PROGRAM and BOOK of ABSTRACTS

JOCLAD2021

9 - 11 DECEMBER

COVILHÃ, PORTUGAL

UNIVERSIDADE BEIRA INTERIOR

CLAD
Associação Portuguesa de Classificação e Análise de Dados

XXVIII MEETING OF THE PORTUGUESE ASSOCIATION FOR CLASSIFICATION AND DATA ANALYSIS

XXVIII JORNADAS DE CLASSIFICAÇÃO E ANÁLISE DE DADOS

Program and Book of Abstracts

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

9–11 December 2021

Covilhã, Portugal

www.joclad.ipt.pt/joclad2021/

Sponsors

Associação Portuguesa de Classificação e Análise de Dados
Universidade da Beira Interior
Instituto Nacional de Estatística
Banco de Portugal
Câmara Municipal da Covilhã
Quinta dos Termos
Queijos Braz
Natura IMB Hotels

Program and Book of Abstracts

XXVIII Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2021)

Editors: José G. Dias, João Cordeiro, M. Paula Brito, Célia Nunes, Sebastião Pais, Conceição Rocha, M. Eugénia Ferrão

Publisher: CLAD

ISBN 978-989-98955-8-4

Preface

Welcome to JOCLAD 2021! The JOCLAD 2021 – Meeting of the Portuguese Association for Classification and Data Analysis aims to bring together researchers and professionals in the field of Data Science. This is already the twenty-eighth meeting of the CLAD. After many meetings throughout Portugal - 2017 in Porto, 2018 in Almada, 2019 in Viseu, 2020 in Lisbon; this year, from 9 to 11 December, JOCLAD 2021 takes place at the University of Beira Interior (UBI), which co-organizes the event. UBI is situated in the beautiful and historical city of Covilhã, near Serra da Estrela, the tallest Portuguese mountain ($\approx$ 2000m). The city roots are anchored to the old traditions of wool production, going back at least to the XII century. Covilhã has just been classified as "Creative City of Design" by UNESCO. From the local organizers, receive a warm welcome to our city and university, wishing you a pleasant discovery of the wonders of this region. The JOCLAD 2021 program includes a mini-course, on December 9, on the topic of "Co-clustering", held by Mohamed Nadif (Université Paris Descartes, Paris), three Keynote Sessions - "Efficient Search for Good Neural Data Processors" (Luís Alexandre, Universidade da Beira Interior, Covilhã), "Clustering with Data Embedding" (Mohamed Nadif, Université Paris Descartes, Paris), and "A Regression Perspective of Binary and Multi-Class Support Vector Machines" (Patrick J.F. Groenen, Erasmus School of Economics, Rotterdam). Their talks present a representative cross-section of research in Data Science. In addition, the program includes the Fernando Nicolau Award, five thematic sessions - CLAD 2021 Scholarship, CLAD-SPE, INE, Banco de Portugal, and CLAD Corporate -, 35 oral communications, and 17 posters. A Thematic Session is aimed at students who have received a CLAD scholarship 2021, whose members of the evaluation committee were José G. Dias (Chair), Irene Oliveira, and Victor Lobo. We also thank the organizers of the other Thematic Sessions: Pedro Campos (INE - Instituto Nacional de Estatística), Luís Teles Dias (Banco de Portugal), Conceição Amado and Ricardo São João (CLAD-SPE), and Pedro Campos (CLAD Corporate).

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 reviewed 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 work with their quick and constructive responses. These procedures contribute to strengthening the overall quality of the JOCLAD 2021 program. Also, we thank all the chairs of these sessions.

Our deep gratitude goes to the members of the CLAD Board, especially Conceição Rocha and Pedro Campos, who have so voluntarily devoted their time to supporting

JOCLAD 2021. Last but not least, we are delighted to thank the sponsors for their support. Our institutional sponsors deserve a special mention: Instituto Nacional de Estatística (INE) and Banco de Portugal.

Finally, thank you all for your support in helping us keep our annual meeting going! With your high-quality work, CLAD will continue its tradition of excellence in the advancement of Data Science!

Lisboa, December 2021

Chair of the Scientific Program
José G. Dias

Conference Chair
João Cordeiro

President of CLAD
Maria Paula Brito

Organization

President of the CLAD

Maria Paula Brito

Chair of the JOCLAD 2021

João Cordeiro (Universidade da Beira Interior)

Local Organizing Committee

Célia Nunes (Universidade da Beira Interior, CMA-UBI)

João Cordeiro (Universidade da Beira Interior, INESC TEC)

Maria Eugénia Ferrão (Universidade da Beira Interior, CEMAPRE)

Sebastião Pais (Universidade da Beira Interior, NOVA LINCS)

Conceição Rocha (INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência)

Chair of the Scientific Program Committee

José G. Dias (ISCTE - Instituto Universitário de Lisboa)

Scientific Program Committee

A. Manuela Gonçalves (Universidade do Minho)

Adelaide Figueiredo (Universidade do Porto)

Adelaide Freitas (Universidade de Aveiro)

Ana Lorga da Silva (Universidade Lusófona)

Ana Sousa Ferreira (Universidade de Lisboa)

Anabela Afonso (Universidade de Évora)

Anabela Marques (Instituto Politécnico de Setúbal)

Carla Henriques (Instituto Politécnico de Viseu)

Carlos Ferreira (Universidade de Aveiro)

Carlos Soares (Universidade do Porto)

Catarina Marques (ISCTE - Instituto Universitário de Lisboa)

Conceição Amado (Universidade de Lisboa)

Conceição Rocha (INESC-TEC)

Fátima Salgueiro (ISCTE - Instituto Universitário de Lisboa)

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)

Luís Miguel Grilo (Instituto Politécnico de Tomar)

Manuela Neves (Universidade de Lisboa)

Margarida Cardoso (ISCTE - Instituto Universitário de Lisboa)

Maria Filomena Teodoro (Escola Naval-Marinha Portuguesa)

Paula Vicente (ISCTE - Instituto Universitário de Lisboa)

Paulo Infante (Universidade de Évora)

Pedro Campos (Universidade do Porto)

Pedro Duarte Silva (Universidade Católica Portuguesa)

Rosário Oliveira (Universidade de Lisboa)

Sónia Dias (Instituto Politécnico de Viana do Castelo)

Susana Faria (Universidade do Minho)

Victor Lobo (Universidade Nova de Lisboa)

Contents

Program Overview

Thursday, 9 December

9:30	Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde
10:30	**Mini-course**
12:30	Lunch Time
13:30	**Mini-course** (cont.)
14:30	**Opening Session**
15:00	**Keynote Session I**
16:00	Coffee Break
16:20	**Fernando Nicolau Award**
16:50	**Contributed Session I - Symbolic and Complex data**
18:30	Visit to the *Real Fábrica Veiga*

Friday, 10 December

8:30	Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde
9:00	**Contributed Session II - Data Science applications in Life Sciences**
10:20	**Poster Session I**
10:40	Coffee Break
11:00	**Thematic Session I - Scholarship CLAD 2021**
13:00	Lunch Time
14:00	**Keynote Session II**
15:00	**Thematic Session II - Statistics Portugal**
16:20	Coffee Break
16:40	**Thematic Session III - Banco de Portugal**
18:00	General Assembly of CLAD
20:00	Meeting Dinner – Hotel Puralã

Saturday, 11 December

8:30	Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde
9:00	**Contributed Session III - Clustering**
10:20	**Poster Session II**
10:40	Coffee Break
11:00	**Contributed Session IV - Data Science applications in Social Sciences**
12:20	**Thematic Session IV - CLAD - SPE**
13:20	Lunch Time
14:20	**Thematic Session V - CLAD Corporate**
15:00	**Keynote Session III**
16:00	**Closing Session**
16:15	Coffee Break

Program

Thursday, 9 December

9:30 Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde

10:30 **Mini-course**
Co-clustering
Mohamed Nadif, p. 5

12:30 **Lunch Time**

13:30 **Mini-course** (cont.)

14:30 **Opening Session**

15:00 **Keynote Session I**
Efficient search for good neural data processors
Luís A. Alexandre, p. 9

Chair: João Cordeiro

16:00 **Coffee Break**

16:20 **Fernando Nicolau Award**

Chair: Helena Bacelar-Nicolau

16:50 **Contributed Session I - Symbolic and Complex data**

Chair: Margarida Cardoso

16:50 **Modelling interval-valued data: a clusterwise regression approach**
Sónia Dias, Paula Brito, and Nikhil Suresh, p. 61

17:10 **Multivariate Parametric Analysis of Distributional Data**
Paula Brito and A. Pedro Duarte Silva, p. 63

17:30 **Qualitative-quantitative synergy**
Ana Lorga da Silva and Artur Parreira, p. 65

18:30 **Visit to the** *Real Fábrica Veiga*

Friday, 10 December

8:30 Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde

9:00 **Contributed Session II - Data Science applications in Life Sciences**
Chair: A. Manuela Gonçalves

9:00 **Adapting the sampling design of research surveys to improve the biomass estimation of non-target species - the case study of *Raja clavata***
Daniela Silva, Raquel Menezes, Ivone Figueiredo, Bárbara Serra-Pereira, and Manuela Azevedo, p. 67

9:20 **Estimating echolocation clicks rates in narwhals (Monondon monoceros)**
Diana Marques, Tiago Marques, Susanna B. Blackwell, Mads Peter Heide-Jørgensen, and Carolina Marques, p. 69

9:40 **A new selection index to perform polyclonal selection in ancient grapevine varieties**
Sónia Surgy, Jorge Cadima, and Elsa Gonçalves, p. 71

10:00 **Waste management on subsurface ships: a case study**
M. Filomena Teodoro, Suzana Lampreia, and Tomás Mendes, p. 73

10:20 **Poster Session I**
Chair: Sebastião Pais

- **Statistical Modeling of User Influx to the HESE's Emergency Service**, Loide Ascenso, Hugo Quintino, Paulo Infante, and Gonçalo Jacinto, p. 93
- **Quality control techniques for tidal data in near real time**, Dora Carinhas, Margarida Alves, Paulo Infante, and António Martinho, p. 95
- **Time series synchronization: An application to COVID-19 data**, José G. Dias, p. 97
- **A Case Study of Pavement Texture Performance using Linear Mixed Models**, Adriana Santos, Susana Faria, and Elisabete Freitas, p. 99
- **Comparing Time Series Models for Forecasting Meteorological Data**, A. Manuela Gonçalves, Marco Costa, and Cláudia Costa, p. 101
- **Determinants for the existence of victims in road accidents in the district of Setúbal**, Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Rodrigo Cesar, Pedro Nogueira, Marcelo Silva, Vitor Nogueira, José Saias, Paulo Quaresma, Patrícia Gois, and Paulo Rebelo Manuel, p. 103
- **Likelihood function approximation through the delta method in mixed SDE models**, Nelson T. Jamba, Patrícia A. Filipe, Gonçalo Jacinto, and Carlos A. Braumann, p. 105
- **The impact of e-government on sustainable development: a logit model**, Cristina Lopes and Conceição Castro, p. 107

10:40 **Coffee Break**

11:00 **Thematic Session I - Scholarship CLAD 2021**

Chair: José G. Dias

11:00 **Bias in Citizen Science: an application to the BioDiversity4All project**
João Alves, Guilherme Correia, Ana Almeida Matos, Ana Subtil, Francisco C. Santos, and M. Rosário Oliveira, p. 17

11:20 **The Shape of Collaboration in Biodiversity Monitoring**
Guilherme Correia, João Alves, Ana Almeida Matos, Ana Subtil, M. Rosário Oliveira, Patrícia Tiago, and Francisco C. Santos, p. 19

11:40 **Modeling the cost associated with body injuries in workplace accidents**
Ana Moreira, Filipe Gonçalves, Luís Maranhão, and Susana Faria, p. 21

12:00 **The use of the EM and the CEM algorithms in variable selection for mixtures of linear models with random effects**
Luísa Novais and Susana Faria, p. 23

12:20 **Short-Term Forecast Models for Meteorological Variables**
F. Catarina Pereira, p. 25

12:40 **Spatio-temporal variability of distribution and abundance of sardine in Portuguese continental coast: environmental effects**
Daniela Silva, Raquel Menezes, Ana Moreno, Ana Teles-Machado, and Susana Garrido, p. 27

13:00 **Lunch Time**

14:00 **Keynote Session II**
Clustering with data embedding
Mohamed Nadif, p. 11

Chair: Paula Brito

15:00 **Thematic Session II - Statistics Portugal**
Challenges in Official Statistics X

Chair: Pedro Campos

15:00 **Second Child: An Uncertain Transition**
Rita Brazão Freitas, Andreia Maciel, and Maria Filomena Mendes, p. 31

15:20 **Covid-19 pandemic: an unprecedented shock to consumer confidence**
Ana Raquel Marques, p. 33

15:40 **Data analysis during Covid time: the e-invoice case**
João Poças and Sofia Rodrigues, p. 35

16:00 **Inter-organizational networks of EuroGroups Register - a Supervised Clustering Algorithm for Network Data**
Bárbara Santos and Pedro Campos, p. 37

16:20	**Coffee Break**

16:40	**Thematic Session III - Banco de Portugal** **New tools for outliers detection in big databases** Chair: Luís Teles Dias
16:40	**Using Isolation Forest in the quality control of the securities database of Banco de Portugal** André Fernandes and Rafael Figueira, p. 41
17:00	**Machine Learning models applied to ITENF data quality control** C. Ferreira, F. Fonseca, L. Pinto, J. A. Neves, N. Azevedo, and V. Lopes, p. 43
17:20	**An application of cluster analysis to the interest rates reported to the Portuguese Central Credit Register** André Costa, Francisco Fonseca, and Susana Maurício, p. 45

18:00	General Assembly of CLAD

20:00	**Meeting Dinner** – Hotel Puralã

Saturday, 11 December

8:30 Registration – Hall of Anfiteatros Verde and Azul, Faculdade de Ciências da Saúde

9:00 **Contributed Session III - Clustering**

Chair: Fernanda Sousa

9:00 **Sparse Divisive Feature Clustering**
Ndèye Niang, Mory Ouattara, and Gilbert Saporta, p. 75

