PROGRAM and BOOK of ABSTRACTS

JOCLAD2019

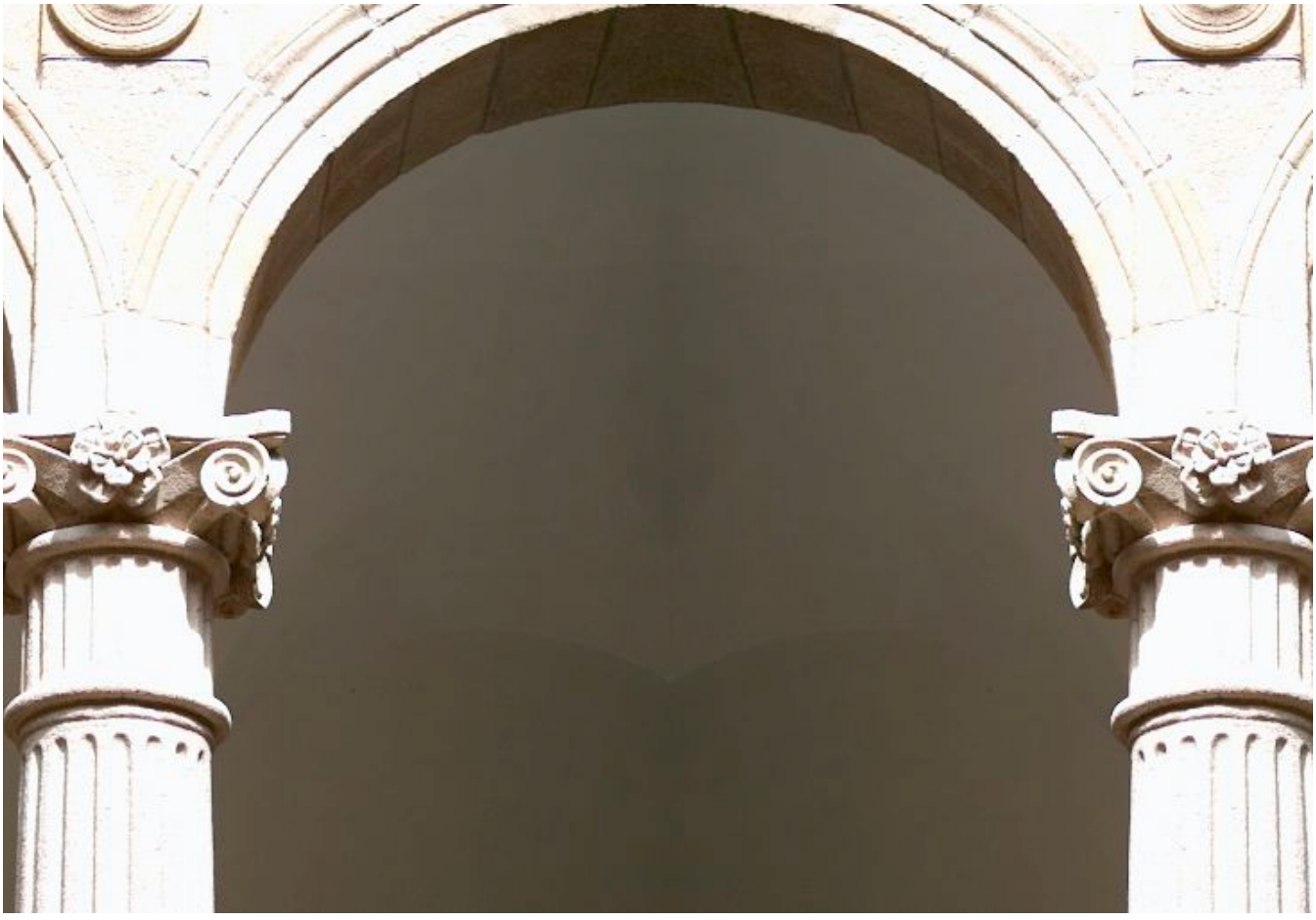
11 - 13 APRIL

VISEU, PORTUGAL

Associação Portuguesa de Classificação e Análise de Dados

XXVI MEETING OF THE PORTUGUESE ASSOCIATION FOR CLASSIFICATION AND DATA ANALYSIS

XXVI JORNADAS DE CLASSIFICAÇÃO E ANÁLISE DE DADOS



Program and Book of Abstracts

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

11–13 April 2019

Viseu, Portugal

www.joclad.ipt.pt/joclad2019/

Sponsors

Associação Portuguesa de Classificação e Análise de Dados
Instituto Politécnico de Viseu
Escola Superior de Tecnologia e Gestão de Viseu
Banco de Portugal
Instituto Nacional de Estatística
PSE – Produtos e Serviços de Estatística

Câmara Municipal de Viseu
Museu Nacional Grão Vasco
Adega Cooperativa de Silgueiros
Chocolateria Delícia
Comissão Vitivinícola Regional do Dão
Confeitaria Amaral
Que Viso Eu? Gastronomia, Arte e Cultura

Program and Book of Abstracts

XXVI Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD2019)

Editors: Conceição Amado, Ana Cristina Matos, José Gonçalves Dias, Conceição Rocha, André Codeço Marques, Carla Manuela Henriques, Nuno Bastos, Isabel Silva

Publisher: CLAD

Printed: Statistics Portugal

ISBN 978-989-98955-6-0

Depósito legal: 454416/19

Preface

Welcome to JOCLAD2019! The JOCLAD2019 – Meeting of the Portuguese Association for Classification and Data Analysis aims to bring together researchers and practitioners interested in Data Science. This is already the twenty-six meeting of the CLAD, the Portuguese Association for Classification and Data Analysis. After many meetings all over Portugal – 2016 in Évora, 2017 in Porto, 2018 in Lisbon – JOCLAD2019 is held in Viseu, at the Escola Superior de Tecnologia e Gestão de Viseu (ESTGV) of the Instituto Politécnico de Viseu (IPV), which co-organizes it. The IPV was created in 1979 and, since then, has played a crucial role in the development of the region of Viseu, namely in the higher education offered by its five integrated schools, in the cooperation with local institutions and companies, and in the promotion and dissemination of scientific research. ESTGV, as an education and research unit of the Instituto Politécnico de Viseu, is a center for the creation, diffusion and transmission of culture, science and technology. Offering courses in the areas of engineering and management and being governed by quality standards, ESTGV ensures appropriate education to the needs of the current labor market. This volume is one of the main outcomes of the JOCLAD2019 and documents the meeting contents.

This twenty-sixth meeting is special as it celebrates the quarter of a century of our association. Indeed, these meetings started before CLAD and somehow triggered its foundation. We have a Thematic Session that celebrates this achievement. This edition also takes a new step in the development of the JOCLAD – Meetings of the Portuguese Association for Classification and Data Analysis. For the first time, this book is edited in English, making its contents available to a wider audience.

The program for this meeting results from the dedicated effort of many people. We thank the invited speakers: M. Salomé Cabral (CEAUL, Departamento de Estatística e Investigação Operacional, FCUL, Portugal), Peter Filzmoser (Vienna University of Technology, Austria), and Agustín Mayo-Iscar (Dpto. Estadística e I.O. & Instituto de Matemáticas, Universidad de Valladolid, Spain). Their talks present a representative cross-section of research in data science. This volume contains the abstracts of the two workshops taught by M. Salomé Cabral on Analysis in Longitudinal Data and Peter Filzmoser on Compositional Data Analysis. A Thematic Session is devoted to the students granted with a 2019 CLAD scholarship, whose members of the evaluation committee were Conceição Amado (Chair), Susana Faria and Catarina Marques. We also thank the organizers of the other Thematic Sessions: Filipa Lima (Banco de Portugal), Carlos Marcelo (INE – Instituto Nacional de Estatística), and Sónia Gouveia and Isabel Silva (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 definitely contribute to reinforce the overall quality of the JOCLAD2019 program. Additionally, we thank all the chairs of these sessions. Our deep gratitude extends to the members of the board of CLAD – Carlos Marcelo, Conceição Rocha, Isabel Silva and Luís Grilo, who volunteered their time to support CLAD activities. 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), Banco de Portugal and PSE.

A successful meeting involves more than the presentation of talks and posters. It is also a meeting of people and exchange of research ideas and collaborations. Thus, a social program – dinner and visit to Viseu – has been arranged in order to promote and facilitate this networking. Our deep thanks extend to the local organizing committee, Ana Matos, André Codeço Marques, Carla Henriques, Conceição Rocha and Nuno Bastos, and to the local sponsors, who made it possible for participants to become more involved with the region of Viseu, creating opportunities of knowing something more about this region through social and tasting moments.

Finally, a big thank you goes to all of you for coming, for your support for JOCLAD2019, and for helping us to make this meeting a success. With your high-quality work, CLAD will continue its tradition of excellence in advancing the data science field for the next 25 years! We wish all of you an unforgettable stay in the city of Viseu. We hope to meet you again for the JOCLAD2020!

Viseu, April 2019

Chair of the Scientific Program
Conceição Amado

Conference Chair
Ana Cristina Matos

President of CLAD
José Gonçalves Dias

Organization

President of the CLAD

José Gonçalves Dias

Chair of the JOCLAD2019

Ana Cristina Matos (Escola Superior de Tecnologia e Gestão de Viseu)

Local Organizing Committee

Ana Cristina Matos (Escola Superior de Tecnologia e Gestão de Viseu)
André Codeço Marques (Escola Superior de Tecnologia e Gestão de Viseu)
Carla Manuela Henriques (Escola Superior de Tecnologia e Gestão de Viseu)
Conceição Rocha (CPES - INESC TEC)
Nuno Bastos (Escola Superior de Tecnologia e Gestão de Viseu)

Chair of the Scientific Program Committee

Conceição Amado (IST-Universidade de Lisboa)

Scientific Program Committee

A. Manuela Gonçalves (Universidade do Minho)
Adelaide Figueiredo (Universidade do Porto)
Adelaide Freitas (Univerdade de Aveiro)
Ana Lorga da Silva (Universidade Lusófona)
Ana Matos (Instituto Politécnico de Viseu)
Ana Sousa Ferreira (Universidade de Lisboa)
Anabela Afonso (Universidade de Évora)
Carla Henriques (Instituto Politécnico de Viseu)
Carlos Ferreira (Universidade de Aveiro)
Carlos Soares (Universidade do Porto)
Catarina Marques (Instituto Universitário de Lisboa)
Conceição Rocha (INESC - TEC and Universidade do Porto)
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 (Instituto Universitário de Lisboa)
Luís Miguel Grilo (Instituto Politécnico de Tomar)
Manuela Neves (Universidade de Lisboa)
Margarida Cardoso (Instituto Universitário de Lisboa)
Maria de Fátima Salgueiro (Instituto Universitário de Lisboa)
Maria Filomena Teodoro (Escola Naval-Marinha Portuguesa)
Paula Brito (Universidade do Porto)
Paula Vicente (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, 11 April

8:30	Registration	Hall of ESTGV Auditorium
9:00	**Mini-course A**	Room SD1
10:15	Coffee Break	
10:30	**Mini-course A** (cont.)	Room SD1
12:00	Lunch Time	
13:00	**Mini-course B**	Room SD1
14:30	Coffee Break	
14:45	**Mini-course B** (cont.)	Room SD1
16:00	**Opening Session of the Meeting**	ESTGV Auditorium
16:30	**Plenary Session I**	ESTGV Auditorium
17:30	**Thematic Session I - CLAD's 25 Years**	ESTGV Auditorium
18:30	**Reception: *Dão de Honra***	Solar do Vinho do Dão

Friday, 12 April

8:30	Registration	Hall of ESTGV Auditorium
9:00	**Parallel Session I**	Room A2 & A3
10:20	Coffee Break	
10:40	**Parallel Session II**	Room A2 & A3
12:00	**Plenary Session II**	ESTGV Auditorium
13:00	Lunch Time	
14:00	**Thematic Session II - Banco de Portugal**	ESTGV Auditorium
15:20	**Thematic Session III - CLAD 2019 Scholarship**	ESTGV Auditorium
16:20	Coffee Break	
16:40	**Thematic Session IV - Statistics Portugal**	ESTGV Auditorium
18:00	**Visit to Historic Centre of Viseu**	
19:30	General Assembly of CLAD	Clube de Viseu
20:30	**Meeting Dinner**	Clube de Viseu

Saturday, 13 April

9:00	Registration	Hall of ESTGV Auditorium
9:30	**Parallel Session III**	Room A2 & A3
10:50	Coffee Break	
11:10	**Thematic Session V - CLAD-SPE**	ESTGV Auditorium
12:30	Lunch Time	
13:30	**Poster Session**	Auditorium atrium
14:10	**Plenary Session III**	ESTGV Auditorium
15:10	**Closing Session of the Meeting**	ESTGV Auditorium
15:30	Coffee Break	

Program

Thursday, 11 April

8:30 Registration - Hall of ESTGV Auditorium

9:00 **Mini-course A** - Room SD1
Longitudinal data analysis
M. Salomé Cabral, p. 5
Chair: Carla Henriques

10:15 **Coffee Break**

10:30 **Mini-course A** (cont.)

12:00 **Lunch Time**

13:00 **Mini-course B** - Room SD1
Compositional data analysis: concepts, software, and examples
Peter Filzmoser, p. 7
Chair: A. Pedro Duarte Silva

14:30 **Coffee Break**

14:45 **Mini-course B** (cont.)

16:00 **Opening Session of the Meeting** - ESTGV Auditorium

16:30 **Plenary Session I** - ESTGV Auditorium
Robust proposals for model based clustering of multivariate data
Agustín Mayo-Iscar, p. 11
Chair: José G. Dias

17:30 **Thematic Session I - CLAD's 25 Years** - ESTGV Auditorium
Helena Bacelar-Nicolau, Fernanda Sousa, José G. Dias, p. 19

18:30 **Reception: *Dão de Honra*** - Solar do Vinho do Dão

Friday, 12 April

8:30 Registration - Hall of ESTGV Auditorium

9:00 **Parallel Session I**

	Room A2 **Clustering and outliers detection methods** Chair: Paulo Infante	Room A3 **Data science modeling** Chair: Conceição Amado
9:00	**From sparse principal components to clustering of variables in high-dimensional data**, Adelaide Freitas, p. 61	**State Space Modeling in Water Quality Monitoring in a River Basin**, A. Manuela Gonçalves, Marco Costa, p. 69
9:20	**Evaluating outlier detection methods: A review of performance measures**, A. Pedro Duarte Silva, p. 63	**Normalization of foot clearance and spatiotemporal gait data using multiple linear regression models**, Flora Ferreira, Carlos Fernandes, Miguel Gago, Nuno Sousa, Wolfram Erlhagen, Estela Bicho, p. 71
9:40	**Looking for atypical groups of distributions in the context of genomic data**,Ana Helena Tavares, Vera Afreixo, Paula Brito, p. 65	**Pediatric arterial hypertension modeling**, M. Filomena Teodoro, Carla Simão, p. 73
10:00	**Internet usage patterns: Segmentation of European users using a multilevel latent class model**, Ana Gomes, José G. Dias, p. 67	**First four order cumulants in Mixed Models**, Patrícia Antunes, Sandra Ferreira, Célia Nunes, Dário Ferreira, João Mexia, p. 75

10:20 **Coffee Break**

10:40 **Parallel Session II**

	Room A2 **Data Science in health and economics** Chair: Fernanda Sousa	Room A3 **Classification and symbolic Data** Chair: Luís Grilo
10:40	**Predictive value in healthcare: a forgotten measure?**, Carina Ferreira, Teresa Abreu, Mário Basto, p. 77	**Multiple-valued symbolic data clustering: a model-based approach**, José G. Dias, p. 85
11:00	**Analysis of administrative data with a binary response variable**, Maria de Fátima Salgueiro, Marcel D.T. Vieira, P.W. F. Smith, p. 79	**Time series clustering using forecast densities based on GAM models**, Maria Almeida Silva, Conceição Amado, Dália Loureiro, p. 87
11:20	**Understanding power at tax investigation - The Portuguese tax inspector's view**, João Marques, Ana Helena Tavares, p. 81	**Clustering Interval Time Series**, Elizabeth Ann Maharaj, Paulo Teles, Paula Brito, p. 89
11:40	**Reduced Social Accounting Matrix for Mozambique**, Eliza Monica A. Magaua, p. 83	**Discriminant factors of website trust**, Ana Andrade, Margarida G. M. S. Cardoso, Vítor V. Lopes, p. 91

12:00 **Plenary Session II** - ESTGV Auditorium
Modelling longitudinal binary data
M. Salomé Cabral, p. 13
Chair: Maria de Fátima Salgueiro

13:00 **Lunch Time**

14:00 **Thematic Session II - Banco de Portugal** - ESTGV Auditorium
Economy and finance
Chair: Filipa Lima

14:00 **Identifying High Growth Enterprises using different criteria**, Ana Filipa Carvalho, Cloé Magalhães, João Meneses, Mário Lourenço, p. 31

14:20 **The Value Relevance of Consolidated Financial Information**, Ana Bárbara Pinto, Diogo Silva, p. 33

14:40 **Loans and debt securities - an analysis of corporate financing**, André Fernandes, José Soares, Pedro Silva, Rafael Figueira, Ricardo Correia, p. 35

15:00 **What are we holding? - households' investments in negotiable financial instruments**, André Fernandes, José Soares, Pedro Silva, Rafael Figueira, Ricardo Correia, p. 37

15:20 **Thematic Session III - CLAD 2019 Scholarship** - ESTGV Auditorium
Chair: Susana Faria

15:20 **Anomaly detection and classification for streaming data**, João Brazuna, p. 23
15:40 **An approach for representing Pareto frontiers on the web**, Marco Marto, p. 25
16:00 **Symbolic clustering and anomaly detection for business analytics**, Ana Teresa Fernandes, p. 27

16:20 **Coffee Break**

16:40 **Thematic Session IV - Statistics Portugal** - ESTGV Auditorium
Challenges in Official Statistics VIII
Chair: Carlos Marcelo

16:40 **Using administrative data to enumerate population**, Sandra Lagarto, Paula Paulino, p. 41
17:00 **Survey on Mobility in the Metropolitan Areas of Porto and Lisboa**, Bárbara Veloso, Rute Cruz Calheiros, p. 43
17:20 **The Well-being Index of Portugal: assessment and outlook**, Sérgio Bacelar, p. 45
17:40 **Speak to Inspire about Statistics**, Carla Farinha, José Pinto Martins, Margarida Rosa, p. 47

18:00 **Visit to Historic Centre of Viseu**

19:30 General Assembly of CLAD - Clube de Viseu

20:30 **Meeting Dinner** - Clube de Viseu

Saturday, 13 April

9:00 Registration - Hall of ESTGV Auditorium

9:30 **Parallel Session III**

	Room A2 **Data science applications** Chair: Margarida Cardoso	Room A3 **Data science in health and educational sciences** Chair: Maria Eduarda Silva
09:30	**Pilgrimage and mobile use**, Ângela Antunes, Carla Henriques, Suzanne Amaro, p. 93	**PLS-SEM in college students' burnout**, Luis M. Grilo, Anuj Mubayi, Katelyn Dinkel, Bechir Amdouni, Joy Ren, Mohini Bhakta, p. 99
09:50	**How social networks influence similarity between examination answers - longitudinal study**, Milton Severo, João Borges, Fernanda Silva-Pereira, p. 95	**Modelling a predator-prey interaction: an in-class exercise**, Inês Bento, Joana Araújo, Joana Pereira, Margarida Marques, Matilde Almodovar, Morgan Ribeiro, Pedro Afonso, Rita Pereira, Tiago Marques, p. 101
10:10	**Prices in the electricity Iberian market - a clustering approach**, Ana Martins, João Lagarto, Margarida Cardoso, p. 97	**Higher Education Students in Viseu Polytechnic - an evolutive study since the Bologna Treaty**, Joana Fialho, Madalena Malva, Paula Sarabando, Paulo Costeira, p. 103
10:30		**Clinical characteristics of patients with chronic obstructive pulmonary disease (COPD): are they different?**, Vera Enes, Ana Helena Tavares, Vera Afreixo, Filipa Machado, Alda Marques, p. 105

10:50 **Coffee Break**

11:10 **Thematic Session V - CLAD - SPE** - ESTGV Auditorium
Analysing time series data: from classical to innovative approaches
Chair: Isabel Silva

12:30 **Lunch Time**

13:30 **Poster Session** - Auditorium atrium

14:10 **Plenary Session III** - ESTGV Auditorium
The log-ratio approach to handle relative information
Peter Filzmoser, p. 15

Chair: Paula Brito

15:10 **Closing Session of the Meeting** - ESTGV Auditorium

15:30 **Coffee Break**

Abstracts

Mini-Courses

11 April, 9:00 - 12:00, Room SD1

Longitudinal data analysis

M. Salomé Cabral[1],
[1] CEAUL, Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal, mscabral@ciencias.ulisboa.pt

---

This course will be about the analysis of longitudinal continuous data and the linear mixed-effects models will be the methodology used. The modelling of the heteroscedasticity and of the correlation present in this kind of data will be also considered. The  packages `nlme` and `lme4` will be used. Examples with real data will illustrate the several topics.

Keywords: Longitudinal continuous data, correlation, random effects, heteroscedasticity, mixed-effects models

---

Longitudinal data are multivariate and are commonly encountered in both experimental and observational studies across all disciplines. In these studies, repeated measurements of response variables are taken over time on each subject in one or more treatment groups, and time itself is, at least in part, a subject of scientific investigation. Studies involving this type of data are called longitudinal studies.
These studies play a fundamental role since they provide valuable information about individual changes and their relationship to a set of factors other than time which make them an important strategy of research in several scientific areas.
In the analysis of longitudinal data there are different features that must be considered: (i) the nature of the outcome of interest, Gaussian and non-Gaussian; (ii) the correlation between repeated measures of each vector response; (iii) the variability among subjects; (iv) the different number of measurements that each subject may have and/or they may have been measured at different time points; (v) the covariates take on time-specific values (i.e., time-varying covariates). All these features pose many challenges in the analysis of longitudinal data and over the last decades several methodologies have been proposed and several books have been published about this subject [3, 4, 2, 1].
In this course the Gaussian outcome will be considered and the mixed-effects models or, more simply, mixed models will be the methodology used. The model building strategy for linear mixed-effects model will be discussed and some examples with real data will illustrate the several steps. The  packages `nlme` and `lme4` will be used.

Acknowledgements This work has been partially funded by FCT-Fundação Nacional para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2019.

References

[1] M.S. Cabral and M.H. Gonçalves. *Análise de Dados Longitudinais.* SPE, Lisboa, 2011.

[2] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis.* John Wiley & Sons, New York, 2004.

[3] J. Pinheiro and B. Bates. *Mixed- Effects Models in S and S-PLUS.* Springer, New York, 2000.

[4] G. Verbeke and G. Bates. *Linear Mixed Models for Longitudinal Data.* Springer, New York, 2000.

11 April, 13:00 - 16:00, Room SD1

Compositional data analysis: concepts, software, and examples

Peter Filzmoser[1]
[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria, P.Filzmoser@tuwien.ac.at

---

Compositional data analysis deals with relative information by making use of log-ratios between the values of the variables. This approach was introduced for constrained data, e.g. for proportional data that sum up to 1. With the log-ratio approach, hoewver, the contraint is not relevant, and this approach can be used for any data set where relative rather than absolute information should be processed. We will present the main concepts of this approach and illustrate these with data examples.

Keywords: compositional data, log-ratio approach, multivariate statistics, software environment R

In his seminal book on compositional data analysis, John Aitchison has introduced the major concepts of the log-ratio approach [1]. The well-known additive and centered log-ratio transformations were treated, together with geometrical insights into the problem. While Aitchison still had contrained data in mind, more recent approaches no longer refer to this restriction, which in fact is not relevant to the log-ratio approach. With the introduction of the isometric log-ratio transformation, the concept of balances and working on coordinates led to a representation of compositional data in the usual Euclidean geometry, with (possibly) interpretable orthonormal coordinates.
In the workshop we will explain these concepts, and show how regression and correlation analysis, principal component analysis, discriminant analysis, clustering, etc., can be carried out with compositional data. Practical examples from geochemistry, demography, and chemometrics will demonstrate the usefulness of this approach [2].

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press), 1986.

[2] P. Filzmoser, K. Hron, and M. Templ. *Applied Compositional Data Analysis: With Worked Examples in R.* Springer, Cham, 2018.

Plenary Sessions

11 April, 16:30 - 17:30, ESTGV Auditorium

Robust proposals for model based clustering of multivariate data

Agustín Mayo-Iscar[1]
[1] Dpto. Estadística I.O., Instituto de Matemáticas Universidad de Valladolid, Spain, agusmayo@med.uva.es

Impartial trimming and constrained approaches have been successfully applied to robustify maximum likelihood procedures when estimating clustering and mixture models during the last 20 years. As usual, trimming methods are useful to diminish the influence of anomalous observations that do not follow the model. However, and for robustness purposes, it is also needed to regularize the estimation due to the presence of singularities in the objective function. Our proposal is to also apply constraints that allow us to derive well-defined estimating procedures and to reduce the prevalence of spurious local likelihood maximizers. These robust procedures, known as TCLUST, initially were developed for normal multivariate distributed components. Now, more flexible procedures are available based on skewed distributions.

The joint application of trimming and constraints also works for identifying regression models when data belong to a mixture of them. In this setting, the corresponding TCLUST estimators for the cluster-weighted model appear highly competitive.
Parsimonious approaches are frequently needed for estimating clusters. Among them are Mixture Factor Analyzers and Celeux and Govaert's collection of models. In a similar fashion, TCLUST proposals are available for estimating clusters robustly. An important issue related with the application of TCLUST methodologies is their input parameters. Users have to provide at least the level of trimming, the strength of the constraints together with the number of clusters/components, which is the classical input parameter in clustering/mixture modelling. There are available exploratory tools and authomatized procedures for assist to the users in choosing these parameters.

TCLUST methodologies are available via the "tclust" package in CRAN and the "FSDA" toolbox in MATLAB.

12 April, 12:00 - 13:00, ESTGV Auditorium

Modelling longitudinal binary data

M. Salomé Cabral[1], **M. Helena Gonçalves**[2],
[1] CEAUL, Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] CEAUL and Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade do Algarve, Portugal

---

Longitudinal binary data studies are a powerful design and they have become increasingly popular in a wide range of applications across all disciplines. Two of the features in these studies are the presence of missing data, since it is difficult to have complete records of all individuals, and the presence of correlation structure in the repeated measures of each response vector. The methodology implemented in the `R` package `bild` will be discussed and two real data sets will be analysed to illustrate how this methodology overcomes those features and how the analysis is carried on.

Keywords: Markov chain, odds-ratio, missing data, marginal models, random effects models

---

The analysis of longitudinal binary data poses two main difficulties. First, the repeated measures of each response vector are likely to be correlated and the autocorrelation structure for the repeated data plays a significant role in the estimation of regression parameters. Second, although most longitudinal studies are designed to collect data on every subject in the sample at each time of follow-up, many studies have missing data, intermittently or dropout, since it is difficult to have complete records of all subjects for a wide variety of reasons.

Generalized linear models have been extended to handle longitudinal binary observations in a number of different ways. Two of them will be considered: marginal models and random effects models. The basic premise of marginal models is to make inference about population means. In contrast, the basic premise of random effects models is that there is a natural heterogeneity across individuals and is used when the goal is to make inferences about individuals. The interpretation of the regressions parameters is not the same in those models. The regression parameters in generalized linear mixed models have subject-specific, rather than population-average, interpretation. The choice between marginal and random effects models for longitudinal data can only be made on subject-matter grounds. When longitudinal binary data are incomplete there are important implications for their analysis and one of the main concerns is to distinguish different reasons of missingness. The nature of missing data mechanism has been classified by [5] as: missing completely at random (MCAR), missing at random (MAR) and non-missing at random (NMAR). Several methods have been proposed for analysing incomplete longitudinal binary responses.

In the `R` package `bild` [3, 4] is implemented the methodology proposed by [2]. In this methodology the inference is based on likelihood and a binary Markov chain model is used to accommodate serial dependence and odds-ratio to measure dependence between successive observations in the same individual. Both marginal and random effects models (intercept model) are considered. The adaptive Gaussian quadrature is used to approximate the log-likelihood using numerical integration when the intercept model is considered. In both cases missing values are allowed in the response, provided they are MAR.
Two real data sets will be analysed to illustrate the use of the `R` package `bild`. The first is a subset of data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in school children from Muscatine (Iowa, USA) available in [4]. The binary response of interest is whether the child is obese (1) or not (0). Since one of the objectives of the study was to determine the effects of sex and age on risk of obesity a marginal model is appropriate. Many data records are incomplete, since not all children have participated in all the surveys, creating, a "genuine" missing data problem. The second data set is from a longitudinal clinical trial of contracepting women, available in [1]. The outcome of interest is a binary response indicating whether a woman experienced amenorrhea (1) or not (0) during the four periods of observation. A random effects model will be used since the goal of the analysis is to determine subject-specific changes in the risk of amenorrhea over the course of the study, and the influence of two dosages of a contraceptive on changes in a woman's risk amenorrhea. A feature of this clinical trial is that there was substantial dropout.

Acknowledgements This work has been partially funded by FCT-Fundação Nacional para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2019.

References

[1] G. Fitzmaurice and J. Laird, N.and Ware. *Applied LongitudinalAnalysis.* John Wiley & Sons, New York, 2004.

[2] M.H. Gonçalves and A. Azzalini. Using Markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach. *Metron*, LXVI:157–181, 2008.

[3] M.H. Gonçalves, M.S. Cabral, and A. Azzalini. The R package bild for the analysis of binary longitudinal data. *Journal of Statistical Software*, 46:1–17, 2012.

[4] M.H. Gonçalves, M.S. Cabral, and A. Azzalini. *bild: A package for BInary Longitudinal Data.* R foundation for statistical computing, version 1.1-5., URL-http://CRAN.R-project.org/package-bild, 2013.

[5] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 1987.

13 April, 14:10 - 15:10, ESTGV Auditorium

The log-ratio approach to handle relative information

Peter Filzmoser[1]
[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria, P.Filzmoser@tuwien.ac.at

---

For many data sets the interest for the analysis is not in the measured or observed values directly, but rather in the relative information. This can be investigated by considering the log-ratios between all pairs of variables, and – to avoid over-parametrization – by constructing an orthonormal basis describing this information. Since (log-)ratios are taken, one could multiply the values of one observation by a positive constant without changing this relative information. This implies that the analysis is invariant with respect to the data scale. The log-ratio methodology is popular in the context of compositional data analysis.

Keywords: compositional data, orthonormal coordinates, multivariate statistics, robust estimation

---

In many applications it is not of interest to directly analyze the reported data values. For example, the number of employees in different economic sectors is not comparable among different countries, since this "absolute information" depends on the total number of employees in the country. One option for making the numbers comparable is to report the values in proportions or percentages, thus to divide by the "total", and possibly multiply by 100. Although the numbers accross the countries are comparable now, there is a problem if the relationships between the sectors are of interest, for instance in terms of correlations. Already Karl Pearson [3] pointed out the problem of *spurious correlations* for proportional data.

