_{.1}\Sigma_{11}^{-1}\Sigma_{12}\right) = 0 \tag{3}$$

stating that the $X_{.2}$ variables do not contribute to the effect under study farther than $X_{.1}$. This additional information hypothesis, and the correspondig matrices, $E_{2|1} = E_{22} - E_{21}E_{11}^{-1}E_{12}$, $T_{2|1} = T_{22} - T_{21}T_{11}^{-1}T_{12}$, $H_{2|1} = T_{2|1} - E_{2|1}$, are the basis for the all-subset methods proposed in [2] and [3], and implemented in the popular *subselect* R package [1].
A problem with most all-subset comparison methods, is that they tend to identify many alternative subsets that perform almost identically. In this presentation, we will discuss some graphical tools that can facilitate the exploration of MANOVA all-subset results.
In particular, we will introduce the MANOVA additional information biplot, where aditional contributions to MANOVA effect will be represented at coordinates $Y = \tilde{X}V_{2|1,r,int}$, where $\tilde{X}$ represents unexplained MANOVA effects, $V_{2|1,r,int}$ is formed by first r columns of $V_{2|1} = E_{2|1}^{-1/2}V_{2|1,temp}$ and the columns of $V_{2|1,temp}$ are normalized eigenvectors of $E_{2|1}^{-1/2}H_{2|1}E_{2|1}^{-1/2}$. Biplot prediction axes can, likewise, be defined from $V_{2|1,pr} = (V_{2|1}^{-1})^T$.
The usefulness of these representations will be illustrated by several relevant applications.

Acknowledgements Financial support from Fundação para a Ciência e Tecnologia (through project UID/GES/00731/2019) is gratefully acknowledged.

References

[1] J. O. Cerdeira, A. P. Duarte Silva, J. Cadima, and M. Minhoto. subselect: Selecting variable subsets. R package, version 0.14. `http://cran.r-project.org/web/packages/MAINT.Data/index.htm`, 2018.

[2] A. P. Duarte Silva. Efficient variable screening for multivariate analysis. *Journal of Multivariate Analysis*, 76:35–62, 2001.

[3] C. J. Huberty and S. Olejnik. *Applied MANOVA and discriminant analysis.* John Wiley and Sons, Hoboken NJ, 2006.

[4] A. La Grange, N. Le Roux, and S. Gardner-Lubbe. Biplotgui: Interactive biplots in r. *Journal of Statistical Software*, 30:1–37, 2009.

XXVII Meeting of the Portuguese Association for Classification and Data Analysis, *Lisboa*, 22-24 October 2020

24 October, 10:30 - 10:50, Zoom Room 1

Community Detection in Interval-Weighted Networks

Hélder Alves[1], **Paula Brito**[2], **Pedro Campos**[3]
[1] FCUP, University of Porto & LIAAD INESC TEC, Portugal, halves2005@gmail.com
[2] FEP, University of Porto & LIAAD INESC TEC, Portugal, mpbrito@fep.up.pt
[3] FEP, University of Porto & LIAAD INESC TEC, Portugal, pcampos@fep.up.pt

---

Although several extensions of modularity to weighted networks have been proposed, none takes into account the variability of link weights. To fill this gap, we extend both Newman's modularity for weighted networks, and one state-of-the-art greedy method to optimize modularity, the Louvain algorithm, to the general case of *Interval-Weighted networks* (IWN). We apply our community detection approach for IWN to a real-world commuter network between the Portuguese mainland municipalities.

Keywords: community detection, interval-weighted networks, networks, Louvain algorithm

---

In order to derive a measure of quality of a partition, even without prior information of the true division of the network into communities, Newman and Girvan–NG (2004) [3] introduced a quality function known as *modularity*. Roughly, modularity compares a given network to a network with the same degree distribution of ties over the vertices placed at random. There are a large number of algorithms to optimize the modularity, e.g., NG applies a basic "fast greedy" approach algorithm [3], which differs from the one used in this work, the Louvain algorithm–LA [1], in the way that the latter includes a community aggregation step allowing its application in large networks. The generalization to *Interval-Weighted Networks* (IWN) of *modularity*, *gain of modularity* and consequently the *Louvain algorithm* [1], was done considering that the IWN can be represented as a *contingency table* whose cells represent the *observed interval–weights* $o_{ij}^I = [\underline{o}_{ij}, \overline{o}_{ij}]$ ($\overline{o}_{ij} \geq \underline{o}_{ij} > 0$; $o_{ij}^I \subseteq \mathbb{R}^+$), if there is an weighted edge between vertices (i, j), and zero otherwise. The *interval total weight/strength* attached to vertex i, is denoted by $s_i^{IO} = \sum_{j=1}^n [\underline{o}_{ij}, \overline{o}_{ij}]$, and the *total weight* is, $\sum_{i=1}^n s_i^{IO} = \sum_{j=1}^n s_j^{IO} = \sum_{i=1}^n \sum_{j=1}^n [\underline{o}_{ij}, \overline{o}_{ij}]$ (to simplify, hereafter we will use the notation $[2\underline{w}, 2\overline{w}]$). Analogously, and assuming independence between the vertices, the *expected interval-weights* e_{ij}^I are defined as the interval-weight that would be obtained if the hypothesis of row-column independence were true (further, these expected frequencies must pass through an "adjustment" of its total lower $2\underline{w}$ and upper $2\overline{w}$ limits)
$e_{ij}^I = \left[\frac{\underline{s}_i^{IO} \underline{s}_j^{IO}}{2\overline{w}}, \frac{\overline{s}_i^{IO} \overline{s}_j^{IO}}{2\underline{w}} \right]$ $(0 \notin [2\underline{w}, 2\overline{w}])$.
Given an undirected weighted network G^W and a partition $\mathcal{C} = \{C_1, C_2, \ldots, C_q\}$ of its vertices into q sets, the generalization of modularity (Q^W) and modularity gain (ΔQ^W) to

interval data was done as follows: *Modularity for IWN* (where "D" represents the difference between the observed o_{rr} and the expected e_{rr} interval-weights of community r), $Q^{IW} = \sum_r^q D(o_{rr}, e_{rr})$; *Modularity Gain for IWN* resulting from the merging of two communities, $\Delta Q^{IW} = Q^{IW}_{new} - Q^{IW}_{last}$; and the *Normalization of modularity for IWN*, $Q^{IW}_{norm} = \frac{Q^{IW}}{Q^{IW}_{\max}} = \frac{\sum_r^q D(o_{rr}, e_{rr})}{D\left([2\underline{w}, 2\overline{w}], \sum_r^q e_{rr}\right)}$.

In the previous generalizations we face two major setbacks: *interval dependency*; and the fact that *the value of the distance between intervals is always positive*. To contour these drawbacks we propose the following measures to evaluate de difference between two intervals $[\underline{x}, \overline{x}]$ and $[\underline{y}, \overline{y}]$: $d_1([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]) = \max\{\underline{x} - \underline{y}, \overline{x} - \overline{y}\}$; $d_2([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]) = \max\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\}$ *sign argmax*$\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\}$, and a "vectorial difference" $\overrightarrow{d}_3([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]) = (\underline{x} - \underline{y}, \overline{x} - \overline{y})$. According to the type of difference used, alternative modularity measures were defined. Similarly, various community detection methods based on the Louvain algorithm have also been developed.

Using the proposed measures and methods, we analyse the community structure that emerges from the movements of daily commuters in mainland Portugal between the twenty three Regions NUTS 3 [2]. The elements o^I_{ij} denote the maximum variability of the *bi–directional* flows ij and ji between the NUTS i and j (Figure 1b): $o^I_{ij} = [\min\{\underline{o}'_{ij}, \underline{o}''_{ji}\}, \max\{\overline{o}'_{ij}, \overline{o}''_{ji}\}] = [\underline{o}_{ij}, \overline{o}_{ij}]$ (flows greater than 50 daily movements).

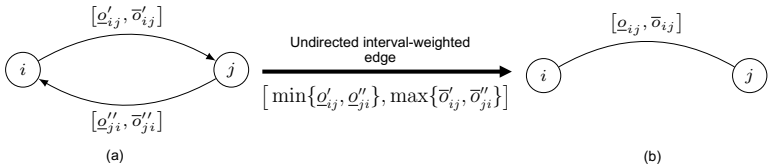

Figure 1: (a) Bidirectional interval flows $i \to j$ and $j \to i$, (b) Undirected interval flow between ij.

The final clustering using *difference* d_2 reveals the existence of three NUTS 3 communities, with $Q^{IW}_{norm} = 0.596$ ($Q^{IW}_{\max} = 10792.1$, and $Q^{IW} = 6371.6$), which means a moderate clustering structure. The Louvain algorithm for IWN reached maximum modularity at the end of the 2nd pass. These communities roughly represent the division of the country into two major regions, the northern and the southern region, and the "interior region center" of Portugal.

Acknowledgements This work was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER, where applicable"

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Lefebvre Etienne. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

24 October, 10:50 - 11:10, Zoom Room 1

Reducing Dimensionality in Multi-Layer Networks through Factorial Techniques

Pedro Campos[1], **Patrícia Gonçalves**[2]
[1] FEP, University of Porto and LIAAD INESC TEC , pcampos@fep.up.pt
[2] LIAAD INESC TEC, patricia.t.goncalves@inesctec.pt

In this work, we explore several methods for reducing the number of layers in multilayer networks and propose a different strategy based on factorial methods. The need to reduce dimensionality in multilayer networks may occur when the number of layers is high and it is important to see how they relate. We test and compare different algorithms and introduce an innovative procedure to reduce the number of layers.

Keywords: reducibility, network science, multilayer network, Multiple Factorial Analysis

We live in a interconnected world. Networks formed by the simultaneous interaction of different channels are called multilayer networks. Multilayer networks explicitly incorporate multiple connectivity channels and constitute the natural environment for describing interconnected systems across different categories of connections. Each channel (relation, activity or category) is represented by a layer and the same node can have different types of interactions, that is, different sets of neighbours in each layer. These networks allow to decode much richer information than networks that use individual layers separately, and can allow the interpretation of phenomena that simple networks cannot. [3] presented a general definition of multi-layer networks that can be used to represent most types of complex systems that consist of multiple networks or include disparate and/or multiple interactions between entities. A multi-layer network has a set of nodes V just like a normal or single-layer network. It is useful to remind that a single-layer network is a tuple G = (V, E), where V is the set of nodes and $E \subseteq V \times V$ is the set of edges that connect pairs of nodes. A multi-layer network can have any number d of aspects (or dimensions) and a sequence $L = \{L_a\}_{a=1}^{d}$ of sets of elementary layers such that there is one set of elementary layer L_a for each aspect a.