9:20 **Clustering domestic water consumption profiles: a Portuguese case study**
Elisa Araújo, Flora Ferreira, Duarte Silva, Estela Bicho, and Wolfram Erlhagen, p. 77

9:40 **Clustering times series of electricity consumption**
Margarida G. M. S. Cardoso, Ana Martins, and João Lagarto, p. 79

10:00 **Clustering of EFCs in European Economies**
Eliana Costa e Silva, Aldina Correia, and Ana Borges, p. 81

10:20 **Poster Session II**

Chair: Célia Nunes

- **A hurdle-gamma regression model for the average number of undeveloped pine nuts per cone**, Anabela Afonso, Dulce G. Pereira, and Ana Cristina Gonçalves, p. 109
- **How the indicators for EU countries in the period 2010-2019 are approaching targets of Europe 2020 agenda?**, Adelaide Figueiredo and Fernanda Otilia Figueiredo, p. 111
- **A state space framework for daily temperature forecasting**, A. Manuela Gonçalves, F. Catarina Pereira, and Marco Costa, p. 113
- **A sociological portrait of the Portuguese, based on their religiosity and values: A cross-time comparison with Europe**, Maria Paula Lousão, Cláudia Vasconcelos Silvestre, and José Luís Casanova, p. 115
- **A longitudinal analysis of the severity of road accidents in the district of Setúbal between 2016 and 2019**, Paulo Infante, Anabela Afonso, Gonçalo Jacinto, Leonor Rego, Pedro Nogueira, Marcelo Silva, Vitor Nogueira, José Saias, Paulo Quaresma, Patrícia Gois, and Paulo Rebelo Manuel, p. 117
- **Data, Technology and Journalism**, Cláudia Silvestre and Pedro Frazão, p. 119
- **Field test validation for predicting VO2max in the Portuguese Military Academy**, Rui Lucena, Lucas Nogueira, Nuno Almeida, Cristiano Almeida, and Paula Simões, p. 121
- **Mechanical Behavior of Skin: an ANOVA Approach**, M. Filomena Teodoro and Teresa Oliveira, p. 123
- **The Relationship of the Human Capital Index with the Level of Education and the Adult Survival Rate**, Alexandra Marques, Ana Pinheiro, Maria Carolina Matos, Cristina Torres, Cristina Lopes, and Isabel Vieira, p. 125

10:40 **Coffee Break**

11:00	**Contributed Session IV - Data Science applications in Social Sciences** Chair: Adelaide Figueiredo
11:00	**Visually supported curriculum development** Rogério Duarte, Ângela Lacerda Nobre, Fernando Pimentel, and Marc Jacquinet, p. 83
11:20	**How to analyze online behavior as a source for political information in the Portuguese 2019 European Parliament election?** Cláudia Silvestre, Rodrigo Pinheiro, and Filipe Montargil, p. 85
11:40	**The role of the pre-university life path in the performance of students who access higher education: the case study of the Master's Degree in Civil Engineering at FEUP** Fernanda Campos de Sousa and Isabel Martins Ribeiro, p. 87
12:00	**Data Analysis to Advance Immersive Technologies: examples from Augmented Reality-based assembly procedures** Carlos Ferreira, Bernardo Marques, João Alves, Paulo Dias, and Beatriz Sousa Santos, p. 89
12:20	**Thematic Session IV - CLAD - SPE** Chairs: Conceição Amado and Ricardo S. João
12:20	**Prediction Models in Medicine** Ana Luisa Papoila, p. 49
12:40	**Is age at menopause decreasing? The consequences of not completing the generational cohort** Rui Martins, Bruno de Sousa, Thomas Kneib, Nadja Klein, Maike Hohberg, Elisa Duarte, and Vítor Rodrigues, p. 51
13:00	**Analysis of cutoff point estimation for determining seropositivity in the context of SARS-CoV-2 infections** Tiago Dias Domingues, Helena Mouriño, and Nuno Sepúlveda, p. 53
13:20	**Lunch Time**
14:20	**Thematic Session V - CLAD Corporate** Chair: Carlos Ferreira
	How to Measure the Outdoor Advertising Audience and the Actual Mobility in Portugal? Paulo Caldeira and João Pequito, p. 57
15:00	**Keynote Session III** **A Regression Perspective of Binary and Multi-Class Support Vector Machines** Patrick J.F. Groenen, p. 13 Chair: A. Pedro Duarte Silva
16:00	**Closing Session**
16:15	**Coffee Break**

Abstracts

Mini-Course

9 December, 10:30 - 12:30, 13:30 - 14:30

Co-clustering

Mohamed Nadif

Université de Paris, CNRS Centre Borelli, 75006 Paris, mohamed.nadif@u-paris.fr

---

In the era of data science, clustering various kinds of objects (documents, genes, customers) has become a key activity and many high quality packaged implementations are provided for this purpose by many popular packages. A natural extension of standard cluster analysis is co-clustering where objects and features are simultaneously grouped into meaningful blocks called co-clusters or biclusters, thus making large datasets easier to handle and interpret. In fact, co-clustering has found applications in many areas such as bio-informatics web mining, text mining and recommender systems. Various co-clustering algorithms have been proposed over the years. The goal of the mini-course is to review popular different approaches to perform co-clustering such as matrix factorization based methods, spectral methods, and model-based methods. Models and algorithms will be presented and illustrated.

Keynote Sessions

9 December, 15:00 - 16:00

Efficient search for good neural data processors

Luís A. Alexandre
NOVA LINCS and Departamento de Informática, Universidade da Beira Interior, Covilhã, Portugal, luis.alexandre@ubi.pt

---

Deep learning approaches have revolutionized most application areas where data needs to be automatically processed. One difficulty that was overcome by such methods was the feature engineering process that became automated, but unfortunately we are now faced with the task of designing ever more complex networks to deal with our data. This design process is not guided by strong theoretical constraints, but ends up being a sort of an "art", reminding us of the times where such an art was required for the, now gone, feature engineering task.

A recent trend tries to automate the full machine learning process, AutoML. A particular subset of AutoML, called Neural Architecture Search (NAS), focus specifically in automatically finding good deep learning networks. Once again such efforts face challenges, such as the need of specific knowledge for developing such models and the heavy computational burden they entail, with some solutions requiring months or even years of GPU computations to find good architectures.

In this talk we will give a brief overview of the area of NAS and discuss ways to optimize the search for good architectures, that 1) do not require specific knowledge of neural architecture details and 2) can make these approaches practically feasible, allowing for the search of a good architecture in a mater of minutes using a single desktop machine.

Keywords: Neural Architecture Search, Deep Neural Networks, Supervised Learning

10 December, 14:00 - 15:00

Clustering with data embedding

Mohamed Nadif
Université de Paris, CNRS Centre Borelli, 75006 Paris, mohamed.nadif@u-paris.fr

Low-dimensional embedding techniques are particularly well suited to embedding high-dimensional data into a space that in most cases will have just two dimensions. Low-dimensional space, in which data samples (data points) can more easily be visualized, is also often used for learning methods such as clustering. Sometimes, however, this space does not necessarily allow to reveal a clustering structure. This contribution reviews recent work in the area of many approaches to address this issue in proposing a simultaneous learning approach for data embedding and clustering that reinforces the relationships between these two tasks.

Keywords: unsupervised learning, data embedding, clustering

1 Introduction

Nowadays, many real-world data sets are high-dimensional. Low-dimensional embedding methods can be used to map a set of high-dimensional data into a low-dimensional space while preserving the intrinsic structure of the data. Principal component analysis (PCA) is the most popular linear approach. In nonlinear cases, however, there are other more efficient approaches to be found [7, 1]. Combining data embedding and clustering is one of the most promising approaches for unsupervised learning. Indeed, in cluster analysis it is generally expected that a clustering of samples can be represented visually in a two-dimensional form. For this purpose, data embedding followed by cluster analysis is often helpful in data science, and `k-means` applied on data embedding, derived from principal components analysis (`PCA`), is the most popular approach. This procedure is carried out sequentially and is referred to as *the tandem approach*. However, `PCA` can be an unsuitable method for reducing the dimension before clustering, as discussed in [5]; it can fail to retain valuable information concerning the clustering that is often contained in components with smaller eigenvalues. Moreover, since `PCA` provides an embedding only for data lying on a linear manifold, the tandem approach fails to cluster data that lie on a curved manifold.

This potential weakness of the tandem approach is now well known, and several authors from both the machine learning and the statistics communities have proposed procedures for executing the two tasks simultaneously. To address this issue, among the popular solutions that have been developed are projection pursuit [4], clustering and disjoint `PCA`

[11], reduced `k-means` [6], factorial `k-means` [10], and a variant, known as `Clust-PCA`, which is tradeoff between reduced and factorial `k-means` that helps to overcome the individual drawbacks of the two methods [12]. In the talk we will see how the weakness of the tandem approach may be also overcome using recent approaches which alternate iteratively between embedding and clustering such as [2, 13, 3, 8, 9].

References

[1] S. Affeldt, L. Labiod, and M. Nadif. Spectral clustering via ensemble deep autoencoder learning (SC-EDAE). *Pattern Recognition*, 108:107522, 2020.

[2] K. Allab, L. Labiod, and M. Nadif. A Semi-NMF-PCA unified framework for data clustering. *IEEE Trans. Knowl. Data Eng.*, 29(1):2–16, 2017.

[3] K. Allab, L. Labiod, and M. Nadif. Simultaneous spectral data embedding and clustering. *IEEE transactions on neural networks and learning systems*, 29(12):6396–6401, 2018.

[4] HH Bock. On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In *Multivariate statistical modeling and data analysis*, pages 17–34. Springer, 1987.

[5] W.C. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, pages 267–275, 1983.

[6] G. De Soete and J.D. Carroll. K-means clustering in a low-dimensional euclidean space. In *New Approaches in Classification and Data Analysis*, pages 212–219. 1994.

[7] D. Engel, L. Hüttenberger, and B. Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization. volume 27, pages 135–149. Schloss Dagstuhl, Leibniz-Zentrum fuer Informatik, 2012.

[8] M. M. Fard, T. Thonet, and E. Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138:185–192, 2020.

[9] L. Labiod and M. Nadif. Efficient regularized spectral data embedding. *Advances in Data Analysis and Classification*, 15(1):99–119, 2021.

[10] M. Vichi and H. Kiers. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37(1):49–64, 2001.

[11] M. Vichi and G. Saporta. Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, 53(8):3194–3208, 2009.

[12] M. Yamamoto and H. Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, 41(1):115–129, 2014.

[13] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, volume 70, pages 3861–3870. PMLR, 2017.

11 December, 15:00 - 16:00

A Regression Perspective of Binary and Multi-Class Support Vector Machines

Patrick J.F. Groenen[1], **Gertjan van den Burg**[2]
[1] Econometric Institute, Erasmus University Rotterdam, groenen@ese.eur.nl
[2] Alan Turing Institute, gertjanvandenburg@gmail.com

Support vector machines are useful for binary and multiclass classification. However, their explanation in machine learning is often done through a dual formalization. Here, we stick to regression perspective. Also, we discuss GenSVM that is a flexible multiclass generalization of the binary SVM. We sketch the outline of their algorithms based on majorization.

Keywords: Classification, Support Vector Machine (SVM), Majorization, MM, CCCP

Support vector machines (SVMs) have become a standard tool for binary classification problems that has become increasingly popular. In the machine learning literature, the SVM is often explained through the dual of a convex optimization problem. Instead, we approach it as a regression problem with a specific error function and a ridge type penalty term [1].

However, in the case that more than two classes need to be predicted often a series of binary SVMs are performed (one-versus-all or between all pairs of classes, one-versus-one). A disadvantage of such methods is that they are heuristics that do not simultaneously estimate all parameters in a single model. We discuss a recent multiclass SVM loss function (GenSVM) that is based on a geometric representation of each class by a vertex of a simplex in $K-1$ dimensional space [2]. As with the binary SVM, an object that is predicted to be nearest to its class receives a zero error and if the object is closer to another class the error consists of a function of the distance to the zero-error region. The present approach is flexible in the hinge function that is used for calculating the error. It builds on the Huberized hinge errors that have as special cases the linear and quadratic hinges. It is also flexible in how these errors are added: we propose to use the L_p norm of the Huberized hinge error. This general loss function has the binary SVM and some existing multiclass SVM loss functions as special cases.

We discuss a majorization algorithm (also named MM or CCCP) that minimizes GenSVM. We present some numerical comparisons showing that for medium sized problems GenSVM compares with the best approaches.

The majorization algorithm for the binary SVM is implemented in SVMMaj [4] R-package and the multiclass SVM in the gensvm R-package [3].

References

[1] Patrick J. F. Groenen, Georgi Nalbantov, and Jan C. Bioch. SVM-Maj: A majorization approach to linear support vector machines with different hinge errors. *Advances in Data Analysis and Classification*, 2(1):17–43, 2008.

[2] Gertjan van den Burg and Patrick J. F. Groenen. GenSVM: A generalized multiclass support vector machine. *Journal of Machine Learning Research*, 17:1–42, 2016.

[3] Gertjan van den Burg and Patrick J. F. Groenen. *GenSVM: A Generalized Multiclass Support Vector Machine*, 2018. R package version 0.1.5.

[4] Hoksan Yip, Patrick J.F. Groenen, and Georgi Nalbantov. *SVMMaj: Implementation of the SVM-Maj Algorithm*, 2019. R package version 0.2.9.

Thematic Session I: CLAD 2021 Scholarship

10 December, 11:00 - 11:20

Bias in Citizen Science: an application to the BioDiversity4All project

João Alves[1], **Guilherme Correia**[1], **Ana Almeida Matos**[1], **Ana Subtil**[1], **Francisco C. Santos**[1], **M. Rosário Oliveira**[1]
[1] Instituto Superior Técnico, Av. Rovisco Pais, 1 1049-001 Lisboa, Portugal

---

Citizen Science projects are growing in number and users, and their data can drive research in numerous fields, such as studies related to biodiversity monitoring and analysis of species distributions. However, most of these projects share a common challenge: the quality of their data depends heavily on their user base's preferences, skills, and observation efforts. To engage a large number of participants, Citizen Science projects raise their accessibility by lowering restrictions and simplifying methods of collecting data, which can lead to biased data. In this work, we provide an analysis of BioDiversity4All's observations. We use a Hurdle model to describe the distribution of observation counts to find possible explanatory variables that influence this distribution. Our results suggest that geographical accessibility is one of the most critical factors for the observers. We also used a SARIMA model to evaluate the impact that the recent Covid-19 restrictions had on BioDiversity4All's observation counts. Our results suggest that the BioDiversity4All project did not suffer from this pandemic; in fact, we observe a substantial increase in observations submitted to the platform during this period, including during lockdown periods.