Another option is to analyze "relative information" by considering log-ratios between the values of the variables. The resulting log-ratio methodology was proposed by John Aitchison [1] for compositional data, where the relative information is of main interest. The aim was to define a family of log-ratio transformations, resulting in new variables which are aggregated pairwise log-ratios, to move compositional data from their original sample space to an unrestricted real space, where standard statistical methods can be applied for their further analysis.

In the recent literature, isometric log-ratio coordinates are proposed to represent compositions in the usual Euclidean space [2]. Specific choices of those coordinates allow for an interpretation of the parameters in statistical models. We will present the major concenpts and some of those choices for multivariate statistical methods. Special attention is given to robust statistical methods. The log-ratio approach will be illustrated with real data examples from geochemistry, metabolomics, and the digital music industry.

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press), 1986.

[2] P. Filzmoser, K. Hron, and M. Templ. *Applied Compositional Data Analysis: With Worked Examples in R.* Springer, Cham, 2018.

[3] K. Pearson. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, LX:489–502, 1897.

Thematic Session: CLAD's 25 Years

11 April, 17:30 - 18:00, ESTGV Auditorium

CLAD's 25 Years

In 2019 the Associação Portuguesa de CLassificação e Análise de Dados (CLAD), the Portuguese Association for Classification and Data Analysis, completes 25 years. In fact, CLAD was born in 1994 after a historical first Jornadas de Classificação e Análise de Dados (JOCLAD) meeting grouping more than 250 people the previous year at LEAD/FPCE at the University of Lisbon. All of them were curious about the new methods and applications related to what we now call data science. The following year, by the hand of the IFCS President Allan Gordon, CLAD became the 10th IFCS member. The 25th CLAD anniversary will then be celebrated during the next 26th JOCLAD Annual Meeting. A special short session is planned, where the participation of the current, the last, and the first CLAD President is expected.

Some of the most important milestones in CLAD's 25 years are:

- 1993, December: JOCLAD'93 - first Meeting on Classification and Data Analysis (LEAD - FPCE, University of Lisbon)
- 1994, June: CLAD foundation (with SFC, BCS, and CSNA support)
- 1994, December: JOCLAD'94 - first Meeting of the CLAD (LEAD - FPCE, University of Lisbon)
- 1995: CLAD becomes Member of IFCS (the 10th "branch" in the IFCS dendrogram)
- 1995: CLAD and SFC sign a co-operation protocol of associate membership (Paris)
- 1995: CLAD and ABE sign a co-operation protocol of associate membership (Universidade de S. Paulo)
- 1996: CLAD and GfKl sign a co-operation protocol of associate membership (Lisbon, at JOCLAD'96)
- 1997: IPM on Classification at ISI'97 Meeting, on behalf of CLAD, ISI's "Sister Society", by invitation of the ISI Director, August 97
- 1998: CLAD becomes Member of ECAS (European Courses in Advanced Statistics)
- 1999: CLAD organizes ASMDA-99 International Conference on Applied Stochastic Models and Data Analysis. ASMDA-IS foundation (Lisbon), June 14th - 17th
- 2001: CLAD organizes EMPG2001, 32nd European Mathematical Psychology Group Meeting (Lisbon) and previous short meetings, Introductory Course on Mathematical Psychology and Data Analysis, and Workshop on Teaching and Training Mathematical Psychology in an Interdisciplinary and International Context, September 26th-29th

- 2002: CLAD sponsors the I Workshop in "Estatística e Análise de Dados", EstAD 2002 (University of Algarve), April 12th
- 2003: CLAD organizes JISS-2003, the IASC-IFCS Joint International Summer School on Classification and Data Mining in Business, Industry and Applied Research - Methodological and Computational Issues (University of Lisbon), July 23th-30th
- 2007: CLAD and SPE sign a co-operation protocol of associate membership
- 2015: CLAD and APD sign a co-operation protocol of associate membership
- 2018: CLAD sponsors the Symbolic Data Analysis Workshop 2018 (Instituto Politécnico de Viana do Castelo), October 18th-20th

Meanwhile many other short courses, round tables, and seminars have been organized, mainly on the scope of JOCLAD and other CLAD organized/sponsored meetings. CLAD has developed partnerships with several institutions and enterprises during these 25 years, especially with INE (CLAD's general partner since the beginning) and Banco de Portugal (since 2012). Both institutions regularly support and/or participate at JOCLAD and other CLAD initiatives.

Helena Bacelar-Nicolau
Universidade de Lisboa (FPUL e ISAMB/FMUL)
hbacelar@psicologia.ulisboa.pt

Fernanda Sousa
Universidade do Porto (FEUP)
fcsousa@fe.up.pt

José Gonçalves Dias
Instituto Universitário de Lisboa (ISCTE-IUL)
jose.dias@iscte-iul.pt

Thematic Session: CLAD 2019 Scholarship

12 April, 15:20 - 15:40, ESTGV Auditorium

Anomaly detection and classification for streaming data

João Brazuna[1], **Conceição Amado**[1], **Paulo Soares**[2]
[1] CEMAT, Instituto Superior Técnico, joao.brazuna@tecnico.ulisboa.pt
[2] CEAUL, Instituto Superior Técnico

Daily business transactions generate continuously growing data streams. On this article, we are provided data related to service requests from a telecommunications company. Our main objective is to develop an anomaly detection and classification procedure, which can bring several business advantages.

The detection step of our proposed solution consists in statistically testing if the distribution of service requests among the servers which process them is uniform as expected. Then, we apply classification methods to allow the system to automatically classify the detected events.

Keywords: anomaly detection, classification, chi-square goodness-of-fit test, random forest, neural network, k-nearest neighbours

---

There are millions of electronic devices continuously generating data. Every second, a simple call or just some portability request may be recorded on the operator's system. Nowadays, the real challenge is learning from this magnitude of data.
Learning from data streams can bring several business advantages related with the system that is used to process daily transactions. A fast detection allows a company to promptly perceive and correct those anomalies. Monitoring the evolution of some indicators makes it possible to be aware of non-conformities or anomalies on the expected system performance. Unfortunately, defining and detecting anomalies are not easy tasks, raising several questions. Which characteristics can be used to distinguish an anomaly from an expected event? Can we find the cause of a specific anomaly? As the amount of generated data is always increasing, it becomes impossible for a human being to detect every anomaly and finding its source. So, automatic anomaly detection and classification algorithms are essential for recognizing non regular events.
The present work is a collaboration with webDisplay Consulting, a Portuguese IT company, arising from their current interest in taking potential benefits of statistical data science methods to expand the Operational Intelligence process. In particular, this article is inserted in the Enterprise Service Intelligence project (ESI), whose purpose is to make use of some statistical procedures to automatically detect anomalies in an system.
We were provided some portions of a data stream corresponding to internal service requests of a company. There are six servers (the hosts) processing the service requests. To make it efficient, there is a load balancing mechanism implemented in such a way that the distribution of service requests among the hosts should always be approximately uniform.

Our main objective is to develop an anomaly detection and classification algorithm for this data stream. It needs to be dynamic and self-learning so that after classifying a certain time period as anomalous or not, that classification can be used to improve its "knowledge".
If there is a departure from uniformity, there are relatively long time periods in which at least one host is processing a lot more or a lot less service requests than the remaining ones. One option to assess that uniformity assumption is to apply chi-square goodness-of-fit tests.

Figure 1: Observed values of the chi-square test statistic applied to 1 second time intervals.

Our proposed algorithm has two steps. The detection phase consists in applying chi-square goodness-of-fit tests for uniformity to small consecutive time intervals. If there is a sequence (larger than a threshold) of consecutive time intervals in which the uniformity assumption was rejected, we add that time period to the potential anomalies list.
In the classification phase, we provide the results from the detection step to a previously trained classifier to decide whether or not the detected period was really anomalous.
After applying the detection step to two months of data, we have detected 89 potentially anomalous time periods. By analysing the number of processed service requests per host and the corresponding mean service duration, we could find possible and reasonable explanations to 78 of them, which were taken as legitimate anomalies. We used the first 65 potential anomalies to train several classifiers (logistic regression, neural networks, random forests and k-nearest neighbours). After a new potentially anomalous time period was detected, it was given to the classifier to predict its label. After that prediction, it joined the training set and the classifier was retrained.
The best results were provided by both random forests and k-nearest neighbours, with only two misclassified potential anomalies among the 24 included in the test set.

References

[1] J. Gama. *Knowledge Discovery From Data Streams*. Chapman & Hall/CRC Press, 2010.

[2] T. K. Ho. Random decision forests. *Proceedings of the Third International Conference on Document Analysis and Recognition*, Vol. 1:pp. 278–282, August 1995.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 8^{th} edition, 2017.

[4] N. G. Pavlidis, D. K. Tasoulis, N. M. Adams, and D. J. Hand. λ-perceptron: An adaptive classifier for data streams. *Pattern Recognition*, 44:pp. 78–96, 2011.

12 April, 15:40 - 16:00, ESTGV Auditorium

An approach for representing Pareto frontiers on the web

Marco Marto[1], **Vladimir A. Bushenkov**[2]
[1] Forest Research Centre, University of Lisbon, marcovmarto@isa.ulisboa.pt
[2] Research Centre for Mathematics and Applications, University of Évora, bushen@uevora.pt

---

During the first decades of the XXI century, some forest decision support systems and decision tools using Pareto frontiers for trade-off analysis with two or more criteria for forest and natural resources management have been developed. Various approaches for representing non-web-based Pareto frontiers were developed. Recently, some approaches dedicated to its web representation have appeared. This work focuses on a tool developed by the authors for representing Pareto frontiers on the web.

Keywords: decision, multicriteria analysis, Pareto frontiers, forest management, web

---

There are various approaches for the multidimensional representation of the set of non-dominated solutions which have been used to support the forest decision and management. [2] propose a three-dimensional representation of efficient solutions. Another possible representation was developed by [1], where a decision map with a triangle (TRIMAP) to represent and compare solutions is used. The solution for multidimensional visualization using interactive decision maps and Pareto frontiers is included in these options of visualization.

Contrarily to the formers, it is not limited to the analysis of three criteria, since it can work with more than three. With two criteria, it represents a unique Pareto frontier, whereas with three or more criteria it represents a three-dimensional map. The first criterion is represented in the x-axis, the second criterion is represented in the y-axis and the third criterion is represented by the set of polygons, each one with a different colour. If the problem has more than three criteria, the values of fourth and fifth criteria can change by the movement of sliders and, through this way, change the entire decision map.

Concerning approaches for representing Pareto frontiers on the web, we can mention some previous works such as [3] and [4]. Our approach was successfully implemented in a forest web-based decision support system, wSADfLOR, and the tool itself is in constant improvement to become more user friendly, both in terms of use and interpretation. The decision tool was developed in PHP and needs as input a flat file with a formatted linear programming problem to be read and interpreted by web graphical user interfaces (wgui) in order for the user to identify which criteria he wants to maximize or minimize. After the decision maker chooses the criteria to be used in the optimization process, the information is

processed by an encapsulated standalone module in order to converge with a discrepancy lower than 10% for the Edgeworth-Pareto hull. Since the server receives the vertices and constraints resulting from the iterative process of the standalone module, it is ready to build the graphical representation of the interactive decision map and it responds to each request of the client by redrawing the Pareto Frontiers accordingly.

Acknowledgements The authors would like to thank the Portuguese Science Foundation for funding the Ph.D. grant of Marco Marto SFRH/BD/108225/2015.

References

[1] J. C. Climaco and C. H. Antunes. Implementation of a user-friendly software package—a guided tour of trimap. *Mathematical and Computer Modelling*, 12(10-11):1299–1309, 1989.

[2] S. F.Tóth , G. J. Ettl and S. S. Rabotyagov. ECOSEL: an auction mechanism for forest ecosystem services. *Mathematical and Computational Forestry and Natural Resource Sciences*, 2(2), 2010.

[3] R. Efremov , D. R. Insua and A. Lotov. A framework for participatory decision support using pareto frontier visualization, goal identification and arbitration. *European Journal of Operational Research*, 199(2):459–467, 2009.

[4] A. V. Lotov , A. A. Kistanov and A. D. Zaitsev. Visualization-based data mining tool and its web application. *Data Mining and Knowledge Management. Springer, Berlin, Heidelberg*, pages 1–10, 2005.

12 April, 16:00 - 16:20, ESTGV Auditorium

Symbolic clustering and anomaly detection for business analytics

Ana Teresa Fernandes[1], **M. Rosário Oliveira**[1], **Conceição Amado**[1], **Sérgio Pinheiro**[2], **Nuno Dias**[2]
[1] CEMAT, Instituto Superior Técnico, University of Lisbon,
ana.assuncao@tecnico.ulisboa.pt
[2] webDisplay Consulting Lda.

Integration platforms are a passage point for many relevant information, provided by several external systems, regarding companies' business process. Our goal is to find homogeneous groups of users/services and detect when they have an anomalous behaviour, in order to improve the quality of the service provided. For this, we analyse information on log-files, continuously created by an integration platform, whenever a process ends. Since we are interested in the users/services and not in the processes per si, it is necessary to summarize each object (user or service). For example, through intervals (symbolic approach) or descriptive statistics (conventional approach).

Keywords: Cluster analysis, outlier detection, symbolic data analysis, stream data

---

Integration platforms play a crucial role in a company infrastructure, since they connect several systems and catch important information regarding the company business process. More specifically, when a process ends, a log-file is produced with the information about it, such as the user that made the request, the service provided, its duration, etc. In the present study, a huge amount of observations are collected. Typically, around 2.5 Gb of data are stored per day.

The understanding of clients and services patterns has a commercial interest, since it can lead to improvements in the quality of service provided. However, for companies, it is not only important to understand these patterns, but also know when an atypical behaviour is occurring, in order to perceive what triggered an eventual problem and correct it as soon as possible. Being so, it is our interest to detect users/services with an outlier pattern in fixed time periods.

To achieve our goals, we need to surpass some difficulties, such as the massive amount of data arriving continuously and what type of features should we use to characterize our users/services. To address these tasks, we decide to aggregate the process durations for each user/service in ten minute periods by six descriptive measures (the features): 10% quantile; mean; median; 90% quantile; maximum; and standard deviation.The aggregation of the information provided by the log-files simplifies our methods and turns them computationally lighter.

The first objective is to obtain homogeneous groups of objects, so we perform a static and a dynamic analysis. In the first case, we collect all of our ten minute features and the clustering techniques are applied to all the available data. In the second case, we update the values of each user/service in all ten minutes periods and perform the clustering methods daily. Since our interest relies on the clients/services per si, it is necessary to aggregate our features. For this, we use two approaches: conventional and symbolic. In the first one, we summarize each user/service process durations by a descriptive measure, and in the second approach, by an interval. The clustering techniques applied are the PAM (partition around medoids, vide [3]) and Sclust (symbolic clustering, vide [1]), for the conventional and symbolic approaches, respectively.
The second objective is to detect users/services with an atypical behaviour. The first step was motivated by the results of dynamic clustering, that is, the same idea as the clustering dynamic analysis is applied, but instead of performing a clustering algorithm at the end of every day, we apply an outlier detection algorithm based on the robust Mahalanobis distance, for both symbolic and conventional approaches. However, for companies it is important to detect anomalies as early as possible, so, therefore, we propose a ten-to-ten minute methodology that returns potential anomalies every ten minutes, based in robust principal component analysis.
To conclude, we believe that, given the complexity of the problem addressed and the potential of the obtained results, more work has to be done. This work allows the company to adapt strategies and resources, tailored to each cluster. Moreover, the line of work followed has the merit of allowing the identification of atypical user/services in a certain period, and serves as a monitoring tool to potentially detect this kind of anomalous behaviour in real time, alerting the process manager almost instantly and allowing him to act quickly, according to the detected anomaly. Furthermore, it is interesting to understand how methods for interval data deal with real data and whether or not results reveal similar patterns to the ones obtained with conventional approaches.

References

[1] F. A. T. De Carvalho, Y. Lechevallier, and R. Verde. *Clustering Methods in Symbolic Data Analysis*, chapter 11, pages 181–203. Wiley-Blackwell, 2008.

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

[3] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons, Inc., New York, 1990.

Thematic Session: Banco de Portugal

12 April, 14:00 - 14:20, ESTGV Auditorium

Identifying High-Growth Enterprises using different criteria

Ana Filipa Carvalho[1], Cloé Magalhães[2], João Meneses[3], Mário Lourenço[4].
[1] Banco de Portugal, afcarvalho@bportugal.pt
[2] Banco de Portugal, clmagalhaes@bportugal.pt
[3] Banco de Portugal, jfmeneses@bportugal.pt
[4] Banco de Portugal, mfllourenco@bportugal.pt

High-growth enterprises can be defined according to different underlying variables, growth criteria and size thresholds, with implications regarding the benchmarking of the most dynamic companies in the economy. Some sectors of activity can be over or under-represented within the population of high-growth enterprises, a situation for which analysts should be aware when conducting this kind of analysis.

Keywords: Enterprises, Growth, Dynamics

Different measures can be considered in order to determine which companies have registered growth within a certain time period and how high growth is. According to EUROSTAT-OECD (2007) [3], high-growth enterprises (HGEs) comprise all enterprises with average annualised growth greater than 20% per annum, over a three year period. Growth can be measured using the number of employees and/or turnover. The same publication suggests that a meaningful size threshold should be set to avoid distortions originated by the growth of small enterprises, while recommending that the size threshold should be low enough to avoid excluding too many enterprises.
More recently, Commission Implementing Regulation (EU) No 439/2014 [1] set the compulsory collection of data regarding HGEs with at least 10 employees at the beginning of the growth period and having average annualized growth in number of employees greater than 10% per annum, over a three year period.
Given the multiple definitions available, it is relevant to be aware of the impact of analysing HGEs using one criterion or another, not only in what concerns the variable under evaluation (turnover or the number of employees) but also regarding different thresholds (size threshold and growth threshold).
Using Banco de Portugal's Central Balance Sheet Database several criteria were implemented (including the methodology stated in BANCO DE PORTUGAL (2019) [2]) leading to the identification of different sets of HGEs for the period between 2013 and 2017. The analysis of the differences between these sets of enterprises led to the following conclusions:

- The number of identified HGEs in 2017, for instance, ranges, according to the implementation of different criteria, from around 2 thousand up to almost 48 thousand.

- A size threshold of 10 employees leads to the exclusion of more than 90% of the NFCs from the set of potential HGEs (40% of NFCs' turnover and number of employees).
- If the growth variable and its threshold are held constant, a size threshold of 10 employees (instead of one employee) implies a drop in the number of HGEs of at least 80%.
- The change of the growth threshold (20% or 10%) is the modification in the criteria which has the strongest impact on the weight of HGEs within total NFCs regarding both turnover and number of employees.
- Considering the number of enterprises, the relevance of microenterprises reduces to less than 5% when the size threshold of 10 employees is applied, while that of small and medium-sized enterprises increases to more than 90%. The share of small and medium sized enterprises, measured in terms of turnover and number of employees, increases when growth is measured using turnover, while large enterprises gain relevance when growth is measured considering the number of employees.
- The share of the manufacturing sector increases when the size threshold of 10 employees is used, while the relevance of the trade sector decreases.
- When the size threshold of 10 employees is applied and growth is evaluated using the number of employees, the share of enterprises with head office in the Lisbon Metropolitan Area rises, in terms of turnover and number of employees.
- The share of HGEs belonging to the export sector is strongly influenced by the number of identified HGEs, which varies significantly when using different criteria. For instance, the implementation of a size threshold of 10 employees leads to a decrease of the number of HGEs in the export sector; however, the weight of the export sector within HGEs increases as the number of HGEs is smaller (which may be linked to the exclusion of the majority of the microenterprises from the set of HGEs when such threshold is used).

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] Commission Implementing Regulation (EU) No 439/2014 of 29 April 2014.

[2] Banco de Portugal. *Análise do dinamismo empresarial em Portugal. (Portuguese version only)*. 2019.

[3] EUROSTAT – OECD. *Manual on Business Demography Statistics, Methodologies and Working Papers*. 2007.

12 April, 14:20 - 14:40, ESTGV Auditorium

The value relevance of consolidated financial information

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

The value relevance measures the ability of financial information to explain investors' decisions, which are reflected in the groups' market value. The study examines the value relevance of IFRS financial information for 8 European countries from 2012 until 2016, throughout sectors and also within group size categories. The approach applied follows the idea that the joint explanatory power of book value and net income gauges the extent to which financial information is relevant to investors. The results point out that the relevance of financial information has been following a negative trend and it appears to be higher in Belgium, whereas Greece displays the lowest value. Also, it is higher for groups from the construction and the energy sectors and tends to increase with size. Robustness tests support the results obtained.

Keywords: Consolidated financial information, IFRS, Value relevance

Financial information is used by investors when making economic decisions if it is relevant. Then market values should be explained to some extent by financial information. The value relevance measures the ability of financial information to explain groups' market value [1]. This study examines the value relevance of IFRS consolidated financial information for 8 European countries: Austria, Belgium, France, Germany, Greece, Italy, Portugal and Spain. The analysis is developed per country, throughout sectors, within group size categories and also for each year from 2012 to 2016. The purpose is to capture the singularities of each of these dimensions. Nowadays groups face innovative and always changing environments which demand adaptable and flexible accounting standards, so they can be applied to different contexts and to particular business sizes and sectors [2]. International standards must also fit countries with different economic and social frameworks.
This analysis uses consolidated annual data available in the European Records of IFRS Consolidated Accounts (ERICA) working group which is part of the European Committee of Central Balance Sheet Data Offices (ECCBSO). A fixed sample of 632 listed groups is considered (3160 observations). Groups that belong to the perimeter of consolidation of other groups included in the sample are excluded. To study the value relevance of financial information this research follows an approach known as the "Price Regression Model". The model assesses the extent to which book value and net income explain groups' market value. The model is estimated through the following equation:

$$MV_{it} = \beta_0 + \beta_1.BV_{it} + \beta_2.NI_{it} + \varepsilon_{it} \quad (1)$$

Where for group i and year t,

MV_{it}: market value;
BV_{it}: book value;
NI_{it}: net income;
ε_{it}: residuals;
$\beta_0, \beta_1, \beta_2$: regression coefficients to be estimated.

The value relevance is measured by the adjusted R-squared of the regression. Equation (1) is applied for each country, year, sector and size, allowing to draw comparisons within dimensions (e.g countries).
Robustness checks are applied. Firstly, for each dimension, the model is augmented with dummy variables that incorporate information regarding all other dimensions to control for heterogeneity at the dimension level. For instance, there may be countries in which groups are widely scattered across sectors and sizes, or other factors, such as investor confidence and countries risk that are more preponderant in specific years. Secondly, the model is also re-estimated but instead of considering the groups' market value at the end of year t, groups' next year $(t+1)$ average market value is applied because it is when annual reports are made available by groups.
The results point out that the relevance of consolidated financial information has been following a negative trend as it has been decreasing since 2013 (R-Squared of 0.84), being statistically lower in 2016 (0.77), which is close to the level of 2012 (0.78). This is consistent with market values increasing at a pace faster that book values. Relevance appears to be higher in Belgium (0.97), whereas Greece displays the lowest value (0.49). Construction (0.91) and energy (0.88) are the sectors where information is more relevant. These sectors include relatively more large groups (0.73) and the value relevance of financial information appears to be an increasing function of group size. Groups from the industry (0.77) and services (0.84) sectors are more heterogeneous and relatively smaller (0.24). Robustness tests support the results obtained. Still, one may bear in mind that factors affecting investors' decisions go beyond financial information, such as investor confidence, countries specific risk or stock market dynamism.

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] J. Ohlson. Earnings, book values, and dividends in equity valuation. *Contemporary accounting research*, 11(2):661–687, 1995.

[2] I. Rubbrecht. Flexibility in classification options within the statement of cash flow. *ERICA series*, 06:1–21, 2017.

12 April, 14:40 - 15:00, ESTGV Auditorium

Loans and debt securities – an analysis of corporate financing

André Fernandes[1], **José Soares**[2], **Pedro Silva**[3], **Rafael Figueira**[4], **Ricardo Correia**[5]
[1] Banco de Portugal, agfernandes@bportugal.pt
[2] Banco de Portugal, jmsoares@bportugal.pt
[3] Banco de Portugal, pmssilva@bportugal.pt
[4] Banco de Portugal, rfigueira@bportugal.pt
[5] Banco de Portugal, rncorreia@bportugal.pt

For the 2008-2017 period, euro area non-financial corporations presented a more pronounced increase in financing through debt securities rather than loans. The indebtedness of the Portuguese non-financial corporations has followed the same trend and is characterized by: (1) a highly concentrated debt securities market in terms of the number of issuers; (2) a relatively higher proportion of short-term debt securities, when compared to euro area.

Keywords: Commercial Paper, Debt Securities, Loans, Non-Financial Corporations

For the last 10 years, euro area non-financial corporations (NFCs) changed their debt structure. As presented in Figure 1, Portuguese and other European peripheral countries (OEPC - Spain, Italy and Greece) NFCs presented a reduction in the amount of loans, contrasting to the euro area pattern. Regarding debt securities, OEPC, and to a lesser degree Portugal, followed the euro area tendency, increasing their issuing amount.

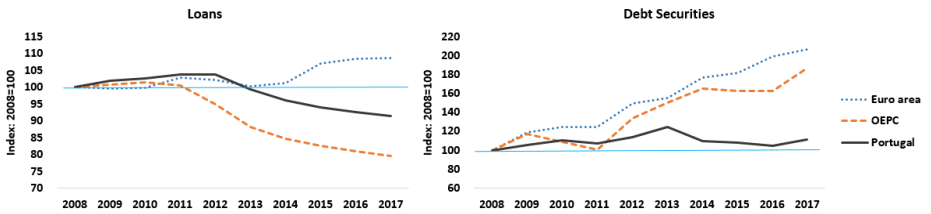


Figure 1: Evolution of liabilities of NFC in loans and debt securities - consolidated stocks. Source: Banco de Portugal and Eurostat.

Despite the fact that loans continued to be the preferred choice for euro area NFCs, the increase of funding through debt securities should be highlighted. In Figure 2, it is observable that the share of euro area NFCs financing via debt securities increased from 9.0% of total debt in 2008 to 15.9% in 2017. This is also observable in the case of Portuguese

NFCs, where debt securities weighed 18.9% in 2017, an increase of almost 3 p.p from 2008. Moreover, even though the total short-term financing (i.e. loans and debt securities) of Portuguese NFCs is in line with euro area throughout the period under analysis (close to 20%), it must be pointed out the relevance of short-term debt securities: 6.3% in Portugal, contrasting to 0.8% in the case of euro area NFCs in 2017. This fact may be explained by the benefits for financing through commercial paper, such as the exemption from stamp duty.

		2008			2013			2017		
	Region	Short-Term	Long-Term	Total	Short-Term	Long-Term	Total	Short-Term	Long-Term	Total
Loans	Euro Area	28.3%	62.6%	91.0%	23.4%	63.3%	86.7%	21.8%	62.3%	84.1%
	OEPC	27.2%	68.8%	96.0%	21.4%	72.0%	93.4%	19.0%	72.2%	91.2%
	Portugal	21.5%	62.5%	84.0%	14.5%	66.2%	80.7%	15.8%	65.3%	81.1%
Securities	Euro Area	1.5%	7.5%	9.0%	0.9%	12.4%	13.3%	0.8%	15.1%	15.9%
	OEPC	0.3%	3.6%	4.0%	0.3%	6.3%	6.6%	0.3%	8.6%	8.8%
	Portugal	10.5%	5.6%	16.0%	8.1%	11.1%	19.3%	6.3%	12.6%	18.9%

Figure 2: NFCs debt structure across euro area. Source: Banco de Portugal and Eurostat.

Using microdata on securities holdings and issues available at Banco de Portugal, it is possible to have a deeper understanding about the Portuguese NFCs debt securities funding. Two characteristics should be noticed: (1) the commercial paper could be viewed as a close substitute of short-term loans; (2) the market players' concentration.
Regarding the former, one could argue that, in Portugal, commercial paper is similar to a loan, given its limited negotiation in the market. In fact, 70.5% of this instrument's outstanding amount is held by Portuguese banks, showed in Figure 3, and it is mostly held by the same institution until final redemption.
For the latter, the number of NFCs which issue debt securities is very limited when compared to the total number of Portuguese NFCs (around 700 in a universe of over 400,000). In addition, taking into consideration the Lorenz curve presented in Figure 3, it can be concluded that the Portuguese NFCs debt securities issues have a high degree of concentration, since 3% of NFCs issuers of debt securities are responsible for 72% of the total amount issued in 2018. Comparing to 2008, there was an increase of concentration, given that for the same percentage of NFCs, the amount outstanding was 59%.

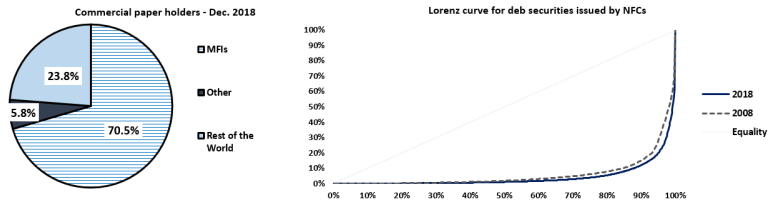


Figure 3: Commercial paper investors and Lorenz curve for Portuguese NFCs issuers. Source: Banco de Portugal.

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.

12 April, 15:00 - 15:20, ESTGV Auditorium

What are we holding? – households' investments in negotiable financial instruments

André Fernandes[1], **José Soares**[2], **Pedro Silva**[3], **Rafael Figueira**[4], **Ricardo Correia**[5]
[1] Banco de Portugal, agfernandes@bportugal.pt
[2] Banco de Portugal, jmsoares@bportugal.pt
[3] Banco de Portugal, pmssilva@bportugal.pt
[4] Banco de Portugal, rfigueira@bportugal.pt
[5] Banco de Portugal, rncorreia@bportugal.pt

Financial negotiable instruments, namely debt securities (F3), listed shares (F511) and investment funds shares (F52), represent 9% of the total financial assets held by Portuguese households in 2018Q3. Investment funds shares are the preferred investment, amounting to half of the households' portfolio. From 2008 to 2017, the Portuguese households' portfolio structure followed euro area pattern, with an increase in investment funds shares and a decrease in debt securities exposure.

Keywords: Households, Investment Portfolio, Portugal, Securities

In 2018Q3, Portuguese households' investment in financial negotiable instruments amounted to 37.0 billion euros, which represented 9% of the total financial assets held by this sector. Currency and deposits was the main financial asset with 45% of the total.

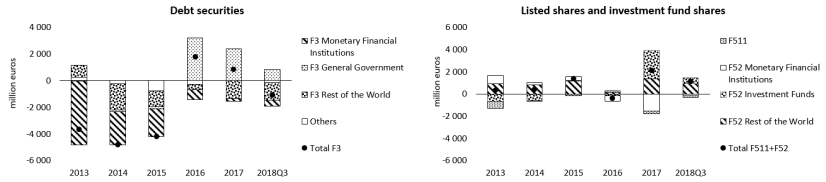


Figure 1: Portuguese households' transactions by financial negotiable instrument - yearly transactions. Source: Banco de Portugal.