The need to reduce dimensionality in multilayer networks may occur when the number of layers is high and it is important to see how they relate. In this work we test and compare different algorithms and introduce an innovative procedure to reduce the number of layers in a multilayer network. We apply the reduciblity to a multi-layer network containing data about food products.

Several authors have studied reduciblity as a way of reducing dimensionality in multilayer networks. [1] , proposed a method to aggregate the layers of a multilayer network while maximizing its distinguishability from the aggregated network. The method is based

on a purely information theoretic perspective, which makes use of the definition of Von Neumann entropy of a graph.
[2] propose the modeling of populations of networks, and suggest cases in which factorial models may naturally capture our intuition about the underlying generative process of the data.
Our proposal for reducibility is based on the approach of [4], where a Multiple Factor Analysis for Contingency Tables (MFACT) is used. MFACT balances the influence of the groups on the first principal dimension by dividing the weights of the variables/columns of a group by the first eigenvalue of the separate analysis of this group. MFACT consists of a classical MFA applied to the multiple table Z assigning the weight $p_{i..}$ to the row i (i = 1, ..., I) and the weight $p_{.jt}$ to the column jt (j = 1, ..., Jt, t = 1, ..., T). Thus the observed proportions are compared to those corresponding to the intra-table independence model. This model neutralizes the differences between the separate average column profiles. We apply this model to the Export network of FAO data, crossing 224 countries with 300 different food products, corresponding to 300 layers in the multi-layer network. The goal is to group the products, reducing the number of layers.

References

[1] M. De Domenico, V. Nicosia, and A. Arenas. Structural reducibility of multilayer networks. *Nat Commun*, 6, 2015.

[2] D. Durante, D. Dunson, and J. Vogelstein. Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112, 2015.

[3] M. Kivelä, A. Arenas, M. Barthelemy, J. Gleeson, Y. Moreno, and M. Porter. Multilayer networks. *Journal of Complex Networks*, 2:203–271, 2014.

[4] B. Kostov, M. Bécue-Bertaut, and F. Husson. Multiple factor analysis for contingency tables in the factominer package. *The R Jou5nal*, 51, 2015.

24 October, 9:30 - 9:50, Zoom Room 2

Performance of Time Series Forecasting Models Applied to Economic Data

A. Manuela Gonçalves[1], **Susana Lima**[2], **Marco Costa**[3]
[1] CMAT-Center of Mathematics, DMAT-Department of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[2] DMAT-Department of Mathematics, University of Minho, Portugal, susanarlima@gmail.pt
[3] CIDMA-Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt

Forecasting economic time series is one of the most important issues that is behind the majority of strategic and planning decisions in effective operations of economic business. Economic time series belong to a special type of time series that present strong trend(s) and seasonal patterns, presenting challenges in developing effective forecasting models. This study compares the forecasting performance of economic time series on the basis of the SARIMA models and their extensions, the decomposition time series associated with multiple regression models with correlated errors, and the exponential smoothing methods (Holt-Winters). This work aims to discuss and compare these model formulations based on the same economic dataset: index of turnover (TOVT) collected in the Eurostat retail databases.

Keywords: forecasting, forecast accuracy, time series models, TOVT

The purpose of this study is to discuss economic time series forecasting on the basis of the SARIMA models and their extensions [4], the decomposition time series associated with multiple regression models with correlated errors [1], and the exponential smoothing methods (Holt-Winters models) [2] when applied to a case study of index of turnover (TOVT), retail sales time series. We first introduce the model with explicit specifications for the components: trend, season, cycle and irregular. We propose different approaches to time series forecasting by combining different models in order to increase the chance of capturing different patterns in the data and thus improve forecasting performance [3]. Therefore, it is required an accurate forecasting system to improve the quality of the decision-making process.
In the business operations of the retail sales segment, forecasting accuracy is even more important to the quality of the decision-making process because retailing is widely recognized as a competitive industry in both mature and developing markets. In this paper we study a set of retail time series in the Eurostat retail databases: the time series of Portugal, Germany, Spain, France, Italy, Netherlands and United Kingdom. The purpose of

this study is to compare the accuracy of retail sales forecasting applied to a monthly retail sales time series from 2000 to 2018. The methods considered in this study are applied to two sets: training data (in-sample data) and testing data (out-of-sample data) in order to testify the accuracy of the proposed forecasting models. The selected training data from January 2000 to December 2016 (the first 204 months) was used in order to fit the models to the data, and the test period from January 20017 to February 2018 (the last 14 months) was used to forecast. In order to evaluate the accuracy of the forecasting capacity of the methodologies adopted, several evaluation measures are used, namely MSE, RMSE, MAPE, MASE, and U.Theil statistic.

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

References

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

[2] R. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, 2014.

[3] S. Makridakis, S. Wheelwright, and R. Hyndman. *Forecasting: Methods and Applications.* John Wiley and Sons, New York, 1998.

[4] R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications With R Examples.* Springer, New York, 2006.

24 October, 9:50 - 10:10, Zoom Room 2

Modelling censored time series of counts

Isabel Silva[1], **Maria Eduarda Silva**[2], **Isabel Pereira**[3], **Brendan McCabe**[4]
[1] Faculdade de Engenharia, Universidade do Porto and CIDMA, Portugal, ims@fe.up.pt
[2] Faculdade de Economia, Universidade do Porto and CIDMA, Portugal, mesilva@fep.up.pt
[3] Departamento de Matemática, Universidade de Aveiro and CIDMA, Portugal, isabel.pereira@ua.pt
[4] Management School, University of Liverpool, UK, Brendan.Mccabe@liverpool.ac.uk

---

Time series under censoring occur when the observations are available for a restricted range only due to detection limits. Ignoring censoring produces biased estimates and unreliable statistical inference. The aim of this work is to contribute to the modelling of time series of counts under censoring using Poisson first-order integer-valued autoregressive (PoINAR(1)) models. The emphasis is on estimation and inference problems.

Keywords: censored count series, parameter estimation, Poisson INAR(1) model

---

Censored data are frequently found in several fields including environmental science, epidemiology, business and social sciences. Censoring (type 1) occurs when measuring devices are not able to detect above and/or below a certain threshold and observations are available for a restricted range only. Censoring may also be imposed by survey design. Disregarding data under censoring gives rise to model misspecification, biased parameter estimation and poor forecasts.

Some statistical methods for dealing with independent censored data have been proposed in the literature. There are also a few studies that fit Gaussian ARMA models to censored time series, e.g. [2]. In the context of time series of counts, modelling censored data has not been investigated previously. Hence, this work considers the analysis of time series of counts under censoring by using first-order integer-valued autoregressive (INAR) models. These models are based on a random operation called thinning (for details, see [4]) coupled with innovations following a discrete distribution.

Thet PoINAR(1) model under censoring is not Conditional Linear Autoregressive (CLAR) and presents an intractable likelihood. Therefore, to tackle the parameter estimation problem we resort to two alternative approaches: Approximate Bayesian Computation (ABC) methodology [3] and Indirect Inference via the so called Efficient Method of Moments (EMM) [1].

ABC is now a popular approach when the likelihood is computationally prohibitive but it is possible to simulate synthetic samples from the model for a given draw of the parameters from a prior. Summary statistics from these simulations are compared with the

corresponding from the observed data and the parameter draw is retained when there is a match between the simulated sample and the observed time series.
In the EMM approach, we need to specify an alternative model (auxiliary model) and corresponding likelihood function, which provides an approximation to the true model and hence the true likelihood function. Additionally, it is required that the true model can be simulated. The EMM estimator is based on the property that the gradient vector estimator of the auxiliary model is zero when evaluated at the actual data. The EMM solution is then given by the set of parameter values of the true model that gives the smallest value of the gradient vector of the auxiliary model evaluated at the simulated data.
Numerical experiments indicate that ABC and EMM procedures produce biased and inconsistent estimates under censoring. Therefore, the analysis of time series of counts under censoring requires alternative approaches such as Gibbs sampler with Data Augmentation.

Acknowledgements The first three authors were 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.

References

[1] V. L. Martin, A. R. Tremayne, and R. C. Jung. Efficient method of moments estimators for integer time series models. *Journal of Time Series Analysis*, 35:491–516, 2014.

[2] J. W. Park, M. G. Genton, and S. K. Ghosh. Censored time series analysis with autoregressive moving average models. *The Canadian Journal of Statistics*, 35:151–168, 2007.

[3] V. Plagnol and S. Tavaré. Approximate bayesian computation and mcmc. In Niederreiter H., editor, *Monte Carlo and quasi-Monte Carlo methods 2002*, pages 99–113. Springer, Heidelberg, 2004.

[4] M. G. Scotto, C. H. Weiß, and S. Gouveia. Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, 15:590–618, 2015.

24 October, 10:10 - 10:30, Zoom Room 2

Study of the Variation of Loans Granted to Families Between December 2009 and July 2019

João Lamy Gil[1], Joana Isabel da Silva Ramalho[2], Vasco Miguel da Silva Barata[3], Ana Lorga da Silva[4]

[1] ANA, Aeroportos de Portugal and Universidade Lusófona de Humanidades e Tecnologias, jcgil@ana.pt

[2] Universidade Lusófona de Humanidades e Tecnologias, joanaisramalho@gmail.com

[3] Universidade Lusófona de Humanidades e Tecnologias, vasco.barata@hotmail.com

[4] CPES, CIPES, Universidade Lusófona de Humanidades e Tecnologias ana.lorga@ulusofona.pt

---

The present work aims to relate the impact of imports, exports, investments and the unemployment rate on the number of loans granted to Portuguese families between 12/31/2009 and 7/31/2019, but also to predict for the next six months the number of loans granted to Portuguese families. For this purpose, a multiple linear regression model was initially used, followed by several univariate forecasting time series models, to determine which model provides a better forecast.

Keywords: loans granted, mutivariate regression models, univariate time series models, error estimation measures

---

This work intends to study not only the impact that imports, exports, investments and unemployment rate have on the number of loans granted to Portuguese family members between December 2009 and July 2019 (https://bpstat.bportugal.pt/dados/explorer), but also the behavior of those loans, and their forecast for the following 6 months).

As known, in 2009, one of the biggest economic and social crises in Europe was taking place. This crisis was a consequence of the economical and financial crisis that took place in 2008 in the USA, when the American Investment Bank Lehman Brothers went bankruptcy. Due to the interdependence of economies and financial systems, Portugal was unable to escape from this crisis and had to ask for financial assistance in April 2011, having received several tranches over 4 years (2011-2014). Their total value was 78 billion euros and to receive them, Portugal was subject to successive evaluations. In 2012 the 7th evaluation of the troika took place where Portugal was not successful. Because of this, it had to adopt additional measures that led to economic, political and social consequences. During 2013, Portugal suffered an increase in unemployment, a decrease in exports and a lack of investment, which led to the indication that it would need more time to be able to reduce the public deficit [2].