Keywords: Citizen Science, Spatial and Temporal bias, Hurdle Model, Time Series forecasting, Covid-19

A Citizen Science project consists in a collaboration between researchers, scientists and the general community, with the common goal of gathering relevant scientific information that can later be used in research or education [3]. A project's participation rate and with it, its success, depends heavily on public engagement. In order to keep current volunteers interested and to capture the attention of new ones, project managers need to lower the project's restrictions and simplify protocols for gathering data. These methods prolong a project's longevity, but with lower restrictions and untrained volunteers, the bias present in gathered data may increase [1].
In this work, we aim to study the observations present in BioDiversity4All,a Portuguese Citizen Science platform that started in 2010 and contains species' observations recorded in Portugal, and understand the impact that different geographical variables have on their distribution. At the time of producing this work, the world fell victim to the Covid-19 pandemic outbreak. In Portugal, confinement laws were established and people were

restricted from going outside. Therefore, we also studied the impact that these pandemic restrictions had on BioDiveristy4All and its observations.

First, we set out to gain some insight on the distribution of observation counts in BioDiversity4All. We decided to create a visual representation of the distribution of observation counts across mainland Portugal, as well as a plot of these counts, grouped by their respective taxonomic groups. This analysis suggests that there may be some bias present in BioDiversity4All's observations. Some taxonomic groups are over represented, and the observations do not seem to be evenly distributed across the country. This may be explained due to the fact that Citizen Science data heavily depends on its participants, their preferences and observational effort.

Then, we set out to understand the relationship between certain geographical variables and the distribution of the observation counts. To this end, we collected several possible explanatory variables and decided to fit Hurdle models [4] under five different scenarios: A scenario where we fitted Hurdle models to the total observation counts, and four scenarios where we fitted Hurdle models for each yearly season. Our results suggest that, for all scenarios, accessibility and higher populated areas seem to be essential factors for users recording their observations. During Spring, areas that have less precipitation and higher artificial area to wetland and water area ratio, tend to be preferred by users. During Summer, underpopulated areas, such as wetlands and water surface areas and forest and scrubland areas, seem most preferred. During Autumn, observations are more likely to occur inside urban areas. Finally, during Winter, urban and warmer areas are most chosen, while areas with higher values of altitude and precipitation tend to be less preferred.

Finally, we intended to evaluate the impact that pandemic restrictions had on BioDiversity4All's observations counts. We decided to analyse the weekly counts as a time-series that ended in 31 Dec 2019. We used a SARIMA model [2] to perform a forecast for counts during the pandemic restrictions, and then we compared the predicted counts with the real counts, obtained from the platform. Our results suggest that, despite the restrictions imposed during Covid-19, it seems that there was no impact in the submission of observations. In fact, they suggest the opposite: We have shown that there was considerable growth in the number of observations during this period, including during national lockdown.

References

[1] Jeffrey P. Cohn. Citizen Science: Can Volunteers Do Real Research? *BioScience*, 58(3):192–197, 03 2008.

[2] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.

[3] Jonathan Silvertown. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467 – 471, 2009.

[4] Alain Zuur, EN Ieno, Neil Walker, Anatoly Saveliev, and GM Smith. *Mixed Effects Models and Extensions in Ecology With R*, volume 1-574. Springer, 01 2009.

10 December, 11:20 - 11:40

The Shape of Collaboration in Biodiversity Monitoring

Guilherme Correia[1], **João Alves**[1], **Ana Almeida Matos**[1], **Ana Subtil**[1], **M. Rosário Oliveira**[1], **Patrícia Tiago**[2], **Francisco C. Santos**[1]
[1] Instituto Superior Técnico, Universidade de Lisboa, Portugal,[2] Faculdade de Ciências, Universidade de Lisboa, Portugal

---

There has been a substantial increase in interest in citizen science, spanning a wide range of areas from biodiversity to water and air quality monitoring. These platforms are particularly efficient in monitoring tasks impossible to be handled by small teams of experts. The acquisition and validation of information emerge as a cooperative effort grounded on large self-organized networks of participants. Despite this, little is known about the structural patterns of these networks of collaboration. Here, we provide a preliminary analysis of a representative collaborative network of a major citizen science platform aiming at mapping and sharing observations of biodiversity. We show that the resulting temporal collaborative network exhibits a power-law dependence on the connectivity that outlasts the entire period investigated, despite significant differences in the number of participants throughout the years. This result suggests the existence of time and scale-invariant topological properties in citizen science platforms. We further show that these collaboration networks portray a well-defined community structure associated with users' taxon preferences. Finally, we show that each participant's role or type of participation tend to evolve in time — the longer at the network, more likely the adoption of the role of Validator of others' observations. The methodology developed here demonstrates the possibility of analyzing, comparing and potentially shaping the time-evolution of social networks associated with collaborative science platforms.

Keywords: Citizen Science, Network Science, Cooperation

Online citizen science projects, such as iNaturalist have hundreds of thousands and even millions of users cooperating to acquire and validate large amounts of information [3, 2, 4]. In these platforms, each user is connected to others she/he cooperated with in a representative self-organized network. However, while citizen science projects (CSPs) are evolving into massive *networks* of users, it remains illusive how volunteers interact to organize into such structures. Here, we perform a *network* analysis on BioDiversity4All [1] - a Portuguese citizen science project started in 2010.
In this platform, users report observations on organisms at a particular time and location [1]. These observations can be identified and commented by other users, creating a network

between the users and the observations they have participated in - we named this network BipartiteCoop. From this network, it is possible to extract the collaboration network of users that have cooperated in the identification of the same observation - we named this network CoopNet.
By analyzing these two networks we found that users organize into a scale and time invariant network, suggesting that despite future variations in the number of users, the platform will always present a power law dependence on the connectivity. Next, we showed that some users, the hubs, bind the structure together, suggesting that the BioDiversity4All highly depends on these to keep users connected - should the platform take measures to keep the hubs engaged?
Regarding community partitioning, users tend to group according to taxon interests and knowledge. This result may be explained by the fact that users with the same interests tend to participate in the same observations, thus becoming more connected and, consequently, the partitioning method groups these users in the same community.
Another interesting result relates to user behaviour, most users tend to either only create or only validate observations, suggesting that users have a strong preference towards the way that they enjoy the BioDiversity4All platform.
Regarding behaviour evolution in the BioDiversity4All, the results show that new users have a very high tendency to create observations rather than validate. In contrast, users that have been in the network longer tend to validate more, while still creating observations.
Overall, the achievements of this work have high implications for the study of citizen science. While formulating a methodology for studying volunteers' interaction and behaviour, we have shown numerous results using network analysis. Mainly, these results have demonstrated that the BioDiversity4All is a highly connected platform presenting time and scale-invariant topological properties. Also, it exhibits a community structure well defined by its users' interests or knowledge. And, (most) users evolve, starting by creating observations but eventually assuming the role of validators. Furthermore, these results prefigure future activity in the BioDiversity4All platform.

References

[1] BioDiversity4All. *https://www.biodiversity4all.org*, Last accessed 25 January 2021.

[2] The eBird About page. *https://ebird.org/about*, Last accessed 25 January 2021.

[3] The iNaturalist Stats page. *https://www.inaturalist.org/stats*, Last accessed 25 January 2021.

[4] Colin J. Torney, David J. Lloyd-Jones, Mark Chevallier, David C. Moyer, Honori T. Maliti, Machoke Mwita, Edward M. Kohi, and Grant C. Hopcraft. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10(6):779–787, 2019.

10 December, 11:40 - 12:00

Modeling the cost associated with body injuries in workplace accidents

Ana Moreira[1], **Filipe Gonçalves**[2], **Luís Maranhão**[2], **Susana Faria**[3]
[1] University of Minho, anaafmoreira98@gmail.com
[2] Actuarial Department, Group AGEAS Portugal - Porto
[3] Mathematics Department, Universidade do Minho

In this study, statistical models are developed to identify the factors associated with the cost of injuries in workplace accidents, based on historical data provided by the insurance company. In this context, generalized linear models are used, in particular the Gamma and Log-Normal regression models, since the response variable *Despesa Total* is continuous taking only positively skewed non-negative values. Also, a cluster analysis is performed as a complement to the treatment and exploratory data analysis.

Keywords: Actuarial Science, Workplace Accidents (Non-Life Insurance), *Clustering*, Generelized Linear Models

---

Insurance companies have to take many factors into account when estimating insurance premiums and, in that respect, they can choose different strategies regarding certain types of insurances. Within the scope of insurance policies associated with accidents in workplace (insurance that is mandatory for all workers), and thus important for the insurance company to identify which factors may be economically unfavourable.
Thus, the main objectives of this study are to develop, compare and validate statistical models to identify factors associated with the cost of injuries within the context of workplace accidents.
The database used in the modeling consists of 57419 observations (claims) and 15 variables, that refers to accidents in the workplace that occurred between 2015 and 2019.
The response variable in this study is the variable related to the total expenditure (in euros) involved with the injuries suffered by the claimants, resulting from an accident in the workplace (variable *Despesa Total*). The total expenditure may include costs with treatments, hospital surgeries, retainer contracts, transportation, absolute temporary disability (associated with sick leave) and others. This variable is a continuous variable taking only positively skewed non-negative values.
Regarding the methodologies considered, initially, a clusters analysis is performed. This analysis was applied for some categorical variables that present a high number of categories, in order to simplify the statistical models estimated, by reducing the number of categories. Thus, various agglometarive hierarchical methods are applied and compared. The non-hierarchical partitioning method K-means is also applied, using the Elbow method to determine the optimal number of clusters ([1]).

Another methodology considered in this study is the application of generalized linear models, in particular, the Gamma regression model and the Log-Normal regression model (classical linear regression model, with previous logarithmic transformation). The Gamma distribution and the Normal distribution belongs to the exponential family. These models are potential candidates to model the data due to the aforementioned characteristics of the response variable *Despesa Total* and due to the approximation that can be confirmed graphically of the estimated density curves to the respective theoretical density curves. ([2] and [4]).
One proceeds to the modeling of the *Despesa Total* while taking into account the models mentioned above. To access the adequacy of the models, regression analysis is performed. Based on the two models obtained, it is possible to highlight a set of factors that may help the insurance company making decisions about certain types of strategies. In particular, the high importance of the number of sick leave days due to absolute temporary disability in total expenditure is confirmed. For example, for individuals who were on sick leave for more than 3 months, the total expenditure is, on average, approximately 11,1 times higher (Gamma regression model) and 15 times higher (Log-Normal regression model) in comparison to individuals who did not need sick leave.
Lastly, regarding the choice of one of the models, the quality of fit of the models is very similar, as well as the quality of prediction, nevertheless, the Gamma regression model indicates slightly better results, in addition to being the more parsimonious model, as it considers fewer parameters.

References

[1] S. Dolnicar, B. Grün and F. Leisch. *Market Segmentation Analysis.* Springer, Singapore, 2018.

[2] K. P. Dunn and K. G. Smyth. *Generelized Linear Models With Examples in R.* Springer, New York, 2018.

[3] P. Jong and Z. G. Heller. *Generelized Linear Models for Insurance Data.* Cambridge, New York, 2008.

[4] A. Turkman and G. Silva. *Modelos Linerares Generalizados. Da teoria à prática.* Edições SPE, Lisboa, 2000.

10 December, 12:00 - 12:20

The use of the EM and the CEM algorithms in variable selection for mixtures of linear models with random effects

Luísa Novais[1], **Susana Faria**[2]
[1] Centre of Molecular and Environmental Biology and Department of Mathematics, University of Minho, Portugal, luisa_novais92@hotmail.com
[2] Centre of Molecular and Environmental Biology and Department of Mathematics, University of Minho, Portugal, sfaria@math.uminho.pt

---

Variable selection plays an important role in any modelling study, requiring the search for the simplest possible model that adequately describes the observed data. In this work, we study the problem of variable selection for mixtures of linear models with random effects. For this, we resort to penalized likelihood estimation using the Expectation-Maximization (EM) and the Classification Expectation-Maximization (CEM) algorithms and we analyse the performance of the different methodologies through a simulation study.

Keywords: EM algorithm, CEM algorithm, mixture models, penalty methods, variable selection

---

In recent years, the computational advances have led to the use of large and greatly complex data, which makes the classic variable selection methods too computationally challenging to be used in practice with the increasing size of the data. In order to overcome the technological complexity, the need to develop new variable selection methods has emerged in the last few years. Among the new developments, the methods based on penalty functions have played an important role, allowing the identification of the subset of the relevant variables and the removal of the non-relevant variables from the model, by estimating their value to be zero, which considerably decreases the computational load.
In this work, we resort to penalized likelihood estimation via the Expectation-Maximization (EM) and the Classification Expectation-Maximization (CEM) algorithms and we compare four different penalty functions that allow for simultaneously variable selection and estimation: the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), the HARD and the Smoothly Clipped Absolute Deviation (SCAD), through an extensive simulation study.
As such, the goal of the simulation study is to compare the performance of a penalized likelihood approach via the EM and CEM algorithms for variable selection in the presence of a large number of variables for mixtures of linear models with random effects and for different penalty methods. To accomplish it, we analyse the computational effort of the

algorithms by studying the mean number of iterations required to reach convergence, we analyse the statistical properties of the estimators (through the bias and the mean squared error (MSE) of the parameter estimates) and we also analyse the average numbers of the correctly and incorrectly estimated zero coefficients, the sensitivity and specificity of the variable selection and the proportion of correctly estimated models.
Based on the simulation study, we conclude that different simulated scenarios influence the performance of the different algorithms and the different penalty functions. As expected, the CEM algorithm demonstrated a superior performance compared to the EM algorithm with respect to the computational effort, also demonstrating a superior performance for the variable selection, something particularly notorious with the increase in the number of components. In particular, regarding the specificity and the sensitivity of the variable selection, once again the CEM algorithm demonstrated its superiority, denoting very high values of specificity and sensitivity for the generality of the simulated scenarios. Finally, with regard to the four penalty functions, the superiority of the ALASSO penalty was demonstrated in the variable selection, while, in the opposite direction, the HARD penalty demonstrated the worst performance.
In conclusion, variable selection in mixture models is not an easy problem, as proved in our simulation study. The use of the CEM algorithm for variable selection is reccommended in any scenario, given its superiority comparatively to the EM algorithm for all the measures of performance studied. In terms of penalty functions, it is advisable to perform variable selection using the ALASSO, given the good results presented for any of the scenarios under study.