Considering Figure 1, one can observe that Portuguese households have made relevant changes in their investment strategy. Between 2013 and 2015, this sector had negative transactions in debt securities, mainly those issued by banks. This disinvestment was partially offset in the following years, through the investment in public debt. It is possible to observe, for the entire period, a positive investment in non-resident investment funds shares, and a positive and significant investment in resident investment funds, in 2013 and from 2017 onward.

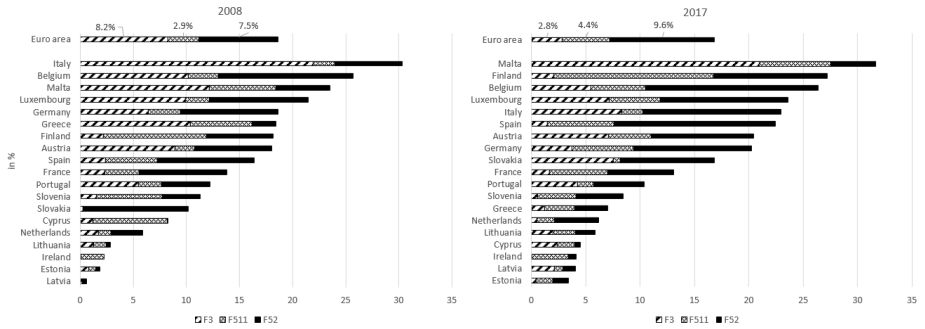


Figure 2: Euro area comparison of negotiable financial instruments load in total financial assets (excluding other accounts receivable) held by households - end of year stocks. Source: Banco Portugal and Eurostat.

When compared with other euro area countries Portuguese households have a lower preference for investing in negotiable instruments, as shown in Figure 2 (10.4% of total financial assets in Portugal compared with 16.9% in the euro area in 2017).
The exposure of Portuguese households to financial negotiable instruments decreased from 12.3% in 2008, to 10.4% in 2017, a pattern presented also by the euro area where this reduction was from 18.7% to 16.9%. There is a relevant reduction in the importance of debt securities (from 5.5% to 4.2%) for Portugal. In the euro area, it is also noticeable a reduction in debt securities importance (from 8.2% to 2.8%), partially offset by an increase in the load of investment fund shares (from 7.5% to 9.6%). This area also presented an increase in listed shares (from 2.9% to 4.4%), which was not followed by Portuguese households. Despite the aforementioned changes, the ranking of Portugal comparing with other countries remains.
Using survey data, one can observe that the preferences for investing in these instruments are dependent on demographic and economic conditions of individuals. According to (Costa, 2016)[2], younger families and families with higher income and wealth show more preference for investing in negotiable instruments. This economic relation is also true for the euro area, despite the older age of the families (ECB, 2016)[1] investing in these instruments when compared to Portuguese case.

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] European Central Bank. The household finance and consumption survey: results from the second wave. *Statistics Paper Series*, 18, 2016.

[2] S. Costa. Financial situation of the households in portugal: an analysis based on the hfcs 2013. *Banco de Portugal Economic Studies*, 2 n4:15–55, 2016.

Thematic Session: Statistics Portugal

12 April, 16:40 - 17:00, ESTGV Auditorium

Using administrative data to enumerate population

Sandra Lagarto[1], **Paula Paulino**[2]
[1] Statistics Portugal, sandra.lagarto@ine.pt
[2] Statistics Portugal, paula.paulino@ine.pt

Since 2014 Statistics Portugal (SP) has been working in a Statistical Population Dataset (SPD) with the country's resident population and a set of demographic and socio-economic variables. The administrative data sources that contributed to build the SPD are presented, as well as the methodological approach and the main results which are encouraging and pave the way to a paradigm change from a traditional to a register based census model.

Keywords: Administrative data, Register-based Census, Statistical Population Dataset

Statistics Portugal has been studying the reliability of using administrative data to enumerate and characterize resident population in Portugal. The use of administrative data sources and register-based census is a general trend among UNECE state member's [5, 3] with obvious advantages: costs reduction, less burdensome to the respondent and increased frequency of outputs. In fact, considering this last topic, EUROSTAT is preparing legislation for annual releases of population statistics after 2024 [1]. For all these reason Statistics Portugal (SP) has developed a framework to the creation of a Portuguese Statistical Population Dataset (SPD), built from administrative data integration and signs of life rules [2, 4].

The reference year for the first SPD exercise was 2011, in order to use the 2011 Census as a benchmark. After that, 2015 and 2016 editions were released and 2017 edition is being prepared. Moreover, the SDP does not provide official statistics: it's a research project and, at this point, all results are considered experimental statistics.

The three main methodological steps to build the SPD are: to link together administrative data and matching records which can potentially be included in the SPD; then, apply the 'Signs of life' rules (a person is considered to be resident in the country if he/she is registered in the Civil Register (CR) or in the Immigration Register (IR) and is also 'active' in at least one more register: studies, works, has been attending healthcare system, pays taxes, etc.); finally, add the relevant socio-economic administrative variables associated to the population in the SPD. In addition to the CR and the IR, eight administrative datasets were used to build the SPD: Social Protection for public servants, State Pension/ Work Fund Register, Education Register, Private Employment Register, Unemployment Register, Social Security Register, Income Taxes Register and the Hospitals attending Register.

For 2016, the SPD estimated 10,3 million residents in Portugal, with a deviation of 0.5 per cent (underestimated 50 thousand persons) from the official Population Estimates for the same year, with population's age structure and sex distribution also consistent.

The administrative variables also provide information on more than 15 census topics, specifically: sex, age, place of usual residence, place of residence one year before, marital status, citizenship, country/place of birth, labour status, occupation, branch of economic activity, status in employment, place of work, number of hours worked, number of employees in the enterprise, educational attainment and school attendance.

The construction of a SPD has made it possible, for the first time in Portugal, to conduct a qualitative and quantitative assessment of the potential of using administrative data for census purposes. The results obtained were encouraging, but not satisfactory enough to undertake a register-based census in 2021. This is primarily because of partial coverage or inexistence/not suitable administrative data, for census key domain like housing, household and family characteristics. The on-going studies on the use of registers should be relevant to add improvements to this project.

References

[1] European Commission (Eurostat). Working group on population and housing censuses – Strategy for the post-2021 census. Technical report, Eurostat, 2017.

[2] Instituto Nacional de Estatística. Estudo de viabilidade da utilização de dados de fontes administrativas no novo modelo censitário para 2021. Relatório QUAR, Gabinete dos Censos 2021 (Documento interno), 2014.

[3] Office for National Statistics (ONS). Beyond 2011 producing population estimates using administrative data: In practice (M7). Technical report, Office for National Statistics, 2013.

[4] UNECE Task Force on Register-based and Combined Censuses. *Portugal Case Study, Guidelines on the use of registers and administrative data for population and housing censuses.* United Nations Publications, 2018.

[5] United Nations Economic Commission for Europe (UNECE). *Conference of European Statisticians, Recommendations for the 2020 Censuses of Population and Housing.* United Nations Publications, 2015.

12 April, 17:00 - 17:20, ESTGV Auditorium

Survey on Mobility in the Metropolitan Areas of Porto and Lisboa

Bárbara Veloso[1], **Rute Cruz Calheiros**[2]
[1] Statistics Portugal, barbara.veloso@ine.pt
[2] Statistics Portugal, rute.cruz@ine.pt

The Survey on Mobility in the Metropolitan Areas of Porto and Lisboa (IMob) was conducted by Statistics Portugal in 2017. It aimed to respond, not only to the national information needs in terms of transport and mobility statistics, but also to the European Statistical System, given the growing importance of this issue in planning and environmental sustainability policies. The main objective was to characterize the trips made by the resident population, as well as to identify its profile, the opinion of the users of individual or collective means of transport and the motivations that led to the means of transport chosen. The results obtained were important to support the decision concerning the transport systems, namely in what concerns the intermodal network definition and price systems, besides monitoring the transition to collective transports and soft transport modes, as well as to produce EU harmonized mobility indicators.

Keywords: mobility, trip-makers, transport, metropolitan area

---

The Survey on Mobility in the Metropolitan Areas of Porto and Lisboa (IMob) focused on resident population in municipalities of both metropolitan areas (nearly 44% of the total Portuguese population), aged between 6 and 84 years.
The sampling base was the total dwellings of usual residence of the National Dwellings Register (composed by the family dwellings). A stratified and multiphase random sample was adopted, based on a previous study of homogeneous areas of accessibility to transport (denominated zones). A *cluster mobility analysis* was carried out to define groups of parishes with similar mobility characteristics to be considered in the IMob survey sampling design, using multivariate data grouping techniques, a hierarchical method (Ward aggregation), complemented by an expert sensitivity analysis in order to guarantee spatial contiguity. In the end, a total of 87 metropolitan homogenous mobility areas below the municipality level were identified: 49 for the Metropolitan Area of Lisboa and 38 for the Metropolitan Area of Porto.
In a first stage, data collection was conducted by self response Web questionnaire (Computer Assisted Web Interview – CAWI) and, in a second stage, a sub-sample was selected between non-responses in the first stage and face-to-face interviews (Computer Assisted Personal Interview - CAPI) were conducted. In the selected dwellings, all the individuals were observed within the age group under the scope.

Data collection ran between October and December 2017 and the reference period was one week day (from Monday to Sunday - previously chosen for each dwelling).
Response rate obtained was 17.1% by CAWI (exceeding the initially expected 5% gross rate for the first stage) and 58.8% by CAPI.
The weighting process took into account demographic information (sex and age groups, total population by zone and job status) and, also, trimming studies were performed.
This survey followed the Guidelines on Passenger Mobility Statistics [1], allowing the harmonisation of results according to the European Statistical System, and intended to answer the following questions: a) How do we move? b) How long do trips take? c) How far do we go? d) What costs do we have?
The results showed that nearly 80% of resident population in both metropolitan areas have done, at least, one trip on the reference day (share of trip-makers). Most of the trips had both origin and destination within the respective metropolitan area.
Considering all means of transport, passenger cars were the most important way of locomotion in the total number of journeys and soft modes (pedestrian or bicycle) appeared as the second most important.
Regarding time and distance, trips made by the residents of the Metropolitan Area of Porto lasted, on average, 22.0 minutes and 10.6 km and, for the ones in the Metropolitan Area of Lisboa, 24.5 minutes and 11.0 km.
In terms of costs, despite fuel was identified as an usual expenditure by a large number of individuals, the majority of the population living in the metropolitan areas also revealed regular expenses with public transport.
In summary, Statistics Portugal conducted the Survey on Mobility in the Metropolitan Areas of Porto and Lisboa, producing results that responded to the information needs of the European Statistical System, Metropolitan Areas and many other national users, with different degrees of specialisation in the field.
The results of IMob enabled an up-to-date knowledge of mobility in metropolitan areas, understanding the main purposes and needs of mobile population, and a profile identification of individual/public transport user and their choices' reasons.
These results also proved to be extremely important for studies developed by the Metropolitan Areas of Porto and Lisboa, specifically in the definition of a new tariff system and the determination of financial impact of new solutions, and also at a level of public transport network, allowing an evaluation of the effectiveness of the existing network and as inputs for a model on demand of public/individual transport, with estimation of the origin/destination matrices by day-type and mode of transportation. Also, Statistics Portugal intends to be aligned with the most recent best practices on this matter, working towards a set of harmonised EU indicators on passenger's mobility.

References

[1] Eurostat. EU Transport Statistics, Eurostat guidelines on Passenger Mobility Statistics. Technical report, Eurostat, July 2016.

12 April, 17:20 - 17:40, ESTGV Auditorium

The Well-being Index of Portugal: assessment and outlook

Sérgio Bacelar

Statistics Portugal, sergio.bacelar@ine.pt

We discuss some of the main priorities for the evolution of the statistical project Well-being Index of Portugal: a new methodology for imputing missing data, a review of possible redundancies of indicators in each domain, studying adequate processes of normalisation of indicators, controlling eventual compensatory effects of indicators in the computation of the indices, by alternative aggregation methods. Moreover, we address the incorporation in the final study of several types of inequalities.

Keywords: Well-being, Composite index

Since 2013, Statistics Portugal has made available, on an annual basis, the Well-being Index (WBI), which is a composite index of a set of indicators derived from information of administrative nature and statistical operations developed in the context of the National Statistical System, and the European Statistical System, among others.
WBI is based on a conceptual framework structured in two analytical perspectives (sub-indices) (*Material living conditions* and *Quality of life*), ten domains (pillars) and 79 baseline indicators.
The selection of indicators was based on criteria such as the preference for outcome indicators, focused on inequality assessment, regular data availability, and international comparability. Indicators of subjective evaluation of well-being are necessarily included.
In this presentation, we will discuss some of the main priorities for the evolution of this project.
At dissemination time (t), some of the indicators of the WBI for the last year ($t-1$) are provisional or even unavailable. This fact explains why it is necessary to project the value of the unavailable indicators for that year to enable the computation of preliminary WBI results. This projection functions similarly to a missing data imputation. With this purpose, we are studying a new method for this projection using an exponential smoothing forecast based on the Holt method.
There is also a need for a more balanced distribution of the number of indicators by domain. The eventual redundancy of indicators by domain has an impact in the implicit weights of each indicator in the computation of the mean domain index [3].
Inspired by the methodology of the *Canadian Index of Wellbeing*, each indicator of the WBI x^j, with $j \in [1..79]$ is transformed on an index $I_t^j = \frac{x_t^j}{x_{t_0}^j}$, where t_0 is the base year

2004. Priority should be given to a previous suitable normalisation method with directional adjustments to obtain indicators with the same range of variation. We have opted to scale data between $[0, 1]$ using a min-max normalisation. To achieve this normalisation we tested for each indicator, goalposts (min and max) using a group of EU reference countries which provide a good frame of reference for Portugal [1]. Besides the well-known influence of dimensions' weights, normalisation, functions as well, as an implicit weighting that can affect the overall results. This fact implies that the choice of the normalisation function should be made as transparent as possible [2].
Aggregation of the indicators is done by using unweighted arithmetic means. Besides the discussion about the weighting procedure, it is well known that the use of arithmetic means, despite being more transparent for the users, has a compensatory effect: poor performing indicators are compensated by good performing ones. We will discuss if compensability between indicators should be allowed.
Finally, and more critical, is the inclusion of a satellite domain of inequalities or even to include a measure of inequality over time in the process of construction of the index. For example, an average life expectancy, which is a well-being indicator, may be very different no only by gender but also by region, education level or any other indicator of social asymmetry. These social and economic asymmetries can be *vertical*, based on the individual values of each population or sample unit, *horizontal*, gaps in average performance between specific population groups and *deprivations*, i.e. the share of people falling below a basic threshold of attainment [4].

References

[1] Auke Rijpma; Michail Moatsos; Martijn Badir; Hans Stegeman. Netherlands beyond a GDP: A Wellbeing Index. Technical Report 78934, Munich Personal RePEc Archive, Munich, 2017.

[2] Ludovico Carrino. The role of normalisation in building composite indicators. rationale and consequences of different strategies, applied to social inclusion. In Filomena Maggino, editor, *Complexity in Society: From Indicators Construction to their Synthesis*, number 70 in Social Indicators Research Series, chapter 11, pages 251–289. Springer International Publishing, 2017.

[3] OECD. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Sep 2008.

[4] OECD. *How's Life? 2017: Measuring Well-Being*. OECD Publishing, Paris, Jan 2018.

12 April, 17:40 - 18:00, ESTGV Auditorium

Speak to Inspire about Statistics

Carla Farinha[1], **José Pinto Martins**[2], **Margarida Rosa**[3]
[1] Statistics Portugal, carla.farinha@ine.pt
[2] Statistics Portugal, pinto.martins@ine.pt
[3] Statistics Portugal, margarida.rosa@ine.pt

Statistics Portugal website was created in 2007 and at the time the aim was to be the main dissemination channel of official statistics. Getting customers closer to Statistics Portugal (INE) in a simple click away! Over the next ten years the paradigm changed: customers yearn for more.

Keywords: channels, customers, dissemination, needs, website

It was revolutionary. Nevertheless, the relationship between INE and the society changed significantly as well as the access to internet (+ 13.5 pp in 4 years (see Table 1)).

Table 1: Proportion of households with at least one person aged between 16 and 74 years old and with broadband connection to Internet at home (%); Annual

2018	2017	2016	2015	2014
76.9	76.4	73.0	68.5	63.4

Consulted in `http://www.ine.pt` and accessed in 22th February 2019

Moreover, the external demand for statistics grew significantly over the last decade and OECD developed new instruments to disseminate statistics as part of a more general renovation of its dissemination and communication policy [4] as did many other organizations. Concerning data release, not only all European statistics are available on Eurostat's website `http://ec.europa.eu/eurostat/data/database` as also a wide range of Eurostat data is also accessible on different mobile apps `http://ec.europa.eu/eurostat/help/first-visit/tools` [3].

The production as well as the dissemination of all this information requires the effort of many people in different areas of the organization working under tight deadlines. INE's business model is aligned with the Generic Statistical Business Process Model (GSBPM) that was developed jointly by UNECE, Eurostat and OECD within the Common Metadata Framework [1]. In the dissemination phase Statistics Portugal website is the main dissemination media, acting as the key source of information spread through all other supports and channels [2]. This phase manages the release of the statistical products to customers, assembling and releasing a range of static and dynamic products via a range of channels.
Considering that Statistics Portugal takes into account customers suggestions and commentaries, as well as the modifications that have taken place on society regarding access to information

and, also that the access through mobile network to mobile network duplicated in 4 years – see Figure 1) the Institution revamped its website. The access to statistics (themes) or products as statistical data (microdata or database), interactive applications (house price in cities is one of the 2018 second semester novelties) or thematic folders is quite intuitive.

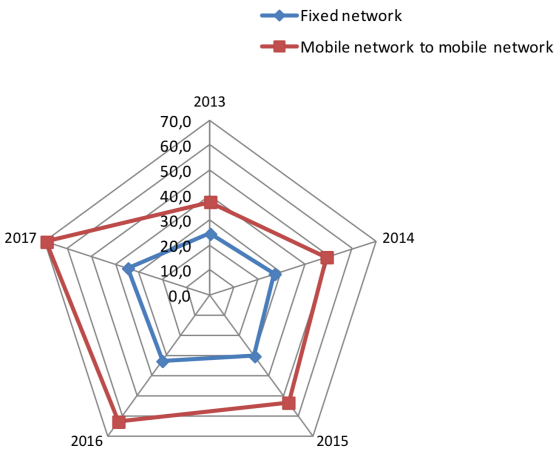


Figure 1: Broadband Internet accesses per 100 inhabitants(%) by Type of network; Annual

One can go further and explore a blog, but a website and a blog are really two different things: a website can contain a blog; a blog doesn't and can't contain a full website [5]. Nevertheless, a blog is another channel.
This paper contributes to the literature by showing that dissemination cannot be apart from communication either by statistical literacy or videos and info graphics and this altogether represents more than information; this is knowledge and Statistics Portugal is on it, close to customer needs.

References

[1] Generic statistical business process model (gsbpm) v. 5.0. `https://gss.civilservice.gov.uk/wp-content/uploads/2016/01/Generic-Statistical-Business-Process-Model.pdf`. Accessed: 2019-02-25.

[2] M. Ribeiro et al. Processo de produção estatística do ine. In *XXV Jornadas de Classificação e Análise de Dados*, Escola Naval, 6th april 2018.

[3] Eurostat. Dissemination. `https://ec.europa.eu/eurostat/about/policies/dissemination`, 2019. Accessed: 2019-02-25.

[4] OECD. *Trends Shaping Education 2019*. OECD Publishing, Paris, 2019.

[5] Lisa Sabin-Wilson. *Wordpress for Dummies*. John Wiley & Sons, Inc., New Jersey, 7th edition, 2015.

Thematic Session: CLAD–SPE

13 April, 11:10 - 11:30, ESTGV Auditorium

Dynamic Principal Component Analysis

Isabel Silva[1], **Maria Eduarda Silva**[2]
[1] Faculdade de Engenharia, Universidade do Porto and CIDMA, ims@fe.up.pt
[2] Faculdade de Economia, Universidade do Porto and CIDMA, mesilva@fep.up.pt

---

Multidimensional time series are observed in the most varied fields of application. Principal Component Analysis (PCA) can be used to reduce dimensionality. However, formal inference procedures based on principal components rely on the independence (and multivariate normality) of the observations, a condition that is violated for time series data. In this work, we describe a frequency domain version of PCA proposed by Brillinger [1] that takes into account the correlation in time. Illustration with real data is presented.

Keywords: dimensionality reduction, principal component analysis, spectral analysis, time series data

---

The multidimensional temporal (and spatio-temporal) series are observed in the most varied fields of application and are characterized by the correlation structure induced by the sequential order of observations. Let $\mathbf{X}_t$ be a $p-$dimensional time series. The process is said to be stationary if $\mathrm{E}[\mathbf{X}_t]$ and $\mathrm{E}[\mathbf{X}_t\mathbf{X}'_{t+h}]$ exist and don't depend of time t. The $p \times p$ autocovariance function is given by

$$\Gamma_{xx}(h) = \mathrm{E}[\mathbf{X}_t\mathbf{X}'_{t+h}] - \mathrm{E}[\mathbf{X}_t]\mathrm{E}[\mathbf{X}'_{t+h}].$$

If $\sum_h \Gamma_{xx}(h) < \infty$ then the spectral density matrix of $\mathbf{X}_t$ is given by

$$f_{xx}(\omega) = \sum_{h=-\infty}^{+\infty} \Gamma_{xx}(h)\exp(2\pi i h\omega).$$

Therefore, the autocovariance function and the spectral density are Fourier transform pairs and therefore contain the same information. As a consequence, there are two approaches (not necessarily mutually exclusive) to analyse time series data. The time domain approach considers the lagged relationships as most important while in the frequency (or spectral) domain, the periodic information is the most important.
In some multidimensional contexts, the number of observations per series exceeds the total number of time series, so it is of great importance to reduce the dimensionality of the data, extracting the most important information and eliminating noise and redundant correlations. By doing this, graphic representation and subsequent statistical analysis of the dataset are facilitated.

One very popular method for dimensionality reduction is Principal Component Analysis, which allows us to obtain a new set of variables, called Principal Components (PC), that are uncorrelated and ordered so that the first few retain most of the variation presented in the dataset (Jolliffe [2]). In some fields of application, PCA not only reduces the dimensionality of the dataset but also allows for reasonable interpretations of the retained PC.
As referred by Jolliffe [2], most of the inference procedures to be performed for PC are based on independence as well as on multivariate normality of the data, condition that are not satisfied for time series data. Several techniques have been proposed to overcome this issue.
One of the developed methodologies is the so called dynamic PCA, proposed by Brillinger [1] for multivariate time series assuming that the underlying process is stationary. As referred by Shumway and Stoffer [3], it can be considered as a PCA in the frequency domain where classical PCA is performed at each frequency, providing a set of principal components series which are uncorrelated at all time lags, thus allowing inferential procedures.
Formally, dynamic PCA approximates a p vector valued time series X_t by a set of k uncorrelated time series Y_t such that Y_t is the best approximation of X_t in mean squared error sense. While the classical ('static') PCA are linear combinations of the original data, the dynamic PC are linear combinations of past, present and future observations.
Note that classical PCA works with a covariance (or correlation) matrix, but in the time series context we can consider (auto)covariance between variables observed at the same time (given in the matrix $\Gamma_{xx}(0)$) but also between variables at different times (given by the matrices $\Gamma_{xx}(k)$ for $k \neq 0$). Therefore, given the equivalence between the autocovariance and the spectral density functions, it is natural to consider PCA in the frequency domain.
The objective of this work is to describe the Dynamic Principal Component Analysis, discussing its strengths and weaknesses and addressing some implementation issues. In addition, the results of the application of this technique to real datasets are exhibited and compared with classical PCA and MSSA (where the original series is decomposed in a small number of independent and interpretable components that can be thought as trend, oscillatory components and a structureless noise).

Acknowledgements The authors were partially supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

References

[1] D. R. Brillinger. *Time Series: Data Analysis and Theory.* Classics in Applied Mathematics, 36, SIAM, 2001.

[2] I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, 2002.

[3] R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications, With R Examples.* Springer, New York, 4th ed. edition, 2017.

13 April, 11:30 - 11:50, ESTGV Auditorium

Time series analysis via complex networks: a first approach

Vanessa Silva[1], **Maria Eduarda Silva**[2], **Pedro Ribeiro**[3]
[1] CRACS and INESC-TEC, Faculdade de Ciências, Universidade do Porto
[2] Faculdade de Economia, Universidade do Porto & CIDMA, mesilva@fep.up.pt
[3] CRACS and INESC-TEC, Faculdade de Ciências, Universidade do Porto

---

Time series data are ubiquitous in the world of data. Mining interesting features from time series has become crucial in multidisciplinary contexts. Here we consider mapping the data to complex networks. The vast arsenal of network science methodologies is then used to characterize the time series. The results indicate that different mappings and a range of network topological features capture specific time series characteristics, opening new avenues in time series analysis.

Keywords: time series, complex networks, clustering

In recent years data indexed in time, time series, have become the norm rather than the exception as a result from technological developments. As an example we may mention that sensors and mobile devices routinely gather data. Summarizing, modelling and inferencing from these multidimensional, temporally dependent and usually large data sets present new challenges to statistical science and require new methodological and computational tools. The set of methods and associated theory for univariate and evenly spaced time series analysis are well developed and understood. Several well-known models are widely used to describe the characeteristics of the data and produce forecasts. However, the multidimensional, temporally dependent and usually large data sets that are being routinely collected as a result from technological development present characteristics that inhibit the application of traditional time series analysis tools. Thus, new methodological and computational tools for time series analysis are required.

Complex networks describe a wide range of systems in nature and society and their analysis has been receiving increasing interest from the research community. The impact has been so big that has led to the emergence of the new field of Network Science [1]. The study of complex networks has advanced in the last few years and there exists a vast set of topological graph measurements available, an established set of problems such as community detection or link predictio, and a large track record of successful application of complex network methodologies to different fields.

Motivated by the success of complex network methodologies and with the objective of acquiring new methods for the analysis of time series, several network-based time series analysis approaches have been recently proposed, based on mapping time series to the

network domain. The resulting in networks capture the structural properties of the series. For instance, periodic series are represented by regular networks, random series by random networks and chaotic series map to scale-free networks. Some mappings result in networks that have as many nodes as the number of observations in the time series, but others, such as a quantile based mapping [2], allow to reduce the dimensionality of the series while preserving the characteristics of the time dynamics. Network-based time series analysis techniques have been showing promising results and have been successful in the description, classification and clustering of time series of real datasets.

This work aims at performing a systematic network based characterization of a large set of linear and nonlinear time series models using global topological features of visibility [3, 4] (NVG, HVG) and quantile graphs [2] (QVG), namely, average degree, average path length, number of communities, cluster coefficient and modularity. To this end we perform a detailed simulation study. Specifically, we generate 100 sample paths of size $T = 10000$ of each of the 11 models, White Noise, AR(1) with two different parameters, AR(2), ARIMA(1,1,0), ARFIMA(1,0.4,0),SETAR with 2 regimes, HMM, INAR(1), GARCH(1,1) and EGARCH(1,1) in a total of 1100 time series. The time series are then mapped into networks using the NVG, HVG and QG (with 100 quantiles) methods.

The results indicate that different mappings and different topological metrics capture different characteristics, complementing each other and providing more information when combined thus improving the results over the use of a single mapping concept, as is common on the literature.

Acknowledgements This research was partially supported by the Portuguese national funding agency for science, research and technology (FCT), within the Center for Research and Development in Mathematics and Applications (CIDMA), project UID/MAT/04106/2019 and SFRH/BD/139630/2018.

References

[1] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[2] Andriana Susana Lopes de Oliveira Campanharo and Fernando Manuel Ramos. Quantile graphs for the characterization of chaotic dynamics in time series. In *Complex Systems (WCCS), 2015 Third World Conference on*, pages 1–4. IEEE, 2015.

[3] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque, and Juan Carlos Nuno. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975, 2008.

[4] Bartolo Luque, Lucas Lacasa, Fernando Ballesteros, and Jordi Luque. Horizontal visibility graphs: Exact results for random time series. *Physical Review E*, 80(4):046103, 2009.

13 April, 11:50 - 12:10, ESTGV Auditorium

Risk stratification of heart failure patients from age-independent thresholds

Sónia Gouveia[1,2],**Manuel G. Scotto**[3], **Paulo J. S. G. Ferreira**[4,1]
[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), UA, sonia.gouveia@ua.pt
[2] Center for R&D in Mathematics and Applications (CIDMA), University of Aveiro (UA)
[3] CEMAT and Instituto Superior Técnico, University of Lisbon
[4] Department of Electronics, Telecommunications and Informatics (DETI), UA

Previous studies have shown that a baroreflex sensitivity value lower than 3 ms/mmHg identifies cardiac patients at higher mortality risk. However, a lower value can also be the result of a process of physiologic senescence besides a sign of cardiac dysfunction. Therefore, the present study aims to assess whether the constant threshold represents a natural partition of a large group of patients and whether its risk stratification capability depends on the age of the patient.

Keywords: heart failure (HF), risk stratification, spontaneous baroreceptor reflex sensitivity (BRS), transfer function (TF), hierarchical clustering

---

Baroreceptor reflex sensitivity (BRS) is an important prognostic factor because a reduced BRS has been associated with an adverse cardiovascular outcome. The threshold for a 'reduced' BRS was established by the ATRAMI study at BRS < 3 ms/mmHg in patients with a previous myocardial infarction [2], and has been shown to improve risk assessment in many other cardiac dysfunctions [3]. The successful application of this cutoff to other populations suggests that it may reflect an inherent property of baroreflex functioning, so our goal is to investigate whether it represents a 'natural' partition of BRS values. As reduced baroreflex responsiveness is also associated with ageing, we investigated whether a BRS estimate < 3 ms/mmHg could be the result of a process of physiological senescence as well as a sign of BRS dysfunction.
This study involved 228 chronic heart failure (HF) patients and 60 age-matched controls. Our novel method combined transfer function BRS estimation and automatic clustering of BRS probability distributions, to define indicative levels of different BRS activities [1]. The analysis produced a fit clustering (cophenetic correlation coefficient of 0.9 out of 1) and, as illustrated in Figure 1, the hierarchical procedure identified one group of homogeneous patients (a) which is well separated from the remaining by the constant cutoff of 3 ms/mmHg (b). Furthermore, the HF patients with BRS < 3 ms/mmHg were shown to exhibit an increased BRS-based mortality risk [hazard ratio (HR): 3.19 (1.73, 5.89), $p < 0.001$] with respect to the remaining HF patients.