There are several authors who argue that the creation of the single currency, the Euro,

and the integration of different countries with different economies was what led to the crisis then installed. The creation of the Euro led to low interest rates and an excess of liquidity in the markets, conducting to an increase in the number of credits. After that, countries like Portugal, which were unable to devalue the currency, were left with a high level of indebtedness leading to a deep crisis [1], these led us to be interested in the study described.
The multiple linear regression model is widely used for empirical analysis in economics and other social sciences, being the ordinary least squares model the most used to estimate the intended parameters. Our model was corrected since we're in presence of residual autocorrelation, using Prais–Winsten estimation. Several univariate models were used to the time series we want to do forecasting, such as Naive, simple average, moving average, exponential smoothing and Box-Jenkins. Using error estimation measures, the best one was found, allowing us to forecast until march 2020. Regarding the selected forecast model, it is concluded that the best suited one to the analyzed data is exponential smoothing, Holt, with alpha parameter equal to one.

Acknowledgements This project was partially funded by ANA, Aeroportos de Portugal and FCT - project SOC 4884/2019. It was carried out due to the collaboration between students of Master degree in Business and Management from ULHT and CPES of ULHT of Portugal.

References

[1] F. Fernandes. As medidas de austeridade debaixo da Troika: Uma análise à cobertura noticiosa dos Orçamentos de Estado de JN e Público. *Eikon Journal on Semiotics and Culture*, 1:37–56, 2017.

[2] M. Pereira. Crise económica e financeira: o enquadramento da sétima avaliação da troika ao programa de ajustamento português no jornal de negócios. *Estudos Em Comunicação*, 1:119–150, 2018.

Poster Session

24 October, 14:00 - 15:30, Zoom Room 1

Variation in abundance of the Azorean Buzzard due to habitat changes

Mónica Lopes[1], **Dulce G. Pereira**[2], **Anabela Afonso**[3], **Fátima Melo**[4]
[1] Department of Biology, University of the Azores, mokitas28@gmail.com
[2] CIMA/IIFA, Department of Mathematics/ECT, University of Evora, dgsp@uevora.pt
[3] CIMA/IIFA, Department of Mathematics/ECT, University of Evora, aafonso@uevora.pt
[4] Department of Biology, University of the Azores, maria.fc.melo@uac.pt

The main aim is to relate Common Buzzard *Buteo buteo rothschildi* abundance to habitat changes within Azores islands. Randomly selected plots were surveyed using the point counts sampling method. Buzzards avoid industrial areas, urbanisation, natural vegetation and agricultural land.

Keywords: gamma generalized mixed model, habitat use, island raptor, point count

The Azorean subspecies of the Common Buzzard *Buteo buteo rothschildi* [5], is an endemic taxon from the Azores and the only resident diurnal raptor there. Raptors are at the top of the food chains and are especially susceptible to habitat disturbance and destruction. Protecting raptor species is one way to improve biodiversity and enhance the sustainable use of natural resources. This study aimed to investigate the reasons explaining the variation in abundance of the Azorean Buzzard in the context of habitat changes.
The survey was conducted in 1998 and 2012 on S. Miguel, and in 1999 on Graciosa. Graciosa island was divided in three plots. In S. Miguel it was selected a random sample of plots (cells of a grid of 5*2km2). In each plot it was used the point count sampling method to survey raptors [2]. In S. Miguel in 1998 eleven plots were sampled four times (temporal replicates), and in 2012 two plots were surveyed just once and four plots four times. In Graciosa four replicates of point count surveys were made. It was recorded the abundance of buzzards in each plot, and encounter rate was computed as the total number of detected individuals divided by the total number of visits to points. Possible resightings of the same birds were discounted in all surveys. The mean of the observed encounter rate was higher in S. Miguel 2012 (2.3 $\pm$ SD 2.0), followed by S. Miguel 1998 (1.9 $\pm$ SD 1.3) and the lowest in Graciosa (1.0 $\pm$ SD 0.8) [3].
At each plot nine habitat variables were recorded each year: mean slope (in degrees) and eight land use classes established by SRAM DROTRh COSAçores [1]. In both islands, bare ground areas, lakes and industrial areas were too small or not observed. In both islands and all years, pasture is the dominant land use occupation, followed by natural vegetaton and forest. Significant differences were found between the percentage cover of all other land use variables (pasture, agriculture, forest, natural vegetation and urban) among islands/years (Kruskal-Wallis test, all $P < 0.05$).

According to the adjusted gamma generalized mixed model [4], pasture, forest, bare ground areas and lakes do not contribute to explaining the encounter rate per point count, nor the second order interactions among island/year and habitat land use variables (Table 1). Natural vegetation, urbanised land, industrial areas, agriculture land and steeper areas tend to be avoided by Azorean Buzzard. Among these variables industrial land was the factor showing the highest effect.
Despite agricultural land use in S. Miguel amounting only to 18.7% [1], avoidance of agricultural expansion should be considered in future management actions focused on this species.

Table 1: Gamma generalized mixed model of the encounter rate per point count + 0.1 with a log link function: fixed parameter estimates (B), estimated standard errors (SE), p-values (P) and variance component estimates for the habitat variables.

Explanatory variable	B	SE	P
Constant	1.178	0.352	0.001
Agriculture (%)	–0.013	0.007	0.064
Natural vegetation (%)	–0.013	0.004	0.002
Urban (%)	–0.023	0.007	0.002
Industrial (%)	–0.052	0.017	0.003
Slope	–0.020	0.007	0.003
$\sigma^2_{Island} = 0.172;\ \sigma^2_{Year} = 0.049;\ \sigma^2_{Residual} = 0.630$			

Acknowledgements A. Afonso and D.G. Pereira acknowledge partial funding by the fCT, Portugal, under the «UIDB/04674/2020 (CIMA)» project. To Regional Government of the Azores for a grant that supported the first year of study.

References

[1] SRAM DROTRh COSAçores. *Carta de ocupação do solo da região Autónoma dos Açores.* Nova Gráfica, Lda, Ponta Delgada, 2007.

[2] M. R. Fuller and J. A. Mosher. *Raptor Survey Techniques*, chapter 4, pages 37–65. National Wildlife Federation, Washingtion, 1987.

[3] M. Lopes, D. G. Pereira, A. Afonso, and F. Melo. Relative abundance of the azorean buzzard buteo buteo rothschildi and its responses to land use. *Ardeola*, 66(2):343–360, 2019.

[4] W. W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications.* CRC press, 2016.

[5] H. K. Swann. *A synopsis of the Accipitres (diurnal birds of prey): comprising species and subspecies described up to 1920, with their characters and distribution.* Wheldon & Wesley, London, 1922.

24 October, 14:00 - 15:30, Zoom Room 1

Survival rate: A non-transparent measure?

Carina Ferreira[1], **Teresa Abreu**[2], **Mário Basto**[3]
[1] Master Student, School of Technology, IPCA, Barcelos, Portugal
[2] Science Department, School of Technology, IPCA, Barcelos, Portugal
[3] Science Department, School of Technology, IPCA, Barcelos, Portugal

---

Understanding basic statistical literacy is necessary for health professionals and patients to understand health information and to allow informed consent to take place. In particular, the meaning behind the five-year survival rate, which is the most commonly used survival statistic in cancer. It consists of the ratio between the number of patients still alive five years later after diagnosis and the total number of patients diagnosed with cancer. Unlike this rate, the mortality rate in a given period, runs as the quotient between the number of cancer patients who die at the end of that period and the total number of people in the population (whether or not they have cancer).

Keywords: survival rate, mortality rate, bias

In a randomized controlled study of lung cancer screening in smoking men, lung cancer survival at 5 years was 35% for screened participants versus 19% for non-screened participants. Mortality rates were 4.4 deaths per 1000 person-years for the intervention group and 3.9 deaths per 1000 person-years for the control group [1]. In other words, screening increased the survival rate and, simultaneously, also increased the mortality rate. This contradiction is only apparent. When the apparent increase in survival comes from the fact that an earlier diagnosis corresponds to an artificially longer survival time, the distortion of the results is called lead-time bias (Figure 1). The lead-time is the period of time that mediates the time of detection of the disease and the appearance of the clinical manifestations.

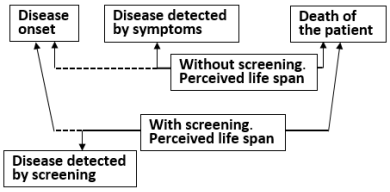

Figure 1: Lead-time bias.

In addition, the survival rate may skew the results in favor of screening even more, if the healthy volunteer bias is also taken into account, since the subjects who volunteer for screening, tend to be of higher social classes, have better health care or eat healthier, as

they can be more in line with the recommendations of health professionals. The most fast progressing cancers, which grow and spread rapidly, have a shorter preclinical period than the slow progressing ones. Screening tends to detect relatively more slow progressing cancers (more likely to be overdiagnosed), which will bias the survival time in favor of those who perform screening, named length time bias (Figure 2). Overdiagnosis bias is a particular form of this bias.

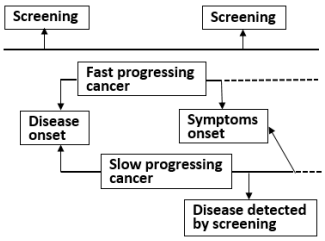

Figure 2: Length time bias.

The 5-year survival rate has other defects as a tool to measure the effectiveness of screening. For example, if a given therapy for cancer detected by early screening is toxic such that it shortens rather than prolongs the patient's life, if that reduction is less than the lead-time, it will still appear to prolong the patient's life.
A questionnaire about statistical literacy was implemented and made available online over a period of 3 months and 10 days, from 02/01/2019 to 12/12/2019, with 154 doctors, 142 nurses and 189 people from other professions responding to it. The question analyzed in this study was the following:
Suppose the survival rate for prostate cancer is 95% in country A and 55% in country B. A friend of yours is diagnosed as having prostate cancer. What recommendation would you give? A) Go for treatment in country A, B) Go for treatment in country B, C) The information is incomplete so that I can give you my opinion.
Actuality the information is incomplete. Nothing is known about the implementation of screening processes in those countries. Also important, nothing is known about mortality in each country, nor about other aspects such as the treatments carried out in each of them or the injuries of those same treatments.
Most physicians, 65.6%, pointed to the proper answer, followed by 50.7% of nurses and 36.0% of the general population ($\chi^2(2) = 30.079, p < 0.001$). These results point to a likely proper understanding of what the survival rate represents by most health professionals, the same not happening with the general population, which it seems to be more easily deceived by the biases inherent to the way the survival rate is defined.