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

References

[1] Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

[2] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1):1–38, 1977.

[3] Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

[4] Luísa Novais and Susana Faria. Selection of the number of components for finite mixtures of linear mixed models. *Journal of Interdisciplinary Mathematics*, pages 1–32, 2021.

10 December, 12:20 - 12:40

Short-Term Forecast Models for Meteorological Variables

F. Catarina Pereira
University of Minho, Department of Mathematics and Center of Mathematics, Portugal, id9976@alunos.uminho.pt

---

Dry periods have been more frequent and prolonged in recent years due to climate change and, therefore, sustainable water management has become essential, particularly in irrigation systems. Thus, we propose state-space models associated with Kalman filter applied to minimum air temperature time series (which have an impact on the evapotranspiration process) to improve the predictive quality of the forecasts. A comparison between state-space and linear regression models is presented.

Keywords: Kalman filter, time series, state-space models, calibration, meteorological variables

---

Population growth, pollution and climate change have contributed to the decrease of water resources on the planet and inefficient management of these resource has become a global concern. According to the report by Associação Natureza Portugal in association with the World Wide Fund for Nature (2019) ([1]), most scenarios point to a reduction of water availability in Portugal caused by the rise of temperature in recent years. It is clear that dry periods cause negative impacts, both at economic level, such as loss of production and reduction of cultivated areas, and at social and environmental levels, as the reduction of the amount and quality of available water resources increases the mortality of many species (loss of fauna and flora). Climate change is, in a global context, an aggravation of the threats associated with natural meteorological phenomena that affect the water cycle and water availability. In addition, agriculture is the largest consumer of water, using on average between 60% and 90% of water resources available annually ([1]). Thus, it has become essential to find the best technical solutions to improve water use efficiency of, particularly in irrigation systems.

This work is carried out in the context of project "The Optimal Challenges in Irrigation", aiming to analyze the behavior of soil humidity through modeling and having as its main objective the efficient management of water resources in irrigation systems.

This work presents a class of models called calibration models, which admit a state-space representation associated with the Kalman filter. The formulation of a problem with this representation intends to evidence a functional, dynamic, and stochastic dependency between components of a system, which can be represented in two equations: the first is called the state equation, which translates the stochastic model underlying the state vector;

the second is called the observation equation, which relates the observed variable to a linear transformation of the state vector plus white noise. The most common procedure to predict the unobservable component, the state, is based on the application of the Kalman Filter, which is a recursive estimation algorithm that allows to obtain both the optimal estimator of the state vector, based on the information available up to time t, when the model is fully specified ([2]), and one-step-ahead predictions by updating and improving the predictions of the state vector, in real time, when new observations become available.
The estimation of the unknown parameters of the proposed state-space model is performed using the maximum likelihood method, assuming normality of the errors. However, even when the assumption of normality of errors is not verified, the Kalman filter still provides optimal estimators within the class of all linear estimators ([3]).
The objective of this study is to improve the accuracy of short-term forecasts in real time (obtained from the `weatherstack.com`), considering a six-day temporal window of minimum temperature forecasts. For this purpose, the statistical analysis was performed using a dataset that includes observations of daily minimum temperature (°C) in the farm Senhora da Ribeira in Bragança between February 20 and October 11, 2019, with a total of 234 observations, and the forecasts obtained from the `weatherstack.com` website.
This study also presents a comparison between state-space and linear regression models, and the results showed that the first approach was able to reduce the root mean square error of the forecasts between 27.88% and 32.34%; the second approach reduced it between 17.02% and 25.72%, compared to the website's initial forecasts, considering the different time horizons.

Acknowledgements This work has received funding from FEDER/COMPETE/NORTE 2020/POCI/FCT funds through grants UID/EEA/-00147/20 13/UID/IEEA/00147/006933-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247. F. Catarina Pereira was financed by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM.

References

[1] ANP/WWF. Relatório da Associação Natureza Portugal em associação com a World Wide Fund for Nature: Vulnerabilidade de Portugal à Seca e Escassez de Água. 2019.

[2] M. Costa and M. Monteiro. Bias-correction of kalman filter estimators associated to a linear state space model with estimated parameters. *Journal of Statistical Planning and Inference*, 176:22–32, 2016.

[3] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 2009.

[4] F. C. Pereira. *Modelos de Previsão a Curto Prazo para Variáveis Meteorológicas*. PhD thesis, Universidade do Minho, 2020.

10 December, 12:40 - 13:00

Spatio-temporal variability of distribution and abundance of sardine in Portuguese continental coast: environmental effects

Daniela Silva[1], **Raquel Menezes**[2], **Ana Moreno**[3], **Ana Teles-Machado**[4], **Susana Garrido**[5],

[1] Centro de Matemática, Universidade do Minho, danyelasylva2@gmail.pt

[2] Centro de Matemática, Universidade do Minho, rmenezes@math.uminho.pt

[3] Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, amoreno@ipma.pt

[4] Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, ana.machado@ipma.pt

[5] Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, susana.garrido@ipma.pt

---

In the last decades, the challenge of sustainability has become a concern due to resource depletion, environmental degradation, and loss of biodiversity. Improving the knowledge of biodiversity and species dynamics represents an essential step for its conservation. This study aims to estimate the spatial distribution of sardine (*Sardina pilchardus*) relating the spatio-temporal variability of fish biomass with environmental conditions using two-part modelling.

Keywords: Environmental effects, Geostatistics, Hurdle model, *Sardina pilchardus*, Species Distribution Model

---

The study of the relationship between the spatial distribution of marine species and environmental changes makes it possible to know the processes regulating the changes in abundance, allowing to develop biomass indicators, to identify potential habitats and to improve the ability to predict trends in the dynamics of these species. This study aims to estimate the abundance and distribution of sardine (*Sardina pilchardus*), relating the spatio-temporal variability of biomass to environmental conditions.

Distribution and abundance data of sardine was obtained from georeferenced fishing hauls estimated during 20 acoustic surveys (PELAGO survey series) conducted every spring (from 2000 to 2020, except 2012). An exploratory analysis allowed the identification of large areas with an absence of the species along the coast and assess the behaviour of sardine biomass over time in the study area. While in the northwestern areas there is an increasing absense of the species over time, the absence remained almost constant in the remaining area. The sectors with higher occurrence do not match the sectors with higher biomass values. These results show that there is spatio-temporal variability in the data,

and that the occurrence and abundance have different behaviours. Daily environmental data was obtained for the region and time of study, particularly satellite-derived sea surface temperature (SST), chlorophyll-a concentration, bathymetry, and surface ocean currents. Species Distribution Models are investigated to relate sardine presence/absence and abundance to the environmental conditions, aiming at predicting sardine distribution in unobserved locations and times. The hurdle model is suitable since allows to incorporate the specificities of the data: complex spatio-temporal dynamics, excess of zeros and the difference between the occurrence and abundance under occurrence processes. Hurdle model is a two-part model where the species abundance is given by the product of two processes: species occurrence and abundance given the occurrence. While occurrence data requires a Bernoulli distribution, Gamma or Lognormal distributions can be used for positive values In addition to considering environmental covariates and the spatio-temporal structure, the impact of environmental conditions with time lag on biomass indicator is evaluated. Data from the west and south Iberian coasts are studied separately due to the shape of the coast and the different oceanographic conditions. For the south coast, all available environmental variables are shown to be important to explain the sardine presence, while SST and bathymetry help to explain the variability of biomass in positive stations. The occurrence is higher in shallow (up to 100 m), colder (up to 14^{o}C) and calmer (intensity up to 0.15 m/s) locations, where, 21 days before, chlorophyll-a and currents direction varied between 10 and 20 mg/m^3 and between 120 and 210^o, respectively. Biomass under occurrence is higher in shallower (up to 80 m) and warmer (between 14^{o}C and 16^{o}C) locations.
This work will allow to understand spatio-temporal dynamics of sardine in Portuguese continental coast and in Gulf of Cádiz as well as relate the biomass indicator with environmental conditions, being able to contribute to a better management of this species.

Acknowledgments The authors acknowledge to FCT Foundation (Fundação para a Ciência e Tecnologia) for funding this research through Individual Scholarship PhD. PD/BD /150535/2019, throught projects PTDC/MAT-STA/28243/2017, UIDB/ 00013/2020 and UIDP/00013/2020; to MAR2020 for funding through SARDINHA 2020 project (MAR-01.04.02-FEAMP-0009) and to all colleagues involved in this work. The data used in this work was collected under the European Commission's Data Collection Framework - PNAB/EU-DCF Programa Nacional de Amostragem Biológica, (Reg. EC 2008/199).

References

[1] James S. Clark and Alan Gelfand. *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*. Oxford University Press, Inc., USA, 2006.

[2] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B*, 71:319–392, 04 2009.

Thematic Session II: Statistics Portugal

10 December, 15:00 - 15:20

Second Child: An Uncertain Transition

Rita Brazão Freitas[1], **Andreia Maciel**[2], **Maria Filomena Mendes**[3]
[1] Instituto Nacional de Estatística, Direção Regional de Estatística da Madeira,Professor of the Universidade da Madeira, rita.freitas@ine.pt
[2] Researcher of the Laboratório de Demografia, CIDEHUS-UE, Universidade de Évora, amaciel@uevora.pt
[3] Professor of the Universidade de Évora, Escola de Ciências Sociais; Researcher of the Laboratório de Demografia, CIDEHUS-UE, Universidade de Évora, mmendes@uevora.pt

---

Portugal remains with one of the lowest fertility levels in Europe. Despite the slight recovery seen in recent years - 1.21 children per woman of childbearing age in 2013 and 1.42 in 2019 - the Portuguese continue to have fewer children than they expected. From the Fertility Survey [2], we know that the majority (60.2%) of residents in Portugal (women aged 18-49 and men aged 18-54) expected to end the reproductive cycle with at least two children, although most of these people had either not yet entered parenthood (42.7%) or had only one child (9.8%), indicating that the Portuguese continue to postpone childbirth or possibly forego their final expected fertility.

Since fertility ideals, desires, and intentions are still associated with two or more children per family, it is important to understand what drives Portuguese who intend to have two or more children to have only one. Using data from the 2019 IFEC [2], we analyze a sample of men and women who expect to have two or more children and, through logistic regression models, we seek to identify factors that best differentiate those who have already transitioned to the second child from those who have not yet done so. To adjust the logistic regression models we resorted to the methodology proposed by Hosmer and Lemeshow [1] and considered the response variable: 0 - individuals with two or more children; 1 - individuals with only one child, but who intend to have at least two children. We took as possible explanatory variables some variables pointed out in the literature as conditioning or enhancing fertility. The final adjusted model showed good adequacy and discriminative ability, measured by the Hosmer and Lemeshow goodness of fit test and the Area Under the Curve (AUC) value of the Receiver Operating Characteristic (ROC) curve, respectively. We conclude that those with lower fertility desires, lower incomes, higher levels of education, who are not married and who had their first child later are more likely to have not (yet) transitioned to the second child, showing the great impact of delayed entry into parenthood on the transition to the second child. In addition to social and economic factors, the possibility that Portuguese may limit the number of children as a way of providing them with fewer restrictions and more opportunities is a determining factor for not having (yet) transitioned to the second child.

References

[1] D. Hosmer and S. Lemeshow. *Applied Logistic Regression, 3rd ed.* John Wiley & Sons, New York, 2006.

[2] IFEC. *Inquérito à Fecundidade, 2019.* Instituto Nacional de Estatística I.P., Departamento de Estatísticas Demográficas e Sociais, Portugal: Lisboa, 2019.

10 December, 15:20 - 15:40

Covid-19 pandemic: an unprecedented shock to consumer confidence

Ana Raquel Marques
Instituto Nacional de Estatística, Departamento de Contas Nacionais,
anaraquel.marques@ine.pt

The consumer confidence indicator is one of the most relevant monthly indicators produced by Statistics Portugal due to its fast availability and strong correlation with economic activity. The consumer confidence indicator exhibits a strong autoregressive behaviour, that allows for the indicator to be successfully modelled through a simple autoregressive process. However, this is not the case for the events of the COVID-19 pandemic, where the shock to consumer confidence was unforeseen and unprecedent in magnitude.

---

10 December, 15:40 - 16:00

Data analysis during Covid time: the e-invoice case

João Poças[1], **Sofia Rodrigues**[2],
[1] Instituto Nacional de Estatística, Departamento de Recolha e Gestão de Dados, joao.pocas@ine.pt
[2] Instituto Nacional de Estatística, Departamento de Recolha e Gestão de Dados, sofia.rodrigues@ine.pt

---

One of the main impacts of the COVID-19 pandemic was the significant decrease in response rates to business surveys, particularly during the second quarter of 2020. The monthly surveys were the most affected, with response rates dropping close to 10% in the collection carried out during April and May.

In this context, information from the e-invoice system became even more relevant, particularly to fill in missing responses to the STS, and contribute to the consistency of the results obtained in the production of statistical indicators. During the first week of April 2020, a very large set of data on company invoicing from January 2018 to February 2020 was received from the Tax Authority. The extensive volume of data (about 80 million records per month) and the urgency of the information in the shortest possible time, required a significant effort by the Information Systems and Data Collection and Analysis teams to make the data received available to internal users. The data received were treated for completeness considering the expected structure and primary validation. Standard processes were defined and developed, at the level of loading, pseudo-encryption of identifiers (when necessary), processing and availability of data that not only ensure data integrity, but mainly the consistency of the information to be used in different statistics. The main tasks performed at the level of data processing, coherence and consistency analysis were the following:

1. Validation of data structure, verification of the number of records, validation of the fiscal identification number and pseudo-encryption of identifiers;
2. Normalization of attributes;
3. Identification and exclusion of outliers very significant and identification of suspicious cases;
4. Consistency tests and comparison with other data sets.