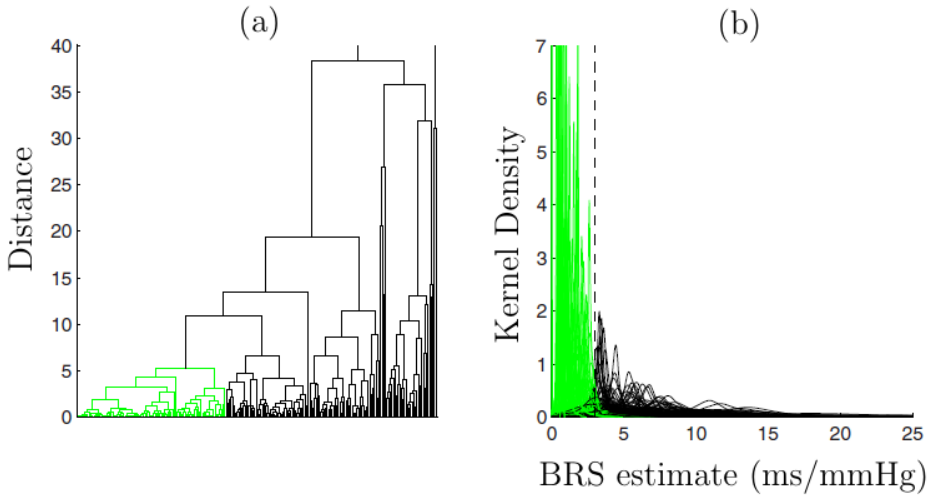


Figure 1: (a) Dendrogram produced by average link and L2-Wasserstein distance on BRS distributions. (b) BRS distribution for each HF subject, where the dashed line positions the empirical cutoff of 3.0 ms/mmHg. The color highlights the group of patients identified by cluster analysis (a) which exhibit the lowest BRS average and variability values (b).

On a subsequent analysis, an age-dependent BRS cutoff (estimated by 5% quantile regression of log(BRS) with age and considering the age-matched controls), provided a similar mortality value [HR: 2.44 (1.37, 4.43), $p = 0.003$]. Therefore, age was found to have no statistical impact on risk assessment, thus suggesting that there is no need to establish age-based cut-offs because 3 ms/mmHg optimally identifies patients at high mortality risk.

Acknowledgements This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) through national funds from Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) and from Fundo Europeu de Desenvolvimento Regional (FEDER), in the scope of the research projects IEETA (UID/CEC/00127/2019), CIDMA (UID/MAT/04106/2019) and CEMAT (UID/Multi/04621/2019).

References

[1] S. Gouveia, M.G. Scotto, G.D. Pinna, R. Maestri, M.T. La Rovere, and P.J.S.G. Ferreira. Spontaneous baroreflex sensitivity for risk stratification of heart failure patients: optimal cutoff and age effects. *Clinical Science*, 129:1163–1172, 2015.

[2] M.T. La Rovere, J.T. Bigger, F.I. Marcus, A. Mortara, P.J. Schwartz, and the group of ATRAMI Investigators. Baroreflex sensitivity and heart-rate variability in prediction of total cardiac mortality after myocardial infarction. *Lancet*, 351:478–484, 1998.

[3] G.D. Pinna, R. Maestri, and M.T. La Rovere. Assessment of baroreflex sensitivity from spontaneous oscillations of blood pressure and heart rate: proven clinical value? *Physiological Measurement*, 36:741–753, 2015.

13 April, 12:10 - 12:30, ESTGV Auditorium

Detection of diseases in heart rate variability

Argentina Leite[1], **Ana Paula Rocha**[2], **Maria Eduarda Silva**[3]
[1] Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro & C-BER & INESC TEC, Portugal, tinucha@utad.pt
[2] Faculdade de Ciências, Universidade do Porto & CMUP, Portugal
[3] Faculdade de Economia, Universidade do Porto & CIDMA, Portugal

This work focus on the application of time series models in the characterization and classification of Heart Rate Variability (HRV) data, considering Fractionally Integrated AutoRegressive Moving Average (ARFIMA) models with Exponential Generalized Autoregressive Conditionally Heteroscedastic (EGARCH) innovations. These models are used to extract measures that best characterize the underlying features of HRV. Then, Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are applied to these sequences of features to automatically detect the abnormality of HRV data.

Keywords: HRV, ARFIMA, EGARCH, LSTM

The characterization and classification of Heart Rate Variability (HRV) data has proved important to assess the integrity of the cardiovascular regulatory system and various methodologies to study HRV may be found in the literature. The most usual approach is based on linear AutoRegressive (AR) spectral analysis, which allows the identification of the autonomic nervous system (sympathetic and parasympathetic) components, namely the Low and High Frequency components (LF and HF). These AR models describe only short memory in the mean. However, it is acknowledged that HRV data display non stationary characteristics and exhibit long memory in mean and time-varying conditional variance (usually designated by volatility) among other nonlinear characteristics. Leite *et al* [1] considered the joint modeling of long memory and heteroscedastic characteristics of HRV using Fractionally Integrated AutoRegressive Moving Average (ARFIMA) models with Generalized Autoregressive Conditionally Heteroscedastic (GARCH) innovations. The ARFIMA-GARCH models which are an extension of the AR models usual in the analysis of HRV and may be used to capture and remove long memory in the mean and estimate the volatility in HRV data. A further empiric characteristic of HRV volatility is asymmetry in response to shocks. A Leite *et al* [2] used Exponential GARCH (ARFIMA-EGARCH) models to capture these effects and found that the parameters of the models are promising in differentiating health and disease. These models satisfy the following equations:

$$\phi(B)(1-B)^d x_t = \epsilon_t \quad (1)$$

$$\epsilon_t = \sigma_t z_t \quad (2)$$

$$\log \sigma_t^2 = u^* + v_1 \log \sigma_{t-1}^2 + u_1 |z_{t-1}| + \xi_1 z_{t-1} \quad (3)$$

where $u^* = u_0 - u_1\sqrt{2/\pi}$, B is the backward-shift operator, $d \in R$ and $z_t = \epsilon_t/\sigma_t$ are independent and identically distributed random variables with zero mean and unit variance. Equation (1) describes the conditional mean of the process with serially uncorrelated residuals ϵ_t and is said an ARFIMA$(p, d, 0)$ where d, the long-memory parameter, determines the long-term behaviour, p and the coefficients in $\phi(B)$ model the short-range properties. Equations (2) and (3) describe the conditional variance of the process where ϵ_t are called shocks and z_t are the standardised shocks. The parameters u_1 and v_1 characterize the volatility clustering phenomena, the short-range properties and the persistence, and the parameter ξ_1 describes the asymmetric effect.
In this work the ARFIMA-EGARCH approach to 24 hour HRV modeling is used to extract measures that best characterize the underlying features of HRV: mean of HRV, long memory in the mean, LF, HF, short range and persistence in volatility and asymmetry effect. Then, Long Short-Term Memory (LSTM) networks [4], a type of Recurrent Neural Network (RNN), are applied to these sequences of features to classify HRV data. LSTM is an improved approach of RNN and is more effective at capturing the long-term dependencies between time steps of feature sequences. The system is trained and tested with 24 hours HRV data from the Noltisalis database [3]. As per our best knowledge, this is the first work in which ARFIMA-EGARCH model and deep learning techniques are employed in distinguishing disease and health HRV data.

Acknowledgements This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation, COMPETE 2020 Programme (project POCI- 01-0145-FEDER-006961) and ERDF-NORTE2020 (project STRIDE-NORTE-01-0145-FEDER-000033), and by National Funds through the Portuguese funding agency, FCT - Fundacao para a Ciencia e a Tecnologia as part of projects UID/EEA/50014/2019, CIDMA UID/MAT/04106/2019 and CMUP UID/MAT/00144/2019, funded by FCT (Portugal) with national (MEC) and European structural funds through the programs FEDER, under the partnership agreement PT2020.

References

[1] A. Leite, A. P. Rocha, and M. E. Silva. Beyond long memory in heart rate variability: an approach based on fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity. *Chaos*, 23:023103, 2013.

[2] A. Leite, M. E. Silva, and A. P. Rocha. Model-based classification of heart rate variability. In *Proceedings of IEEE-EMBS International Conference*, pages 518–521, 2018.

[3] M. G. Signorini, R. Sassi, and S. Cerutti. Working on the noltisalis database: Measurement of nonlinear properties in heart rate variability signals. In *Proceedings of IEEE-EMBS International Conference*, pages 547–550, 2001.

[4] K. Taylor. *Deep Learning Using Matlab. Neural Network Applications.* CreateSpace Independent Publishing Platform, 2017.

Contributed Sessions

12 April, 9:00 - 9:20, Room A2

From sparse principal components to clustering of variables in high-dimensional data

Adelaide Freitas[1]
[1] Department of Mathematics & CIDMA, University of Aveiro, adelaide@ua.pt

Clustering and Disjoint Principal Component Analysis (CDPCA) is a sparse PCA methodology aimed at identifying clusters of objects and, simultaneously, describing the data in terms of sparse and disjoint components. An iterative alternating least squares algorithm (with random initialization step) was suggested to implement CDPCA. In this work, the ability of this algorithm for producing similar clusterings of variables when multiple runs are applied on high dimensional data sets is analyzed.

Keywords: principal component analysis, sparsity, clustering

Sparse Principal Component Analysis is an appealing area of research due to its usefulness for interpretation purposes, in particular, in high-dimensional data sets. A constrained principal component analysis, called clustering and disjoint principal component analysis (CDPCA), which is aimed at a non-overlapped clustering of variables and, simultaneously, a clustering of objects, on the reduced set of these CDPCA components, was proposed by [2]. In this context, variable clusters will be determined by the variables that define each sparse and disjoint CDPCA component.
Given a data matrix $\mathbf{X} = [x_{ij}]_{I\times J}$, the CDPCA procedure applies Principal Component Analysis (PCA) on a matrix obtained from $\mathbf{X}$ where each original object of $\mathbf{X}$ has been replaced by its cluster centroid given by the application the k-means algorithm to the original data matrix $\mathbf{X}$. Consequentely, the CDPCA model defines the data matrix $\mathbf{X}$ as follows ([2]):

$$\begin{aligned}\mathbf{X} &= \mathbf{U}\bar{\mathbf{X}} + \mathbf{E}_1 && \text{(k-means on } \mathbf{X}\text{)}\\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T + \mathbf{E}_1 + \mathbf{E}_2 && \text{(PCA on } \mathbf{U}\bar{\mathbf{X}}\text{)}\\ &= \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T + \mathbf{E} && \text{(CDPCA model)}\end{aligned}$$

where $\mathbf{U} = [u_{ip}]_{I\times P}$ and $\mathbf{V} = [v_{jq}]_{J\times Q}$ are the binary matrix of the assignment of the objects into P clusters and the binary matrix of the assignment of the variables into Q components, respectively, $\mathbf{A} = [a_{jp}]_{J\times Q}$ is the component loading matrix where the nonzero loadings of each component are determined by the nonzero elements of the correspondent column of $\mathbf{V}$, $\bar{\mathbf{Y}} := \bar{\mathbf{X}}\mathbf{A}$ is a $(P \times Q)$ object centroid matrix in the reduced space of the components and $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ with $\mathbf{E}_1$, $\mathbf{E}_2$ the $(I \times J)$ error matrices arising from k-means and PCA, respectively.

In order to estimate the parameters $\mathbf{U}$, $\bar{\mathbf{Y}}$ and $\mathbf{A}$ of the CDPCA model, a iterative alternating least-squares (ALS) algorithm was proposed by [2] and rewrite in two basic steps by [1].
In the initialization step of the ALS algorithm, random procedures for the first assignments of the matrices $\mathbf{U}$ and $\mathbf{V}$ are required. To increase the possibility to achieve a stable solution (expectably, the optimal solution) using ALS, it has been suggested to run the algorithm several times for different initial assignments of these matrices. However, there are no studies for evaluating whether (final) CDPCA components, and consequently, their correspondent clusterings of variables obtained from different applications of CDPCA on the same data set are similar. In this work, we compare the outcomes of CDPCA for different numbers of CDPCA components retained in the model. In particular, we consider applications of CDPCA for three real gene expression data sets already available in `R` packages (`plsgenomics` and `spls`) for which the number of variables (genes) is higher than the number of the objects (samples), namely:

- *leukemia*: 3051 genes and 38 samples extracted from two types of tumor (dimension of each type-group: 11/27);
- *lymphoma*: 4026 genes and 62 samples extracted from three types of cancer (dimension of each type-group: 42/9/11);
- *SRBCT*:2308 genes and 83 samples extracted from four different groups (dimension of each type-group: 29/11/18/25).

Also, we compare results obtained using CDPCA with outcomes provided by other variable clustering methods.

Acknowledgements The author was supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

References

[1] E. Macedo and A. Freitas. The alternating least-squares algorithm for cdpca. Plakhov, A. et al (Eds.) Optimization in the Natural Sciences, Communications in Computer and Information Science, Springer Verlag, 2015.

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

12 April, 9:20 - 9:40, Room A2

Evaluating outlier detection methods: A review of performance measures

A. Pedro Duarte Silva

Católica Porto Business School and CEGE / Universidade Católica Portuguesa, psilva@porto.ucp.pt

---

The comparison of alternative outlier detection methods is discussed. Common performance measures are reviewed, and it is argued that they should be complemented by methodologies that explicitly take into the differences between type I and type II errors. The issues discussed are illustrated by comparisons of outlier detection techniques for interval-valued data in an Internet security application.

Keywords: Multivariate outlier detection, Decision curves, Expected utilities, Error costs, Interval-valued data

---

The detection of multivariate outliers is a major topic in robust statistics, and nowadays many alternative methodologies exist for this purpose [2]. However, when comparing different methodologies it is not always clear which criteria should be employed, and some studies use different, and sometimes conflicting, performance measures.

In particular, one common approach uses measures related to hypothesis testing theory, such as estimates of test size, power and false discovery rates. Other studies employ simple proportion measures such as precision, recall, or their harmonic mean known as the F-measure. However, none of these approaches takes into account the different costs of type I (wrongly flagging false outliers) and type II (missing true outliers) errors, although these costs are not irrelevant for comparing methodologies, and vary widely from application to application.

In this presentation, it will be argued that tools of statistical decision theory commonly used in classification analysis applications, may and should be used in the evaluation of outlier detection methodologies. Specifically, approaches that take type I and type II costs (explicit, implicit or even imprecise) into account, such as expected utilities or decision curves [1] [4], give invaluable insights into the relative merits of different techniques. These approaches, by focusing on error estimates close to an application dependent optimal threshold, often lead to more relevant comparisons.

The issues discussed above will be illustrated by comparisons of alternative outlier detection techniques for interval-valued data [3] in an Internet security application. This application requires outlier detection methodologies for identifying Internet attacks, and creating security profiles for Internet entities.

References

[1] S.G. Baker, N.R. Cook, A. Vickers, and B.S Kramer. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society. A*, 172:729–748, 2009.

[2] A. Cerioli and A. Farcomeni. Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis*, 55:544–553, 2011.

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

[4] D. Hand. Assessing the performance of classifiication methods. *International Statistical Review*, 80:400–414, 2012.

12 April, 9:40 - 10:00, Room A2

Looking for atypical groups of distributions in the context of genomic data

Ana Helena Tavares[1], **Vera Afreixo**[1], **Paula Brito**[2]
[1] CIDMA, University of Aveiro, ahtavares@ua.pt
[2] FEP & LIAAD-INESC TEC, University of Porto

This work addresses the problem of detecting groups of observations (distributions) and flagging those that differ abnormally from the majority of the groups, termed as atypical groups. The proposed method combines a hierarchical classification technique, to identify groups of similar distributions, with a functional outlier detection method, to identify those groups that contain outliers. Groups with outlying observations are forwarded for sub clustering. Once the final partition is obtained, each cluster is represented by a class prototype, whose outlyingness is evaluated according to a functional approach. Clusters with atypical class labels are flagged as atypical groups. The method is applied for the detection of groups of atypical genomic words, based on their distances distributions.

Keywords: clustering, outlying distribution, atypical group

The identification of outliers can lead to the discovery of truly unexpected knowledge in several areas, e.g. electronic commerce, video surveillance and health care. A widely reported definition of outlier observation is the one proposed by Grubbs in 1969 and quoted in Barnett and Lewis [1]: *An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.* This general definition of outlier is vague and becomes meaningful only under a given context or application.
In this work, we introduce the concept of *atypical group* and propose a procedure for its identification. We focus our work in the detection of such groups in data that can be represented by a distribution, in particular, in distances distributions between genomic words [3]. We are convinced that large heterogeneous datasets, where distinct patterns coexist, may exhibit one or more atypical groups, meaning groups of observations whose 'mean' pattern stands out from the majority of the 'mean' patterns.
If large heterogeneous datasets where distinct patterns coexist can validly be clustered, then the class prototypes may provide a meaningful description of similarities and differences in the data. By representing each group by a prototype, the inicial dataset is reduced to a given number of representative distributions. By applying a functional outlier procedure over the set of prototypes, it is possible to identify those groups whose prototype is flagged as outlier. Such group is then termed as an atypical group.

To identify distinct patterns in a set of distributions we have combined a hierarchical clustering method with a functional outlying detection method. The first creates a hierarchy of clusters according to a dissimilarity measure, while the second flags observations with atypical curves in the set of group members. In this second step, a measure of outlyingness is used that privileges the shape of the distributions and not only the magnitude of their values [2]. Groups in which atypical observations are identified, are forwarded for (sub)clustering, and the procedure is repeated until no outliers are identified. Once the final partition is obtained, each cluster is represented by a class prototype and its outlyingness is evaluated according to the same functional approach [2]. The key idea of our proposal is to use a functional outlyingness criterion as indicator of the cluster homogeneity and then use it again to identify the atypical class prototypes.
We are particularly interested in developing a method that recovers groups of genomic words with similar distribution patterns along the genome sequence and, in particular, those very small groups with a distribution pattern which is markedly different from the majority. We analyze the dataset of the inter-word distance distributions of words of length $k = 5$, which contains 1024 distributions. To form the clusters an agglomerative hierarchical method is applied, considering the Mallows L^1 distance and average linkage. To decide on the number of clusters to retain in each step of the procedure we resort to two validity indexes, the Calinski-Harabasz index and the Silhouette score.
The application of this new procedure allowed identifying three groups of distributions with homogeneous patterns and very different from the others. These groups are of small dimension, and the words belonging to the identified groups are rich in CG dinucleotides. The groups of genomic words identified may have a potential biological interest, since atypical distribution patterns may be related to words that have biological meaning.

Acknowledgements This work was partially supported by Fundação para a Ciência e a Tecnologia, within projects UID/MAT/04106/2019 (CIDMA) and UID/EEA/50014/2013 (INESC TEC), and by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961.

References

[1] Vic Barnett and Tomy Lewis. *Outliers in Statistical Data*. Wiley, 1994.

[2] Peter J Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, pages 1–15, 2018.

[3] Ana Helena Tavares, Vera Afreixo, João MOS Rodrigues, and Carlos AC Bastos. The symmetry of oligonucleotide distance distributions in the human genome. In *Proc. ICPRAM (2)*, pages 256–263, 2015.

12 April, 10:00 - 10:20, Room A2

Internet usage patterns: Segmentation of European users using a multilevel latent class model

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 use of the Internet in the 28 countries of the European Union, based on the usage patterns and characteristics of the users. It aims to identify typologies of Internet use – the frequency of use, means of access, and activities by individuals – using data set from the Eurobarometer. A two-level latent class analysis was specified: in the first level the individuals within each country are grouped according to their characteristics of use; and in parallel in the second level, countries are grouped based on the similar structure of individual segments. At the first level of analysis (individuals), four segments were found: Non Users, Instrumental Users, Socializers, and Advanced Users. At the second level, countries were grouped into three segments based on their similarities.

Keywords: European Union, Internet, Latent class Models, Multilevel Analysis

---

Multilevel data structures are quite common in the social and behavioral sciences and new analytical techniques have been applied to these specific data sets. The Multilevel Latent Class Model (MLCM) generalizes the conventional Latent Class Model (LCM) by taking the multilevel structure into account, i.e., the fact that individuals living in the same country share specific characteristics [3].

The Multilevel Latent Class Model (MLCM) considers not only the individual level (Level 1), but also an upper level (Level 2) that defines a nesting or hierarchical structure [1]. The MLCM decomposes the existing heterogeneity between countries and within countries (individuals), resulting into homogeneous segments of countries and individuals. Thus, by using the MLCM instead of the LCM, the analysis is conducted at two distinct levels with the simultaneous clustering at each level: individual level, i.e., individuals' profile within each country in terms of their internet usage; country level, i.e., the similarities and differences between European countries in this context.

The data set comes from the Eurobarometer 87.4/2017 [2] and contains information on the 28 countries of the European Union (with 27812 citizens). The average age of the respondents is 48.49 years (s.d. = 18.75) and varies between 15 and 99 years old.

At the individual level (Level 1), four variables were used to identify individual segments in Europe, taking their Internet usage pattern into account: frequency of internet access

at work, frequency of internet access at home, means of access, and online activities. Six sociodemographic variables were introduced to characterize the latent classes, namely: gender, age, literacy, marital status, occupation, and type of community. At the second level of analysis (Level 2), countries were introduced as contextual predictors, allowing the grouping of individuals into segments based on the similarities found.
Based on the Bayesian Information Criterion (BIC), the best model contains four classes at the individual level and three classes at the country level. At the individual level (Internet usage patterns), four segments of Internet users were identified, each with a distinct sociodemographic profile: Class 1 - Non Users (21%) reveals no internet use at all; Class 2 - Instrumental Users (23.2%) shows a widespread of utilization in what regards the frequency, varying from an occasional weekly utilization to a regular daily utilization; Class 3 - Socializers (22.4%) presents an higher use of social networks; Advanced Users (33.5%) are always online. At the second level, countries were grouped into three segments. Most countries have a concentrated probability (maximum probability) of belonging to a specific cluster. Bulgaria, Croatia, Cyprus, Czech Republic, Greece, Hungary, Italy, Poland, Portugal, Romania, and Slovakia present a maximum probability of belonging to Class 1. The same happens with Austria, Belgium, Estonia, France, Germany, Ireland, Latvia, Lithuania, Luxembourg, Malta, Northern Ireland, Spain and Slovenia in Class 2. The most developed countries, Denmark, Finland, Great Britain, Sweden, and the Netherlands present a higher probability of belonging to Class 3.

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

References

[1] H. Kimberly and B. Muthén. Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, 17(2):193–215, 2010.

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

[3] J. K. Vermunt. Multilevel latent class models. *Sociological Methodology*, 33:213–239, 2003.

12 April, 9:00 - 9:20, Room A3

State space modeling in water quality monitoring in a river basin

A. Manuela Gonçalves[1], **Marco Costa**[2]
[1] CMAT-Center of Mathematics, DMA-Department of Mathematics and Applications, University of Minho, Portugal, mneves@math.uminho.pt
[2] CIDMA-Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal

---

This study is conducted within the context of surface water quality monitoring in a river basin, and it is proposed an approach for the structural time series analysis based on the state space models associated to the Kalman filter. The main goal is to analyze and evaluate the temporal evolution of the environmental time series, and to identify trends, seasonality or possible changes in water quality within a dynamic monitoring procedure.

Keywords: State Space Models, Kalman Filter, Dissolved Oxygen, River Basin, Change points

---

State space models constitute a significantly important class of models in time series analysis due to their flexibility in dynamic phenomena analysis and of variable systems evolution, randomly and with meaningful variability throughout time. State space models have significantly contributed to extending the classic domains of application of statistical time series analysis. They allow a natural interpretation of a time series as the combination of several components, such as trend, seasonal or regressive components. A structural model can therefore not only provide forecasts but also, through estimates of the components, present a set of stylized facts, and this formulation will allow making some useful interpretations.

In this study, it is proposed a dynamic modeling procedure based on the state space approach (associated to the Kalman filter) in time series of water quality variables [2] and [4]. The data concerns the River Ave's hydrological basin located in the northwest of Portugal, where monitoring has become a priority in water quality planning and management because its water has been in a state of obvious environmental degradation for many years. As a result, the watershed is monitored by seven monitoring sites distributed along the River Ave and its main streams. For the modeling process we consider time series relating to the Dissolved Oxygen water variable measured on a monthly basis over a 15-year period (January 1999–January 2014).

State space models show the versatility of the incorporation of unobserved components (states), of stochastic nature, that describe the variation of time series (such as trends and seasonality), which are updated in real time in a recursive way as new observations

become available and help to improve the forecasts by reflecting the dynamic nature of the process under study. These components have a natural interpretation, representing the salient features of the environmental time series under investigation [3] and [1] .
From an environmental point of view, the proposed approach allows to obtain pertinent findings concerning water surface quality interpretation and change point analysis, thus highlighting the potential value of this type of analysis, by identifying unexpected changes that are important for the process of water quality management and evaluation.

Acknowledgements This work is financed by FEDER funds through the Competitivity Factors Operational Programme - COMPETE, and by national funds through FCT (Fundação para a Ciência e a Tecnologia) within the framework of Project POCI-01-0145-FEDER-007136, and Project UID/MAT/00013/2013.

References

[1] 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.

[2] A.M. Gonçalves, O. Baturin and M. Costa. *Time series analysis by state space models applied to a water quality data in Portugal.* American Institute of Physics, Volume 1978, 470101-1–470101-4, 2018.

[3] M. Costa and A.M. Gonçalves. Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, 25(2):151–163, 2011.

[4] M. Costa and A.M. Gonçalves. *Combining Statistical Methodologies in Water Quality Monitoring in a Hydrological Basin - Space and Time Approaches.* Editors Kostas Voudouris and Dimitra Voutsa. Croacia: Intech, 2012.

12 April, 9:20 - 9:40, Room A3

Normalization of foot clearance and spatiotemporal gait data using multiple linear regression models

Flora Ferreira[1,2], **Carlos Fernandes**[1], **Miguel Gago**[3,4], **Nuno Sousa**[4], **Wolfram Erlhagen**[5], **Estela Bicho**[1]
[1]Algoritmi Center, University of Minho, flora.ferreira@gmail.com[2]ESTG, Polytechnic of Porto, [3]Neurology Service, Hospital Senhora da Oliveira, [4]ICVS, School of Medicine, University of Minho, [5]Center of Mathematics, University of Minho, Portugal

The aim of this study is to use a multiple regression normalization strategy that accounts for subject age, height, weight, sex, and walking speed or stride length to identify differences in foot clearance and spatiotemporal gait variables between patients with parkinsonism and controls. The results show that the multiple regression approach reduced the correlations between gait measures and physical properties, speed, and stride length, and has the potential to improve gait related analysis.

Keywords: Multiple regression models, gait analysis

Foot clearance and spatiotemporal gait measurements are important in distinguishing dysfunctional gait. However, the differences in subject physical properties, as well as intersubject variations on walking speed, may limit the capacity to detect between-group differences in a given gait variable [2]. In order to minimize the effect of between-subject physical differences on gait data different methods such as dimensionless equations, detrending method, and multiple regression (MR) approaches [2] have been proposed. By comparing these three approaches Wahid et al. [2] showed that the MR normalization method better reduces the correlations between subject-specific physical properties and gait variables. Furthermore, the MR normalization has the potential to improve the ability to differentiate Parkinsonian gait from healthy controls [2]. In [2] and more recently in [1], normalization was employed on spatiotemporal gait data. However, foot clearance measures (toe and heel height during swing phase), have not been yet explored. The aim of this study is to employ the MR normalization approach [2] to identify differences in spatiotemporal gait as well in foot clearance variables between patients with parkinsonism and controls. Recently, foot clearance was reported to be inherently influenced by the stride length. Then, we include the stride length as an independent variable.
Gait measurements (see Figure 1) of 30 patients with parkinsonism and 15 age-matched healthy controls were collected using foot-worn inertial sensors while the subjects walking a 60-meter continuous course at a self-selected walking speed. Using the control dataset, different MR models were found for each gait variable considering different combinations

of the independent variables. The best regression model was selected based on adjusted R-square and Akaike's information criterion (AIC) values. Statistical assumptions for linear regression were met. Finally, robust fitted models were computed using a "bisquare" weight function, and then each gait variable are normalized by dividing the original value by the value estimated according to the MR model (for more detail, see [2]). To assess the influence of physical properties, speed and stride length on the gait variables before and after normalizing Spearman's rank order correlation coefficient (r_s) were computed. Speed and stride length was strongly correlated ($r_s \approx 0.95$) and then for each gait variable only the one which presents a higher correlation coefficient was considered as an independent variable. Speed was significantly correlated with stride time and cadence ($r_s > 0.52$). Stride length was strongly correlated with maximum toe late swing ($r_s \approx 0.65$). The remaining foot clearance variables were weakly correlated with speed and stride length ($r_s < 0.28$). Weak to moderate correlations between subjects' physical properties were observed. After normalization, all correlation coefficients were reduced ($r_s < 0.30$), with exception for the correlation between speed and stride length that persisted high ($r_s \approx 0.62$).

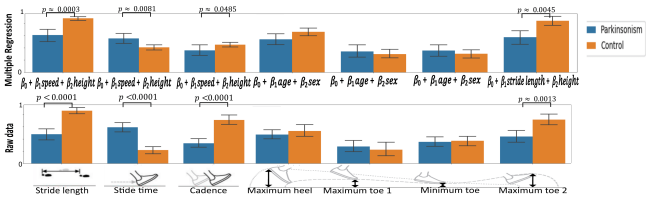


Figure 1: Comparison of mean gait data in patients with parkinsonism and controls. Results are presented as the mean±standard deviation. P-values for significant differences ($p < 0.05$) obtained by independent t-test and final MR model of each gait variable are displayed. For visualization purposes, data were scaled between 0 and 1.

Consistent with previous studies [2, 1] normalization using the MR approach reduced the correlations between spatiotemporal and foot clearance gait measures and subject physical properties, speed, and stride length. Although the significant differences in gait variables founded on raw data were also observed after normalization, the differences on the evidence measure reflect that raw data are influenced by between-subject differences in physical properties and variations on speed or stride length (Figure 1). Further studies with larger sample size are required to improve the reliability of the MR model's coefficient estimation.

Acknowledgements This work was partially supported by the projects NORTE-01-0145-FEDER-000026(DeM-Deus Ex Machina) financed by NORTE2020 and FEDER, and FP7 NETT project.

References

[1] V. Mikos and et al. Regression analysis of gait parameters and mobility measures in a healthy cohort for subject-specific normative values. *PloS one*, 13(6):e0199215, 2018.

[2] F. Wahid and et al. A multiple regression approach to normalization of spatiotemporal gait features. *Journal of applied biomechanics*, 32(2):128–139, 2016.

12 April, 9:40 - 10:00, Room A3

Pediatric arterial hypertension modeling

M. Filomena Teodoro[1,2], **Carla Simão**[3,4],
[1] CEMAT - Center of Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Avenida Rovisco Pais, n. 1, 1048-001 Lisboa, Portugal, maria.alves.teodoro@marinha.pt
[2] CINAV - Center of Naval Research, Portuguese Naval Academy, Base Naval de Lisboa, Alfeite, 2810-001 Almada, Portugal
[3] Faculty of Medicine, Lisbon University, Av. Professor Egas Moniz, 1600-190 Lisboa, Portugal
[4] Department of Pediatrics, Santa Maria's Hospital, Centro Hospitalar Lisboa Norte, Avenida Professor Egas Moniz, 1600-190 Lisboa, Portugal

---

The objective of the present study is to characterize the blood pressure (BP) profile of the Portuguese pediatric population at school age and to assess the prevalence of pediatric arterial hypertension (PAH), normal BP, high-normal BP, and analyze the relationship between normal-BP, high-normal BP, PAH and some demographic characteristics. A representative sample of the pediatric population was drawn up at the national level and data collection was completed recently. The statistical approach evidences that the results obtained are in agreement with some literature confirming a high prevalence of PAH among children and adolescents of the Portuguese population.