References

[1] P.M. Marcus, E.J. Bergstralh, R.M. Fagerstrom, D.E. Williams, R. Fontana, W.F. Taylor, and P.C. Prorok. Lung cancer mortality in the mayo lung project: impact of extended follow-up. *JNCI: Journal of the National Cancer Institute*, 92:1308–1316, 2000.

24 October, 14:00 - 15:30, Zoom Room 1

Photointerpretation as a Tool to Support the Creation of an Ontology for Dolmens

Ariele Camara[1], **Ana de Almeida**[2], **João P. Oliveira**[3], **Matheus Silveira**[4]
[1] ISCTE-IUL Instituto Universitário de Lisboa, ISTAR-IUL-Information Sciences and Technologies and Architecture Research Centre, Portugal, ariele.camara@gmail.com
[2] ISCTE-IUL Instituto Universitário de Lisboa, CISUC- Centre for Informatics and Systems of the University of Coimbra and ISTAR-IUL Information Sciences and Technologies and Architecture Research Centre, Portugal, Ana.Almeida@iscte-iul.pt
[3] ISCTE-IUL Instituto Universitário de Lisboa, IT-IUL Instituto de Telecomunicações and ISTAR-IUL- Information Sciences and Technologies and Architecture Research Centre, Portugal, Joao.P.Oliveira@iscte-iul.pt
[4] Paragon Labs, matheus@paragonlabs.com

The use of ontologies can provide a general and standardised knowledge base, which allows to represent patterns logically. This work aims at the construction of an ontology for representing the information about dolmens from aerial and satellite photo interpretation techniques. The use of this ontology will provide a consistent basis for knowledge-oriented systems, allowing, in addition to an improvement, the development of new approaches in pattern recognition systems for computational vision.

Keywords: archaeology, dolmens, ontology, computational vision.

Methods to work with large amounts of data, from different locations and formats, have been of great importance in the last decade, due to the exponential increase in the information that people generate. This increase in data has led to the need of analysing the semantic connections between concepts and data relationships, and ontologies can be used to model and structure existing data appropriately [4].
The architecture of the ontology (knowledge base) is similar to that of the knowledge graph (Knowledge Graph, KG), and its conceptual separation is difficult since there is no consensus among existing state of the art literature. However, according to Ehrlinger and Wöß (2016) "A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge." [2].
Ontologies allow organising all the existing knowledge logically on various themes, functioning as a methodology that allows integrating and representing all information from different domains. Moreover, ontologies are useful for those who work with Geographical Information Systems (GIS), as they make explicit and formal declarations of how phenomena are represented [3]. The archaeological field lacks tools to communicate, share and reuse knowledge that can be understood by humans and machines. However, there already

exist databases with archaeological information that can be used to populate an ontology, such as Direção Geral do Património Cultural (DGPC). For this research, we use the information provided by the portal of DGPC and the visual interpretation keys extracted previously by Câmara and Batista [1].
Additionally, understanding the vegetation aspects is also important in archaeologic scope. For identifying and visualising archaeological monuments, a set of features must be identified by their surrounding landscape. To collect these data we extracted information from Land Use Map (COS) from Portugal.
In this work, we show the processes for implementing a graph-based knowledge representation tool that represent the information within the dolmens domain. To accomplish the construction of a domain ontology to represent information about dolmens we used a graph where:

- the vertices (nodes) of the graph are the entities we want to represent; labels are used here to group the entities in their own class, that is, dolmens, water lines, relief, state of conservation, etc;
- the edges represent the relationships between the entities, i.e., how they interact;
- each node and vertices have their own properties: values to identify each attribute, like the presence or absence of a hat, size, shape, etc.

Dolmens are monuments with a vast architectonic polymorphism and the landscape around them is not linear. The use of graphs to represent the information on this domain allowed us to identify both the patterns of dolmens' features and the relationship between these buildings and the landscape.
This ontology will also serve to assign features to the area where these monuments are inserted, thus helping in the development of an automatic classification system to detect dolmens in remote sensing images.

References

[1] A. Câmara and T. Batista. Photo interpretation and gis as a support tool for archaeology. *Journal on Advances in Theoretical and Applied Informatics*, 3(1):116–120, 2017.

[2] L. Ehrlinger and W. Wöß. Towards a definition of knowledge graphs. *SEMANTICS (Posters, Demos, SuCCESS)*, 48, 2016.

[3] F. T. Fonseca, M. J. Egenhofer, P. Agouris, and G. Câmara. Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6(3):231–257, 2002.

[4] I. Robinson, J. Webber, and E. Eifrem. *Graph databases*. O'Reilly Media, Inc., 2013.

24 October, 14:00 - 15:30, Zoom Room 1

Profiling clusters of European electricity markets

Margarida G. M. S. Cardoso[1], **Ana Martins**[2], **João Lagarto**[3]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal
[2] Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal
[3] Instituto Superior de Engenharia de Lisboa and INESC-ID, Lisboa, Portugal

---

We aim to characterize clusters of European regions with similar electricity prices behavior. Since electricity prices are influenced by many drivers such as demand, fuel and CO_2 emission allowances prices and renewable and non-renewable energy production, we integrate in the analysis, time series data regarding multiple related predictors. The insights on the clusters are obtained via descriptive and exploratory techniques including classification trees. Auxiliary analysis includes building adequate features from time series (e.g. Discrete Fourier transform are considered).

Keywords: electricity markets, time series, classification trees

---

The creation of an integrated electricity market at the European Union (EU) level is seen as one important step to increase the competitiveness of the EU economy, as well as, to contribute to the security of energy supply of the EU member states and to the sustainability goals. For this reason, the European Commission through its 2009/72/EC directive has established common rules to attain an internal market of electricity in the EU. This directive enables European citizens and businesses to choose their supplier and creates new business opportunities while enhancing cross-border trade.

In the present analysis we aim to characterize clusters of European regions with similar electricity prices behavior – [2]. K-medoids [3] is used to constitute the clusters, based on the combination of several dissimilarity measures – namely Euclidean distance and measures based on Pearson correlation, periodogram and autocorrelation. Clustering data refers to hourly prices of electricity (in €/MWh), observed in the day-ahead in 2018, for 26 regions of Europe (regions in the MIBEL, the Italian, the Nordpool, in the French and German markets).

Electricity market prices are extremely volatile and influenced by many drivers such as demand, fuels and CO_2 emission allowances prices, and renewable and non-renewable production.

The influence of demand on market prices, stems from the fact that, for the same set of technologies put in place to produce electricity, as demand increases there is a need to switch on more costly power plants. Since demand presents a daily seasonal behavior, lower between midnight and early hours of the day (off-peak hours) and higher during the

day and evening hours (peak hours), prices also tend to present the same daily seasonal behavior.
Also, fuel and CO_2 emission allowances prices play an important role in determining electricity market prices, since to produce electricity burning fossil fuels and emitting CO_2 might be required. This is the case when, to satisfy demand, there is a need to switch on power plants that use coal, natural gas or fuel oil. Since these fuels and CO_2 emissions allowances must be acquired in the respective market, the prices at which these fuels are purchased influences electricity production costs and, thus, electricity market prices.
Other important driver to influence electricity market prices is the amount of electricity produced from the different technologies, since different technologies present different costs, from the near zero costs of renewable technologies (hydro, wind, solar, etc.) to the more expensive non-renewable technologies (coal, natural gas and fuel oil).
Taking into account the data available regarding the variables referred that may influence electricity prices we profile the 6 clusters derived from K-Medoids. First, descriptive statistics using correlation and auto correlation indicators provide useful insights into the clusters. Furthermore, the use of classification trees - using R package "rpart" [4], - enables to gain some understanding of the relationship between multiple correlated predictors and the clusters. Regarding the use of different technologies, for example, we find that clusters mainly differ in Biomass, Solar and Wind production (the tree produced has a 100% precision which is adequate for descriptive purposes). On the other hand, analyzing a 247x 8762 data matrix, referring to hourly production data during 2018, we can identify some periods that contribute more to distinguish between the clusters (e.g. some time periods in June 20). Finally, we also build some classifiers based on features we first extract from time series data, namely using Discrete Fourier transform, enabling dimension reduction [1].

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grants UIDB/00315/2020 and UIDB/50021/2020.

References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In *Internacional conference on foundations of data organizations and algorithms*, pages 69–84. Springer, 1993.

[2] M. G. M. S. Cardoso, A. Martins, and J. Lagarto. Combining various dissimilarity measures for clustering electricity market prices. *Book of Abstracts of XXIV Congresso da Sociedade Portuguesa de Estatística, Amarante, 2019*, pages 248–249, 2019.

[3] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons, 2009.

[4] T. Therneau and B. Atkinson. Package 'rpart'. *URL: https://cran.r-project.org/web/packages/rpart/index.html (available 18.02.2020)*, 2019.

24 October, 14:00 - 15:30, Zoom Room 1

Prediction of tides using data in near-real time

Dora Carinhas[1], **Paulo Infante**[2], **António Martinho**[3]
[1] Instituto Hidrográfico, IIFA/Universidade de Évora
[2] CIMA/IIFA and DMAT/ECT, Universidade de Évora
[3] Marinha Portuguesa

Accurate analysis and forecasting of tidal level are very important tasks for human activities in oceanic and coastal areas. They can be crucial in catastrophic situations like occurrences of storms or tsunamis. Conventional tidal prediction methods are based on harmonic analyses using the least squares method to determine harmonic parameters. Harmonic tidal predition methods are often problematic when the contribution of non-astronomical components, such as weather, is significant. A procedure for correcting harmonic method prediction using recent observations proved most effective and was also successfully aplied to sea-level records.

Keywords: forecasting, tide, tide gauge, time series

The classical harmonic method of tidal analysis and prediction is long-established, having beem developed by Laplace, Lord Kelvin and George Darwin [2] and further advanced by Doodson [3] and Cartwright and Tayler [1] among others.
Intense coastal floods can occur when extreme weather phenomena such as tropical storms or typhoons are coincidental. Weather conditions are the main cause of differences between predicted and observed tide heights (Figure 1 shows these differences between the forecasts and the observations), with greater intensity in the winter periods [4].
When data for the observed periods are lost or incomplete, methods like harmonic analysis are not effective in supplementing the lost data. Therefore, in such cases it is important to find an accurate tidal level prediction technique. For this reason, recently artificial neural networks have been used in the literature as an alternative approach. Based on limited field data, the neural network method can predict hourly, daily, weekly or monthly tidal level more accurately than, for example, harmonic analysis methods.
In recent years Singular Spectrum Analysis, used as a powerful technique in time series analysis, has been developed and applied to many practical problems in such diverse areas, such as meteorology and oceanography.
A new approach, correcting the forecast based on error modeling, was also applied to correct in near-real time the predictions.
This work aims to apply three different methodologies (singular spectrum analysis, artificial neural networks and correct the forecast with error models) to tidal forecasts on the Portuguese coast and evaluate their quality by comparing to the forecasts made by the Hydrographic Institute, using harmonic analysis.