The definition of these processes also made it possible to streamline the entire process of manipulating, exploring, and analyzing the data received.

10 December, 16:00 - 16:20

Inter-organizational networks of EuroGroups Register - a Supervised Clustering Algorithm for Network Data

Bárbara Santos[1], **Pedro Campos**[2],
[1] Instituto Nacional de Estatística, Departamento de Metodologia e Sistemas de Infomação e Faculdade de Economia do Porto,barbara.santos@ine.pt
[2] Instituto Nacional de Estatística, Departamento de Metodologia e Sistemas de Infomação e Faculdade de Economia do Porto,pedro.campos@ine.pt

---

Eurostat has set up a unique database - together with EU Member States and EFTA countries - called the EuroGroups Register which contains a large amount of information on multinational enterprise (MNE) groups operating in Europe. Together with the national statistical business registers (NSBRs), the EuroGroups Register is part of the European Framework of business registers.

In this work, we introduce a new method of supervised clustering with attributed networks, and the proposed methodology is applied to an inter-organizational network, obtained from the EuroGroups Register. The goal of the new method is to obtain class-uniform clusters, while minimizing the number of clusters. This method deals with representative-based supervised clustering, where a set of initial representatives is randomly chosen. By assigning each observation to the closest representative, clusters are obtained. With the new methodology, the way nodes are associated to clusters does not only depend on their network distance, but also on the distances between their attributes.
As a benchmark, we use the Subgroup Discovery perspective of Atzmueller [1], using Network data. This view is based on the fact that classical community detection techniques focus only on finding subgroup of nodes with a dense structure, lacking an interpretable description [2]. For this matter, Subgroup Discovery, a data mining technique that focus on discovering interesting relationships between different objects [4], can provide insights beyond connectivity within communities, and the relationships between subgroups of nodes as well. Subgroup Discovery focuses on detecting subgroups described by specific patterns that are interesting with respect to some target concept and a set of explaining features. Therefore, interesting patterns among subgroups can be revealed, for example, by inductive and exploratory data analysis tasks that find relations between a dependent and (several) independent variables [1], considering the compositional aspect of the networks. This way, with the additional information supplied by attributed networks, Subgroup Discovery method can be applied in order to combine both structural and compositional characteristics of the network. For this work, SD-Map, a fast algorithm for exhaustive subgroup discovery [3], will be used to perform Subgroup Discovery on attributed networks.

References

[1] M. Atzmueller. Subgroup discovery. *WIREs Data Mining Knowledge Discovery*, 5(1):35–49, 2015.

[2] M. Atzmueller, S. Doerfel, and F. Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, 329:965–984, 2016.

[3] M. Atzmueller and F. Puppe. *SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery.* Springer, Berlin, Heidelberg, 2006.

[4] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge Information Systems*, 29(3):495–525, 2011.

Thematic Session III: Banco de Portugal

10 December, 16:40 - 17:00

Using Isolation Forest in the quality control of the securities database of Banco de Portugal

André Fernandes[1], **Rafael Figueira**[2],
[1] Banco de Portugal, agfernandes@bportugal.pt
[2] Banco de Portugal, rfigueira@bportugal.pt

Machine learning techniques are being used to surpass challenges posed by a large increase in data volume (big data). The Isolation Forest (IF) technique was applied to improve the data quality control process in the Securities Statistics Integrated System database of Banco de Portugal. Two different models were implemented – for securities issues and holdings – and they have shown to be able to detect outliers with higher precision and efficiency than the status quo, thus highlighting the potential of IF technique in microdata management.

Keywords: Isolation Forest, Microdata, Securities, Database management

The Securities Statistics Integrated System (SSIS) is a security-by-security and entity-by-entity database managed by Banco de Portugal, and presents several challenges in the field of data management. These challenges range from data collection to data dissemination (Aguiar and Martins, 2011[1]). Across the entire database management process, one of the areas with an increasing need for robust techniques is quality control (QC). To address this need, the Isolation Forest (IF) technique was considered.

The IF technique is an unsupervised algorithm that works through a process of isolation, computing an isolation score for each data point, allowing to rank them by level of severity (likeliness to be an outlier). IF relies on a training dataset to learn the usual and predictable behaviour of data points in a specific dataset. Hereafter, IF relies on this "learning" to understand, for a particular test dataset, if a specific data point is an unexpected one, or in other words, an outlier. IF computes an outlier score which ranges between 0.5 (complete normality) and -0.5 (highest possible severity). This technique was first developed by Liu et al. (2008) [2], who show several advantages of its usage, including very low processing time for large datasets.

Two sets of models were developed based on IF technique for application to the SSIS database. These models, described in Figure 1, focus in microdata quality control of OCVP [1], a residual quantitative variable in the SISS database, using quantitative and qualitative variables needed to characterize the data in hands.

To evaluate the models' performance, a set of metrics was considered and an analysis was ran on subsets of the outliers flagged by the models. The results of the two principal

[1]OCVP (Other Changes in Volume and Price) = Other Changes in Volume + Price changes (gains and losses). For more information, please refer to the Handbook on Securities Statistics.

Details by:	Securities issues			Securities holdings			
Group of analysis	**Instrument Type**	**Issuer Sector**	**Instrument Type + Issuer Sector**	**Holder Sector**	**Holder Sector + Issuer Sector**	**Holder Sector + Instrument Type**	**Holder Sector + Issuer Sector + Instrument Type**
OCVP	value and %	value and %	value and %	Value, % and Issues OCVP %	Value, % and Issues OCVP %	Value, % and Issues OCVP %	Value, % and Issues OCVP %
Instrument	Currency	Currency, Type	Currency	Currency, Type	Currency, Type	Currency	Currency
Issuer	sector, size, country, economic activity	size, country, economic activity	size, country, economic activity	sector, size, country, economic activity	size, country, economic activity	sector, size, country, economic activity	size, country, economic activity
Holder				country	country	country	Country
Output	By ID of the micro data under analysis						

Figure 1: Models developed for Banco de Portugal securities database.

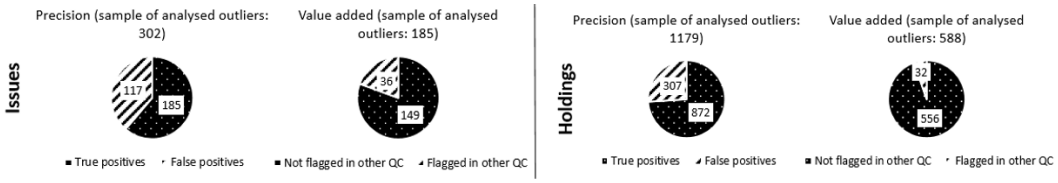

Figure 2: Precison and value added from the models // data: march 2021

evaluation dimensions considered, precision and value added, are presented in Figure 2. The precision was measured by the number of true positives, i.e., true errors, against the total severe outliers flagged by the models. To assess the value added against the status quo, the number of outliers flagged by these models and not flagged in other QCs procedures already implemented in the SSIS database was analysed.
The precision metric is relatively high for both sets of models (61% for issues and 74% for holdings). Moreover, despite the fact that the flagged false positives are not incoherencies in the database, they are in fact outliers in SSIS information, meaning that these observations may be relevant for economic/financial analysis. Regarding the value added metric, both sets of models were able to flag new observations which were not flagged in other QC tasks (149 for issues and 556 for holdings).
In sum, the application of IF technique allowed the creation of models to be applied in automatic QC process in the SSIS database. With low computation requirements and processing time, these models revealed higher values for precision and value added metrics.

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] M. Aguiar and C. Martins. Adding business intelligence to statistical systems – the experience of Banco de Portugal. NTTS, Eurostat, Brussels, Belgium, FEB 22-24, 2011.

[2] F. Liu, K. Ting, and Z. Zhou. Isolation forest. 8th IEEE International Conference on Data Mining, Pisa, ITALY, DEC 15-19, 2008.

10 December, 17:00 - 17:20

Machine Learning models applied to ITENF data quality control

C. Ferreira[1], F. Fonseca[2], L. Pinto[3], J. A. Neves[4], N. Azevedo[5], V. Lopes[6]
[1] Banco de Portugal, ciferreira@bportugal.pt
[2] Banco de Portugal, ffonseca@bportugal.pt
[3] Banco de Portugal, lmpinto@bportugal.pt
[4] Banco de Portugal, janeves@bportugal.pt
[5] Banco de Portugal and NIPE, Universidade do Minho, ncazevedo@bportugal.pt
[6] Banco de Portugal, vlopes@bportugal.pt

Good raw data quality is an essential asset to achieve accurate and timely decisions. This paper proposes a more efficient way of selecting firms that require a deeper manual validation of their reported data. We applied two unsupervised machine learning (ML) algorithms to a 12 period sample of quarterly data collected from the Quarterly Survey on Non-Financial Corporations (ITENF): DBSCAN and Isolation Forest. Our results indicate that these methodologies could contribute to a remarkable increase in the efficiency of the data quality control process, while ensuring minimal loss of quality of the data. Lastly, we discuss the decision making process that led to our final model choice.

Keywords: Machine Learning, Isolation Forest, DBSCAN, ITENF, Corporations

---

The Quarterly Survey on Non-Financial Corporations (ITENF) is a joint statistical operation of Banco de Portugal and INE, which collects quarterly accounting information from approximately 4000 non-financial corporations, relating to its activity and financial soundness. ITENF is the main data source used by Banco do Portugal's Central Balance Sheet Office (CBSO) to compile and publish quarterly statistics, requiring for that matter extensive data quality control at firm level of detail ([1, 2]).
Current data quality controls in place at CBSO mostly relate to traditional evaluation of quarter-on-quarter changes for each accounting variable at firm level, which may trigger warnings requiring further manual checking by a human analyst.
In this paper, we show how we applied two distinct ML models to ITENF data quality control for the reference quarter of $2020Q4$, using training data since $2018Q1$ and up to $2020Q3$. The main goal of our research is to replace traditional quality control procedures currently in place with a more efficient machine learning outlier detection technique. A human analyst would then manually analyze outliers detected by the ML model and proceed with corrections to the firms' reports, if needed. We tested two different unsupervised ML models: DBSCAN and Isolation Forest.

DBSCAN (Density–based spatial clustering of applications with noise) is a clustering algorithm that is widely used for the purpose of outlier detection ([3]). Given two parameters, ϵ and n, a point A is considered a neighbor of a point B if they are separated by a distance smaller than ϵ. If the point A has n neighbors, it is a core point (i.e. a point of high density). Any point that is not a core point, and is not in the neighborhood of a core point, is an outlier. Hence, we must pick the values of both ϵ and n that yield the best results for the dataset at hand, which is, in general, a challenging task. Additionally, parameters must be evaluated periodically to assure that they remain optimal when new data arrives. The Isolation Forest algorithm is a tree-based algorithm built specifically for anomaly detection ([4]). At each step, the algorithm randomly selects a feature and a split value between the maximum and minimum for that feature. The branching process continues until either all observations are isolated or a predetermined maximal tree depth is reached. Observations are attributed a score based on the number of splits required to isolate them – a lower number of splits represents a higher likelihood of being anomalous, hence a higher anomaly score.

Results obtained for the two different models were very similar, using relevant firms selected by current traditional criteria as a benchmark of model precision. The software used was Python. The DBSCAN model highlighted 73% of relevant firms as outliers (for 976 outliers detected), against the Isolation Forest model which selected 72% of relevant firms as outliers (for 967 outliers detected). Bearing in mind the need for a reasonable number of outliers, since each outlier selection requires further human analysis, we compared results for a streamlined list of outliers. For this reduced list, results remained similar.
In spite of DBSCAN's relevance, we selected the Isolation Forest model as our final model, mainly due to its parameter calibration simplicity, computational speed and the usefulness of a final outlier score, which critically assists in outlier prioritization during regular quarterly statistical production.

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

References

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

[2] Banco de Portugal. Quarterly economic and financial indicators of the non-financial corporations, 2021.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231, 1996.

[4] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

10 December, 17:20 - 17:40

An application of cluster analysis to the interest rates reported to the Portuguese Central Credit Register

André Costa[1], **Francisco Fonseca**[2], **Susana Maurício**[3]
[1] Banco de Portugal, anfcosta@bportugal.pt
[2] Banco de Portugal, ffonseca@bportugal.pt
[3] Banco de Portugal, scmauricio@bportugal.pt

The data covering credit contracts reported to the Portuguese Central Credit Register (CCR) has increased quite substantially in both volume and complexity over the last years. This work focuses on the application of unsupervised machine-learning techniques to the interest rate variables reported to the CCR. This is a two-step process, starting with the application of a K-means clustering algorithm to group similar contracts, followed by an Isolation Forest algorithm to isolate the potentially anomalous observations.

Keywords: Interest Rates, K-means, Central Credit Register, Cluster Analysis, Isolation Forest, Unsupervised Machine-learning

---

The Portuguese CCR is a system managed by Banco de Portugal, which gathers on a monthly basis a wide range of data provided by the credit-granting institutions associated with actual and potential credit liabilities of their costumers (natural or legal persons). The main purpose of the CCR is to provide support to the credit-granting institutions in their assessment of counterparty risk.

The data reported to the CCR is quite rich in both volume and complexity. It covers over 18 million contracts reported on a monthly basis with up to 200 variables (both qualitative and quantitative), making unfeasible the traditional approach to detect subtle behaviors and creates the need for new tools to increase the efficiency and the effectiveness of the data quality assessment process. This new approach embodies a shift in focus from the analysis of the aggregates, which are in general far more stable, towards exploring the individual contract dynamics over time.

This context motivated the application of unsupervised machine-learning techniques to the interest rate variables, reported to the CCR, based on the K-means clustering and Isolation Forest algorithms. This process starts with the K-means algorithm ([1]) which is one of the most widely used clustering methodologies. It splits the space of observations into K clusters, in a way that minimizes intra-cluster variance. In our case, we make use of the scikit-learn implementation of the algorithm ([2]).