Keywords: pediatric hypertension, questionnaire, statistical approach, generalized linear models

---

High pediatric blood pressure has serious risk factors [1, 2, 3] being its prevention mandatory. In order to evaluate the caregivers' knowledge of the existence and details related to PAH, in [4] a preliminary study was carried out analyzing a simple and experimental questionnaire with 5 questions applied to caregivers that have attended to regular consultation at Santa Marias's regular pediatrics Consultation. Later, was analyzed an improved and more complete questionnaire was filled online by caregivers of children and adolescents who attended some public schools from Lisbon region. The objective of the present study is to characterize the BP profile of the Portuguese pediatric population at school age and to assess the prevalence of PAH, normal BP, high-normal BP, and analyze the relationship between normal-BP, high-normal BP, PAH and some demographic characteristics. A preliminary approach was initially undertaken using data collected during PAH outreach and screening activities on the day of hypertension. The results provided significant differences in the prevalence of hypertension among boys and girls. The age of the children was also influential in BP. Continuing the study, a representative sample of the pediatric population

was drawn up at the national level and data collection was completed very recently. Using this new data set, the prevalence of PAH was estimated and some associated factors were identified. Some statistical techniques were considered,for example, analysis of variance, generalized linear models, mixed models and factorial analysis. The results obtained are in agreement with those expected by health professionals confirming a high prevalence of PAH among children and adolescents of the Portuguese population.

Acknowledgements This work was supported by Portuguese funds through the FCT, *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2019, and *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal.

References

[1] P Muntner and J He. Trends in blood pressure among children and adolescents. *Journal of American Medical Association*, 291:1719—-1742, 2009.

[2] National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. The fourth report of the diagnosis, evaluation and treatment of high blood pressure in children and adolescents. *Pediatrics*, 114:555–576, 2004.

[3] S Stabouli and V Kotsis. The fourth report of the diagnosis, Adolescent obesity is associated with high ambulatory blood pressure and increased carotid intimal-medial thickness. *Journal of Pediatrics*, 147:651–656, 2005.

[4] M Filomena Teodoro and Carla Simão. Perception about Pediatric Hypertension. *Journal of Computational and Applied Mathematics*, 312:209–215, 2017.

12 April, 10:00 - 10:20, Room A3

First four order cumulants in Mixed Models

Patrícia Antunes[1], **Sandra Ferreira**[1,2], **Célia Nunes**[1,2], **Dário Ferreira**[1,2] **and João Mexia**[3]

[1] Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal

[2] Department of Mathematics, University Beira Interior, Covilhã, Portugal

[3] Center of Mathematics and its Applications, Faculty of Science and Technology, New University of Lisbon, Monte da Caparica, Portugal

---

In this presentation we focus on a new approach for describing the expressions for the first four order cumulants in Mixed Models. We will present results on Cumulant Generation Function to obtain estimators for the variances, third and fourth central moments and the remaining estimable vectors.
The usefulness of the proposed approach is assessed through a numerical simulation.

Keywords: Cumulants, Mixed Models, Moments, Parameter Estimation

---

In this presentation, the authors propose a new least-squares based method for estimating the cumulants, up to the 4th order, in the linear (additive) models - including the linear mixed models.
This topic is of interest in general and the proper methods for estimation of the higher-order moments and/or cumulants in linear (mixed) models are desired.
There has been much interest in deriving expressions for moments and cumulants using available computer technology and many authors have provided methods for expressing moments in terms of cumulants and vice versa, because cumulants are one of the importance characteristics alternative of moments of a distribution, in other words the moments can define cumulants.
The cumulants of order higher than two have many interesting properties. For most excellent accounts of the literature, we refer the readers to [4] and [1].
We will show that additive models have interesting properties and, besides this, we will estimate the parameters of models

$$Y = X_0\beta_0 + XL, \tag{1}$$

where X_0 and X are design matrices, β_0 is fixed and $L = (L_1^\top, ..., L_m^\top)\top$ is a random vector with independent components with null mean value and variances $\sigma_1^2, ..., \sigma_m^2$. These components will be grouped in m equivalence classes, given by $L_h d L_{h'}$ if $E(L_h^r) = E(L_{h'}^r), r = 2, 3, 4$. The components will, in each equivalence class, be grouped into a vector $\dot{L}_l, l =$

$1, ..., m$, and the corresponding columns of X can now be rewritten into sub-matrices $\dot{X}_l, l = 1, ..., m$, and we can decompose the random part, so that we will have

$$XL = \sum_{l=1}^{m} \dot{X}_l \dot{L}_l, \tag{2}$$

where the sub-vectors $\dot{L}_l, l = 1, ..., m$, will have null mean vectors and variance covariance matrices $\sigma_l^2 I_{c_l}, l = 1, ..., m$, coming

$$Y = X_0\beta_0 + \sum_{l=1}^{m} \dot{X}_l \dot{L}_l. \tag{3}$$

In this presentation we will recall certain results on CGF and on cumulants which will be useful to show how to obtain estimators for the cumulants of mixed models.
These models are easy to implement, not requiring structural conditions to be fulfilled. Such conditions, such as blocks with the same size, and orthogonal block structure, have played an important part in the study of models, see for instance [2] and [3]. The fact that these conditions are no longer required makes additive models much more comprehensive. Practical performance and relevance of theoretical results in this presentation is illustrated by some simulation results, considering an application with two crossed random vectors (A with two levels and B with three levels). All computation were performed using R software.

Acknowledgements This work was partially supported by the Center of Mathematics, University of Beira Interior through the project PEst-OE/MAT/00212/2019 and CMA through the project PEst-OE/MAT/00297/2019.

References

[1] D. F. Andrews. Asymptotic expansions of moments and cumulants. *Stat Comput*, 11:7–16, 2001.

[2] T. Caliński and S. Kageyama. *Block Designs: A Randomization Approach. Vol. I: Analysis. Lecture Notes in Statistics, 150.* Springer-Verlag, New York, 2000.

[3] T. Caliński and S. Kageyama. *Block Designs: A Randomization Approach. Vol. II: Design. Lecture Notes in Statistics, 170.* Springer-Verlag, New York, 2003.

[4] P. McCullagh. *Tensor methods in statistics.* Chapman and Hall, London, 1987.

12 April, 10:40 - 11:00, Room A2

Predictive value in healthcare: a forgotten measure?

Carina Ferreira[1], **Teresa Abreu**[2], **Mário Basto**[2],
[1] Master Student, School of Technology, IPCA, Barcelos, Portugal,
carinafcferreira@gmail.com
[2] Science Department, School of Technology, IPCA, Barcelos, Portugal

Basic statistical literacy is necessary for health professionals and patients to understand health information. The implications are many, among them are the informed consent and the adequate joint decision-making between doctor and patient, with the consequent increase in the quality of the services provided, the possibility to decrease the number of interventions and treatments performed and the consequent reduction in healthcare spending. In particular, the positive predictive value is an important statistic for both the physician and the patient.

Keywords: predictive positive value, prevalence, sensitivity, specificity

Many procedures and medical treatments are based on weak or non-existent statistical evidence, but nonetheless they are often performed around the world. This phenomenon leads to the so-called overtreatment. This problem is aggravated by the existence of excessive medical diagnoses for situations that would never cause symptoms or harm the patients [2], a phenomenon known as overdiagnosis. These situations have obvious negative consequences, not only for the patient, but also for healthcare spending in general. According to several authors [1], a large part of the population, and, in particular, physicians and patients, do not understand the really mean of various statistical concepts in health. Besides, the non-transparent form as information is often transmitted to physicians and patients, that exaggerate the benefits and underestimate the harms [1], makes the situation worse. Information in pamphlets, websites, and even medical journals often transmits information in a biased form, suggesting great benefits and little harms. Key points in pamphlets summarize results of published studies, but are often distorted and several important details are omitted [1], in addition to any conflicts of interest that may exist.
All medical procedures involve risks, hence a correct informed consent can only occur if the physician and patient know adequately the risks and the size of the associated benefits. In particular, in mammography screening for breast cancer, health authorities and health professionals have the duty to provide patients with the best estimates of the benefits and associated risks. Thus, the necessity of the computation of the positive predictive value, a statistic based on the sensitivity, specificity, and prevalence, so that physicians and patients can estimate how likely a woman is truly sick in light of a positive mammography.
Gigerenzer [1] posted a multiple choice question (4 options) to 160 gynecologists to ascertain if they had the basic knowledge necessary to estimate the predictive positive value, which establishes the proportion of true patients in the set of all positive tests.

This work intends to verify if the health professionals (physicians and nurses) and population in general, know how to interpret a positive result of a diagnostic test, in this case a mammography, from the current available values of sensitivity (true positive rate), specificity (true negative rate) and prevalence (rate of sick patients of a population at a specific time). Since January 2nd of 2019, an online survey was implemented with a question almost identical to [1]. The present data is the one obtained until January 20th of 2019. The question made considers a particular region where the probability of a woman of a given age group to have breast cancer is 1%, the sensitivity and the specificity of the mammography is 90% and 91% respectively (all concepts were explained). In the survey it was asked to choose the best answer for the probability of a woman with a positive mammography to have breast cancer. Four options were given: 1%, 9%, 81% and 90%. 180 responses were analysed, of which 54 were from physicians and 33 from nurses.
Among health professionals, 79.6% of the physicians, and 39.4% of the nurses, grossly overestimated the probability of a woman to have cancer, which is equivalent to 64.4% of health professionals. From those, 90.7% of the physicians and 61.5% of the nurses pointed to the maximum estimation available. Unexpectedly, about 21.8% of health professionals underestimated the probability, 13.0% of physicians and 36.4% of nurses, pointing to this probability being equal to the prevalence of breast cancer in the population. Only 7.4% of the physicians and 24.2% of the nurses gave the correct answer, which corresponds to only 13.8% of health professionals.
Among the general population, about 53.8% overestimated the probability, 17.2% underestimated it, and 29.0% gave the correct answer, a global better performance.
Physicians were the ones who overestimated more the benefits, and nurses the ones who gave the most unexpected response, that is, the answer 1%. The physicians were also the ones who least identified the right answer, only 7.4%, followed by the nurses, 24.2%. The general population had a higher correct hit rate, 29.0%, which is intriguing.
These results show that, most likely, many physicians and nurses do not know what is the relevant probability against a positive test, making it impossible to correctly assess the probability of a person having the disease when in the presence of a positive test. This way, the transmission of the complete information to the patient by the professionals is compromised. Can this be explained, at least in part, by the fact that health professionals are the main propaganda victims?

References

[1] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L.M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8:53–96, 2007.

[2] H.G. Welch, L. Schwartz, and S. Woloshin. *Overdiagnosed: Making People Sick in the Pursuit of Health*. MA: Beacon Press, Boston, 2012.

12 April, 11:00 - 11:20, Room A2

Analysis of administrative data with a binary response variable

Maria de Fátima Salgueiro[1], **Marcel D.T. Vieira**[2], **P.W.F. Smith**[3]
[1]Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, fatima.salgueiro@iscte-iul.pt
[2]Universidade Federal de Juiz de Fora, Department of Statistics, Juíz de Fora, Brasil, marcel.vieira@ice.ufjf.br
[3]University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, UK, P.W.Smith@soton.ac.uk

---

In this paper we estimate a binary logistic regression model to big register data, using population and simple random samples of alternative sizes. Results suggest a 5% sample is enough to reproduce the odds ratio structure of the chosen population model. Moreover, a considerable reduction of computational time was achieved, allowing for a faster decision-making process.

Keywords: big data, binary logistic regression, Bolsa Família Programme, sampling

---

Conditional Cash Transfers (CCTs) Programs are anti-poverty devices (public policies), with the aim of alleviating poverty in the short term and investing in human capital in the long term. The "Bolsa Família" Programme (BFP) was created in 2003 ([1] and serves about 14 million families ($>$ 50 million individuals), $\simeq$ 1/4 of the Brazilian population. Eligible families are those living in poverty and extreme poverty.
CadÚnico is a Brazilian administrative data source, which serves as a means of selection of low income families for the BFP. Information provided are self-declared through an interview. It should be noted that just being registered does not guarantee a family access to the BFP; the selection is made by the federal government based on the budgetary limit available for the program. In 2015, 27 192 314 families (our target population) were registered in CadÚnico.
A binary logistic regression model was estimated to explain the probability of a family receiving the benefit, as a function of 16 covariates ([3]): i) income group ($<$ R\$45; 46-89; 90-178; 179-358; $\geq$ R\$359; ii) ethnic group (other; indigeneous; quilombola); iii) type of residence (permanent residence; collective or improvised residence) iv) residence area (urban; rural); v) number of rooms in the residence (0-2; 3-4; 5-6; $\geq$ 7); vi) number of bedrooms in the residence (1; 2; 3; $\geq$ 4); vii) number of people in the household ($<$2; 3; 4; $\geq$ 5); viii) number of families in the household (1; $\geq$ 2) ; ix) total monthly expenditure of the household (in Reals); x) whether there is a piped network (yes; no); xi) whether there is a toilet in the residence (yes; no); xii) house floor material (earth; cement; ceramic/stone; wood/other); xiii) wall type (coated bricks; uncoated bricks; other); xiv) sanitary drainage

(yes; no); xv) whether there is garbage collection (yes; no); xvi) whether there is electricity (yes; no). In total 31 parameters were estimated, and 324 seconds were required to achieve convergence (i5 processor, with 16 GB of RAM memory). Some of the values obtained for the odds ratios (OR) in the population are displayed in Table 1 (column 2).

Simple random samples of 1% ($n_{1\%} = 271\,924$) and 5% ($n_{5\%} = 1\,359\,616$) were selected and the chosen population model was estimated. Results show computer burden was much reduced ([2]). The population odds ratio structure was preserved with just a 5% sample size. Deviations from the population model for a 1% sample are highlighted in Table 1: deviation of at least 2% ($\diamond$); failure to reject H_0 ($\diamond\diamond$).

Table 1: OR for the binary logistic regression model in the population (column 2) and estimated OR and $p-$values for two simple random samples of 1% (3-4) and 5% (5-6)

	(1)	(2)	(3)		(4)	(5)	(6)
	Covariate	OR	OR		p-value	OR	p-value
i)	Income Group						
	46-89	0.773	0.778		0.000	0.769	0.000
	90-178	0.313	0.323	$\diamond$	0.000	0.314	0.000
	179-358	0.071	0.074	$\diamond$	0.000	0.071	0.000
	$\geq$359 Reals	0.005	0.005	$\diamond$	0.000	0.005	0.000
vi)	Number of Bedrooms						
	2	1.037	1.021	$\diamond\diamond$	0.183	0.047	0.000
	3	0.942	0.956		0.049	0.935	0.000
	≥ 4	0.946	0.917	$\diamond\diamond$	0.065	0.930	0.001
xiii)	Wall Type						
	uncoated bricks	1.027	1.011	$\diamond\diamond$	0.475	1.020	0.006
	other	1.074	1.033	$\diamond\diamond$	0.113	1.082	0.000
	Computing time (s)	324			5		19

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grant UID/GES/00315/2019, and FAPEMIG grant APQ-02032-15.

References

[1] B.J. Fried. Distributive politics and conditional cash transfers: the case of Brazil's Bolsa Família. *World Development*, 40 (5):1042–1053, 2012.

[2] D.J. Hand. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society, Series A*, 183 (3):555–605, 2018.

[3] A.P. Kern, M.D.T. Vieira, and R.S. Freguglia. Impactos do programa bolsa família na imunização das crianças. Rio de Janeiro, 2018.

12 April, 11:20 - 11:40, Room A2

Understanding power at tax investigation - The Portuguese tax inspector's view

João Marques[1], **Ana Helena Tavares**[2],
[1] School of Criminology, Faculty of Law, University of Porto,
joaoaraujomarques@gmail.com
[2] Center for Research & Development in Mathematics and Applications, University of Aveiro

---

The tax inspection procedure depends on a set of interactions between the tax inspector and the taxpayer. Tax inspectors are responsible for conducting this relationship and they have a large set of legal prerogatives to use - power actions. The way those power actions are used may define the taxpayers propensity to regularize their tax situation or, on the contrary, adopt a position of resistance or confrontation. From a questionnaire carried out with 85 Portuguese Tax Inspectors (PTIs), 40 power actions were analyzed. We identify five different levels of power actions according to their harshness. We have also analyzed the correlations between frequency of use and perceived efficiency. We conclude that PTIs do not use more invasive power actions very often, even though these are perceived as more efficient.

Keywords: tax inspection, tax compliance, power actions, clustering

---

The tax inspection procedure is a very particular stage of the tax procedure. There are many "slippery" zones where the taxpayer's decision to comply or not comply may depend on how the tax inspector manages the conduct of the inspection procedure. Following Kirchler's Slippery Slope Framework (SFF) model [2] and Braithwaite's pyramid regulatory model [1], a specific regulatory model was developed - Tax Investigation Diamond (TID). According to the SFF, the two determining factors to mediate this relationship are Power and Trust.

Tax inspectors' comprehension of legal prerogatives and the power they have at their disposal are unknown. The implementation of a regulatory tool such as TID, based on a reasoned use of power, depends on the way power is assessed, perceived and used by the tax inspectors. This work intends to understand how the use of power actions in a tax inspection is perceived. Power actions are defined as any legal procedure that a tax inspector might use within a tax investigation (e.g., "access to taxpayers facilities" and "start a criminal investigation").

A questionnaire was carried out to PTIs (n = 85) allowing to obtain their opinion about 40 power actions. The intention is to verify if the different power actions can be clustered

according to their harshness, and if there is a relation between the frequency of use of a certain power action and the perception about its efficiency by the PTIs.
Each one of the 40 power actions is analyzed under four perspectives: Frequency of use (V1); Perceived efficiency (V2); The degree of Invasiveness (V3); and Proportionality (V4). We apply a clustering algorithm over V3 in order to identify power actions profiles in relation to their Invasiveness, and over V4 to explore the influence of the missing tax amount for the decision to use such action. Combining both clusters of V3 and clusters of V4, we are able to distinguish five levels of power actions: Extremely aggressive measures; Very aggressive measures; Aggressive measures; Slightly-aggressive measures; Non-aggressive measures. We conclude that power actions can be used in escalation, within a regulatory approach to be performed within the tax inspection procedure.
The association between V1 and V2 is analysed in different ways, depending on whether the power actions are more or less used by the PTIs. We intend to understand if the power actions perceived as being the most efficient are also those that are most used. For the power actions that are most frequently used, we analyse their correlation with the corresponding efficiency. The analysis indicates that PTIs base their action using less aggressive measures, following a more administrative orientation of the concept of tax inspection. The most aggressive measures aimed at tackling tax crimes or abusive tax planning are much less frequent used. This suggests that an increase in the use of more aggressive measures, always within the legal limits, might lead to a higher rate of detection of tax crimes and avoidance situations.

Acknowledgements This work had the collaboration of the Association of Professionals of the Tax and Customs Inspection (APIT), which proceeded to disseminate the questionnaire to the Tax Inspectors. A. Tavares was supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

References

[1] I. Ayres and J. Braithwaite. *Responsive Regulation: Transcending the Deregulation Debate*. Oxford Socio-Legal Studies. Oxford University Press, 1992.

[2] E. Kirchler and V. Braithwaite. *The Economic Psychology of Tax Behaviour*. Cambridge University Press, 2007.

12 April, 10:40 - 11:00, Room A3

Multiple-valued symbolic data clustering: a model-based approach

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

This research discusses model-based clustering of multiple-valued symbolic data. A mixture of conditionally independent Dirichlet distributions is specified to account for this type of data characteristics. Model estimation and selection is based on direct optimization and BIC, respectively. The clustering of country-based population pyramids illustrates the method.

Keywords: Multiple-valued symbolic data, Model-based clustering, Mixture models

Symbolic data analysis (SDA) has been developed as an extension to data analysis that handles 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 model-based clustering of multiple-valued symbolic data.
Consider a sample of n individuals (observations). An individual will be denoted by i ($i = 1, ..., n$) and is characterized by K variables or attributes. The kth attribute of individual i is denoted by Y_{ik} and the sample value is y_{ik}. The vector $\mathbf{Y}_i$ consists of elements Y_{ik}; $\mathbf{y}_i$ is defined similarly, and $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ denotes the sample data. The finite mixture (FM) model with S components or latent classes for $\mathbf{y}_i$ is defined by the composite density

$$f(\mathbf{y}_i; \boldsymbol{\varphi}) = \sum_{s=1}^{S} \pi_s f_s(\mathbf{y}_i; \boldsymbol{\theta}_s). \quad (1)$$

The mixture proportion, π_s, is the *a priori* probability that the data for an individual comes from component or subpopulation s, and can be interpreted as the components relative size. These mixing proportions, $\pi_1, ..., \pi_S$, satisfy $\pi_s > 0$ and $\sum_{s=1}^{S} \pi_s = 1$. Within each component (*i.e.*, conditional on belonging to component s), observation $\mathbf{y}_i$ is characterized by the density $f_s(\mathbf{y}_i; \boldsymbol{\theta}_s)$, and $\boldsymbol{\varphi}$ represents all parameters in the model. For a detailed statistical analysis of FM models, see [3]. In a recent work Brito et al. [1] developed a finite mixture of Gaussian distributions that can handle interval-value data.
For multiple-valued symbolic data, y_{ikl} with $l \in \{1, ..., L_k\}$ represents the proportions for observation i in variable k, where L_k is the number of categories of variable k. Let

$f_{sk}(\mathbf{y}_{ik};\boldsymbol{\alpha}_{sk})$ with $\mathbf{y}_{ik}=(y_{ik1},\cdots,y_{ikL_k})$ and $\boldsymbol{\alpha}_{sk}=(\alpha_{sk1},\cdots,\alpha_{skL_k})$ be the Dirichlet distribution defined by $\mathcal{D}(\mathbf{y}_{ik};\boldsymbol{\alpha}_{sk})$. By local independence of the K variables, the finite mixture model is

$$f(\mathbf{y}_i;\boldsymbol{\varphi})=\sum_{s=1}^{S}\pi_s\prod_{k=1}^{K}\mathcal{D}(\mathbf{y}_{ik};\boldsymbol{\alpha}_{sk}). \tag{2}$$

The EM algorithm has been the core algorithm for maximum likelihood estimation of mixture models. In the case of Dirichlet distributions, the maximization step cannot be computed using a close-form equation [4]. Thus, the EM algorithm needs an iterative optimization procedure in M-step to compute the estimates of the parameters conditional on posterior probabilities. In recent years the development of general procedures to maximize functions have added new alternative algorithms to the use of the EM. For instance, MacDonald [2] shows that the direct optimization of the likelihood function can be six times faster than using the EM algorithm. Moreover, direct optimization using Newton-Raphson provides standard errors of the estimates. In this research, given the EM algorithm had to include an iterative M-step, a direct optimization (Newton-Raphson) procedure was applied.

The model-based clustering model is illustrated with a demographic (population pyramids) data set that contains the population structure for males and females for 220 countries. Hence, the population structure is summarized by three variables: proportion of males and females, and age distribution of male and female populations. Results show a two-component solution. The first component contains 42.4% of the countries and presents an aged population structure. Remaining countries belong to the second component and show a young population structure. For instance, all European countries belong to the first component, whereas all African countries are in the second component.

This new model-based clustering of multiple-valued symbolic data based on Dirichlet distributions can be extended to the setting of mixture of regression and mixture-of-experts.

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

References

[1] M. P. Brito, A. P. Duarte Silva, and J. G. Dias. Probabilistic clustering of interval data. *Intelligent Data Analysis*, 19(2):293–313, 2015.

[2] I. L. MacDonald. Numerical maximisation of likelihood: A neglected alternative to EM? *International Statistical Review*, 82(2):296–308, 2014.

[3] G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[4] A. Narayanan. Algorithm AS 266: Maximum likelihood estimation of the parameters of the Dirichlet distribution. *Journal of the Royal Statistical Society C*, 40(2):365–374, 1991.

12 April, 11:00 - 11:20, Room A3

Time series clustering using forecast densities based on GAM models

Maria Almeida Silva[1,2], **Conceição Amado**[1], **Dália Loureiro**[2]
[1] CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, maria.jose.silva@tecnico.ulisboa.pt
[2] National Laboratory for Civil Engineering (LNEC), Lisbon, Portugal

---

In the last decade, a new distance based on forecast densities was proposed for time series clustering. A smoothed bootstrap procedure is used to estimate predictions assuming a generic autoregressive model. In this study, the same procedure is applied but considering a Generalized Additive Model (GAM). These two approaches were applied to two time series data sets, one unlabelled and one labelled. The results with the dynamic time warping or the Euclidean distances were better than those obtained with the forecast distance in both data sets, regardless the model used.

Keywords: clustering, forecast densities, generalized additive model, time series

---

Clustering is one of the most common data mining tools, aiming to identify patterns that contain valuable information about the data. In order to perform a clustering analysis, a distance (or dissimilarity) between the objects needs to be defined. In the case of time series, given the order and the temporal dependence between the observations, the distance should preserve this structure. The most popular are the Dynamic Time Warping (DTW) and the distances based on the estimated (partial) autocorrelations. These distances are defined based on different goals with regard to the characteristics that will be used to perform the clustering.

The definition of a distance that uses forecasts is related to the situations when it is intended to group time series according to their future behaviour. Instead of using only the point forecasts, the forecast densities can be used, allowing to incorporate the forecasts variability [1, 2]. To estimate the forecast densities, an approach based on a smoothed sieve bootstrap procedure is proposed by [2], assuming that a real value stationary process $\mathbf{X} = \{X_t : t \in \mathbb{N}\}$ can be defined as:

$$X_t = m(\mathbf{X}_{t-1}) + \epsilon_t,$$

where $\{\epsilon_t : t \in \mathbb{N}\}$ is a sequence of i.i.d. random variables, $\mathbf{X}_{t-1}$ is a d-dimensional vector with the known variables at the previous instants, and $m(\cdot)$ is a smooth function. For each pair of series, $\mathbf{X}$ and $\mathbf{Y}$, the L_p distance, $p = 1, 2$, among the forecast densities is calculated:

$$D_{\mathbf{X},\mathbf{Y}} = \int \left| f_{X_{T+h}}(x) - f_{Y_{T+h}}(x) \right|^p dx.$$

The goal of this distance is to find the dissimilarities in the forecasts at a specific time instant $T + h$. In the present work, besides assuming a generic autoregressive model, as originally proposed by [2], this distance was also applied using a Generalized Additive Model (GAM). These two approaches of the forecast distance are compared between each other, such as with DTW and the Euclidean distance.
The first application was in an unlabelled set of flow time series. Since the true label of each time series is not known, a labelled set of heartbeat (ECG) time series was also used. Knowing the true class label, the accuracy can be computed, allowing a better evaluation of the distance. In both cases, the GAM model needs to be defined. The previous time instants that should be included were selected analysing the estimated partial autocorrelations. In the case of flow time series, the GAM obtained for a time series $(y_1, y_2, \ldots, y_{365})$ is as follows:
$E[Y_t|y_{t-1}, y_{t-2}, y_{t-7}, D_t, W_t, M_t] = s_0+s_1(y_{t-1})+s_2(y_{t-2})+s_3(y_{t-7})+s_4(D_t)+s_5(W_t)+s_6(M_t)$,
where $t \geq 8$; D_t is the day of the week of the time instant t, coded from 1 (Sunday) to 7 (Saturday); W_t is the week of the year, with values between 1 and 53; M_t is the month, with values between 1 and 12; s_0 is an unknown constant and $s_i(\cdot)$, $i \in \{1, \ldots, 6\}$, are smooth functions to estimate. The variables D, W and M were included due to the flow time series seasonality. For ECG data, the GAM model only includes the time instants $t - 1$, $t - 2$ and $t - 3$ to forecast the time series at time instant t. To apply the distance based on the forecast densities, four possibilities for the future time instants to forecast were studied (1, 3, 8 and 15) for flow time series, and two possibilities (1 and 3) for ECG series. For flow time series data set the partitions obtained by DTW and forecast distances are not similar. In the labelled data set case, the accuracy was computed, as the computational time needed to run each analysis. DTW and the Euclidean distances were computationally faster and also more accurate than the forecast distance. Although the accuracy was not very high (DTW: 56.4%; Euclidean: 61.8%), it was higher than those obtained with the forecast distance, regardless the model used and the future time instant.
These results allow concluding that the distance based on forecast densities can be useful in some situations [2], but there are also some, like those presented in this study, where this distance is not advantageous.

Acknowledgements This work was partially supported by the Portuguese FCT - Fundação para a Ciência e a Tecnologia, through the project UID/Multi/04621/2019 of CEMAT/IST-ID, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, University of Lisbon.

References

[1] A.M. Alonso, J.R. Berrendero, A. Hernández, and A. Justel. Time series clustering based on forecast densities. *Computational Statistics & Data Analysis*, 51(2):762–776, 2006.

[2] J. A. Vilar, A. M. Alonso, and J. M. Vilar. Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis*, 54(11):2850–2865, 2010.

12 April, 11:20 - 11:40, Room A3

Clustering interval time series

Elizabeth Ann Maharaj[1], **Paulo Teles**[2], **Paula Brito**[2]
[1] Monash University, Melbourne, Australia
[2] Faculty of Economics, University of Porto & LIAAD INESC TEC, Portugal, mpbrito@fep.up.pt

We address the problem of clustering interval time series (ITS), exploring different approaches. ITS may be clustered based on point-to-point comparisons, or else using time domain or wavelet features. Autocorrelation matrix functions, which gather the autocorrelation and cross-correlation functions of the ITS upper and lower bounds, as well as a new autocorrelation function of ITS, are compared by adequate distances and employed for clustering. The different approaches are compared for ITS simulated under different set-ups, and illustrated with an application to sea level daily ranges, observed at different locations in Australia.

Keywords: Interval Autocorrelation, Interval Data, Interval Time Series, Time Series Clustering

The clustering of time series has received considerable attention in recent years, given its importance in many domains such as Medicine, Ecology or Finance. Different approaches have been proposed, viz., in the time and spectral domains and using wavelets. A recent survey may be found in [1]. When an interval rather than a single value is recorded at each point in time, we have an Interval Time Series (ITS) which arises, e.g., when we record minimum and maximum temperature values along time, or the daily range of sea levels in different locations, or low and high values of asset prices in consecutive sessions. In this paper, we explore a number of methods for the clustering of ITS. All the methods used are appropriate for discrete time series observed at equally spaced time intervals recorded at the same periods and therefore synchronous and with equal length. Furthermore, they can exhibit variability both in levels and after appropriate differencing if required.