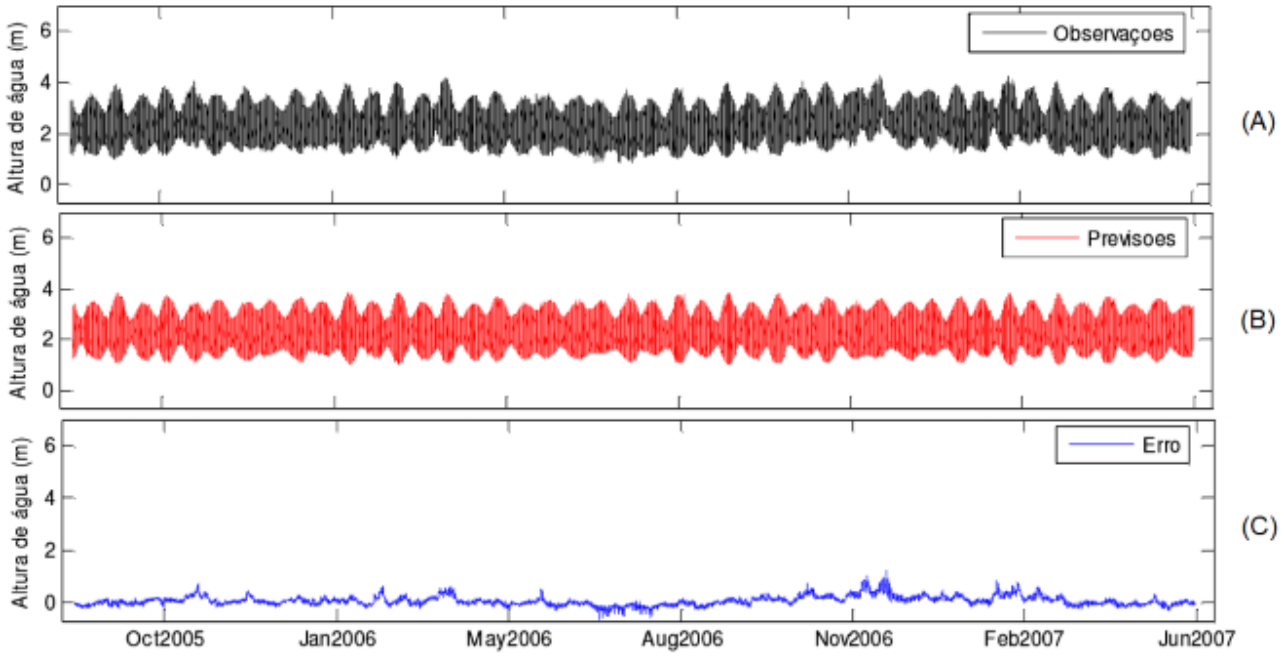

Figure 1: Comparison between observations and hourly forecasts in the Port of Caminha. (A) Tidal observations; (B) Predictions obtained through harmonic analysis; (C) Weather error. (source: Hydrographic Institute)

References

[1] D.E. Cartwrigth and R.J. Tayler. New computations of the tide-generating potential. *Geophys J R Astron Soc.*, 23:45–74, 1971.

[2] G.H. Darwin. *The tides and Kindred Phenomena in the solar system.* London: John Murray, 1911.

[3] A.T. Doodson. Harmonic developments of tue tide-generating potential. *Proc R Soc London*, A100:305–329, 1921.

[4] D.T. Pugh. *Tides, Surges and Mean Sea Leve.* Wiley, 1987.

24 October, 14:00 - 15:30, Zoom Room 1

A study of Aging and Cognitive performance using Symbolic Data Analysis

Sónia Dias[1], **Marta Neiva**[2], **Alice Bastos**[3]
[1] Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Viana do Castelo & LIAAD - INESC TEC, Universidade do Porto, sdias@estg.ipvc.pt
[2] Escola Superior de Educação - Instituto Politécnico de Viana do Castelo, marta.s.o.neiva@gmail.com
[3] Escola Superior de Educação - Instituto Politécnico de Viana do Castelo & CINTESIS, abastos@ese.ipvc.pt

---

Present societies are confronted with one of humanities major transformations - longevity, resultant from growing life expectancy and decreased birth rate. The aging process is associated to the deterioration of functional capacity, which is progressive with age and influences cognitive performance and the degree of independence to execute activities of daily living. So, it is important to study the cognitive performance of the older adults and analyze the influence of the social characteristics in their ability to perform activities of daily living and cognitive performance. In this study, the focus is not on understanding what happens with each older adults but rather to analyze the behavior of certain groups of individuals. As such, in here we employ a recent statistical approach, named Symbolic Data Analysis, that allows to analyze groups of individuals with certain characteristics without loss of information concerning the data associated with each group. Results indicate that in the third and fourth ages, and despite educational level, the percentage of men with cognitive deficit is lower than that of women. This study demonstrates the usefulness of using advanced statistical techniques such as Symbolic Data Analysis in the field of Psychology, or more specifically Gerontology.

Keywords: Symbolic Data Analysis, Modal-valued variables, Aging, Cognitive performance, Social Gerontology

---

The growth of the elderly population is a worldwide phenomenon. The proportion of people over 60 years of age grows faster than any other age group, and it is expected that in the year 2025 there will be about 1.2 billion people over 60 years of age, corresponding to an increase of 223% since 1970. Population aging arises mainly due to improvements in technology and health conditions in society. We are thus facing an increasingly aging world, where there is a decline in the proportion of children and young people and an increase in the proportion of people over 60, making the age pyramid become similar to that of a cylinder [1]. Consequently, there is a growing concern to study the aging process to try

to understand the factors that can influence the deterioration of physical and cognitive abilities and to understand the characteristics of the elderly who present better or worse cognitive performances and better or worse capacities to their activities of daily living. Providing successful aging is one of the concerns of today's societies.
The aims of this study are 1) to evaluate the cognitive performance of groups of individuals with certain characteristics and 2) to know the social characteristics of the groups with different levels of dependency to perform activities of daily living and cognitive capacities. The initial database with 324 individuals - microdata, was collected through a sociodemographic sheet, the Lawton Instrumental Activities of Daily Living Scale and the Mini Mental State Examination (MMSE). In accordance with the Symbolic Data Analysis approach [2, 3], in the first study the cognitive performance was analyzed in groups of third and fourth age (65-79 ages and 80+ ages) separated by gender and level of schooling. In the second study, the social characteristics are analyzed in groups with and without cognitive capacities separated in different dependence levels in daily living activities. In symbolic data tables - macrodata, the information of the variables for each group is not lost. In these tables the values of the variables associated to each group are not single values or categories, but the distribution of all records associated to such variable, for that group. As values of the single values associated to the variables in microdata are categorical and discrete, the type of symbolic variables used in this work are the modal-valued variables [2].
The symbolic data analysis allowed the description and comparison of different groups segmented by gender, age group and educational level, regarding their cognitive capacities evaluated using the MMSE [4]. Results indicate that in the third and fourth ages, and despite educational level, the percentage of men with cognitive deficit is lower than that of women. Taking into account the groups obtained through aggregation of cognitive performance and level of dependence in daily living, the variables that appear to influence the most the change from groups without cognitive deficit to groups with cognitive deficit are mainly the variables gender, occupation and educational level.
In summary, the study of group behavior using symbolic data analysis grants the possibility to analyze in more detail information concerning every variable in each group. Even though this is mostly a descriptive study, symbolic data analysis appears to have significant potential to gerontological practice considering the level of specification it generates.

References

[1] V. Barnett. *Active Ageing, a Policy Framework.* World Health Organization., Geneva, 2002.

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

[3] M. Noirhomme Fraiture and P. Brito. Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining*, 4(2):157–170, 2011.

[4] M. Simões and I. Santana. *Escalas e Testes de Demência.* Novartis., Lisboa, 2015.

24 October, 14:00 - 15:30, Zoom Room 1

The Sustainable Society Index: its reliability and validity

Nikolai Witulski[1], **José G. Dias**[2]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, nwiii@iscte-iul.pt
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, jose.dias@iscte-iul.pt

---

The Sustainable Society Index (SSI) is known as a comprehensive index that contains substantive aspects of all three dimensions of sustainable development (SD): social, environmental, and economic. This paper assesses the reliability (internal consistency) and external validity of the SSI for 154 developing and developed countries for the year 2016 using confirmatory factor analysis and standard measures of reliability. Our results show that a simpler version of the SSI achieves construct reliability, i.e., the three modified indices of the SSI show strong internal consistency. The external validity of the modified indices is supported as the country rankings are similar to those of the HDI and EPI and show a high correlation in 2014 and 2016.

Keywords: Sustainable Society Index (SSI), Human Development Index (HDI), Environmental Performance Index (EPI), reliability, validity, factor analysis

---

The improvements in the social, environmental, and economic dimensions are important and highlight progress toward a more balanced future, but further advancements are required in many areas. The selection of indicators to measure these dimensions, i.e., sustainability, was addressed and critically reviewed by its own authors [3]. In 2012 the Joint Research Centre (JRC) of the European Commission audited the SSI and confirmed that it is conceptually coherent and meets the requirements of the JRC. They concluded that the SSI is "suited to assess nations' development towards sustainability in its broad sense: Human, Environmental and Economic Wellbeing" [2].
This study provides a statistical analysis of the reliability and validity analysis of the three SSI dimensions for 154 developed and developing countries. We apply confirmatory factor analysis (CFA) and propose modified dimensions to increase its statistical reliability. CFA is therefore adequate as it is used to determine whether a set of indicators has a common underlying construct (latent variable) [1]. To assess the model fit of each dimension, we apply the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Standardized Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). To assess the internal consistency of the three dimensions, we use three standard measures: Cronbach's alpha, composite reliability (CR), and average variance

extraction (AVE). The external validity is analyzed by comparing the country rankings of each modified dimension with well acknowledged and widely used indices that focus on these specific dimensions (HDI - Human Development Index and EPI - Environmental Performance Index) and by calculating Kendall rank correlation coefficients in 2014 and 2016.
Results show that the internal consistency of the dimensions is supported by the good model fit of the overall model (three dimensions of the SSI) and by the standard measure of reliability (for the social and environmental dimensions). The external validity is confirmed by comparing the country rankings of each dimension (based on their scores) with the HDI and the EPI rankings and Kendall rank correlation coefficients.
The SSI conceptual framework can capture the overall picture of sustainability and its modified measurement version that estimates the three dimensions simultaneously improves the convergence with well-known partial indices: the HDI for social and economic and the EPI for the environmental component. Thus, the statistical analysis supports the modification of the original set of indicators, which increases the quality of the measurement of the underlying dimensions.