To initiate the process the observations are divided into groups considering the following contract variables: type of financial product, type of interest rate and the nature of entities involved. After that, we apply a K-means clustering technique using as variables the age

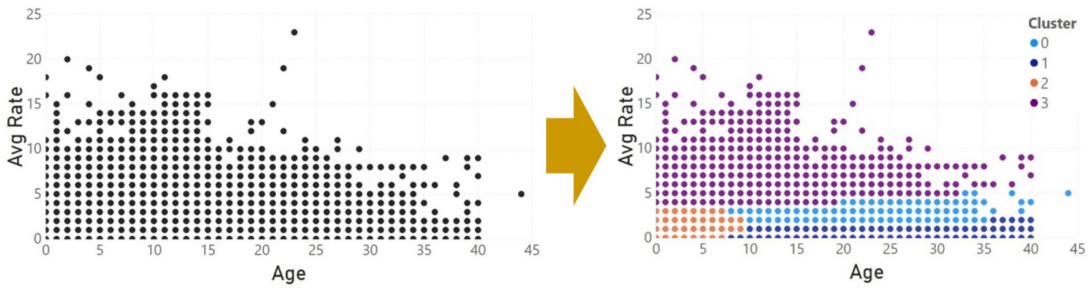

Figure 1: Cluster formation for housing credit contracts with variable rate

of the contract and the average of the reported rates. Figure 1 displays an example of the application of this procedure to housing credit contracts with variable rates.
This approach allows us to include, within the same model, interest rates which are reported only once, when the contract is celebrated, and rates which are updated on a monthly basis, despite their very distinct dynamics.
Subsequently, the differences between the observations of each of the four interest rates[1] and their medians are calculated, for each cluster, setting up the data needed for our following outlier detection process, which is an application of the Isolation Forest algorithm.
The purpose of the Isolation Forest is two-fold:

- To detect cases where the reported rates are consistent but significantly different from their cluster medians;
- To identify cases with higher dispersion between the differences, which usually indicates that the reported rates are considerably distant from each other.

These tools are already being used and have proven to be an important enhancement to the analysts' tool set, allowing for the identification of a set of anomalies that previously would not be detected or would require complex and time-consuming ad hoc analyses.

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. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[2] Pedregosa *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[1] Currently, four different interest rate variables are reported – annual interest rate, nominal annual interest rate, annual percentage rate and effective annual rate.

Thematic Session IV: CLAD–SPE

11 December, 12:20 - 12:40

Prediction Models in Medicine

Ana Luisa Papoila
NOVA Medical School/Faculdade de Ciências Médicas da Universidade Nova de Lisboa, CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa
ana.papoila@nms.unl.pt

The aim of this presentation is to focus on several methodological options within the scope of prediction in the field of Medicine, including all the mandatory steps to obtain valid prediction models. Therefore, it is intended to present several possible approaches to a prediction problem, as well as to talk about the necessary precautions to obtain quality results.

Keywords: Prediction, medicine, statistical approach, artificial neural networks

---

In Medicine, both patients and physicians are daily faced with decision-making considering the estimated risk of a certain disease (or of any other condition) or of the future occurrence of a certain event. In the first case, knowledge of the probability estimate of a particular disease can be used, for example, to start treatment immediately or to refer patients for further testing. On the other hand, in the second case, the estimates can help to plan the future regarding lifestyle or, even, to make certain therapeutic decisions based on the estimated risk of reaching, in the future, a specific state of health. In fact, this information about the disease evolution is precious to both the physician and patient, and also to the family members. Therefore, there is a need to implement models that estimate probabilities of the presence or occurrence in the future of a given condition.

To obtain these estimates, as the outcomes in this context usually depend on several explanatory variables, it is necessary to use multivariable models, commonly called prediction models. Due to their importance, these have been developed not only in a context of Health Sciences but also in areas such as Meteorology, Finance and Ecology. It is, however, in Medicine that these models have gained greater relevance and where, in the recent years, their application has been increasing. In fact, areas such as clinical practice, medical research and public health have benefited from the information provided by these models. However, given the level of dissemination achieved, there was a need to define some rules for its implementation and dissemination of results ([1]; [4]). Accordingly, several methodological options will be presented within the scope of prediction models in the field of Medicine, including all the mandatory steps to obtain valid prediction models.

Regression models represent the usual choice in this context although other approaches such as regression trees and artificial neural networks are widely used alternatives. Some of the regression models belong to the class of generalized linear models, that have as an extension the corresponding generalized additive models. These are more flexible insofar

as they allow to model the influence of continuous predictor variables on the response through smoothers, overcoming the linearity assumption of many regression models. Even more flexible are the generalized additive models for location, scale and shape [3] that allow for the outcome more than 100 distributions characterized by parameters that can be modelled as additive functions of the explanatory variables. Still in a regression context, and not forgetting that prognostic studies are longitudinal in nature, it is mandatory to refer methods that are more appropriate to analyze data from these types of studies. Effectively, given the growing number of longitudinal studies (in Medicine and beyond), new statistical methodologies emerged that take into account the autocorrelation structure inherent to the several observations of the same individual obtained over the follow-up period. On this subject, namely on mixed-effects regression models, the literature is already quite extensive. Survival regression models are also widely used when modelling longitudinal data. Frank Harrel's book ([2]) is a very good reference to those who wish to use regression models.
After fitting a particular prediction model, we need to evaluate its performance taking into account the difference between the observed values and the values predicted by the model. This assessment can be performed globally and, in the case of binary responses, go further and try to quantify not only the predictive ability through calibration measures (agreement between observed and predicted values) but also the discriminative power (ability to distinguish between those who have and those who have not suffered the event of interest). Finally, as we want to use our prediction models to obtain valid outcome estimates on new individuals (who were not involved in the modelling process) the validation procedure follows. There are several techniques essentially grouped into two types: internal and external validation. In the former case, the model is tested with the sample that was used in the modelling process. In the latter, individuals that, in some way, differ from those used during the model fitting, are considered. When possible, external validation is more advisable as it increases the external validity of the study.
Some of the explained methodologies will be implemented using real medical data.

References

[1] G.S. Collins, J.B. Reitsma, D.G. Altman and K.G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, 162:55–63, 2015.

[2] Frank E. Harrell Jr. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer Series in Statistics, Switzerland, 2016.

[3] D.M. Stasinopoulos and R.A.Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23 (Issue 7):1–46, 2008.

[4] E.W. Steyerberg. *Clinical Prediction Models-A Practical Approach to Development, Validation, and Updating.* Springer, Switzerland, 2019.

11 December, 12:40 - 13:00

Is age at menopause decreasing? The consequences of not completing the generational cohort

Rui Martins[1], **Bruno de Sousa**[2], **Thomas Kneib**[3], **Nadja Klein**[4], **Maike Hohberg**[5], **Elisa Duarte**[6], **Vítor Rodrigues**[7]

[1] Departamento de Estatística e Investigação Operacional, Universidade de Lisboa; Centro de Estatística e Aplicações da Universidade de Lisboa, rmmartins@fc.ul.pt

[2] Faculty of Psychology and Education Sciences; Center for Research in Neuropsychology and Cognitive and Behavioral Intervention, University of Coimbra, bruno.desousa@fpce.uc.pt

[3] Georg-August-Universität Göttingen, tkneib@uni-goettingen.de

[4] Georg-August-Universität Göttingen, mhohber@uni-goettingen.de

[5] Humboldt-University of Berlin, School of Bus. Econ., nadja.klein@hu-berlin.de

[6] Independent Research Fellow, Paris, duarte.elisa@gmail.com

[7] Faculty of Medicine, University of Coimbra, Portugal, vrodrigues@netcabo.pt

---

Breast cancer screening programs inherently suffers from missing values for the age at menopause since not all observed women have already reached that state. The uncertainty of whether age at menopause is currently increasing or decreasing in Western countries, lead us to apply and compare two methods for handling this type of unavailable data: (i) a multiple imputation methodology based on a truncated distribution but ignoring the mechanism of missingness (ii) a copula-based multiple imputation method that simultaneously handles the age at menopause and the missing mechanism.

Keywords: Copula, Distributional regression, Imputations, Menopause age, Missing values

---

The starting point for this work is the dataset on the Breast Cancer Screening Program provided by the Portuguese Cancer League (LPCC) [1]. The records have the follow-up of 278 282 women, aged 45–69, attending the screening program between 1990 and 2010. At the time of the last screening date, 65 765 women (23.6%) stated they had not yet reached menopause.

The age of menopause plays an important role when investigating breast cancer risk factors. Although it is a variable prone to many missing values, because the screening program's onset overlaps the time when women are most likely to enter menopause, thus younger women in the program tend not to have that information. Not taking into account the censored nature of the information might lead to bias in the results since evidence suggests a spatial pattern of the missing mechanism in the Central region of Portugal, implying a violation of the missing at random assumption. Since the percentage of missing data in the

variable menopause is high (23.6%), consider simply the deletion of those individuals might be very inefficient, because the missing values might be reflecting a different population's group behavior.
A simultaneous modeling approach of the missing information and age at menopause as the response variable was considered through two generalized regression models [2]. Spatial information (the municipalities where these women live), viewed as a Markov random field, will be taken into account in the model in order to see how the age at menopause differs between regions. Penalized splines will be used to estimate the effects of continuous covariates. The two models will be linked with the introduction of a bivariate copula [3], allowing us to model simultaneously the two responses, conditional on some covariates.
The models were fitted considering a Generalized Joint Regression Model within the GJRM package in R [3]. The results from this approach are clearly different from the results obtained only with the complete dataset. The latter evidences a decreasing trend in time for a woman's year of birth, while the former reverses this and shows an increasing effect of the menopause age as a function of the birth year.
Both imputation methods unveiled an increasing trend of age at menopause when viewed as a function of the birth year for the youngest generation. This trend is hidden if we model only women with an observed age at menopause.
The results of this work show that when studying age at menopause, missing ages must be recovered with an adequate procedure for incomplete data. Imputing these missing ages avoids excluding the younger generation cohort of the screening program in breast cancer risk analyses and hence reduces the bias stemming from this exclusion.

Acknowledgements This work was co-financed by *Fundação para a Ciência e Tecnologia* (FCT), project UIDB/00006/2020 and the European Regional Development Fund (ERDF - FEDER) through Portugal 2020 - *Programa Operacional Competitividade e Internacionalização* [POCI-01-0145-FEDER-029443].

References

[1] E. Duarte, B. de Sousa, C. Cadarso-Suarez, V. Rodrigues, and T. Kneib. Structured additive regression modeling of age of menarche and menopause in a breast cancer screening program. *Biom J*, 56(3):416–427, 2014.

[2] M. Gomes, R. Radice, J. Camarena Brenes, and G. Marra. Copula selection models for non-gaussian outcomes that are missing not at random. *Stat Med*, 38(3):480–496, 2019.

[3] G. Marra and R. Radice. Bivariate copula additive models for location, scale and shape. *Comput Stat Data An*, 112:99–113, 2017.

11 December, 13:00 - 13:20

Analysis of cutoff point estimation for determining seropositivity in the context of SARS-CoV-2 infections

Tiago Dias Domingues[1], **Helena Mouriño**[2], **Nuno Sepúlveda**[1,3]
[1] CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, tmdomingues@fc.ul.pt
[2] CMAFcIO, Faculdade de Ciências, Universidade de Lisboa, Portugal, mhnunes@fc.ul.pt
[3] Faculty of Mathematics and Information Science, Warsaw University of Technology, N.Sepulveda@mini.pw.edu.pl

---

Gaussian mixture models are popular in serological data analysis to help determining seropositive and seronegative individuals. In this work, we propose the finite mixture models based on scale mixtures of Skew-Normal distributions as an alternative to model serological data. We illustrate the application of these models to antibody data against four SARS-CoV-2 virus antigens. Methods for cutoff point estimation are presented. Sensitivity, specificity and accuracy are calculated. The results of a simulation study will also be presented.

Keywords: serology, finite mixture models, skew-normal distribution, skew-t distribution, cutoff point

---

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection that causes the devastating and often lethal COVID-19 disease was first detected in China, province of Wuhan in December 2019. The detection of the virus is so far done by the so-called reverse quantitative PCR reverse transcriptase (RT-qPCR) on samples from nasopharyngeal or throat swabs. Nowadays, serological studies are done in order to identify antibodies against the virus. The detection of antibodies in the serum samples is classically done via enzyme linked immunosorbent assays (ELISA), where the resulting data are light intensities, also called optical density, which reflects the underlying antibody concentration in the samples ([1]). For statistical convenience, the analysis of serological data proceeds by dichotomizing the amount of antibodies present in the serum of an individual using an arbitrary cutoff point in the antibody distribution. One of the traditional methods to establish the cutoff point in serological assays is to consider the logarithmic transformation of the antibody concentration of a known seronegative population and proceed to calculate the mean plus 2 or 3 standard deviations. This method is more adequately when the antibody distribution of the seronegative population is normally distributed. However, our previous studies of different serological data ([1]) showed evidence against a normality assumption for the antibody levels associated with a putative seronegative population. Alternatively, finite mixture models can be used to determine the seropositivity

cutoff directly from the data ([1]). We proposed three methods for determining seropositivity cutoff using the so-called scale mixtures of Skew-Normal distributions - of which the Skew-Normal and Skew-t distributions are particular cases - in the case where the true infection status is unknown: M1 - based on the 99.9%-quantile associated with the estimated seronegative population; M2 - based on the minimum of the density mixture functions; M3 - imposes a threshold in the the so-called conditional classification curves (e.g. 90%). Since the true serological status is known *a priori* in this study, we performed the ROC curve and evaluate the performance of the proposed methods in freely available serological data (https://github.com/MWhite-InstitutPasteur/SARSCoV2SeroDXphase2). We analyzed IgG antibody responses against four SARS-CoV-2 spike or nucleoprotein antigens: RBD - glycoprotein receptor-binding domain; S^{tri} - S trimeric spike protein; S1 - spike glycoprotein S1 domain; S2 - SARS-CoV-2 spike glycoprotein S2 domain. Estimation of the cutoff point based on the M2 method proved to be the method with the highest sensitivity (*sens*) for classifying seropositive individuals, as well as the one that produces the highest proportion of correct results (accuracy-*ACC*) for the RBD antigen (cutoff=2.49, *sens*=86.45%, ACC=92.89%), S1 (cutoff=2.27, *sens*=71.03%, ACC=86.89%) and S2 (cutoff=2.39, *sens*=83.64%, ACC=90.89%). In the case of the S^{tri} antigen, it was not possible to calculate the sensitivity and accuracy of the method M1 given the high values that the quantile assumes leading to the seropositive population being fully absorbed by it. Thus, for comparison purposes, the application of each methods to the Skew-Normal distribution was considered, again verifying that the method M1 produces the highest sensitivity (cutoff=2.46, *sens*=90.19%). However, for this antigen, the method with the highest accuracy is the method M3 (ACC=93.44%). We also performed a simulation study to assess the performance of cutoff points proposed by each method for the RBD and S^{tri} antigens, considering the Skew-Normal and Skew-t distributions, respectively. We simulated 1000 samples with dimensions 100, 500 and 1000, varying the proportions of seronegative individuals in $\pi = 0.3, 0.6$ and 0.9. With the results of our simulation study we found that as the sample size increases, both the relative error and the mean square error tend to decrease. It is also found that for small samples and extreme values ($\pi = 0.3$ or $\pi = 0.9$), the models tend to have some difficulty in identifying a seronegative and seropositive population. This is a result that alerts to the existence of possible false positives and false negatives in the case of small samples.