The first method uses distance measures based on point-to-point comparisons for every pair of ITS under consideration, and averaged over the time record. A distance matrix with these measures is used as an input to hierarchical and dynamical (non-hierarchical) clustering. This approach consists of comparing the ITS on the basis of the actual recorded minimum and maximum values at each time point and therefore not directly using their autocorrelation structure.

Another method involves using time domain features of the radius and centre series as clustering variables, following [2], and wavelet features of the radius and centre series as

variables for clustering, following [3]. The use of the time domain features provides an insight into the similarity of the ITS based on the first four moments of the radius and centre series, whereas the wavelet features provides an insight of the similarities of the dynamics of the radius and centre series at different frequency levels.

A further new method involves fitting space-time models to each of the ITS under consideration and using the parameter estimates of the fitted models as clustering variables (see [4]). Such estimates are expected to reflect the joint distribution or dependency between the two interval bounds and therefore can be used as clustering variables. However, the success or failure of this approach relies on the estimation accuracy of the model parameters which can be a serious drawback.

An ITS may be regarded as a bivariate time series, considering its upper and lower bounds. A further approach is then based on the autocorrelation structure of such bivariate time series as a clustering tool and requires using matrix distance measures, such as the Frobenius distance.
Distance measures based on autocorrelation functions are commonly used for time series clustering and therefore considering interval autocorrelations is a natural extension. We propose an improved approach to determine the autocorrelation function of an ITS based on its upper and lower bounds; clustering may then also be performed using distances based on the interval autocorrelation measures.

An extensive simulation study allows assessing the performance of the alternative approaches. An application to sea-level interval time series recorded at different locations in Australia illustrates the proposed methods.

Acknowledgements The work of P. Teles and P. Brito is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project "POCI-01-0145-FEDER-006961", and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

References

[1] J. Caiado, E.A. Maharaj, and P. D'Urso. Time series clustering. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci, editors, *Handbook of Cluster Analysis*. Chapman and Hall, 2015.

[2] P. D'Urso and E.A. Maharaj. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160:3565–3589, 2009.

[3] P. D'Urso and E.A. Maharaj. Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems*, 193:33–61, 2012.

[4] E.A. Maharaj, P. Teles, and P. Brito. Clustering of interval time series. *Statistics and Computing*, (in press), 2019. http://doi:10.1007/s11222-018-09851-z.

12 April, 11:40 - 12:00, Room A3

Discriminant factors of website trust

Ana A. Andrade[1], **Margarida G. M. S. Cardoso**[2], **Vítor V. Lopes**[3],
[1] Research, Studies and Renewables Division / Directorate-General for Energy and Geology (DGEG), Ana.Andrade@dgeg.pt
[2] Business Research Unit (BRU-IUL) / Instituto Universitário de Lisboa (ISCTE-IUL)
[3] CMAF-CIO / Faculdade de Ciências, Universidade de Lisboa

E-commerce sellers aim to better understand trust formation mechanisms and assess the level of initial trust created by their sites on potential clients. In this work, a survey is conducted to assess websites characteristics as viewed by the respondents and reveal their perceptions. Based on the survey data, we resort to rules classifiers to provide new insights on website trust and suggest specific recommendations for e-commerce vendors. Decision trees and rough sets are used to this end. In addition, we propose a heuristic aiming to derive simpler classifiers, taking into account their predictive ability and the parsimony of rules' sets.

Keywords: trust, e-commerce, rough sets, decision trees

E-commerce generally enhances access to a high number of potential clients and consumers. Lack of direct contact may, however, prevent potential consumers of accessing the usual physical nature cues of trustworthiness - e.g vendor's body language or store appearance - making on-line trust a major issue. Providing site/vendor's ratings may help the potential customer to develop trust, even without a previous history of interaction. However, since these measures don't take into account consumers' first reactions to sites, they have limited relevance for e-commerce sellers.

This work proposes building sets of logical rules that e-commerce sellers can use to assess the level of initial trust created by their sites on potential clients. Ultimately, these rules may be used as guidance to improve website design.

The data was collected through an on-line survey. One out of six real e-commerce sites was randomly attributed to each respondent, none being from a well-known vendor. The first questions refer to site objective characteristics related to appearance, design, functionality and information. The second part of the survey refers to perceptions regarding main trust constructs such as appearance, reputation, fulfillment and security - [4].

Initially, survey data is used to constitute several sets of attributes considering their level of measurement and the degree of relationship with levels of revealed website trust (target classes) – e.g. Mutual Information is used. Recodification is also essential to deal with dimensionality issues.

In order to generate sets of logical rules, expressed in the form of "if... , then..." propositions, we resort to rough sets and decision trees. Ordinal nature in data is acknowledged.

So, besides using a rough sets algorithm for nominal data, corresponding to a Classical Rough Sets Approach (CRSA), implemented through Rosetta software, [2], an algorithm for ordinal data corresponding to a Dominance-based Rough Sets Approach (DRSA) is also used and implemented through jMAF software - [1]. For the induction of propositional rules based on decision trees, the See5/C5.0 algorithm is adopted – [3].
Classifiers predictive ability and the final number of rules are used to evaluate the performance of rules sets. Finally, the classifier generating the best results is selected to perform a wrapper procedure, SSA-Successive Selection of Attributes, based on a repeated cross-validation procedure.
Decision Trees results outperform Rough sets results, advantages in predictive ability of some DRSA models being offset by an excessive number of rules.
The higher relative performance obtained for models based on consumer opinions around trust constructs, when compared to those based on objective characteristics, suggests they are a valid proposal. It is however worthwhile to note that they are also a less practical one, since data referred to perceptions is more difficult to collect.
The logical rules obtained confirm the relevance of signs and perceptions related to security, and point to a number of other factors characterizing the formation of initial trust. It is generally found that negative-valent perceptions are more useful in predicting trust, than positive perceptions.
A limitation of this work regards the reduced sample size. If, on the methodological perspective, this fact potentiated new developments, namely SSA, increasing sample dimension should allow creating better models.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grant UID/GES/00315/2019.

References

[1] Jerzy Błaszczyński, Salvatore Greco, and Roman Słowiński. Multi-criteria classification–a new scheme for application of dominance-based decision rules. *European Journal of Operational Research*, 181(3):1030–1044, 2007.

[2] Aleksander Øhrn. *Discernibility and rough sets in medicine: tools and applications*. PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, 2000.

[3] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[4] Shumaila Y Yousafzai, John G Pallister, and Gordon R Foxall. Strategies for building and communicating trust in electronic banking: A field experiment. *Psychology & Marketing*, 22(2):181–201, 2005.

13 April, 9:30 - 9:50, Room A2

Pilgrimage and mobile use

Ângela Antunes[1], **Carla Henriques**[2], **Suzanne Amaro**[3]
[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu
[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseun, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS),carlahenriq@estv.ipv.pt
[3] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS)

It is undeniable that smartphones have become part of our daily lives. They are, indeed, useful to perform daily tasks, but they can be even more useful in travelling contexts and, in particular, for pilgrimages. This study focuses on the use of mobile technology by pilgrims of Santiago, in an effort to better understand modern pilgrims and to characterize them regarding their use of mobile technology. A cluster analysis was carried out and four different segments of pilgrims were identified, which were further profiled using other variables that were not used in the cluster analysis.

Keywords: cluster analysis, factor analysis, Santiago's pilgrims

---

Data for this study include 1140 responses to an online survey conducted to pilgrims of Santiago's way in August and September of 2015. Questions regarding the use of mobile devices during pilgrimage, expectations regarding an eventual app for pilgrims, as well as demographic and other characteristics, were included in the questionnaire. Exploratory factor analysis, with Varimax rotation, was applied three times, in order to reduce the number of items to analyze regarding the activities carried out with a mobile device, the features valued in an app regarding Santiago's way and the attitude towards the use of technologies. Cluster analysis was applied to the three factors obtained for the attitude towards the use of technologies (Innovativeness trait, Addiction behaviour and Regular use). Four differentiated segments of pilgrims were found concerning the attitude towards the use of technologies (Figure 1):

- Very Regular Users (VRU) – Pilgrims who use regularly mobile apps but do not consider themselves addicted nor show much interest in technological novelties;
- Addicted (Ad) – Pilgrims who consider themselves addicted to applications, hence use them regularly, and are attracted to new information technologies;
- Apathetic (Ap) – Pilgrims who do not use or even care about mobile apps;

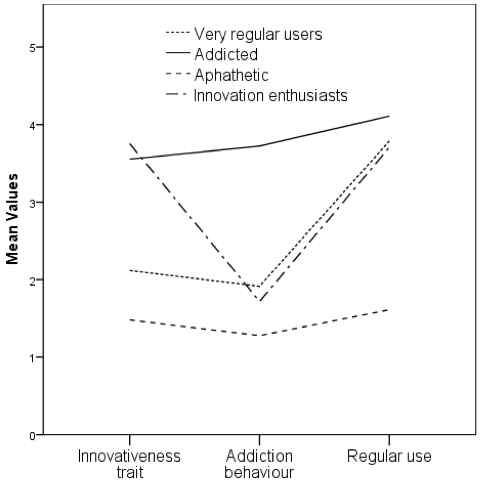


Figure 1: Cluster centroids

- Innovation Enthusiasts (IE) – Pilgrims who are very interested in new information technologies and regularly use mobile apps but do not consider themselves addicted.

Cluster analysis involved the application of two hierarchical methodologies (Average Linkage and Ward's Method) whose solutions were submitted to the k-means method. The four-group solution figured out to be an optimal solution both in terms of interpretation and stability. In fact, the four groups were clearly distinguishable and had a meaningful profile. The stability was assessed by comparing the two solutions derived from applying K-means to the four-group solutions of Average Linkage and Ward's method. Only 1.9% of the observations were assigned to different groups, which supports stability of a four-cluster solution. Additionally, a procedure suggested by Dolnicar and Leisch [1] was also conducted and the results supported the four-group solution. This procedure analyses the similarity of clusters solutions for k clusters when applying k-means to 100 bootstrap samples. The similarity between cluster solutions was assessed through the Rand index.
The study found differences between the groups regarding the main use of smartphone, the importance given to characteristics of a pilgrim app, age, education and the willingness to pay for a pilgrim app. The findings also show that the Very Regular Users are very similar to Innovation Enthusiasts and both reveal interest in acquiring and using a pilgrim app.
The results of this study will be useful to technology researchers and stakeholders, as they provide a better understanding of modern pilgrims regarding their use of mobile technology.

References

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

13 April, 9:50 - 10:10, Room A2

How social networks influence similarity between examination answers - longitudinal study

Milton Severo[1], **João Borges**[1], **Fernanda Silva-Pereira**[1]
[1] Department of Public Health and Forensic Sciences, and Medical Education, University of Porto Medical School, milton@med.up.pt

The objective of this study is to compare longitudinally the similarity of the students' responses belonging to the cohort from 2013 to 2019 of the medical course of the FMUP, as well as to try to understand the influence of the social network on similarity. The indicators of the social work network of each student were associated with the final classifications and with the answers similarity. In conclusion, the study of social networks will optimize classifications and improve students' academic integrity.

Keywords: Social Networks, Academic integrity

Nowadays, social relations are more important than ever. They influence the way we feel, the way we work, our success and failure. Working as a team allows you to gain more knowledge and make better decisions. In this way, the main argument for the existence of similar answers among students in examinations is to have studied together. Despite this a previous cross-sectional study showed that the prevalence of responses similarity increased over the academic course and pointed as the main argument the copy during the examination[1]. The objective of this study is to compare longitudinally the similarity of the students' responses belonging to the cohort from 2013 to 2019 of the medical course of the FMUP, as well as to try to understand the influence of the social network on similarity. To evaluate the responses similarity among students, a total of 124 multiple-choice examinations (regular and remedy) were used during the first 5 years of the medical course. The Angoff A-index was used to classify student pairs with similar responses. Whereas R_i represents the number of correct answers by the student i, R_j the number of correct answers by the student j and R_{ij} the number of correct answers shared by both. The probability of agreement between student i and j is determined by calculating the residue of the regression of R_{ij} on the $\sqrt{R_i \times R_j}$. Considering the multi-comparisons we applied the correction of Sidák. Figure 1 shows an example of detection in an examination. A questionnaire was constructed in which each student had to identify within a list the students with whom he usually studies. The list consisted of 250 students from the 2013 to 2019 student cohort who performed at least 10 of the tests within the 124. A total of 127 (51%) responses from the 250 students. We obtained the social network represented in Figure 2. The following indicators were used to characterize the position of each individual in the social network: (1) *In-degree* which indicates the number of students who reported that

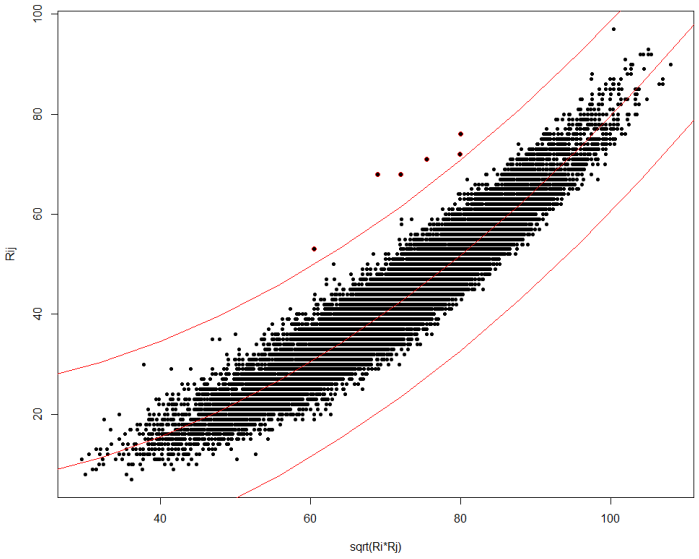


Figure 1: The number of common responses observed for each pair of students in an examination (R_{ij}) versus $\sqrt{R_i \times R_j}$ for one of the examinations.

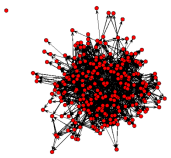


Figure 2: Social network of the study

they studied with that particular student (represents the number of vertices that point to the vertex); (2) *Closeness score* which indicates the degree of proximity of a student to the other students of the social network defined bt the inverse of average length of the shortest paths to/from all other vertices in the graph; (3) *Betweenness centrality* which indicates the likelihood of a student being part of the more direct path between two students. It was found that the indicators difference in absolute *In-degree* and *Closeness score* for each pair were associated negatively and positively, respectively, with the responses similarity in the examinations. In addition, it was found that the indicators *In-degree* and *Closeness score* were positively associated with students' final classifications.
In conclusion, the study of social networks will optimize classifications and improve students' academic integrity by increasing the cohesion between all students.

References

[1] Jorge Monteiro, Fernanda Silva-Pereira, and Milton Severo. Investigating the existence of social networks in cheating behaviors in medical students. *BMC medical education*, 18(1):193, 2018.

13 April,10:10 - 10:30, Room A2

Prices in the electricity Iberian market– a clustering approach

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

---

Aiming to better understand daily patterns of price in the Iberian Electricity Market (MIBEL) we resort to a clustering analysis. The data regards the hourly prices of electricity, in €/MWh, observed in the day-ahead MIBEL market between 2016 and 2018. We propose clustering daily time series to obtain groups of similar days according to the variation of the price throughout the day. Alternative distance measures between time series are considered: Euclidean, Pearson correlation and Periodogram based measures. K-Medoids algorithm is used to form the clusters. The clustering solutions are selected by resorting to several cohesion-separation measures. Finally, clusters are profiled.

Keywords: Clustering, Electricity markets, Time series

The Iberian Electricity Market (MIBEL) was formed in July 2007, by the integration and cooperation between the Portuguese and Spanish electricity markets. In day-ahead electricity markets, price and quantity exhibit daily fluctuations. Due to the electricity characteristics, its prices are extremely volatile and influenced by many variables such as demand, fuel prices, hydro and other renewable production, market agents behavior and CO2 emission prices. Understanding the complex electricity price behavior is a major concern for investment decisions and power plants management.

In this work we propose clustering daily time series to obtain groups of similar days according to the hourly electricity price. The characterization of daily variations of the price can provide an useful information to improve the price forecasting. The data analyzed regard the hourly prices of electricity (in €/MWh) for Portugal, observed in the day-ahead MIBEL market between 2016 and 2018 and obtained from the Iberian Market Operator – Spanish pole (OMIE). The input data matrix has in each column the daily information regarding the hourly prices, with a total of 1096 days (columns) and 24 rows. Alternative distance measures between time series are used: Euclidean, Pearson correlation and Periodogram based measures. K-Medoids algorithm [3] - implemented in R package "cluster" - is used for clustering. It generalizes K-means using arbitrary-defined distance measures; it aims at the minimization of the distance of objects belonging to a cluster from the cluster's medoid; it is somewhat more flexible in terms of cluster shapes and more robust to outliers

and noise. In what concerns time-series clustering, the fact that a medoid (a member of the data set) is considered overcomes the need to define a centroid, which can be a problematic issue [4]. To find the "best" clustering solutions we resort to several cohesion-separation measures such as Average Silhouette, Calinski and Harabasz or Dunn modified index [1] (implemented in the "fpc" R package [2]). The use of alternative dissimilarity measures results in different data partitions but with similar number of clusters. Finally, groups are characterized considering variables that influence the electricity prices, namely demand and prices of commodities such as natural gas, coal, oil, and CO2 emissions.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grant UID/GES/00315/2019.

References

[1] James C Bezdek and Nikhil R Pal. Some new indices of cluster validity. 1998.

[2] Christian Hennig. Package 'fpc': Flexible Procedures for Clustering. *URL: http://cran. r-project. org/web/packages/fpc/fpc. pdf (available 08.07. 2017)*, 2018.

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

[4] Pablo Montero, José A Vilar, et al. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014.

13 April, 9:50 - 10:10, Room A3

Modelling a predator-prey interaction: an in-class exercise

Inês Bento[1] **Joana Araújo**[1] **Joana Pereira**[1] **Margarida Marques**[1] **Matilde Almodovar**[1] **Morgan Ribeiro**[1] **Pedro Afonso**[1] **Rita Pereira**[1] **Tiago Marques**[1,2,3]

[1] Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, inesbento23@gmail.com
[2] Centro de Estatística e Aplicações, Faculdade de Ciências da Universidade de Lisboa
[3] Centre for Research into Ecological and Environmental Modelling, University of St Andrews

---

Ecological modelling can be challenging for many researchers, especially for students without a strong statistical background. To bridge the gap between these two fields, the MSc students from the Ecological Modelling course from the Faculty of Sciences of the University of Lisbon took part in an experiment that intended to model a typical predator-prey situation. In this study, we focus on three key aspects: (1) ascertaining variables that influence predation success; (2) searching for a possible predator learning process throughout the experiment and (3) assessing if individual heterogeneity affects predation success. These types of exercises could be used as tools to better understand a complex natural phenomenon, helping students from different areas of expertise to embrace statistical models in their work.

Keywords: predator-prey, predator efficiency, ecological modelling, education

---

Ecological modelling can be challenging for many scientists, especially for students without a strong statistical background. To bridge the gap between these two fields, the MSc students from the Ecological Modelling course from the Faculty of Sciences of the University of Lisbon took part in an experiment that intended to model a typical predator-prey situation. These students played the role of predators, while prey was represented by animal shaped pasta (each measuring approximately 1 cm2) displayed on separate tables of the same size. In total there were 34 predators (each assigned a number), divided into groups of 3 to 5. Two main methods were used: (1) Capture Attempts and (2) Capture time, as a function of different abundances of prey (N=15,30,50,60,70,80,100,140,160). The first method assessed how many attempts (C1, C2, C3) were needed for predators to capture a total of three preys, while blindfolded and using only their fingertips. In the second method, predators moved a single finger along each table until they touched 3 preys, timing each capture cumulatively (T1, T2, T3). Each predator performed both methods on all prey groups, i.e. on all tables.

Explanatory variables were defined as follows: prey abundance, the order in which the different tables of prey were preyed upon, predator number, predator size (represented by student's height and hand size) and predator's eye colour. Response variables (capture attempts and capture time) were modeled as a function of the previous explanatory variables. Modelling predation phenomena that occur in nature implies simplifying complex relationships. As such, in this study we focus on three key aspects: (1) ascertaining variables that influence predation success; (2) searching for a possible predator learning process throughout the experiment; (3) assessing if individual heterogeneity affects predation success, and if so how.

Firstly, we expect that out of all tested variables only predator eye colour does not have a significant effect on the response variables. Secondly, we predict a decline in capture attempts and capture time according to the order in which the different tables of prey were preyed upon, which may suggest a learning curve during the procedure. Finally, and often ignored in modelling exercises, here we anticipate that individual heterogeneity between predators is an underlying factor.

In conclusion, despite not being a primary field for many scientists, statistical modelling does not need to be considered rocket science or something to shy away from. These types of exercises should be used as tools to better understand complex natural phenomena, as is the case of predator-prey interactions. These relationships, whether more obvious or cryptic, can be better understood by students and researchers alike using easy approaches to teach ecological models.

13 April, 10:10 - 10:30, Room A3

Higher education students in Viseu Polytechnic - an evolutive study since the Bologna Treaty

Joana Fialho[1,2], **Madalena Malva**[1], **Paula Sarabando**[1,3], **Paulo Costeira**[1]
[1] Institute Polytechnique of Viseu, malva@estv.ipv.pt
[2] CI&DETS
[3] INESC Coimbra

This work intends to characterize the students who enrolled in the Institute Polytechnic of Viseu (IPV) since the beginning of Bologna Treaty, that is, since the school year 2006/2007. This study helps to perceive and characterize IPV students in general, and to characterize the students of each organic unit, in particular.

Keywords: data analysis, descriptive statistics, statistic inference, higher education students

---

There are divers studies that characterize the portuguese higher education, produced by portuguese agency for the evaluation and accreditation of higher education. These works gave the idea of analyzing the IPV in particular, namely the IPV students. This anlysis allows to perceive the origin of the students, as well as their course preferences. IPV can use this information, on the one hand, to adapt its formative offer, on the other hand, to understand which localities are more important to promote IPV in order to attract new students. Furthermore, with the information collected, it is possible to relate different aspects and realize if those relations have some significance.

In order to perform this work, it was necessary to collect data of IPV students. In this sense, for all new students, from the 2006/2007 school year to the 2018/2019 school year, several aspects were collected, such as the nationality, city of born, age, gender, type of course chosen, among others.

For the data analysis, it was used, firstly, descriptive statistics: if the variables in study were qualitative, their description is made using absolute, relative and relative cumulative frequencies; if the variables were quantitative, their description include mean, maximum, minimum and standard deviation. It was important to analyze if there were significant relationships between the variables under study. For that, statistical tests were used, considering, in all, a level of significance of 5%. To relate qualitative variables, two by

two, the Chi-Square test was used, which rejects independence of the variables and, therefore, the relation between them can be assumed, if the p-value associated with the test is lower than the level of significance. In these cases, the intensity of the relationship was quantified through the contingency coefficient. In the analysis of qualitative variables, a cross-sectional table is presented. Age is a quantitative variable and, in order to relate age with qualitative variables, the t-test was used, if the qualitative variable had two attributes, or the ANOVA test, if the qualitative variable had three or more attributes. In both cases, it is considered that age differs between groups of the qualitative variable, if the p-value associated with the test is lower than the level of significance In some cases, age was divided into classes, reason why was considered as a qualitative variable.

References

[1] Bases de dados da direcção-geral de estatísticas de educação e ciência (dgeec/gpeari) do ministério da educação e da ciência.

[2] Bases de dados do acesso ao ensino superior público da direcção geral do ensino superior.

[3] M. Fonseca and S. Encarnação. *O Sistema de Ensino Superior em Portugal - Perfis Institucionais: Os Institutos Politécnicos Públicos*. Agência de Avaliação e Acreditação do Ensino Superior, Lisboa, 2012.

[4] J. Maroco. *Análise Estatística com o SPSS Statistics*. Lisboa: Report Number, Lisboa, 2014.

13 April, 10:30 - 10:50, Room A3

Clinical characteristics of patients with chronic obstructive pulmonary disease (COPD): are they different?

Vera Enes[1], **Ana Helena Tavares**[2], **Vera Afreixo**[2], **Filipa Machado**[3], **Alda Marques**[1,3]

[1] Institute of Biomedicine (iBiMED), University of Aveiro, vera.enes@ua.pt

[2] Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro

[3] Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health Sciences (ESSUA), University of Aveiro

---

Chronic Obstructive Pulmonary Disease is a major public health problem known to affect 800.000 people in Portugal. Symptoms include breathing difficulty, cough, fatigue and sputum. Although known that the disease progresses differently in patients with the same level of airway obstruction, the clinical characteristics of patients that may be associated with different disease phenotypes are not fully understood. This study aims to enhance our knowledge on the clinical characteristics of patients with COPD. A clustering procedure was performed, based on lung function, oxygen saturation, muscle strength and impact of the disease on patients' daily life and well-being.

Keywords: COPD, Clustering, Principal Component Analysis

Chronic Obstructive Pulmonary Disease (COPD) is a condition characterized by progressive and persistent airflow limitation resulting from a chronic inflammatory response of the airways and lungs in response to inhaled harmful gases and particles. Clinical diagnosis is based on airflow obstruction (assessed with lung function test-spirometry) and symptoms. Its prognosis depends on several factors including acute exacerbations (define as worsening of symptoms that result in additional therapy), environmental exposures, comorbidities and genetic predisposition [2]. COPD is burdensome not only for economic and social systems but most importantly to patients since it significantly affect their quality of life. It is known that the disease does not progress in the same way in all patients and that lung function, symptoms and reduction of quality of life may not be correlated. In fact, the interplay between patients' clinical characteristics and different disease phenotypes is not fully understood.

This study aims to enhance our knowledge on the clinical characteristics of patients with COPD. We retrospectively reviewed 394 patients with COPD. A clustering procedure is designed to stratify patients with COPD. From the 70 registered variables, we focus

on the most commonly assessed clinical variables: body mass index (BMI), age (AGE), the modified British Medical Research Council questionnaire (mMRC), number of acute exacerbations (AECOPD), number of hospitalization by respiratory cause (nHosp), the Charlson comorbidity index (CCI), Peripheral oxygen saturation (SpO2), forced expiratory volume in one second (FEV1pp), quadriceps muscle strength (QMSpp), 1-minute sit-to-stand test (1STS), COPD assessment test total score (CAT), St. George's Respiratory Questionnaire (SGRQ), dyspnoea and fatigue Borg scores (dysp.Borg and fat.Borg), and Hospital Anxiety and Depression Scale (anx.HADS and dep.HADS). Other variables were excluded due to missing values.

Clustering aims to find groups in a dataset. Since k-means looks for spherical clusters, it works best when the input variables are uncorrelated and have similar scales. In our dataset several variables were strongly correlated, e.g. the Pearson correlation between CAT and SGRQ was 0.79. We apply Principal Component (PC) Analysis on these vectors, obtaining a set of values of linearly uncorrelated variables. The number of components to retain is selected such that at least a given percentage of the variance is explained. The scores associated to those first PCs yield a data matrix, on which the k-means clustering algorithm is applied. The result of k-means depends on the number of clusters k, which is often hard to choose a priori. Therefore it is common practice to run the method for several values of k, and then select the 'best' value of k as the one which optimizes a certain criterion called a validity index. Many such indices have been proposed in the literature. Here we consider the Calinski-Harabasz index [1] and the GAP statistic [3].

Our procedure retained 6 principal components that explained 70% of the total variance of the dataset. Carrying out k-means clustering for different numbers of clusters yields and evaluating the obtained validation indices it appears that 3 or 5 clusters are appropriate. By looking at the composition of each cluster we define a patient prototype of each cluster.

Acknowledgements This work was funded by Programa Operacional de Competitividade e Internacionalização – POCI, through Fundo Europeu de Desenvolvimento Regional - FEDER (POCI-01-0145-FEDER-007628 and POCI-01-0145-FEDER-028806), Fundação para a Ciência e Tecnologia (PTDC/DTP-PIC/2284/2014 and PTDC/SAU-SER/28806/2017). Moreover, the costs resulting from the FCT hirings is funded by national funds (OE), through FCT, I.P., in the scope of the framework contract foreseen in the numbers 4, 5 and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19. The work of VA and AT is partially funded by FCT under project UID/MAT/04106/2019.

References

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

[2] Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2019 report), 2019.

[3] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Poster Session

13 April, 13:30 - 14:10, Auditorium atrium

Statistical modeling: a study on customer retention in health & fitness industry

A. Manuela Gonçalves[1], **Guadalupe Costa**[2], **Alexandre Freitas**[3]
[1] CMAT-Center of Mathematics, DMA-Department of Mathematics and Applications, University of Minho, Portugal, mneves@math.uminho.pt
[2] DMA-Department of Mathematics and Applications, University of Minho, Portugal
[3] University of Porto, Faculty of Economy, Portugal

This study is conducted within the context of a Portuguese Health and Fitness company and its main proposal is to identify the factors that influenced customers' behaviour by analysing customer retention and customer lifecycle, which are the most important key performance indicators (KPI) in this industry. Thus, it is developed statistical models in the areas of generalized linear models and survival analysis to predict and forecast customer retention (continuing (active) or non-continuing (dropout) customer). The data were collected on February 1, 2018.

Keywords: Health & Fitness Industry, Customer Retention, Sampling, Generalized Linear Model, Survival Analysis

---

Given the increasing competitiveness of the health fitness industry, implementing strategies and tactics to prevent loss of customers is highly important, particularly because the acquisition of new customers entails high costs for companies. Their retention and loyalty may be vital in the medium- to long-term financial health of a company. It is important for marketing planners to develop initial strategies to attract customers and engage them to patronize the products in the long run. It is up to the companies and their professionals to be constantly on alert in order to anticipate what customers expect from their services, and thus be able to provide a service that meets their expectations and needs, hence generating client satisfaction and loyalty [2].
This Portuguese Health & Fitness company is one of the largest and most prestigious fitness chains in Portugal, with an average number of active members in excess of 48 thousand. Its activity began in 1995, having expanded since 1997 with the opening of other clubs. In 2018 there were 20 clubs scattered throughout the country. For each member was collected information regarding their profile and behaviour throughout their membership (several variables regarding customer retention and customer lifecycle). The most important data were: Number of accessions (per month); Number of contract cancelations (per month); Months (duration of the contract in months); Schedule (type of schedule that the member attends (Limited or Total)); Age (member's age); Gender (member's gender); Number of Visits (number of customer visits during the contract); PT (the customer is accompanied

by a personal trainer (Yes or No)); Number of Group Classes; and Tax Value (monthly fee charged (in Euros)). In this work, it is considered the Health & Fitness Company (Global) and, in particular, 10 Health & Fitness clubs. The database used includes registers of members who have joined the clubs between January 1, 2013 and December 31, 2015. In this period, the company had successfully recruited 62183 new members. Of these 62183 members, 9725 remain active in the clubs as of February 1, 2018, and 52458 are dropout. We may be able to obtain more precise estimates of population quantities by taking a stratified random sample (proportional allocation). We draw an independent probability sample from each stratum (two strata: active and dropout customers) in the Global sample (total) and in each of the 10 clubs, generating random samples. For the Linear Regression models, we focus on the relationship between the dependent variable (Months, a continuous variable measured in months) and the independent variables or predictors, and in separated groups (samples) active and dropout customers. Logistic Regression is used to explain the relationship between the dependent binary variable (active and dropout customers) and one or more independent variables [4]. To establish a Survival Analysis in this study, the survival time is the time until the cancellation of the contract by the customer (right-censoring) [3]. For the estimation of the survival function, we used the Kaplan-Meier estimator. Also, we established Cox Regression models in order to investigate the effect of several variables on the time a specified event takes to happen – contract cancellation [1]. All models were established for the Global sample, and for each of the 10 clubs.
With the present study it was possible to create a knowledge base for this industry on the determinants that predict client retention and loyalty and contributes to the adoption of management strategies for client retention, in order to prevent customer cancellations by improving customer touch points, and learning about customer retention as a team effort, but especially as a fitness team effort.