Acknowledgements Funding from Fundação para a Ciência e Tecnologia (Portugal), UID/GES/00315/2019.

References

[1] T. A. Brown. *Confirmatory Factor Analysis for Applied Research.* Guilford Publications, New York, second edition, 2015.

[2] M. Saisana and D. Philippas. Sustainable society index (SSI): taking societies' pulse along social, environmental and economic issues. the joint research centre audit on the ssi report eur 25578. Technical report, Publications Office of the European Union, Luxembourg, 2012.

[3] G. van de Kerk and A. R. Manuel. A comprehensive index for a sustainable society: the SSI - the sustainable society index. *Ecological Economics*, 66:228–242, 2008.

24 October, 14:00 - 15:30, Zoom Room 1

Statistical Models for Environmental Processes

Carla Silva[1], **Susana Faria**[2], **A. Manuela Gonçalves**[3]
[1] University of Minho, Mathematics Department, Guimarães, Portugal, cscgsilva@gmail.com
[2] University of Minho, Mathematics Department and CBMA, Guimarães, Portugal, sfaria@math.uminho.pt
[3] University of Minho, Mathematics Department and CMAT, Guimarães, Portugal, mneves@math.uminho.pt

---

In the context of a surface water quality monitoring problem in a river basin, it is proposed an approach based on spatial and temporal models in order to analyze and evaluate the temporal evolution of the time series relative to the Dissolved Oxygen (DO) quality variable. This quality variable was measured monthly from March 2002 to February 2013. Linear Mixed Effects Models were proposed since there are repeated measurements over time in experimental units, with great variability between them.

Keywords: watershed, Douro river, water quality, geostatistics, mixed effects models

---

Environmental degradation is nowadays a critical issue, both due to the difficulty of restoration and rehabilitation and to the serious social and economic consequences. The environmental crisis is partially the result of many man-made mistakes that still remain visible today. Investigations aimed at curbing or estimating environmental problems have led to a more in-depth study of methods to better understand the data associated with these problems.

There has been an increase in the number of methodologies in the area of Statistics for modeling environmental processes and, in particular, in processes of modeling Water Quality in the surface of a hydrographic basin.

This study investigates a problem in the context of surface water quality monitoring in a watershed, and we propose an approach based on spatial and temporal models in order to analyze and evaluate the time series evolution of environmental variables.

The data refer to the Douro watershed located in northern Portugal and for the modeling process we considered time series relative to the Dissolved Oxygen quality variable measured monthly from March 2002 to February 2013.

In order to obtain estimates of monthly precipitation values, in area, in the quality sampling stations (where there are no precipitation measurements), we developed a methodology using spatial stochastic processes (Kriging) to be applied to the precipitation data extant in this basin. The estimated values will represent the hydrometeorological factor in the quality sampling stations for the Dissolved Oxygen modeling process.

For the Dissolved Oxygen modeling process we established Linear Mixed Effects Models, as they show versatility and flexibility in including random effects, in incorporating trend and seasonality components, covariates (such as the hydrometeorological factor and other surface water quality variables), as well as the temporal correlation structure typical of the environmental series. Trend was modeled using a linear term or a polynomial, while seasonality was modeled using sine and cosine terms with period of one year.
The random structure was identified through the estimates of the confidence intervals for the individual adjustment parameters, and the significance of the included covariates was evaluated based on the Likelihood Ratio Test, in the case of nested models, and by the AIC / BIC criterion, in models not nested. In modeling the correlation structure of random errors, an order 1 autoregressive model was used.
The final model was the model with only a random effect on the constant term and a significant and positive association was observed between the biochemical oxygen demand, pH value, hydrometeorological factor and the concentration of Dissolved Oxygen, but a significant and negative association between the temperature and the concentration of Dissolved Oxygen.

Acknowledgements This research was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020, UIDP/00013/2020 and UID/BIA/ 04050/2019.

References

[1] M. Costa and A. M. Gonçalves. Combining statistical methodologies in water quality monitoring in a hydrological basin-space and time approaches. *Water Quality Monitoring and Assessment*, pages 121–142, 2012.

[2] N. Cressie, A. Zammit-Mangion, and C. K. Wikle. *Spatio-Temporal Statistics with R.* Chapman & Hall/CRC, 2019.

[3] A. M. Gonçalves and M. Costa. Predicting seasonal and hydro-meteorological impact in environmental variables modelling via kalman filtering. *Stochastic Environmental Research and Risk Assessment*, 27(5):1021 –1038, 2013.

[4] J.C. Pinheiro and D.M. Bates. *Mixed-Effects Models in S and S-Plus.* Springer, 2000.

24 October, 14:00 - 15:30, Zoom Room 1

Performance of Portuguese Students: a bivariate multilevel analysis

Susana Faria[1], **Carla Salgado**[2]
[1] University of Minho, Mathematics Department and CBMA, Guimarães, Portugal , sfaria@math.uminho.pt
[2] University of Minho, Mathematics Department, salgcarla@gmail.com

---

In this work, we illustrate how to perform a bivariate multilevel analysis in the complex setting of large-scale assessment surveys. In particular, a bivariate multilevel model is applied to the Portuguese data from Program for International Student Assessment (PISA) 2015 with the aim to identify a relationship between students' mathematics and science test and the characteristics of students.

Keywords: students' achievement, multilevel regression models, Programme for International Student Assessment (PISA) 2015.

Several studies have been conducted to identify the different factors that infuence students' school performance. International programmes of educational achievement, such as the Programme for International Student Assessment (PISA) run by the Organisation for Economic Cooperation and Development (OECD) are being developed, which makes it easier to obtain information to carry out these studies (see [4]).
The PISA survey collects information in three areas of competency (mathematics, reading and science) based on test scores. This survey is a self-administred questionnaire that tests student skills and gathers information on several facets of each student's family, home and school background.
Studies based on PISA data to explain student performance are very common in literature (see [1] and [3]).
The PISA data are hierarchically structured, in which students are nested within schools, in turn, schools are nested within regions and regions are nested within countries. Given the hierarchical structure of data, the models adopted for statistical analysis were multilevel regression models, which can take into account data variability within and among the hierarchical levels (see Goldstein [2]).
The purpose of this study was to identify a relationship between students' mathematics and science test scores and the characteristics of students themselves. Data on about 7325 Portuguese students and 246 schools who participated in PISA-2015 were used to accomplish our objectives. A bivariate two-level linear models in which students (level 1) are nested in schools (level 2) was fitted.

The PISA-2015 computed 10 plausible values (PVs) of mathematics and science test scores to measure student's performance and all PVs were simultaneously considered as the dependent variables. The analysis considered the average coefficients derived from using all the ten plausible values for each competence. As PISA prescribes, we used the student weight provided by PISA in our models.
All the analysis were carried out using R statistical software.
The results obtained by this approach are in line with the existing research: the index of the socioeconomic status of the student, being a male student, the total number of students in school and the proportion of girls in school positively influence the student's performance in mathematics and science. On the other hand, the grade repetition had a negative influence on the performance of the Portuguese student in Mathematics and Science.

Acknowledgements This work was supported by the strategic programme UID/BIA/04050/2019 funded by national funds through the FCT I.P.

References

[1] Tommaso Agasisti and Jose M. Cordero-Ferrera. Educational disparities across regions: A multilevel analysis for italy and spain. *Journal of Policy Modeling*, 35(6):1079–1102, 2013.

[2] H. Goldstein. *Multilevel Statistical Models*. John Wiley and Sons, 2011.

[3] C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni. Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics*, 44(7):1296–1317, 2017.

[4] OECD. *PISA 2015 Results (Volume I)*. Paris: Organisation for Economic Co-operation and Development, 2016.

24 October, 14:00 - 15:30, Zoom Room 1

Bootstrap method in the Analysis of Variance for data from von Mises-Fisher distributions

Adelaide Figueiredo[1]
[1] Faculty of Economics of University of Porto and LIAAD-INESC TEC Porto, adelaide@fep.up.pt

An important problem in directional statistics is to test the null hypothesis of a common mean vector across several populations. In this study we consider the analysis of variance to test the equality of the mean vectors of several von Mises-Fisher distributions. As this test is only valid for large concentration parameters, we suggest to use an alternative test such as a bootstrap test in the analysis of variance. We compare the empirical power of the tests for data from two von Mises-Fisher populations with equal or different concentration parameters.

Keywords: Bootstrap, Hypersphere, Monte-Carlo methods, von-Mises Fisher distribution

Directional statistics deals with observations that are directions, i.e., unit vectors in $\mathbb{R}^q$ or points of the unit sphere in $\mathbb{R}^q$ denoted by S_{q-1}. In most cases, the observations lie on the circumference of the unit circle S_1 or on the surface of the unit sphere S_2 (see for instance, [4]), but the observations can also lie on the surface of the unit hypersphere S_{q-1}, with $q \geq 4$. There are recent applications of directional data on the hypersphere in text analysis, gene expression profiles, neuroscience, bioinformatics (see for instance, [1], [2], [3]).

In this study we consider one of the most used distributions to model vectorial data, the von Mises-Fisher distribution. This distribution has two parameters: a mean vector and a concentration parameter that measures the concentration around the mean vector. It is a rotationally symmetric distribution about the mean vector and its probability density function is relatively simple, only the normalising constant is based on the modified Bessel function of the first kind. We focus on the important problem of testing the null hypothesis of the equality of the mean vectors of several von Mises-Fisher populations. We refer the analysis of variance test for this case and as this test is valid only for large concentrations, we suggest using an alternative test based on a bootstrap approach in the analysis of variance. We carried out a simulation study to compare the empirical power of the tests to investigate whether the bootstrap test is preferable to the analysis of variance test for small concentrations.

Acknowledgements This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme and by National Funds through the FCT within project FCOMP-01-0124-FEDER-037281.

References

[1] A. Banerjee, I.S. Dhillon, J. Ghosh, and Sra S. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345 – 1382, 2005.

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

[3] C. Ley and T. Verdebout. *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman and Hall, CRC, 2018.

[4] G.S. Watson. *Statistics on spheres*. Wiley-Interscience, 1983.