Acknowledgements NS acknowledges funding from Polish National Agency for Academic Exchange (ref. grant: PPN/ULM/2020/1/00069/U/00001). This work was partially funded by FCT - Fundação para Ciência e Tecnologia (ref:UIDB/00006/2020 and UIDB/04561/2020).

References

[1] T. Dias Domingues, H. Mouriño, and N. Sepúlveda. Analysis of antibody data using finite mixture models based on scale mixtures of skew-normal distributions. *medRxiv*, 2021.

Thematic Session V: CLAD Corporate

11 December, 14:20 - 15:00

How to Measure the Outdoor Advertising Audience and the Actual Mobility in Portugal?

Paulo Caldeira[1], **João Pequito**[2]
[1] PSE Mobility & OOH Panel Manager, pcaldeira@pse.pt
[2] PSE Chairman and CEO, jpequito@pse.pt

An innovation of PSE in mobile research allows to measure the effective mobility of the Portuguese and also, in a official way since January 1st 2021, the actual audience of outdoor advertising. The construction of a representative panel sample of the population, combined with the installation of an APP on the cell phones of its participants, allows to have daily information about people's movements. And therefore, getting information about the likelihood of the outdoor advertising supports audience. The present pandemic was an unexpected challenge for this tool of PSE that was useful to monitor the daily rate of confinement, as well as the actual mobility and behavior of the population. The current transformations, in particular teleworking, imply important changes in mobility habits, which we are currently monitoring and classify.

Contributed Sessions

9 December, 16:50 - 17:10

Modelling interval-valued data: a clusterwise regression approach

Sónia Dias[1], **Paula Brito**[2], **Nikhil Suresh**[3]
[1] ESTG, Instituto Politécnico de Viana do Castelo & LIAAD - INESC TEC, Portugal, sdias@estg.ipvc.pt
[2] Faculdade de Economia, Universidade do Porto & LIAAD - INESC TEC, Portugal, mpbrito@fep.up.pt
[3] Faculdade de Economia, Universidade do Porto, Portugal, prodigionikhil@gmail.com

---

We focus on regression analysis of interval-valued data. Regression models that allow explaining one interval-valued variable from other interval-valued variables have been proposed in the literature. However, a single model may not be sufficient to adequately explain some phenomena, and it may be necessary to identify classes in the observed set and fit a specific model in each class. We present a clusterwise regression model for interval-valued data, developed as a combination of the partitioning dynamic clustering method, and the Interval Distributional (ID) regression model.

Keywords: Interval-valued variable, Interval Distributional regression model, Clusterwise regression

---

The need to analyse big data makes it necessary to innovate and develop new statistical methods. Data tables where the cells contain a single quantitative or categorical value are no longer sufficient. In the Symbolic Data Analysis (SDA) framework the cells of data arrays may contain finite sets of values/categories, intervals or distributions, representing the variability associated with each unit.

We focus on the case of interval-valued variables, i.e. variables whose observations are intervals of real numbers. Such interval-valued observations may also be represented by quantile functions, the inverse of the cumulative distribution functions, which are non-decreasing functions defined in the unit interval. Assuming the Uniform distribution within each interval, quantile functions are linear functions. Descriptive statistics and statistical methods for interval-valued data have been proposed in the literature (see, e.g. [1]). The Interval Distributional (ID) Model [2] considers intervals represented by the corresponding quantile functions $\psi(t), t \in [0,1]$. The error between predicted and observed intervals, for each unit $i \in \{1,\ldots,n\}$, is evaluated by the Mallows Distance $D_M(Y(i), \widehat{Y}(i)(t)) = \sqrt{\int_0^1 \left(\psi_{Y(i)}(t) - \psi_{\widehat{Y}(i)}(t)\right)^2 dt}$.

However, sometimes when a single regression model is not appropriate, and it may be necessary to cluster the units and fit a regression model in each cluster (see, e.g. [4]). The

goal of this work is to propose a Clusterwise Regression model for interval-valued variables that finds the best partition of the data in clusters and provides a linear regression model for each cluster. The corresponding algorithm, combining the dynamical clustering algorithm [3] and the Interval Distributional (ID) regression model [2], is as follows [5]:

Step 1: Consider an initial partition of the given units.

Step 2: Fit a regression model in each cluster using the ID Model.

Step 3: Re-assign each unit to the cluster that provides the best fit, as measured by the squared Mallows distance.

Steps 2 and 3 are repeated until convergence is attained, and a local minimum of the sum of squares of the errors W is obtained (or the fixed maximum number of iterations is reached), with $W = \sum_{k=1}^{K} \sum_{i \in C_k} D_M^2(Y(i), \widehat{Y}^k(i))$, where $\widehat{Y}^k(i)$ is the estimated interval of $Y(i)$ obtained by the (local) regression model in cluster C_k, $k \in \{1, \ldots, K\}$.
The process may be applied repeatedly varying the number of clusters K; for each fixed K, the algorithm allows for different initial partitions and selects the solution with lowest Total Error W. To select the best solution across different K, we use, among other measures, the Weighted Coefficient of Determination [2, 5], $\Omega = \sum_{k=1}^{K} \frac{n_k}{n} \Omega_k$ with $\Omega_k = \frac{\sum_{i \in C_k} D_M^2 \left(\widehat{Y}^k(i), \overline{Y}_k\right)}{\sum_{i \in C_k} D_M^2 \left(Y(i), \overline{Y}_k\right)}$,
where $n_k = |C_k|$ and $\bar{Y_k}$ is the (local) symbolic mean of Y.
The final clusters may then be used to predict target intervals for new observations.
The developed model is applied to real interval-valued data to illustrate its behaviour.

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] P. Brito. Symbolic Data Analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4):281–295, 2014.

[2] S. Dias and P. Brito. Off the beaten track: a new linear model for interval data. *European Journal of Operational Research*, pages 47–94, 2017.

[3] E. Diday and J.C. Simon. Clustering analysis. In *Digital Pattern Recognition*, pages 47–94. Springer, 1976.

[4] H. Späth. A fast algorithm for clusterwise linear regression. *Computing*, 29(2):175–181, 1982.

[5] N. Suresh. Clusterwise Linear Regression for Interval Data - An Extension of Interval Distributional Model. Master's thesis, Faculdade de Economia, Universidade do Porto, 2020.

9 December, 17:10 - 17:30

Multivariate Parametric Analysis of Distributional Data

P. Brito[1], **A.P. Duarte Silva**[2]
[1] Fac. Economia, Universidade do Porto & LIAAD-INESC TEC mpbrito@fep.up.pt
[2] Católica Porto Business School & CEGE, Universidade Católica Portuguesa psilva@ucp.pt

---

We present parametric probabilistic models for numerical distributional variables building upon previous models for interval-valued variables. The proposed models are based on the representation of each distribution by a location measure and inter-quantile ranges, for given quantiles. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. For all cases, maximum likelihood estimators of the corresponding parameters are derived. This framework is applied to discriminant analysis of distributional internet traffic data.

Keywords: discriminant analysis, histogram data, symbolic data

---

In this work we consider data where individual units are described by distributions. Distributional data may result from the aggregation of large amounts of open/collected/generated data, or may be directly available in a structured or unstructured form, describing the variability of some features. In recent years, different approaches have been investigated and methods proposed for the analysis of such data. However, most existing methods rely on non-parametric descriptive approaches.
In [1], parametric inference methodologies based on probabilistic models for interval variables are developed. Following a similar approach, we propose parametric models for numerical distributional variables based on the representation of each distribution by a central statistic C, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles $\psi_1, \ldots, \psi_k$. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix.
Let $Y_1, \ldots, Y_p$ be the p distributional variables, defined on a set of units $S = \{s_1, \ldots, s_n\}$. The model consists in representing $Y_j(s_i)$ by

- a central statistic C_{ij}, typically the Median Med_{ij} or the MidPoint $\frac{Max_{ij}-Min_{ij}}{2}$
- the $[Min, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - Min_{ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- ...

- the $[\psi_k, Max[$ range: $R_{mij} = Max_{ij} - \psi_{kij}$

Assume that the joint distribution of the central statistic C and the logarithms of the ranges $R^*_\ell, \ell = 1, \ldots, m$, is multivariate Normal:

$(C, R^*_1, \ldots, R^*_m) \sim N_{(m+1)p}(\mu, \Sigma)$

$$\mu = \left[\mu^t_C, \mu^t_{R^*_1}, \ldots, \mu^t_{R^*_m}\right]^t ; \Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*_1} & \ldots & \Sigma_{CR^*_m} \\ \Sigma_{R^*_1 C} & \Sigma_{R^*_1 R^*_1} & \ldots & \Sigma_{R^*_1 R^*_m} \\ \ldots & \ldots & \ldots & \ldots \\ \Sigma_{R^*_m C} & \Sigma_{R^*_m R^*_1} & \ldots & \Sigma_{R^*_m R^*_m} \end{pmatrix}$$

In the most general formulation (configuration 1) we allow for non-zero correlations among all centres and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables Y_j are non-correlated, but for each variable, the centre and all its log-ranges may be correlated among themselves (configuration 2); centres (respectively, log-ranges) of different variables may be correlated, but no correlation between centres and log-ranges is allowed (configuration 3); centres (respectively, each log-range) of different variables may be correlated, but no correlation between centres and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all centres and log-ranges are non-correlated (configuration 5).

We note that in cases 2, 3, 4 and 5, Σ can be written as a block diagonal matrix: in configuration 2 there are p blocks, all $(m+1) \times (m+1)$; in configuration 3 there are two blocks, one is $p \times p$, and the other is $mp \times mp$; in configuration 4 there are $m+1$ blocks, all $p \times p$, and in configuration 5 the $(m+1)p$ blocks are single real elements. For all cases, maximum likelihood estimators of the corresponding parameters are easily derived taking advantage of the covariance matrix block structure.

This model allows for parametric discriminant analysis of distributional data. For each configuration, an estimate of the optimum classification rule can be obtained with the corresponding Σ, by directly generalising the classical linear and quadratic discriminant classification rules.

This framework has been applied to the problem of identifying internet traffic re-direction ("attacks"), both in a two-class (regular vs irregular traffic) and a five-class (regular, and four distinct relays) formulations. Results show that the proposed approach has an excellent classification performance in the problem at hand. Furthermore, the restricted covariance configurations allow for parsimonious classification rules.

Acknowledgements This work was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within projects UIDB/50014/2020, UIDB/00731/2020, and PTDC/EEI-TEL/32454/2017.

References

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

9 December, 17:30 - 17:50

Qualitative-quantitative synergy

Ana Lorga da Silva[1], **Artur Parreira**[2]
[1] Lusófona University and ENIDH, ana.lorga@ulusofona.pt
[2] Santa Úrsula University, arturmparreira@gmail.com

The present work aims to relate the impact of the value of the adverbs in a continuous scale to persons with more than 17 years old that speaks Portuguese but live in several countries. It is a primary approach to the results obtained by a questionnaire applied using Google Forms. Here we present mainly descriptive statistics, also identifying the outliers (moderate and severe) and justifying why they appeared, considering the diverse factors such age, level of education, between others.

Keywords: The value of adverbs, Statistical analysis, Interval scales, Qualitative and quantitative synergy

---

One of the areas of our research, since 2003 [2] and [3], has been to study the numerical value of quantity and frequency adverbs, with the aim of replacing ordinal scales by interval scales [4].
The present investigation continues these studies, using an adverb scale to measure the impact of the use of power on the use of information, in situations of analysis and problem solving. We chose to apply the scale to concrete research because in previous studies we found that a considerable number of participants had difficulty in understanding the numerical value of each adverb as a scale value.
In this research, respondents assessed five power situations - from first situation extreme use of power to the fifth situation no use of power - with the adverbs scale, and then assigned a numerical value to the adverbs used in the scale (Figure 1). A first analysis of the results highlighted some interesting points: even with the adverbs of quantity forming a scale, a certain number of respondents assigned them a numerical value detached from their scale position. This was visible in the numerical evaluation of the adverb extremely and expressions nothing and none in which 10.5% of respondents (in the first case) and 11% (in the second case) assigned numerical values visibly associated with their daily experiences and not with the adverb's scale position on the used scale. This behavior was extended to other adverbs, although in a smaller percentage: 9.68% in the adverb much; 5.99% in the adverb averagely; 8.75% in the adverb little (Figure 2). This behavior of respondents corresponds to the attraction of responses to the central values of the scale, a phenomenon well referenced in behavioral sciences [1]. Still within the scope of this study, we will continue to find a better solution to eliminate this problem, increasing respondents' awareness of the scale condition of the used adverbs.