Acknowledgements The research of A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/MAT/00013/2013.

References

[1] D.R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[2] P. Jain and S.S. Singh. Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing*, 16(2):34–46, 2002.

[3] S. Lemeshow, D. Hosmer Jr. and S. May. *Applied Survival Analysis: Regression Modelling of Time Event Data.* Wiley-Interscience, New Jersey, 2008.

[4] S.N. Wood. *Generalized Additive Models: An Introductin with R.* Chapman & Hall/CRC Texts in Statistical Science, New York, 2017.

13 April, 13:30 - 14:10, Auditorium atrium

Application of principal components analysis to life cycle analysis for environmental assessment in production systems in Mexico - case studies of maize and porcine production

Miriam Paulino Flores[1], Maria del Rosario Villavicencio[1], Angel Roberto Martínez Campos[2], Francisco Ernesto Martínez Castañeda[2], Ana Lorga da Silva[3]

[1] CPES – Universidade Lusófona de Humanidades e Tecnologias de Lisboa, Portugal and Livestock and Agricultural Sciences Institute, Autonomous University of the State of Mexico, Mexico

[2] Livestock and Agricultural Sciences Institute, Autonomous University of the State of Mexico, Mexico

[3] CPES, ECEO, and FCSEA – Universidade Lusófona de Humanidades e Tecnologias de Lisboa, Portugal, ana.lorga@ulusofona.pt

---

Maize production is Mexico's most representative activity because of its economic, social and cultural importance; part of the national territory is suitable for the production of maize, also the porcine Mexican sector has an important participation worldwide, dynamics of growth has positioned it inside the principal producing countries, in such a way that, not only competes in satisfying the needs of the market, also in the creation of social value. In this work we intend to analyze the results obtained from environmental evaluation of production systems in Mexico, using multivariate data analysis, in particular Principal Component Analysis.

Keywords: Environmental impacts, Maize production, Porcine Production, LCA, PCA.

The production of food, like any activity, has implications for environmental quality, depletion of resources, soil degradation, emissions to the atmosphere, contamination of bodies of water, generation of waste, are some effects associated with this productive activity. This research work conducted an assessment of potential environmental impacts associated with maize production and porcine production through a Life Cycle Analysis (LCA).

The international standard ISO 14040: 2006 [1],[2], defines the LCA as "a technique to determine the environmental aspects and the potential impacts associated with a product: compiling an inventory of the relevant inputs and outputs of the system; evaluating the potential impacts associated with these inputs and outputs, and interpreting the results of the inventory and impact phases.

The objective of this study is to analyze the results of the environmental evaluation of maize and porcine production systems, by applying multivariate data analysis applied to the LCA results. Principal Component Analysis (PCA) is a method that allows the reduction of the number of variables to a small set of independent variables (main components), being a linear combination of the original ones, which represent most of the information of the original variables [3]. The analysis of data considering PCA, allows to explain in a more concise and clear way the results of the LCA, establishing a complementary data analysis base for the application to other research works focused on environmental assessment.
In this study, both the production of maize and pig production are analyzed two levels of environmental assessment designated by Midpoint and Endpoit; both levels are composed of three environmental aspects: Ecosystem, Human health and Resources, for each of which we obtain components that are described and analyzed.
At each level, several scenarios were evaluated; six in maize production and nine in pig production.The results of each scenario were compared in each type of production.
Finally, for each one of the levels of environmental evaluation, the scenarios were gathered, obtaining for each of them a database, which allowed to carry out a global analysis of midpoint and endpoint in each type of production.
It was concluded that the environmental impacts obtained as a consequence of each of these types of production are different.

Acknowledgements This project was partially funded by CONACYT and FCT - project SOC 4884/2016. It was carried out due to the collaboration between researchers from ICAR of UAMex of Mexico and CPES of ULHT of Portugal.

References

[1] ISO. *Environmental Management-Life Cycle Assessment-Principles and framework.* International S. Organization, Geneva, 2006.

[2] ISO. *Environmental Management-Life Cycle Assessment-Requirements and Guidelines.* International S. Organization, Geneva, 2006.

[3] D. W. Johnson, R. A. & Wichern. *Applied Multivariate Statistical Analysis.* Pearson, Prentice Hall, New Jersey, 2007.

13 April, 13:30 - 14:10, Auditorium atrium

Pavement friction performance model

Adriana Santos[1], **Susana Faria**[2], **Elisabete Freitas**[3]
[1] Universidade do Minho, Departamento de Engenharia Civil, CTAC – Centro de Território, Ambiente e Construção, Guimarães, Portugal
[2] Universidade do Minho, Departamento de Matemática e Aplicações, CBMA – Centro de Biologia Molecular e Ambiental, Guimarães, Portugal, sfaria@math.uminho.pt
[3] Universidade do Minho, Departamento de Engenharia Civil, CTAC – Centro de Território, Ambiente e Construção, Guimarães, Portugal

---

Degradation models of the pavement allow the asset manager to guarantee the safety of its users. A linear mixed effect model was developed to describe the pavement friction performance as a function of the weather conditions, the traffic volume, the pavement age, the pavement structure and the geometric characteristics of the road. This study is based on real database obtained over 8 years in the Ascendi network, in six different districts, in a total of 720 pavement sections of 1km.

Keywords: Performance model, friction, linear mixed effects models, longitudinal data

---

Pavement degradation models play a crucial role in pavements management systems. The main goal of these models is to characterize the response-variable throughout time, as well as, to determine whether this is related with a set of factors, such as, traffic volume, pavement structure, weather conditions, among others.
Many pavement performance models have been developed to describe the evolution of pavement performance indicators. One of these indicators is friction that should be taken into account due to its important effect on user safety. Friction has been acknowledged as one of the main factors contributing to the number of traffic accidents and is therefore essential to the assessment of the pavements quality, integrated within management systems.
The purpose of this study is to identify the most influencial factors associated with pavement friction performance over time.
Mixed effect models are recommended for modelling a wide variety of pavement performance data, to account for the correlation between repeated observations on the same pavement section. These models are useful for modelling the dependence among responses inherent in longitudinal or repeated measures data by incorporating random effects ([4]).
In order to study the pavement friction performance, we develop a linear mixed effect model by monitoring and analysing the conditions of road pavements in a period of eight years. Data were collected on highways, in six different districts in the north and centre of Portugal, in a total of 720 pavement sections of 1km.
The maximum likelihood method is used to estimate the parameters of the model and the pavement section is included in the model as a random effect. Likelihood ratio tests are

applied for choosing between two models, to select the final model. Akaike information criterion (AIC) are also used to compare several alternative models.
The main conclusion of this work is that annual average daily traffic, precipitation, maximum temperature, humidity, and pavement age have a significant effect on pavement friction performance.
These results may help to assist the network manager in conducting effective maintenance and/or rehabilitation measures in order to promote the better quality of the surface characteristics of the pavement and, consequently, optimize the overall level of road.

Acknowledgements This work was supported by the strategic programmes UID/BIA/04050/2019 and UID/ECI/04047/2019 funded by national funds through the FCT I.P.

References

[1] A. Gałecki and T. Burzykowski. *Linear Mixed-Effects Models Using R, A Step-by-Step Approach.* Springer - Verlag, New York, 2013.

[2] J. C. Pinheiro and D. M. Bates. *Mixed-effects Models in S and S-Plus.* Springer, New York, 2000.

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

[4] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer - Verlag, New York, 2000.

13 April, 13:30 - 14:10, Auditorium atrium

The effect of incubation on the companies' performance: a study with companies from the central region of Portugal

Carla Henriques[1], **Pedro Pinto**[2], **Rita Almeida**[3]
[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), carlahenriq@estv.ipv.pt
[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseun), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS)
[3] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu

The support offered by business incubators assumes growing relevance, helping entrepreneurs who need to develop their projects. The objective of this work is to evaluate if there are benefits in the incubation of companies with respect to the performance, that is, to verify if the incubated companies present better performance than those that were not incubated. Multivariate regression models were applied to evaluate the effect of incubation while controlling for other variables that are well-known determinants of companies' performance.

Keywords: regression modeling, robust standard errors, incubators

The number of business incubators has grown significantly in recent years, as a strong allied to the creation of new businesses, through legal, financial and technological support, as well as providing facilities for the establishment of new companies [1]. In order to understand whether there is an incubation benefit on the companies' performance, the Return Operational Asset (ROA) and the Increase in Turnover (IT) were selected as dependent variables.
The sample considered for this study includes companies of the central region of Portugal, 221 incubated and 2.959 non-incubated companies.
The comparison between incubated and non-incubated companies was first carried out through the t-test and the Mann-Whitney test. Then, linear regression models were estimated in order to evaluate the impact of incubation on the dependent variables, adjusting for the effect of control variables, when significant. The significance of the variables was evaluated by estimating consistent heteroscedastic standard errors.
The final models estimated for ROA and IT are the following:

$$\widehat{ROA}_i = 22,891 + 13,142Inc_i - 0,551Age_i - 0,107AI_i - 0,002AT_i - 0,944Inc \times Age_i$$

$$\widehat{IT}_i = 55,862 + 82,456Inc_i - 1,505Age_i + 0,005AT_i - 5,842Inc \times Age_i$$

where Inc is the dummy variable for incubated companies, Age is the age of the companies, AI is the percentage of Intangible Assets and AT stands for the total assets.
Both the estimated model for the ROA and for IT provide evidence that the effect of incubation depends on age. More precisely, incubated companies are more profitable (have higher ROA and IT) when they are young, however, this effect decreases as the company matures. This can be understood since when incubated companies reach the maturity they stay abreast with the other companies present in the market, that is, the incubation stops being relevant.
The study therefore provides evidence that there is a benefit in the incubation of companies. In fact, incubated companies present better performance in terms of operational assets and increase in turnover, however, this difference fades away with the age of the company.

References

[1] K. Aerts, P. Matthyssens, and K.. Vandenbempt. Critical role and screening practices of european business incubators. *Technovations*, 27:254–267, 2007.

13 April, 13:30 - 14:10, Auditorium atrium

Corporate social responsibility: What about Portugal?

Cláudia Silvestre[1], **Mafalda Eiró-Gomes**[2], **Ana Raposo**[2], **João Simão**[2], **Tatiana Nunes**[2]
[1] Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt
[2] Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa

---

From the private to the public sectors organizations are being called to improve their practices in accordance to social, economical and environmental standards. The present study tries to explore how organizations belonging to GRACE (an association concerned with these issues) understand and practice what is in general called as Corporate Social Responsibility.

Companies have been grouped into 4 clusters according to their policies and activities about this topic. The main practices have also been ranked in accordance with their relevance to the clusters results.

Keywords: Cluster Analysis, Factor Analysis, Mutual Information, CSR, Communications

---

In the first decades of the 21st century companies have been confronted with new challenges and risks. No one expects anymore to hear from a CEO that the only responsibility of a business is to improve profits but we don't seem to hear often which are the specific responsibilities organizations are willing to take in our contemporaneity. Do they go beyond philanthropic and voluntary work? How concerned are they with Internal Corporate Social Responsibility (CSR) activities? What constraints do they impose on the supply chain practices? What are their environmental practices? What about innovation and specially responsible innovation?
The main research question authors tried to adress was precisely what do the organizations that belong to the portuguese association for the development of corporate citizenship - GRACE, define as being their main social responsible principles, policies and practices. Are they mainly concerned with economical, social or environmental issues? Does CSR belong to the enterprises DNA?
In this work which is based on a sample of 44 companies (response rate of 28%), the authors are focused on (1) how they understand CSR (e.g. How does your organization define CSR?; What subjects are managed within CSR scope?; or Who are the stakeholders involved in CSR actions?), and on (2) their policies (e.g. Does your organization have: a CSR policy; a strategic plan for CSR; a code of ethics?; or Does your organization:

promote recycling; use natural resources efficiently; follow the guidelines of the Global Report Initiative?).
Based on 18 Likert scale questions of 6 points, a factor analysis was conducted. There were identified 4 latent variables that explain 73% of the total variance, and their estimated internal consistency varies between 0.71 and 0.91 (Cronbach's coefficient alpha). Since the goal was to identify groups of companies with similar policies and activities according to CSR, to select the number of latent variables, a cluster analysis was carried out with 2, 3, 4, and 5 latent variables. In each case the authors could identify 3 or 4 clusters. To compare all these solutions, the confusion matrix was calculated. There were large differences between solutions with 2, 3 and 4 latent variables. However, the solution with 4 and 5 latent variables produced the similar group structure when companies were grouped in 3 clusters and the same structure when 4 clusters were considered. So, the Ward hierarchical clustering method based on 4 latent variables was performed. To measure similarities Euclidean distance was used and as a result, 4 companies segments were obtained.
To profile the segments, Qui-Squared test was conducted for nominal variables and Kruskal-Wallis test for ordinal ones. When the null hypothesis was rejected, the authors have considered that variables were relevant to clustering. For those which were ordinal, they have calculated the mutual information that has allowed ranking relevant variables to companies segmentation according their policies and CSR activities. The most relevant were: monitoring of the suppliers' ecological footprint, reduction of waste produced, using LED bulbs, promoting recycling of waste and making investments in energy efficiency.

Acknowledgements This research was partially supported by Instituto Politécnico de Lisboa, IDI&CA - ref. IPL / 2018 / 3Cs_ESCS.

References

[1] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, 1991.

[2] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[3] D. Steinley and M. Brusco. Selection of Variables in Cluster Analysis an Empirical Comparison of Eight Procedures. *Psychometrika*, 1:125–144, 2008.

13 April, 13:30 - 14:10, Auditorium atrium

Comparison of tides in real time

Dora Carinhas[1], **Paulo Infante**[2], **António Martinho**[3], **Pedro Santos**[4]
[1] Instituto Hidrográfico; IIFA/Universidade de Évora, paulino.carinhas@gmail.com
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora
[3] Marinha Portuguesa
[4] Instituto Hidrográfico

Issues related to the quality of tide gauge measurements has become more important with the modernization of equipment and the recent concerns about the rise in the average sea level. This paper allowed to assess the performance of tide gauge installed in Setubal Peninsula - Troia. The Global Sea Level Observing System (GLOSS) target of 1 cm accuracy in the individual sea level measurement.

Keywords: accuracy, regression analysis, tide, tide gauge, time series

---

In the last decades much attention has been paid to the performance of the tide, especially in the context of the GLOSS program ([2], [3]). The tide data are of particular interest when studying climate change and, consequently, increasing the mean sea level.
Since march 23, 2017, the tide heights of two tide gauges on the ferry docks installed in the Setubal Peninsula are being recorded; the location and the tide records of january 18, 2018 are shown in Figure 1.
Classical methods applied to analyze the data of such on-site experiments include (e.g., [4], [1]):

1. examination of the time series of the computed differences between the tide gauge measurement and the standard or reference gauge measurement;
2. computation of the root-mean-square-error (rmse) of the time series of the differences;
3. visualization of one tide gauges data against the other (scatterplot) and computation of the slope of the linear regression trend between both sea level series. This slope expresses the distinct sensitivities of the gauges to the tidal range;
4. inspection of the spectral power of non-tidal residuals after tidal variations have been removed by means of the harmonic analysis;
5. comparison of the tidal constituents obtained from the harmonic analysis.

This paper also presents a technique to compare two measurement systems to evaluate if the tide gauges, installed in the Setubal Peninsula, are compatible.

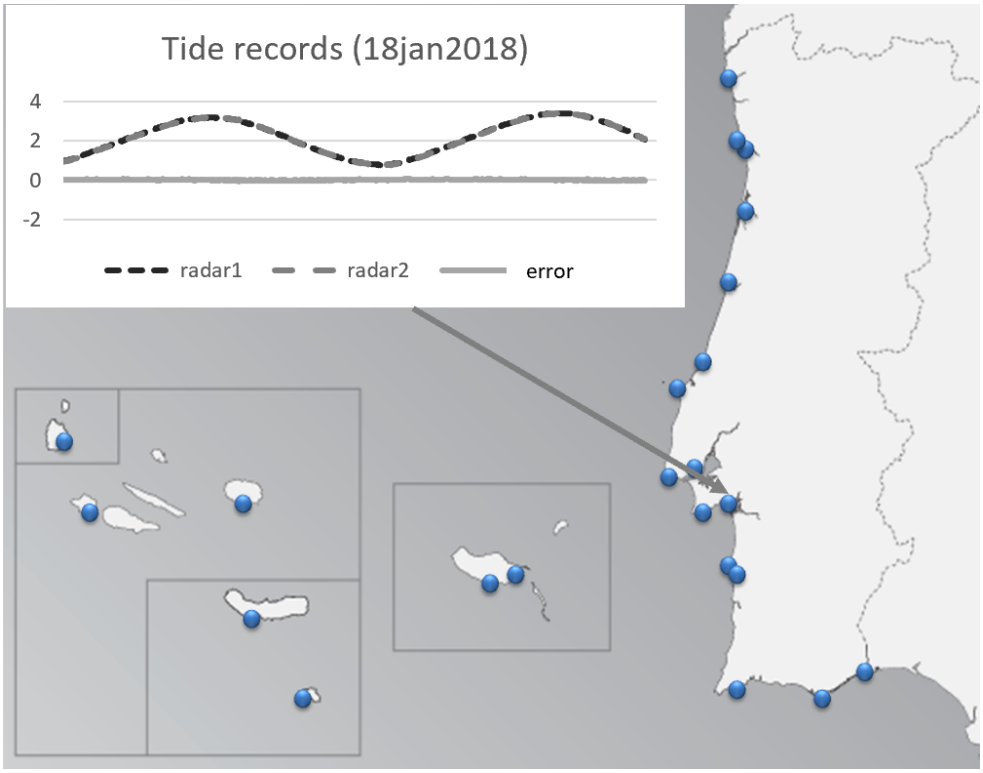


Figure 1: Tide heights of two tide gauges installed in the Setubal Peninsula. (source: Hydrographic Institute)

References

[1] E. Alvarez Fanjul B. Martín, B. Pérez. The eseas-ri sea level test station: reliability and accuracy of different tide gauges. *Int. Hydrogr. Rev*, 6:44–53, 2005.

[2] IOC. Global sea level observing system (gloss) - implementation plan. Technical Series N.50, Paris, 1997.

[3] A. Allen A. Aman E. Bradshaw P. Caldwell R.M. Fernandes H. Hayashibara F. Hernandez B. Kilonsky B. Martin Miguez G. Mitchum B. Pérez Gómez L. Rickards D. Rosen T. Schöne M. Szabados L. Testut P. Woodworth G.Wöppelmann J. Zavala M. Merrifield, T. Aarup. The global sea level observing system (gloss). in: J. hall, d.e. harrison and d. stammer (eds.). *Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society Conference*, 2, 2019.

[4] P.L. Woodworth and D.E. Smith. A one-year comparison of radar and bubbler tide gauges at liverpool. *Int. Hydrogr. Rev*, 4:2–9, 2003.

13 April, 13:30 - 14:10, Auditorium atrium

Nonparametric two-way ANOVA: A simulation study to compare results from balanced and unbalanced designs

Dulce G. Pereira[1], **Anabela Afonso**[1]
[1] Centro de Investigação em Matemática e Aplicações/IIFA, Departamento de Matemática/ECT, Universidade de Évora, dgsp@uevora.pt

Several alternatives to parametric ANOVA with two factors have been proposed in the last years. In this work, we conduct a simulation study to compare the performance of some of these alternatives. We considerer balanced and unbalanced homocedastic designs, with fixed effects, with different total samples sizes and discrete distributions. We concluded that the Wald-type statistic is the most powerful, but with high rate Type I error. In the presence of interaction, the test L of Puri & Sen and van der Waerden do not present a good performance.

Keywords: permutation tests, rank transform, ties, Wald statistic.

Analysis of variance (ANOVA) is frequently used in experimental science to study the influence of one or more factors on a given dependent variable [2]. However, the underlying assumptions are difficult to hold true when real data sets are analyzed.
Since the second half of the last century, several alternatives to parametric ANOVA were proposed to be used in case of serious violations of the ANOVA assumptions or with ordinal data [1]. These approaches are essentially divided into semi and non-parametric methods. In the literature we can found some works that study the performance of these methods which considered data from continuous distributions, with different weights in the tails and different degrees of skewness. However, when data is drawn from discrete distributions it can often produce ties. Few studies studied the impact of ties in the performance of these methods. With balanced designs, when interaction in not present, empirical error Type I is not affected by the number of ties [3]. In the presence of interaction, L of Puri & Sen and van der Waerden tests did not show a consistent behaviour with the decrease in the number of ties. Their performance depends on the distribution, sample size and size effect. In this work we intend to extend the studies of Afonso and Pereira [1, 3] to unbalanced designs. In agricultural and biological sciences it is usual to have unbalanced designs, i.e., the number of observations per factor levels combination is not the same. This situation can occur due to several reasons, which may be due to design convenience, for instance due to costs, but also because the researcher can not control the experience at all.

The Wald-type statistic is the most powerful, but with high rate Type I error. In the presence of interaction, the test L of Puri & Sen and van der Waerden do not present a good performance.

Acknowledgements A. Afonso and D. G. Pereira acknowledge partial funding by the FCT, Portugal, under the project «UID/MAT/04674/2019 (CIMA)».

References

[1] A. Afonso and D. G. Pereira. Comparação entre métodos não paramétricos para a análise de variância com dois fatores: Um estudo de simulação. In *In Classicação e Análise de Dados - Métodos e Aplicações III.* Instituto Nacional de Estatística, in press.

[2] D. G. Montgomery. *Design and Analysis of Experiments.* John Wiley & Sons, Hoboken, 8 edition, 2013.

[3] D. G. Pereira and A. Afonso. Potência e erro de Tipo I das alternativas não paramétricas à ANOVA com dois fatores. In *In Atas do XXIII Congresso da Sociedade Portuguesa Estatística.* Sociedade Portuguesa Estatística, in press.

13 April, 13:30 - 14:10, Auditorium atrium

Chemical hazard pictograms and safety signs taught in higher education: a statistical approach

Fernando Sebastião[1], **Lizete Heleno**[2], **Sílvia Monteiro**[1]
[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] School of Technology and Management (ESTG), Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal

In this work we performed some comparisons between the identification of pictograms of dangerous substances and safety signs by students of several courses in higher education. The knowledge of understanding safety pictograms is essential to prevent accidents.
To compare some results between the assessments before and after the contents have been taught in the classroom we used statistical inference to highlight the significant differences in the learning process. The results can be used to improve the contents of the courses related to safety pictograms in higher education.

Keywords: chemical hazard substances, safety signs, ghs pictograms, higher education, hypothesis tests

In quotidian life and in the workplace it is crucial to understand the safety pictograms since they allow to know the risks involved in activities, promoting preventive measures and decrease accident's occurrence. Therefore, safety pictograms are a way to call attention, quickly and clearly, for objects and risk situations [4]. According to international rules, namely, the International Organization for Standardization (ISO), the safety pictograms are recognized by the colour, type and shape, with different means, like "chemical hazard", "firefighting equipment", "obligation", "prohibition", "rescue or emergency" and "warning" [2, 4].
The safety pictograms can be divided in two groups, the safety signs defined by Portuguese national legislation [1], and the chemical hazard pictograms that classify hazardous substances according to the Globally Harmonized System (GHS) [3]. That classification was adopted by the European Commission through Regulation (EC) No. 1272/2008 on the classification, labelling and packaging of substances and mixtures (CLP Regulation), effective in all member states since 2010.
We have intentions to present some interesting statistical results of our study. For example, in Figure 1 we represent some statistics of the sum of scores (on the scale of 0% to 100%), resulting from the global assessments, when we compare the obtained values before

and after the introduction of the contents in the classroom related to pictograms, in two different and relevant teaching areas: Business and Legal Sciences (CEJ) and Engineering and Technology (ET). We can observe that before occurring the formation, in the ET area, the dispersion is higher than in the CEJ area, while for the case after occurring the formation the dispersion is slightly smaller in the ET area, which can be associated to a big sensibility to learn the subjects about safety pictograms, once the ET students spend more time in laboratorys during the course.

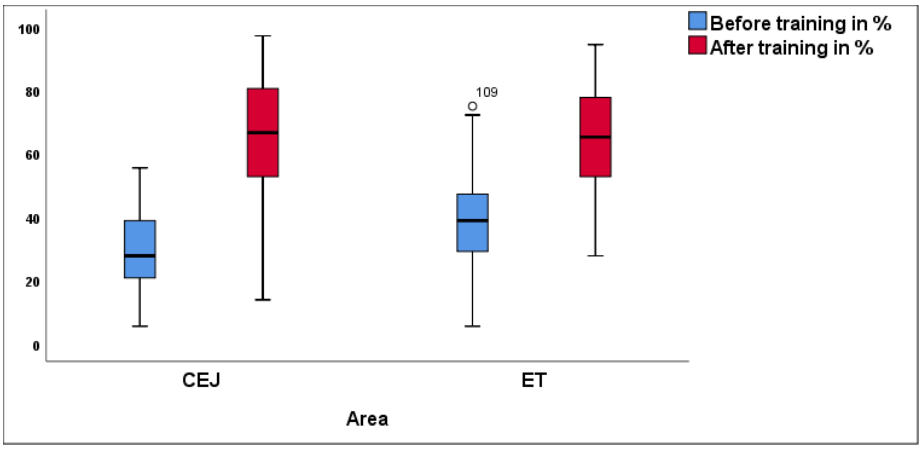


Figure 1: Boxplots of sum of scores compared before and after the training between two teaching areas: Business and Legal Sciences (CEJ) and Engineering and Technology (ET)

When we use 95% confidence intervals for mean of sum of scores for the comparison between before and after occurring the training for all teaching areas involved in the study, we conclude that the mean of sum of scores had a significant increase after the training. We have used other relevant statistical analysis based in hypothesis tests to detect the student's performance when they are assessed in the learning process, by other factors as gender, course and age. In this work we will present the main conclusions about these topics.

References

[1] Portaria 1456-a/1995, de 11 de dezembro. Diário da República n.º 284/1995 – I série-B. Ministério do Emprego e da Segurança Social. Lisboa, 1995.

[2] J. Teles E. Duarte, F. Rebelo and M.S. Wogalter. Safety sign comprehension by students, adult workers and disabled persons with cerebral palsy. *Safety Science*, 62:66–77, 2014.

[3] United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*. 6th Edition eISBN 978-92-1-057320-7, New York and Geneva, 2015.

[4] K. Ispolnov S. Monteiro, L. Heleno and M. Ribeiro. Safety pictograms perception analysis. inted2016. *Conference 7th-9th March 2016, Valencia, Spain*, 7385–7392, 2016.

13 April, 13:30 - 14:10, Auditorium atrium

Maximum likelihood method by logistic regression in the evaluation of lifestyles, anthropometric and lipid indicators in young university students with and without family support

João Paulo Figueiredo[1], **Mariana Pratas**[2], **Mariana Pereira**[2], **Daniela Correia**[2], **Nádia Osório**[2], **Armando Caseiro**[2], **António Gabriel**[2], **Andreia Costa**[3], **Ana Ferreira**[3]

[1] Instituto Politécnico de Coimbra, Coimbra Health School – ESTeSC, Departamento das Ciências Complementares (Bioestatística e Epidemiologia), jpfigueiredo1974@gmail.com

[2] Instituto Politécnico de Coimbra, Coimbra Health School - ESTeSC, Ciências Biomédicas Laboratoriais

[3]Instituto Politécnico de Coimbra, Coimbra Health School - ESTeSC, Saúde Ambiental

---

The entrance to higher education involves several changes, being that both the family and the university environment play an important role in the health of those same individuals. The purpose was to assess the impact of the type of family support on the prevalence of certain risk behaviors in young college adults. Conclude that family support did not have a significant impact on tobacco and alcohol-related risk behaviors. The same occurred for dietary behaviors and cholesterol indexes, and youngsters who did not have family support presented changes in cholesterol (Total, LDL, HDL).

Keywords: life style, family support, body mass index, triglycerides, cholesterol

---

Admission into adolescence and/or adult life is considered a critical period for changing behaviors that affect, positively or negatively, lifelong health. In addition to contact with psychoactive substances (tobacco, alcohol and other drugs), new practices and attitudes towards food, physical and recreational activity and changes resulting from emotional management and stress, university students are, in the great majority, confronted with the "outside world" associated with greater independence from parents. According to some authors, young adults integrate/adopt/develop some healthy practices and behaviors or of risky that may have a significant impact on their health in the future [1] [2]. There are many factors that can interfere in the choices and options of these individuals, namely the influence of the family, which is the first level of socialization, defining rules and limits and allowing the autonomy and self-expression of young people, protecting them against risk behaviors [3].

The aim of this research was to study the prevalence of health and risk behaviors among young university students, with and without family support, and their relationship with

their lipid and anthropometric conditions. The target population of this study were young adults aged 18-29 (N = 155). Regarding the type and technique of sampling, it was a non-probabilistic sample. Some of the parameters controlled were Body Mass Index (BMI), HDL-cholesterol, LDL-cholesterol, triglycerides (Trig.), Blood Pressure (BP) and Lifestyle (smoking habits, alcohol habits, eating habits, physical activity).
The statistical method applied was Binary Logistic Regression by the Maximum Likelihood Method. We evaluated the adequacy of the models: Calculation of Odds Ratio (% of individuals correctly predicted), Omnibus Test and Hosmer and Lemeshow Test. Regarding the results, 70% of the participants were normal weight, 87.7% had adequate BP, 62.2% had total cholesterol at adequate levels as well LDL levels (71.1%). Risk behaviors: 18.1% were habitual smokers, 74.2% consumed alcoholic beverages, 69.0% did not practice regular physical activity. Lastly, the majority were living alone or in the company of other students, without the presence of parents and/or siblings and/or other relatives (74.2%).
When we tried to evaluate the explained probability of the impact of the family support in the explanation of the anthropometric and biochemical values, these revealed little differentiation (ORIMC: 0.991, CI [0.777-1.265], ORHDL: 1.001, CI [0.933-1.075], ORLDL: 0.999; IC [0.969-1.031], OR_{Trig}: 0.998; CI [0.972-1.024]; ORBP: 1.621; CI [0.319-8.246]). On the other hand, young adults who did not have family support during their academic lives were also the most sedentary (physical activity) (OR: 2.771; CI [1.288-5.963]) compared to those living with the family. However, were not observed any explanatory effects of the presence / absence of family support in the remaining behaviors ($OR_{Hábitos\ Tabágicos}$: 0.688; CI [0.263-1.797]; $OR_{Hábitos\ Alcoólicos}$: CI [0.477-2.590]; $OR_{Hábitos\ Alimentares}$: 1.018; CI [0.382-2.716]). The present study allowed us to conclude that, although family support does not have a direct and/or immediate significant impact on the lifestyle of young adults, this is still of major research interest, given the increase in prevalence estimates of students who have moved to outside their area of residence upon entering higher education, as well as increasing the adoption of risk-taking behaviors detrimental to their health [1] [3]. The control of risk factors is the best way to prevent cardiovascular diseases, and most of them coincide with the characteristics analyzed in the present study [4].