24 October, 14:00 - 15:30, Zoom Room 1

Statistical Modeling in the Pay-As-You-Throw System in a Local Public Company

A. Manuela Gonçalves[1], **Vítor Silva**[2], **Laura Jota**[3], **Vítor Pinheiro**[4]
[1] CMAT-Center of Mathematics, DMAT-Department of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[2] DMAT-Department of Mathematics, University of Minho, VITRUS AMBIENTE, EM, S.A., Portugal, vitor.silva@vitrusambiente.pt
[3] VITRUS AMBIENTE, EM, S.A., Portugal, laura.jota@vitrusambiente.pt
[3] VITRUS AMBIENTE, EM, S.A., Portugal, vitor.pinheiro@vitrusambiente.pt

---

A pioneering project in Portugal called Pay-As-You-Throw (PAYT) was implemented in the Portuguese city of Guimarães and is managed by the municipal company VITRUS AMBIENTE. This work focused on the analysis and modeling of data using Linear Regression models in the study of the factors that influence municipal solid waste production produced in the area of the implementation of the PAYT system. The modeling processes were carried out based on data collected between April 2016 and May 2019.

Keywords: linear regression, modeling, PAYT, recycling, waste management

---

Due to the economic and social development in general and to population growth, the amount of waste, particularly municipal waste, has been significantly increasing in recent years. It is one of the major problems, both at a national and global levels, and action is urgently needed to ensure that waste is recovered and its volume reduced. VITRUS AMBIENTE, EM, S.A. is a public company that operates at various levels in local business management, namely in Urban Waste Management, and ensures the collection of waste in the municipality of Guimarães.
In 2016 a pioneer project called Pay-As-You-Throw (PAYT) was implemented in the Historic Center of the city of Guimarães [1]. This system is based on the polluter-payer principle and on the concept of shared responsibility, according to which those who generate less waste pay less. VITRUS is the managing entity and the Urban Hygiene Service is responsible for implementing the necessary measures to ensure the success of this project. This work focuses specifically on Waste Management and aims at modeling and predicting the behavior of Urban Waste production within the company's areas of activity. Thus, statistical models are developed in the context of Linear Regression Models (in a single and multiple modeling approach) to predict, in the observed periods, the waste production in the areas of the undifferentiated collection circuits and in the pilot zone of the PAYT system implementation.

The main objective of this work is both to evaluate the influence of factors related to the amount of waste collected in the areas covered by the Urban Hygiene Service and to analyze the evolution of the respective quantities produced. Thus, we use Linear Regression Models ([2], [3]) to both identify the factors that, from a business perspective, do influence the amount of waste produced in the pilot zone of the PAYT system implementation and possible seasonal trends and patterns, in order to further improve management actions to be implemented by the company.
The methodologies used support the company's management and decision-making process regarding Urban Waste Management, aiming at improving the services provided to the population and always having the preservation of the environment as its cornerstone.

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

References

[1] B. Bilitewski. Pay-as-you-throw–A tool for urban waste management. *Waste management*, 12(28):2759, 2008.

[2] L. Fahrmeir, Th. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications.* Springer, 2013.

[3] A. Sen and M. Srivastava. *Regression analysis: theory, methods, and applications.* Springer Texts in Statistics, 2012.

24 October, 14:00 - 15:30, Zoom Room 1

Detection of solar production in smart grids

Conceição Rocha[1], **Ricardo Bessa**[1]
[1] INESC TEC, conceicao.n.rocha@inesctec.pt and ricardo.j.bessa@inesctec.pt

The planning and management of the distribution grid has faced several challenges with the introduction of renewable energies in the network. In particular, the costumers production of energy with photovoltaic panels, has increased the variability associated with the net power consumed at low voltage transformer stations. In this work we propose a method to identify if a transformer has or not production of photovoltaic energy in the smart grid based on the aggregated consumption recorded every 15 minutes during a period of four years. The method allows to correctly identify the transformers where the presence of photovoltaic production is high and achieves an accuracy of 86,5% when tested in a data set with photovoltaic production in a range between 1kW and 50 kW.

Keywords: clustering, functional data analysis, Mahalanobis-Wassertein distance, photovoltaic energy

The introduction of renewable energy sources in the production of electrical energy has posed several challenges in the distribution and storage of energy. However, with self-consumption production[1] the greatest concern is to be able to identify whether there is self-consumption in a given network and, if there is one, to determine the installed power. This information is crucial to plan the network in order to deal with transformer overloads and high voltages.

In particular, the use of photovoltaic energy to self-consumption is aggravated by the fact that it is not mandatory to have a meter of the energy produced, in the case of systems whose connection power is less than 1.5 kW, and if it is less than 200 W it is not even mandatory registration. In addition to this concern, there is also a concern with the connection to the grid of illegal photovoltaic solar panels.

Self-consumption, through photovoltaic energy, has increased the variability associated with the net power consumed at low voltage transformer stations. In addition to the variability associated with the profile of each customer, the day of the week, the time of day or the season, etc., we now have the variability associated with climatic factors, such as the radiation of the sun and the temperature throughout the day that is felt on site. These factors govern the amount of energy produced in the photovoltaic panels and, consequently, the power consumed and the power injected into the grid.

[1]Self-consumption production means that the electric energy produced by the system is used to supply the needs of the producer and, in the event of an excess, this is injected into the public service electricity grid.

As far as our knowledge goes, no work has been carried out to identify the presence of self-consumption photovoltaic production in smart grids. The published works are essentially focused on studying the effect of their presence, either on the distribution network or on transformers, or on measures to be taken to prevent possible network overloads or voltage drops in the network. Hence, in this work we present a method for identifying the presence of self-consumption in low voltage transformers.
For this study we have:

- photovoltaic production of several solar panels, recorded at 15-minute intervals during three years.
- hourly radiation forecast for the three years.
- power consumed at a low voltage station during a period of four years. Values recorded every 15 minutes.

Taking into account the characteristics of the phenomenon, the method we propose is based on the comparison of the consumption observed on days whose forecast is of high radiation and clear skies, with the consumption observed on days when the forecast is of low radiation and cloudy skies.
As pre-processing, functional data analysis and unsupervised classification techniques are applied to the daily radiation series to obtain the days of the two data sets to be used. To attenuate the variability of the consumed power, the daily series of the same are standardized by the daily mean and standard deviation and later the percentiles of the power consumed for the 31 moments of the day are determined, which correspond to the period from 9:45 to 17:15, for both sets. It is the difference between the percentiles of the 31 variables in the two sets that are used as variables in the hierarchical classification. The Mahalanobis-Wassertein distance and Ward's aggregation criterion are used to classify the transformers in one of the two groups, presence of self-consumption or absence of self-consumption.
To investigate the performance of our method we simulated a data set with near 800 aggregated power consumption. All them have self-consumption in a range between 1 kW and 50 kW. We identify 86,5% of the cases with self-consumption. Furthermore, we observe that the cases that were incorrectly classified are the cases in which the photovoltaic production is relatively low, almost always less than 11% of the mean power consumed.

Acknowledgements This project was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER.

References

[1] M. A. Awadallah, B. Venkatesh, and B. N. Singh. Impact of solar panels on power quality of distribution networks and transformers. *Canadian Journal of Electrical and Computer Engineering*, 38(1):45–51, 2015.

[2] A. Verde, R.and Irpino. Comparing histogram data using a mahalanobis–wasserstein distance. In Paula Brito, editor, *COMPSTAT 2008*, pages 77–89, Heidelberg, 2008. Physica-Verlag HD.

24 October, 14:00 - 15:30, Zoom Room 1

Using common exploratory statistical tools to interpret acoustic suspended sediment response in the Portuguese inner shelf

Ana Isabel Santos[1], **Dora Carinhas**[2], **Anabela Oliveira**[3]
[1] Instituto Hidrográfico; IDL/Faculdade de Ciências da Universidade de Lisboa, ana.santos@hidrografico.pt
[2] Instituto Hidrográfico; IIFA/Universidade de Évora, dora.carinhas@hidrografico.pt
[3] Instituto Hidrográfico, anabela.oliveira@hidrografico.pt

In the present work, an evaluation of the use of ADCP (acoustic Doppler current profiler) backscatter data as a sediment profiler is made, based on concurrent ADCP and LISST time series. Multivariate statistical techniques like cluster analysis was applied to the resulting variables in order to separate different populations of ADCP acoustic response through time and better understand how changeable suspended particle attributes (concentration and grain size) affect ADCP response.

Keywords: acoustics, ADCP, cluster analysis, entropy analysis, suspended sediments

---

The assessment of suspended sediment parameters from an ADCP (acoustic Doppler current profiler) relies on the premise that a relation exists between its response in the form of acoustic backscatter and the suspended sediment signature present in the measured water column ([1], [4]). Therefore, it is expectable that different suspended sediment signatures (concentration and particle size distributions) will yield different ADCP acoustic responses. Under this assumption, in this work, common exploratory statistical tools are applied to ADCP acoustic intensity time series and concurrent instrumental suspended sediment measurements (LISST – Laser in-situ Scattering and Transmissiometry) in order to determine if statistically distinct ADCP responses match statistically distinct suspended sediment signatures.
To this end, two ADCP datasets and concurrent LISST100 measurements (S. Pedro de Moel and Costa da Caparica) were subject to a series of exploratory multivariate statistical analyses in order to:

1. Verify the relations between ADCP acoustic intensity and suspended sediment parameters (Spearman correlation coefficients);

2. Identify and separate populations of acoustic response and characterized them in terms of concurrent suspended sediment signatures (k-means cluster analysis);

3. Identify and separate LISST grain size distributions (entropy analysis) to verify if the two sets of populations (clusters and entropy groups) match ([2], [3]).

The results of the statistical analysis and characterization of the identified populations in these two first datasets were then used to apply a similar analysis to a third ADCP dataset where no LISST or suspended sediment data was available to characterize the suspended sediment signature (Douro estuary). This cluster and entropy separation and characterization was then validated applying Mann-Whitney-Wilcoxon (MWW - for 2 populations) and Kruskal-Wallis (KW – for 3 or more populations) tests.
Three distinct ADCP cluster populations were identified with similar characteristics in both datasets: Population I ("higher fine"), characterized by higher than average acoustic intensity, higher than average sediment concentrations, higher than average fine acoustic class content and lower than average mean grain size; and Population II ("lower coarse") characterized by lower than average acoustic intensity, lower than average sediment concentrations, lower than average fine acoustic class content and higher than average mean grain size. A third population was identified and baptized as Population III ("standard") and characterized by average acoustic response, average sediment concentrations, average mean grain diameters and average fine acoustic class content. Non-parametric statistical tests (Kruskall-Wallis) were then applied to the resulting partitions proving that K-means clustering effectively separated statistically distinct suspended sediment populations.
LISST grain size spectra were grouped into populations using entropy analysis which has been recognized by several authors as the best method to analyze suspended sediment in-situ measurements. Entropy grouping only takes into account the grain size distribution of the measured sediments, disregarding its quantity or concentration. No definitive statistically significant match could be established between the resulting cluster and entropy populations (via MWW and KW tests); however, the identified populations yield similar interpretations of the two analyzed time series, in terms of suspended sediment response to different forcing scenarios.

References

[1] J.W. Gartner. Estimating suspended solids concentrations from backscatter intensity measured by acoustic doppler current profiler in san francisco bay, california. *Mar. Geol.*, 211:169–187, 2004.

[2] O.A. Mikkelse, P.S. Hill, T.G. Milligan, and R.J. Chant. In situ particle size distributions and volume concentrations from a lisst-100 laser particle sizer and a digital floc camera. *Cont. Shelf Res.*, 25:1959–1978, 2005.

[3] W.E. Sharp and P.F. Fan. A sorting index. *J. Geol.*, 71:76–84, 1963.

[4] P.D. Thorne and D.M. Hanes. A review of acoustic measurement of small-scale sediment processes. *Cont. Shelf Res.*, 22:603–632, 2002.

24 October, 14:00 - 15:30, Zoom Room 1

Willingness to pay for environmental quality in Portugal: an application of SEM

Paula Vicente[1], **Catarina Marques**[2], **Elizabeth Reis**[3]
[1] ISCTE – Instituto Universitário de Lisboa, BRU-ISCTE, paula.vicente@iscte-iul.pt
[2] ISCTE – Instituto Universitário de Lisboa, BRU-ISCTE, catarina.marques@iscte-iul.pt
[3] ISCTE – Instituto Universitário de Lisboa, BRU-ISCTE, elizabeth.reis@iscte-iul.pt

---

Environmental protection is a major concern in contemporary societies. The main goal of this paper is to investigate the willingness of Portuguese citizens to pay for environmental protection using a structural equation modeling approach. Results suggest that willingness to pay more for environmental quality is positively associated with environmental locus of control and civic participation.

Keywords: civic participation, environmental locus of control, environmental protection, pro-environmental behaviour, willingness to pay

---

Economics literature provides the most research on the factors that drive people to make financial contributions toward the provision of a public good such as the environment. This is because collecting taxes or increasing prices to favour the environment is regarded as a public policy with a fiscal framework. The extant literature reveals that the individual's willingness to pay for environment quality can be explained by a combination of socio-economic and demographic variables. The models underlying the analysis tend to be regression models, either logistic, ordered probit or Ordinary Least Squares depending on how the dependent variable is measured ([1], [2]). However, this methodological approach is simplistic since it assumes that the various factors influencing the willingness to pay for environment quality are not interrelated. In this paper we use structural equation modeling to explore the associations between several constructs in order to understand what drives individuals to contribute monetarily toward environmental quality; specifically, a multigroup analysis is conducted to assess the invariance of two educational level segments. Figure 1 presents the model under test.
Data was collected by means of a household survey covering the south of Portugal (territory below the Tagus River). Strata was defined by region according to NUTSIII. The sampling procedure was very like a random-route procedure. A sample of 595 respondents was obtained: 42.9% males, 26.7% aged 25-34 years, 56.8% highly educated (graduates or postgraduates), and 51.1% living in families with no children. The results show that Environmental Locus of Control and Civic Participation have the greatest (positive) impact on citizens' intention to pay for environmental quality, which validates the use of these

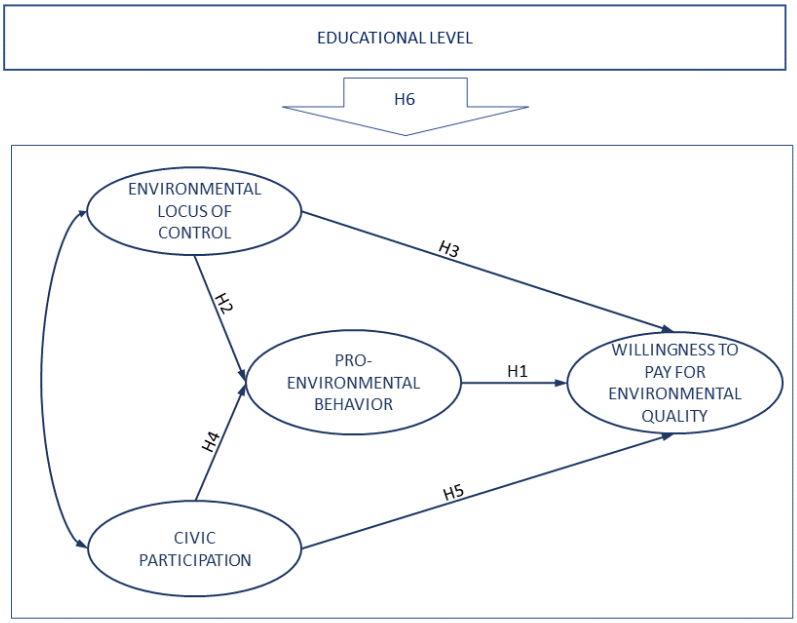

Figure 1: Conceptual model for willingness to pay for environment quality

constructs to influence citizens' behaviour and adherence to environmental protection policies; educational level was shown to influence Willingness to Pay model; specifically, the relationships identified as significant were stronger in the without university education subgroup than in the university education group.

References

[1] Potrafke N. Kauder, B. and H. Ursprung. Behavioral determinants of proclaimed support for environment protection policies. *European Journal of Political Economy*, 54:26–41, 2018.

[2] G. Marbuah. Willingness to pay for environmental quality and social capital influence in sweden. *FAERE Working Paper*, 2016.13, 2016.

24 October, 14:00 - 15:30, Zoom Room 1

Literacy About Waste Management in a Maritime Environment

M. Filomena Teodoro[1,2], **José B. Rebelo**[2], **Suzana Lampreia**[2]
[1] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Av. Rovisco Pais, 1, 1048-001 Lisboa, Portugal
[2]CINAV, Center of Naval Research, Portuguese Naval Academy, Portuguese Navy, Base Naval de Lisboa, Alfeite, 1910-001 Almada, Portugal

---

We pretend to characterize the profile of waste management in ships. With such purpose, was implemented a questionnaire to evaluate Knowledge, Attitudes and Practice (KAP) about the waste management during the boarded period. As first approach, the authors performed a preliminary statistical analysis using an incomplete sample of boarded population. The results evidenced that staff have extra care about the correct storage, but some times the knowledge how to handle certain kind of waste is not adequate. Using the complete sample, this study was extended, a statistical approach using an exploratory factorial analysis (EFA) and one factor analysis of variance are developed. The results using parametric and non parametric approach were similar. In this article we use a multivariate technique, the MANOVA approach.

Keywords: Waste management, environmental guidelines, questionnaire, environmental literacy, statistical approach

---

The earth surface is covered mostly by water in the liquid state, being a very important medium to balance of the entire climate system of planet. The sea biodiversity and natural resources contribute largely to global economic development. The growth of the global economy has a direct consequence: the increment of the sea pollution. This growth created a significant pressure at sea, specifically, the marine litter is a major cause of intense human activity. About 80% of the world trade volume transportation is maritime [6], with intense traffic of ships, being one of the main sources of pollution, generating solid wastes, sewage and wastes from hydrocarbons, also a contribute to atmospheric pollution.
This study started in [3, 1] where was compiled specific regulation about the waste management in ships. To characterize the profile of embarked staff in ships from Portuguese Navy was implemented a questionnaire [3] so we could evaluate their KAP. The first approach, using statistical descriptive techniques, the questionnaire results were summarized and analyzed in [5] using an incomplete sample.
The exploratory factorial analysis is adequate when we want to identify variables to create indexes or new variables without inter-correlated components. We consider the case

using raw data, taking the first 4 factors (with the same percentage of explained variance we can get less factors with raw data than when we consider scaled data). In the present case, we could get the 'meaning' for the first 4 factors. The interpretation of such meaning is done analyzing the rotated factors scores. We can identify a meaning for each one: F_1 combines variables that characterize $Awareness$, F_2 combines variables from $Hygiene\ and\ Safety$, F_3 combines variables from $Practice$, F_4 takes into account variables associated to $Knowledge$. Each of the selected factors (factors that have a bigger variance explanation) was used as independent variable in simple predictive models. The univariate ANOVA technique was also applied. Kind of ship, the attendance of training courses and the military hierarchical posts, were statistically significant [4]. Some times was used the parametric ANOVA modeling [2], in other cases, when the homogeneous variance test was rejected, was necessary to use the nonparametric approach.
Extending this approach, in the present work we use a multivariate technique, the MANOVA approach, were we can relate more than one dependent variable simultaneously with several factors so we can identify the most determinant variables to KAP. The kind of ship, the military hierarchical posts (category), attendance of an environmental training course, or the space to store waste have significant effects in KAP.
It was built an indicator of good practices combining the information of some questions. The results appear to be adequate and in accordance with what is expected.
The use of the results obtained in the present work gives a good contribution to support the decisions about improving the waste management done by embarked staff.

Acknowledgements This work was supported by Portuguese funds through the CEMAT, *The Portuguese Foundation for Science and Technology*, University of Lisbon, project UID/Multi/04621/2019, and CINAV, Portuguese Naval Academy.

References

[1] J. Monteiro. *Poluição. Marítima; Normas e Controlo nos meios Navais das Forças Armadas.* Instituto Universitário Militar, Lisboa, 2016. Monography.

[2] D.C. Montgomery. *Design and Analysis of Experiments. 5th ed.* Wiley, New York, 2001.

[3] J.B. Rebelo. *Impacto Ambiental da Marinha Portuguesa. Análise e resoluçcão da Gestão de Resíduos no mar.* Escola Naval, Almada, 2019. Master Thesis.

[4] J.B. Rebelo, J.S. Jerónimo, M.F. Teodoro, and S.P. Lampreia. Modeling the waste management in nrp ships. pages 183–188, 2019. Entrepreneurial Ecosystems and Sustainability. Proceedings of Regional Helix 2019.

[5] J.B. Rebelo, J.S. Jerónimo, M.F. Teodoro, and S.P. Lampreia. Preliminary reflexion about waste management planning in nrp ships. 2186, 2019. Computational Methods in Science and Engineering. T. Simos et al. (eds), ICCMSE 2019.

[6] UN. United nations conference on trade and development, 2018. `https://unctad.org/en/PublicationsLibrary/dtl2018d1_en.pdf`. Accessed at 25/01/2019.

Author Index

SPONSORS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


UNIVERSIDADE
LUSÓFONA


CPES
Centro de Pesquisa e Estudos Sociais
Universidade Lusófona


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


BANCO DE PORTUGAL
EUROSISTEMA


sas


museubordalo pinheiro


Garrocha
ESTATE WINES
SINCE 1945