References

[1] D.M. Alves, L.M. Almeida, and H.M. Fernandes. Estilos de vida e autoconceito: Um estudo comparativo em adolescentes. *Revista Iberoamericana de Psicología del Ejercicio y el Deporte*, 12(2):237–247, 2017.

[2] C. Balsa, C. Vital, and C. Urbano. IV Inquérito Nacional ao Consumo de Substâncias Psicoativas na População Geral, Portugal 2016/2017. SICAD -Serviço de Intervenção nos Comportamentos Aditivos e nas Dependências, 2018.

[3] C. Lopes, D. Torres, and et al. Inquérito Alimentar Nacional e de Atividade Física (IAN-AF 2015-2016)-Relatório Parte II. Universidade do Porto, Porto, Portugal, 2017.

[4] C. Mariano, M. Antunes, Q. Rato, and M. Bourbon. Caraterização do perfil lipídico da População Portuguesa. Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal, 2015.

13 April, 13:30 - 14:10, Auditorium atrium

Evaluation of potential biomarkers in the development of chronic complications in diabetes mellitus using the binary logistic regression model

João Paulo Figueiredo[1], **Andreia Almeida**[2], **Ana Cristina Alves**[2], **Cláudia Silva**[2], **Tatiana Varandas**[2], **Amélia Pereira**[3], **Élio Rodrigues**[3], **Marta Amaral**[3], **Ana Valado**[2], **Nádia Osório**[2], **António Gabriel**[2], **Armando Caseiro**[2]

[1] Instituto Politécnico de Coimbra, Coimbra Health School – ESTeSC, Departamento das Ciências Complementares (Bioestatística e Epidemiologia), jpfigueiredo1974@gmail.com

[2] Instituto Politécnico de Coimbra, Coimbra Health School - ESTeSC, Ciências Biomédicas Laboratoriais,[3]Hospital Distrital da Figueira da Foz, Serviço de Medicina Interna

---

Diabetes Mellitus (DM) is a metabolic disease in association with changes in metabolism of carbohydrates, lipids and proteins. Aims: To assess serum levels (MMP-10, VEGF, TIMP-1) in serum samples from type 1 diabetic and healthy subjects. The population of the study consisted of 10 healthy controls and 12 type 1 diabetic patients. Results: TIMP-1 and VEGF levels tended to be lower in DM patients compared to the control group. Regarding the levels of MMP-10, no significant differences were observed, however a trend of increase in the MMP-10/TIMP-1 ratio was observed in patients.

Keywords: Type 1 Diabetes mellitus, retinopathy, nephropathy, MMP-10, TIMP-1

Diabetes Mellitus (DM) is a metabolic disease in which there is a chronic increase in serum glucose concentration, in association with changes in carbohydrate, lipid and protein metabolism, secondary to deficient total or partial secretion of insulin and / or resistance to it's action [3]. It is a risk factor for the development of chronic microvascular complications, such as retinopathy, nephropathy and neuropathy, and macrovascular [1]. Vascular endothelial growth factor (VEGF) is one of the major angiogenic factors that acts in processes such as endothelial cell proliferation and increased vascular permeability. MMPs are zinc and calcium dependent endopeptidases, which participate in several physiological processes such as extracelluar matrix (ECM) remodeling, healing, angiogenesis and apoptosis. They are secreted by various cell types, including neoplastic, epithelial and inflammatory cells, in the form of proenzymes, which require activation by proteolysis, a process controlled by tissue-specific inhibitors of metalloproteinases (TIMPs). Increased MMP activity may contribute to the pathological reorganization of ECM in atherosclerosis, aneurysms and diabetic nephropathy. Thus, it is extremely important to investigate new biomarkers that allow the diagnosis and monitoring of DM-related complications in their early stages enabling the clinical intervention in advance. The aim of the study was to evaluate the levels of MMP-10, VEGF and TIMP-1 in serum samples from individuals with

type 1 DM compared to the control group composed of healthy individuals. The study population consisted of 22 individuals with type 1 DM (with 7 or more years of disease) and a healthy control group at the Hospital District of Figueira da Foz. Statistical methods applied: Bivariate and Multivariate Statistics (Binary Logistic Regression). Regarding the main results the serum levels of MMP-10 from the DM group compared to the control group were similar to each other. Serum VEGF levels tended to be lower in the DM group compared to the control group, although the differences were not significant ($p > 0.05$). Similar profile occurred with serum levels of TIMP-1: DM compared to the control group. We proposed to understand, in a multivariate way, the effect of predictors (biomarkers) that best explain the expression of pathology (Table 1).

Table 1 - Predictors of Diabetes Mellitus

			95% C.I.for OR		Nagelkerke	H-L test
	Wald Test;(df);p-value	OR	Lower	Upper	R^2	(x^2; df; p-value)
MMP 10	0.232;(1); p:0.630	189.567	0.001	35096.55	0.310	6.149;8;0.631
VEGF	1.804;(1) p: 0.179	0.0001	0.001	50338.84		
TIMP	0.074;(1) p: 0.786	3.312	0.001	18958.91		
Razão MMP/TIMP	1.452;(1) p: 0.228	1.848	0.681	5.01		
Constant	0.669;(1) p: 0.413	8.555				

Legend: predicted variable: Pathology (Diabetic or Non-Diabetic); H-L test: Hosmer and Lemeshow Test.

There was no impact of the significant and differentiating potential biomarkers on DM patient classification versus controls ($p > 0.05$). However, it can be stated that patients with DM presented significantly higher risk estimates in the biomarkers MMP-10, MMP / TIMP ratio and TIMP compared to the non-disease group. Our initial (exploratory) study allowed us to conclude that serum levels of MMP-10 per slot blot did not show significant differences between the diabetic group and the control group. Serum TIMP-1 levels also tended to be lower in DM patients compared to the control group. In agreement, Mohammad et al. found a decrease in TIMP-1 levels in the retinas of diabetic rats [2]. VEGF increases the permeability of vascular endothelial cells by altering glomerular filtration. Kornel et al. found an increase in serum levels of VEGF in diabetic patients compared to control group [4]. In the present study, serum VEGF levels tended to be lower in patients compared to the control group, although differences were not significant. Such outcome may be explained by the therapy to which DM patients are subjected.

References

[1] O.F. Leal and F. Soares. Impacto da diabetes mellitus tipo 1 e tipo 2 na doença cardiovascular e a sua avaliação por ecodoppler codificado a cores. *Revista de Ciências da Saúde da ESSCVP*, 7:22–31, 2015.

[2] G. Mohammad, M.M. Siddiquei, and et al. The ERK1/2 inhibitor u0126 attenuates diabetes-induced Upregulation of MMP-9 and biomarkers of inflammation in the retina. *Journal of diabetes research*, 658548, 2013.

[3] T. Ângelo. Diabetes mellitus e doença periodontal. Universidade Católica Portuguesa, Viseu, 2013.

[4] K. Semeran, P. Pawlowski, and et al. Plasma levels of IL-17, VEGF, and adrenomedullin and s-cone dysfunction of the retina in children and adolescents without signs of retinopathy and with varied duration of diabetes. *Mediators of inflammation*, 274726, 2013.

13 April, 13:30 - 14:10, Auditorium atrium

Detection of outliers municipalities in Portugal: a compositional analysis of occupational status and academic qualification

Letícia Leite[1], **Adelaide Freitas**[2], **Cristina Gomes**[3]
[1] Departamento de Matemática, Universidade de Aveiro, c.leite.leticia@ua.pt
[2] Departamento de Matemática & CIDMA, Universidade de Aveiro
[3] Departamento de Ciências Sociais Políticas e do Território & Govcopp, Universidade de Aveiro

---

Based on robust multivariate statistical methods in the context of compositional data, we explore data extracted from the 2011 Census related to internal migration flows considering the occupational status and the academic qualification of the residents of the 308 municipalities in Portugal. Regarding the multivariate compositions of these data, our analysis identified some municipalities as being atypical. These municipalities tend to be in the interior regions of Portugal.

Keywords: compositional data, robust biplot, ternary diagram, outliers

---

Compositional data are multivariate observations of positive values which sum results in a constant. They represent quantitative descriptions of the parts of a whole, conveying relative rather than absolute information, usually proportions or percentages, being the sum of those values equal to 1 or 100, respectively.
Portugal lacks the necessary instruments for a direct analysis of the features related to population migrations dynamics. Due to the absence of available information, the present study corresponds to an indirect analysis based on data extracted from the 2011 Census concerning the 308 municipalities of Portugal internal flows. The compositions of academic qualification (A) and occupational status (B) of the residents of each municipality are analyzed. The academic qualification of the residents was divided in ten parts (None, 1st Cycle of Basic Education (grades 1-4, ages 6-10), 2nd Cycle of Basic Education (grades 5-6, ages 10-12), 3rd Cycle of Basic Education (grades 7-9, ages 12-16), High School, Post-Secondary School, Bachelor, Graduation, Master's, PhD). The occupational status of the residents for each municipality was partitioned in three parts (Unemployed, Employed, Inactive).
The main goal of the present study is the identification of municipalities outliers from the point of view of a compositional analysis in order to pick up particularities that may exist between the residents' qualifications and their occupational status. Only the first dataset contains zero counts (e.g. municipality of Aguiar da Beira counts zero residents

with a PhD). In our analysis, all these zeros were imputed using the robust model-based procedure.
Hence, robust multivariate statistical techniques implemented in the package mvoutlier [1] will be applied to the two datasets (A and B) using RStudio. Two types of exploratory plots, Biplot and Ternary Diagram, will be used.
Our analysis identified some municipalities as being atypical. These municipalities tend to be in the interior regions of Portugal as it is shown, for instance, in Figure 1 for the occupational status.

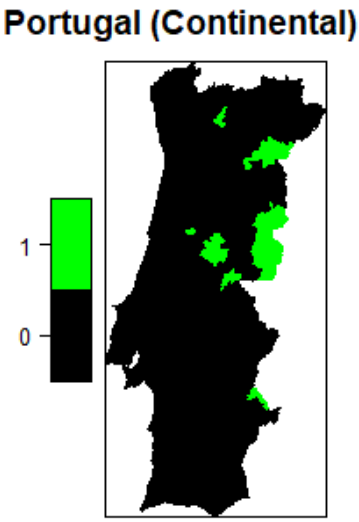


Figure 1: Map of Portugal (Continental) with the thirteen municipalities outliers highlighted (label 1 - Freixo de Espada à Cinta, Idanha-a-Nova, Mourão, Oleiros, Pampilhosa da Serra, Penamacor, Penedono, Ribeira de Pena, Sabugal, Torre de Moncorvo, Vila Nova de Foz Côa, Vila Nova de Poiares, Vila Velha de Ródão). Source: INE Census 2011.

Acknowledgements The second author was supported by Fundação para a Ciência e Tecnologia (FCT), within the project UID/MAT/04106/2019 (CIDMA).

References

[1] Peter Filzmoser and Moritz Gschwandtner. *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*, 2018.

13 April, 13:30 - 14:10, Auditorium atrium

A simulation study for robustly estimate the number of components for finite mixtures of linear mixed models

Luísa Novais[1], **Susana Faria**[2]
[1,2] Department of Mathematics and Applications, University of Minho, Portugal, luisa_novais92@hotmail.com

---

Choosing the number of components for mixture models has long been considered an important and unresolved research problem. In this study we investigate a robust estimation of the number of components for mixtures of linear mixed models by comparing the performances of trimmed and traditional information criteria through a simulation study.

Keywords: Finite mixtures of linear mixed models, Model selection, Trimmed information criteria, Robustness, Simulation study

---

Finite mixture models are a widely known method for modelling data that arise from a heterogeneous population. In regression analysis, it has been a popular practice to model unobserved population heterogeneity through finite mixtures of regression models.
Within the family of mixtures of regression models, finite mixtures of linear mixed models have also been applied in different areas of application since, besides taking into consideration the heterogeneity in the population, they also allow to take into account the correlation between observations from the same individual, which makes them particularly used in longitudinal data.
One of the main issues in mixture models is related to the estimation of the parameters. A pertinent subject concerns the robustness in the estimation of mixture models given that the parameter estimates, calculated using the EM algorithm, are sensitive to outliers. In particular, the estimation of mixtures of linear mixed models is very sensitive to outliers, since it is generally considered that not only the errors but also the random effects follow a Normal distribution.
As a consequence, although information criteria have been popularly used to select the number of components for mixture models due to their simplicity, information criteria are also sensitive to outliers and the presence of a single outlier may cause the estimated number of components to change, which may compromise the use of these criteria to select the number of components of a mixture model. Therefore, one of the main difficulties in mixture models arises in the selection of the correct number of components for each data set.

In order to overcome the problem, in this study we provide a simulation study to compare a robust information criteria with the traditional information criteria in the selection of the number of components for finite mixtures of linear mixed models.
The robust version of the information criteria is based on trimmed maximum likelihood estimates (TLE). Hence, assuming that $\alpha \times 100\%$ of the observations in a sample are outliers, the calculation of the trimmed maximum likelihood estimates for mixture models, proposed by Neykov *et al.* [4], only uses $(1-\alpha)\times 100\%$ of the observations to fit the model, removing the remaining observations.
Thus, to compute the trimmed maximum likelihood estimate is necessary to fit all partitions of the data and then, among the resulting estimates, choose the one that maximizes the log-likelihood function, which causes the computation of the trimmed maximum likelihood estimate to be very complex for large samples.
In order to avoid adjusting all partitions, Neykov *et al.* [4] proposed the *FAST-TLE* algorithm, which we used to compute the robust information criteria. The main idea behind this algorithm is to repeatedly iterate a two-step procedure consisting of a trial step and a refinement step. Therefore, the *FAST-TLE* algorithm allows an approximate solution of the *TLE*, being computationally much less demanding, particularly for large samples.
In the simulation study it was clear that both versions of the criteria yield similar results when there are no outliers present, but the presence of outliers clearly diminishes the performance of the traditional criteria since these criteria tend to overestimate the number of components in almost every case. On the other hand, the presence of outliers does not affect the performance of the robust information criteria given that most of the criteria performed well for the majority of the scenarios.
Therefore, selecting the correct number of components in a mixture model is not an easy problem and different configurations clearly influence the performance of the information criteria. Despite the high computational time, which can be a drawback to its use, the superiority of the robust information criteria was evident in the presence of outliers so its use is recommended whenever there are outliers present.

Acknowledgements This research was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference SFRH/BD/139121/2018.

References

[1] N. Depraetere and M. Vandebroek. Order selection in finite mixtures of linear regressions. *Statistical Papers*, 55(3):871–911, 2014.

[2] M. Li, S. Xiang, and W. Yao. Robust estimation of the number of components for mixtures of linear regression models. *Computational Statistics*, 31(4):1539–1555, 2016.

[3] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2000.

13 April, 13:30 - 14:10, Auditorium atrium

Zika: literacy and behavior of individuals on board ships. A preliminary analysis

João Faria[1], **Rosa Teodósio**[1], **M. Filomena Teodoro**[2,3], **Claudia Valete**[1]
[1] Institute of Hygiene and Tropical Medicine, New University of Lisbon, Lisboa, Portugal
[2] CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Avenida Rovisco Pais, n. 1, 1048-001 Lisboa, Portugal
[3] CINAV - Center for Naval Research, Naval Academy, Base Naval de Lisboa, Alfeite, 2810-001 Almada, Portugal, maria.alves.teodoro@marinha.pt

The objective of this study is to describe the knowledge, attitudes and preventive practices regarding the infection by the Zika virus (ZIKV) among the population embarked on Portuguese Navy ships. We performed a statistical analysis, a cross-sectional study that, besides allowing us to describe knowledge, attitudes and practices related to ZIKV infection, also let us to stratify the different groups under study: those who will navigate in endemic areas of Zika virus, those that have traveled to endemic areas of ZIKV and navigators in non-endemic areas of ZIKV. The data collection is still in progress. The knowledge level about ZIKV reveals significant differences between the distinct groups. The preliminary results obtained with the provisional data set are in agreement with similar performed studies.

Keywords: Zika virus, questionnaire, statistical approach, generalized linear models

During a study of yellow fever in the Zika forest of Uganda, Zika virus was detected in a rhesus monkey in 1947. Between 1960 and 1980, few cases of ZIKV infection were identified by serological methods, being mainly benign. However, the expansion of urban centers, transatlantic travel and increased airflow, as well as the movement of asymptomatic carriers between countries and continents, contributed to spread Zika virus. This spread has seen a huge increase [4, 3]. All these factors potentiated the increase in ZIKV transmission rate, as well as the possibility of genetic mutations in certain pathogenic microorganisms, allowing the existence of more resistant viruses with greater epidemic potential. The rapid expansion of the disease and importation into several countries on opposite sides of the globe, in addition to the constant intercontinental migratory flows, the prolonged time of viremia and the persistence of the virus in certain body fluids allows for a relationship with a high number of asymptomatic cases [2, 1]. Knowing that a military ship's garrison can visit endemic ZIKV sites where virus exposure can occur, disease prevention and health promotion of on-board personnel is an important issue. It is pertinent to analyze the knowledge, attitudes and practices regarding this issue, in order to develop intervention strategies, through health education actions. The objective of this study is to describe the

knowledge, attitudes and preventive practices regarding the infection by the Zika virus, among population on Portuguese Navy ships. We performed a statistical analysis (first applying some descriptive techniques, secondly applying some traditional comparison tests, using general linear models to obtain predictive models), a cross-sectional study that, besides allowing us to describe knowledge, attitudes and practices related to infection by ZIKV will allow us to compare the different groups under study: those who will navigate in areas endemic to Zika virus, those that have navigated to endemic areas of ZIKV and navigators in non-endemic ZIKV areas. A questionnaire was applied to each of these three groups. Data collection is still ongoing, but the preliminary results evidences that distinct groups have a different level of Zika virus knowledge. This issue is in line with similar studies already conducted.

Acknowledgements This work was supported by Portuguese funds through the FCT, *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2019, and *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal.

References

[1] Van-Mai Cao-Lormeau, Alexandre Blake, Sandrine Mons, Stéphane Lastère, Claudine Roche, Jessica Vanhomwegen, Timothée Dub, Laure Baudouin, Anita Teissier, Philippe Larre, Anne-Laure Vial, Christophe Decam, Valérie Choumet, Susan K Halstead, Hugh J Willison, Lucile Musset, Jean-Claude Manuguerra, Philippe Despres, Emmanuel Fournier, Henri Pierre Mallet, Didier Musso, Arnaud Fontanet, Jean Neil, and Frédéric Ghawché. Guillain barré syndrome outbreak associated with zika virus infection in french polynesia: a case control study. *The Lancet*, 387(10027):1531–1539, 2016.

[2] European Centre for Disease Prevention and Control. Rapid risk assessment. zika virus disease epidemic. potential association with microcephaly and guillain barreé syndrome. Second update, 8 February 2016. European Centre for Disease Prevention and Control (ECDP), Stockholm, 2016.

[3] Lisa Walddel and Judy Greig. Scoping review of the zika virus literature. *PLOS ONE*, 11(5), 2016.

[4] Camila Zanluca, Vanessa Melo, Ana Mosimann, Glauco Santos, Claudia Santos, and Kleber Luz. First report of autochthonous transmission of zika virus in brazil. *Memórias Instituto Oswaldo Cruz*, 110(4):569–572, 2015.

13 April, 13:30 - 14:10, Auditorium atrium

Perception of business corruption in EU28: A multilevel application

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

---

A two-level latent-variable model is applied to Eurobarometer data to study the perception of managers on business corruption in EU28. Company- and country-level covariates are added to the model (multiple indicator, multiple cause model) to account for their possible influence on these perceptions. Results show that both levels influence these perceptions.

Keywords: Multilevel analysis, Structural equation modeling, Latent variables, Corruption perception

---

This study investigates the under-researched area of business corruption within the EU28 by analyzing corruption perception of managers and the effects of company and country level on their perception [2, 4, 1]. We use a unique representative European Union survey from 2017, covering the EU28 (7746 responses by managers), and that collects indicators on the perception of business corruption and characteristics of the companies. This data set provides specific insights into an area that typically uses aggregate indicators at country level.

We apply a multilevel framework (multilevel factor model - multiple indicator, multiple cause) to analyze the influence of country-level indicators and specific company characteristics on managers' perception of corruption [3]. In particular, we assume a two-level model for the perception, as the (overall) corruption perception is defined by different first-order perception categories f^P (the perception about seriousness, widespread, and agreement of corruption) (Figure 1). The first-order perception categories (f^P) are measured by different items (Y_k, $k = 1, \cdots, K$). The items represent the responses of the managers to different questions about corruption corresponding to each category. The macro variables W_l, $l = 1, \cdots, L$ explain the country background, which influences the perception categories of the managers. Company variables (X_a), $a = 1, \cdots, A$ are added to control for the impact of the company at the overall perception of corruption.

The goodness of fit of the specified model is confirmed. The SRMR and RMSEA are below 0.1 and the CFI and TLI are above 0.9. Moreover, the company characteristics such as sectors (healthcare & pharmaceutical, engineering & electronics & motor vehicles, and construction & building), number of employees, turnover, and participation on a public tender are statistically significant and explain the corruption perception of managers in the EU28. The country-level covariates – the economic dimension and two dimensions

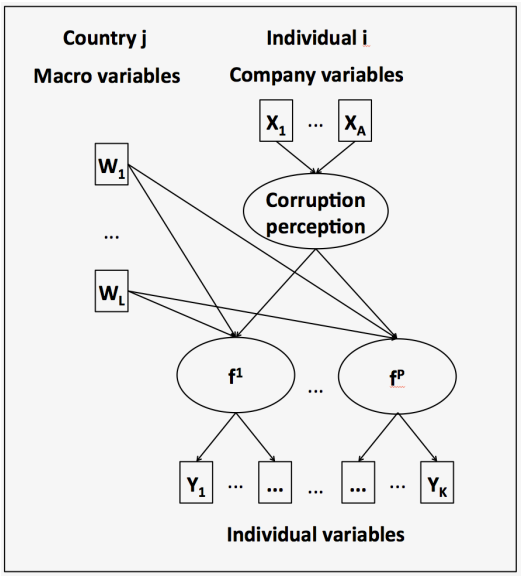


Figure 1: Conceptual Model

of Hofstede's national culture framework (power distance, and individualism) – are all significant control variables throughout all three first-order factors.
These findings show that managers' perceptions are not only explained by company characteristics (level 1), but also by the national setting (level 2) that plays a crucial role. Hence, politicians should focus on company and national policies to fight and prevent corruption.

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

References

[1] B. Bosco. Old and new factors affecting corruption in Europe: Evidence from panel data. *Economic Analysis and Policy*, 51:66–85, 2016.

[2] M.A. Cole. Corruption, income and the environment: An empirical analysis. *Ecological Economics*, 62:637 – 647, 2007.

[3] J. Hox. *Multilevel Analysis: Techniques and applications*. Mahwah: Lawrence Erlbaum Associates, 2002.

[4] Y.M. Mensah. An analysis of the effect of culture and religion on perceived corruption in a global context. *Journal of Business Ethics*, 121:255–282, 2014.

13 April, 13:30 - 14:10, Auditorium atrium

Desires, fears and degree of satisfaction with life of young students of secondary education in a county in the interior of Portugal

Paulo Infante[1], **Anabela Afonso**[1], **Gonçalo Jacinto**[1], **Rosalina Pisco Costa**[2], **José Conde**[3], **Luísa Policarpo**[3]
[1] CIMA/IIFA and DMAT/ECT, Universidade de Évora, pinfante@uevora.pt
[2] CICS.NOVA.UÉvora and DSOC/ECS, Universidade de Évora
[3] Secção de Juventude e Desporto, Câmara Municipal de Évora

The Municipality Évora (a town from the interior of Portugal) is preparing a Municipal Youth Plan that will allow, on the one hand, to respond to the various challenges to youth; on the other hand, to plan the development and implementation of more innovative youth policies of a global and transversal nature. Based on a questionnaire survey applied to a random sample of secondary schools students in the municipality of Évora, we present some factors that lead to a greater satisfaction with life, characterize some ideas for the future of these students and study some associations and correlations between some experiences that they desire and fear for the next 10-15 years.

Keywords: associations, correlation, logistic regression, young people

The Municipality of Évora is preparing the Municipal Youth Plan, a document that aims to plan the development and implementation of innovative youth policies with a global and transversal character. Given the differentiated reality of young people, a questionnaire survey was designed specifically for the population aged 15-29 in the municipality of Évora. The general objective of the study is to characterize different dimensions of the life of young people who study, work or live in the municipality of Évora. The main specific objectives are: (i) to outline the socio-demographic profile of young people in the municipality of Évora; (ii) to describe ways of school participation (and also professional insertion, when applicable according to age); (iii) to characterize socio-cultural practices; (iv) characterize civic intervention practices; (v) to identify risk behaviors; (vi) to know the level of satisfaction with life and ideas for the future.
In this work the subpopulation under analysis comprises the secondary school students in the municipality of Évora. The students were selected through a multistage probabilistic sampling process. In each secondary school, for each year, groups of classes were randomly selected. The questionnaire was answered by the secondary school students of the selected groups who had their informed consent signed by the parents (in case of minority of the students). The data were collected through a face to face questionnaire, applied between

October and November 2017 in secondary schools in the municipality of Évora and in the Professional School of the Alentejo Region. The application was authorized by the Schools Diretors and by the Direcção Geral de Educação. It was surveyed a sample of 674 students representative of the population. Overall, the questionnaire had a high level of adherence of the respondents, with a response rate of over 98% for almost all questions.
Results shows that almost all students are satisfied or very satisfied with life and about half of them indicated at least 8, on a scale of 0 to 10 (10 represents the maximum satisfaction). Based on a logistic regression model, the obtained results seem to support the idea that the profile that maximizes the probability of a secondary school student being very satisfied with life is based on two main dimensions, one respecting to the present situation and the other with respect to the future. On the one hand, at the present time, this profile gives an account of a young person who seems to be well integrated, from an academic, family and social point of view. School performance is positive, in terms of sociability such student prefers to take advantage of his free time rather than being alone and enjoy being with the family, an indirect indicator of positive family integration. The integration between peers seems to be equally positive and salutary, and those students shows no signs of negative treatment due to his personality. He does not regularly or occasionally consume cannabinoids and derivatives. He does not take too much medication without a prescription. On the other hand, this student seems to have a clear orientation towards the desired future. In the horizon of the next 10-15 years, the family projects are crossed by the desire for a stable relationship to be achieved through marriage, and also the professional projects, which confirmation becomes more visible with the increase on his fear of not come to be professionally recognized.
In the next 10-15 year, to have health, to be happy in life, to have a stable job and to have a stable relationship are the experiences that almost all secondary school students wish to see accomplished. To have children or to get married are the experiences with the highest percentage of students who said they did not want to happen.
The death of a significant other, unemployment and being unhappy in life are experiences that students (at least 3 out of 4) most fear in the next 10-15 years. The experiences that students least fear are divorce (in which 1 in 3 students are not afraid of iy) and not to be professionally recognized or experiencing political instability (in which only about 1 in 3 students are very much afraid of).
Some significant relations between desires and fears for the next 10-15 years and other important variables, such as gender and self-assessment of student performance and leisure time are also presented. Finally, we study the correlation between the experiences that the students fear and want to see happen in the next 10-15 years.

Acknowledgements This work is partially funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia under the projects «UID / MAT / 04674/2019 (CIMA)» and «UID / SOC / 04647/2019 (CICS.NOVA)».

13 April, 13:30 - 14:10, Auditorium atrium

Handling overdispersion count data

Susana Faria[1]
[1] Universidade do Minho, Departamento de Matemática e Aplicações, CBMA – Centro de Biologia Molecular e Ambiental, Guimarães, Portugal, sfaria@math.uminho.pt

---

Poisson regression models are widely used in the analysis of count data. However, it is well known that count regression data often exhibit overdispersion or extra-Poisson variation, i.e, a situation where the variance of the response variable exceeds the mean. Several regression models have been proposed in literature to handle overdispersed count data.

In this work, different regression models will be discussed and applied on different sets of overdispersed count data.

Keywords: Count data, generalized Poisson regression model negative binomial regression model, overdispersion

---

For count data, Poisson regression models have been widely used to explain the relationship between the outcome variable of interest and a set of explanatory variables.
A major drawback of Poisson regression is the model restricts the variance of the data to be equal to the mean, conditional on explanatory variables. This equal mean-variance relationship rarely occurs in observational data and in most cases, the observed variance is larger than the mean, which is called overdispersion.
Various reasons, e.g. missing covariates or interactions, neglected or unobserved heterogeneity, violations in the distributional assumptions of the data, outliers in the response variable or correlation between responses, make counts overdispersed (see [2]).
Two main problems are associated with overdispersion: a possible loss of efficiency in the estimations under different conditions and incorrect inferences on the regression parameters (a variable may appear to be a significant predictor when it is in fact not significant) (see [4]).
To model overdispersion, many alternatives to Poisson regression models have been suggested in literature. Among them, we consider the negative binomial regression model [2] (which have been approached frequently to model overdispersion) and the generalized Poisson regression model introduced by Consul and Famoye [1].
In this paper, we model the occurrence of daily road accidents in Britain applying Poisson regression models. However, the models developed show an overdispersion problem and the alternatives are negative binomial and generalized Poisson regression models.

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

References

[1] P. C. Consul and F. Famoye. Generalized poisson regression model. *Comm. Statist. Theory Methods*, 21 (1):89–109, 1992.

[2] J. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, 2012.

[3] J. Hinde and C. Demétrio. Overdispersion: Models and estimation. *Comput. Statist. Data Anal.*, 27 (2):151–170, 1998.

[4] A. Quintero-Sarmiento, E. Cepeda-Cuervo, and V. Núñez-Antón. Estimating infant mortality in colombia: some overdispersion modelling approaches. *Journal of Applied Statistics*, 39(5):1011–1036, 2012.

Author Index

SPONSORS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

BANCO DE PORTUGAL
EUROSISTEMA

PSE
PRODUTOS E SERVIÇOS DE ESTATÍSTICA

DÃO
VINHOS WINES


Que Viseu?

ADEGA COOPERATIVA
SILGUEIROS

MUNICÍPIO DE
VISEU

Confeitaria
Amaral

Museu Nacional
Grão Vasco