		Extremely	Much	Averagely	Few	Little
N	Valid	216	216	216	216	216
	Missing	0	0	0	0	0
Mean		8,72	7,42	5,35	3,61	1,34
Median		10,00	8,00	5,00	3,00	,00
Mode		10	8,0	5	3,00	0
Standard error		2,480	1,464	1,346	1,725	2,544
Mínimum		0	,0	0	,00	0
Máximum		10	10,0	10	10,00	10
Percentiles	25	9,00	7,00	5,00	3,00	,00
	50	10,00	8,00	5,00	3,00	,00
	75	10,00	8,00	6,00	4,00	1,00

Figure 1: Descriptive Statistics of the numerical value of the adverbs

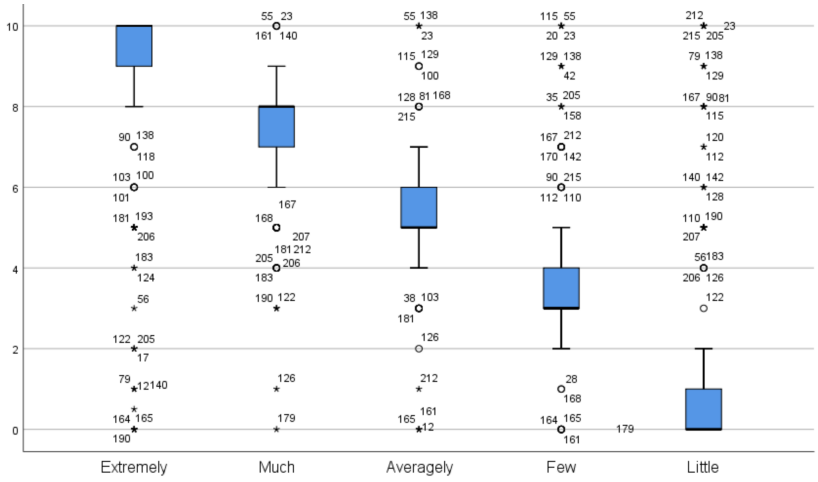

Figure 2: Outliers of the numerical value of the adverbs

References

[1] R. A. Cummins and E. Gullone. Why we should not use 5-point likert scales: The case for subjective quality of life measurement. Singapore, National University of Singapore, 2000.

[2] A. Parreira. *Gestão do Stress e da Qualidade de Vida.* Monitor, Lisboa, 2006.

[3] A. Parreira and A. S. Lorga. The use of numerical value of adverbs of quantity and frequency in the measurement of behavior patterns: Transforming ordinal scales into interval scales. *Revista Ensaio – Avaliação e Políticas Públicas em Educação*, 190:109–126, 2016.

[4] Sharma S.and Niedrich R.W. Weathers, D. The impact of the number of scale points, dispositional factors, and the status quo heuristic on scale reliability response accuracy. *Journal of Business Research*, 58(11):1516–1524, 2005.

10 December, 9:00 - 9:20

Adapting the sampling design of research surveys to improve the biomass estimation of non-target species - the case study of *Raja clavata*

Daniela Silva[1], **Raquel Menezes**[2], **Ivone Figueiredo**[3], **Bárbara Serra-Pereira**[4], **Manuela Azevedo**[5]
[1] Centro de Matemática, Universidade do Minho, danyelasylva2@gmail.com
[2] Centro de Matemática, Universidade do Minho, rmenezes@math.uminho.pt
[3] Divisão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, ifigueiredo@ipma.pt
[4] Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, bpereira@ipma.pt
[5] Divisão de Modelação e Gestão de Recursos da Pesca, Instituto Português do Mar e da Atmosfera, mazevedo@ipma.pt

---

Research surveys are important to monitor the spatial distribution and abundance of fishery resources. Their sampling design is usually focused on specific species, however an efficient design may consider non-target species data. This study evaluates the adequacy of sampling designs for the bottom trawl survey off the Portuguese continental coast, maximizing the accuracy of *Raja clavata* biomass estimates and maintaining the accuracy of *Merluccius merluccius* (target species) abundance estimates.

Keywords: Spatial sampling design, Species distribution models, Hurdle models, Research surveys, *Raja clavata*

Research surveys are important to monitor the spatial distribution and abundance of fishery resources. Their sampling design is usually conceived with the focus on specific species. However, an efficient design may reconcile this objective with the collection of non-target species data. This study evaluates the adequacy of different sampling designs for the bottom trawl survey off the Portuguese continental coast, aiming to maximize the accuracy of *Raja clavata* biomass estimates while maintaining the accuracy of the abundance estimates of *M. merluccius*, one of the target species of the survey.
Samples have been collected from fishing hauls performed during IPMA demersal research surveys which are held along the Portuguese continental coast. Fishing hauls were performed using a bottom trawl fishing gear at depth ranging from 20-750m. From 2013 to 2016, *R. clavata* was caught in a total of 212 fishing hauls, being observed at least in 49 different locations each year. Each fishing haul is identified by a pair of coordinates (longitude and latitude); total and exploited biomass (i.e., considering only specimens with

total length larger than 50 cm) in Kg per hour and total abundance (in number of specimens per hour) of *R. clavata*; sector (CAM: Caminha, MAT: Matosinhos, AVE: Aveiro, FIG: Figueira da Foz, BER: Berlengas, LIS: Lisboa, SIN: Sines, MIL: Vila Nova de Mil Fontes, ARR: Arrifana) and stratum (combination of sector and bathymetry level (0-100m or 100-200m)). In addition to biomass and abundance data, we have considered two environmental variables, bathymetry and type of substratum as they are known to impact the habitat selection of *R. clavata* across the area of study.

A common feature in abundance datasets is the semi-continuous nature of the response variable, and it is also present in our datasets of study. Another evidence on the analysis of the biomass data of *R. clavata* is the high number of zero values [2]. Thus, a model-based geostatistics is developed taking into account the specificities of the data. Species abundance is modelled as two independent processes: one dealing with the presence/absence data and the other with the intensity given a nonzero response, supported by the SPDE approach [1]. The results of the assumed model showed that the probability of occurrence of *R. clavata* increases for locations corresponding to the mixed sediment substratum; and the bathymetry has a negative impact on *R. clavata* occurrence at mixed sediment substratum and on biomass values for all substrates.

The selection of sampling designs rely on the prediction process results and eight optimization measures. Eight survey designs are evaluated, each adding 11 stations to the 54 stations identified as fixed according to the standard survey protocol. Species estimates are compared to those obtained from a baseline design with the location of stations randomly selected from a homogeneous spatial Poisson process. Three sampling designs resulted in higher accuracy of *R. clavata* biomass and *M. merluccius* abundance estimates than that obtained with a random selection of stations. The sampling design maximizing the accuracy of *R. clavata* biomass estimates shows an acceptable trade-off between bias and variance of *M. merluccius* abundance estimates. The approach presented in this study is easily replicated to other group of species caught by the research survey.

Acknowledgements: Supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within projects UIDB/00013/2020, UIDP/ 00013/2020 and PTDC/MAT-STA/28243/2017. The first author also acknowledges Foundation FCT for funding this research through Individual Scholarship PhD. PD/BD/150535/2019.

References

[1] Finn Lindgren, Havard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Statistical Methodology*, 73(4):423–498, 2011.

[2] Alain F. Zuur, Elena N Ieno, and Anatoly A. Saveliev. *Beginner's Guide to Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA*, volume I: Using GLM and GLMM, chapter 20: GAM with correlation and zero-inflation in R-INLA using owl data. Highland Statistics Ltd., 2017.

10 December, 9:20 - 9:40

Estimating echolocation clicks rates in narwhals (Monondon monoceros)

Diana Marques[1], **Tiago Marques**[2,3], **Susanna B. Blackwell**[4], **Mads Peter Heide-Jørgensen**[5], **Carolina Marques**[6]
[1] Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal, marquesdiana98@gmail.com
[2] Centre for Research into Ecological and Environmental Modelling, University of St Andrews, The Observatory, Buchanan Gardens, Fife, KY16 9LZ, UK, tiago.marques@st-andrews.ac.uk
[3] Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Centro de Estatística e Aplicações da Universidade de Lisboa, Bloco C6, Piso 4, 1749-016, Lisboa, Portugal, tamarques@ciencias.ulisboa.pt
[4] Greeneridge Sciences, Inc., Santa Barbara, California, United States of America, susanna@greeneridge.com
[5] Greenland Institute of Natural Resources, Copenhagen, Denmark, mhj@ghsdk.dk
[6] Centro de Estatística e Aplicações da Universidade de Lisboa, Bloco C6, Piso 4, 1749-016, Lisboa, Portugal, carolinasegmarques@gmail.com

---

In this work we estimate Narwhal acoustic cue production rates, using data from 8 acoustic tags deployed in East Greenland narwhals. Using a generalize additive modelling framework we predict sound production as a function of depth, which can be used to estimate cue production when tags with depth but without acoustics are available. We provide average cue rates and respective precision measures that can be used to inform passive acoustic density estimation exercises.

Keywords: acoustic behavior, cue rate, density estimation, East Greenland, tags

---

Narwhals in East Greenland are declining and particularly sensitive to climate change [3]. Density estimation is fundamental for narwhal's management and conservation. If passive acoustic approaches are to be considered, knowledge on the species' acoustic behavior is fundamental [1]. Consequently, it is important to obtain a mean cue rate estimate that could be used for turning a density of cues into a population density [2].
In this work, we use Acousonde tag data from eight whales, six females and two males, collected in Scoresby Sound, East Greenland, during the August months of the years 2013 to 2016, and in August 2019, to provide a first estimate of click production rate of narwhals and a corresponding precision measure. Further, by modelling withing a generalized additive model framework (1) the probability of an animal being clicking as a function of depth and (2) the number of clicks per second while clicking, we propose a

method for estimating click rates based on tags providing depth profiles alone, i.e. without an acoustic record [4]. This is illustrated for one of our males, for which we estimate the cue rate for the entire duration of the tag while only the first half of the tag was processed for the sound data. We explored differences in click rates being dependent on sex or depth. The sound was originally processed for clicking and non-clicking periods, but the exact number of clicks per time unit during clicking periods was not available. We developed a sampling scheme and counted echolocation clicks per second with the help of the Software MT Viewer, from which the cue rate can be calculated.
The data showed that narwhals were regularly clicking at depths deeper than 400m. The data, not surprisingly, showed that narwhals echolocate regularly over time and that there was a slight tendency to produce more clicks per second at deeper depths. The estimated click rate depends on whether an unweighted or weighted (by tag duration) click rate was considered, and we discuss the implications of said decision. The estimated unweighted average was 1,33 clicks/s (with a coefficient of variation equal to 10,27%) and the value obtained for the weighted average was 1,23 clicks/s (with a coefficient of variation equal to 9,45%). Evidence for differences in click rates being animal sex- or length-dependent could not be found. Nonetheless, the considerably small sample size associated with this proof-of-concept approach implies that further data should be collected before any definite conclusion regarding which factors might affect cue rate production is made. In the future, it would be interesting to increase sample size available to evaluate whether sound production rates are influenced by either sex or animal length, and other variables, including narwhals' prey distribution, ambient noise or animal density itself.

References

[1] Susanna B. Blackwell, Outi M. Tervo, Alexander S. Conrad, Mikkel H. S. Sinding, Rikke G. Hansen, Susanne Ditlevsen, and Mads Peter Heide-Jørgensen. Spatial and temporal patterns of sound production in east greenland narwhals. *PLOS ONE*, 13(6):e0198295, 2018.

[2] T. A. Marques, Len Thomas, Stephen W Martin, David K Mellinger, Jessica A Ward, David J Moretti, Danielle Harris, and Peter L Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2):287–309, 2013.

[3] Terrie M. Williams, Shawn R. Noren, and Mike Glenn. Extreme physiological adaptations as predictors of climate-change sensitivity in the narwhal, monodon monoceros. *Marine Mammal Science*, 27(2):334–349, 2011.

[4] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

10 December, 9:40 - 10:00

A new selection index to perform polyclonal selection in ancient grapevine varieties

Sónia Surgy[1], **Jorge Cadima**[2], **Elsa Gonçalves**[3]
[1] LEAF—Linking Landscape, Environment, Agriculture and Food—Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, soniasurgy@isa.ulisboa.pt
[2] CEAUL - Centre of Statistics and its Applications, Faculdade de Ciências; Instituto Superior de Agronomia, Universidade de Lisboa jcadima@isa.ulisboa.pt
[3] LEAF, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, elsagoncalves@isa.ulisboa.pt

---

When genetic selection is based on several traits simultaneously, superior genotypes are usually identified using a selection index. Several selection indices have been proposed in the context of plant breeding. However, some disadvantages have been identified. In this work a new selection index is proposed to perform polyclonal selection in ancient grapevine varieties.

Keywords: Selection indices, Linear Mixed Models, Genetic selection

---

The genetic selection of ancient grapevine varieties seeks genetic gains in the most important agronomic and oenological quantitative traits. Quantitative genetics relies on linear mixed models theory to estimate the variance components of the random variables in the model and to predict the Empirical Best Linear Unbiased Predictor (EBLUP) of the genotypic effect of each genotype/clone for a given trait. The selection of the best genotypes is based on these predictors. When working with several traits, a selection index is the most practical way to identify the genotypes that can satisfy the selection purposes. A selection index integrates the information related to several traits of each genotype into a single value. The Smith-Hazel index [1] is the most frequently used, among the existing classic indices. When adapted to the context of grapevine, for a given genotype j, it can be written as:

$$I_{SH_j} = \sum_{k=1}^{p} w_k h_k^2 y_{k_j} \tag{1}$$

where w_k is the economic weight for trait k; h_k^2 is the broad-sense heritability, at genotype mean level, for trait k (the ratio between the estimate of the genotypic variance component and the phenotypic variance component, at genotype mean level, for trait k); y_{k_j} is the phenotypic value of the genotype j for the trait k; and p is the number of traits.
When classical univariate linear mixed models are fitted (balanced designs, with one random effects factor - genotype, and with diagonal variance-covariance matrices), the Smith-Hazel index can be applied according to the equation 1. In this form, this index is very

susceptible to the scale differences between the units in which the different traits are expressed and its application to more complex models is not adviseable. Thus, a new selection index was developed taking into account the disadvantages previously mentioned. The new index proposed in this work (I_j), for a given genotype j, is defined as:

$$I_j = \frac{\sum_{k=1}^{p} \left(\frac{w_k EBLUP_{k_j}}{\overline{y}_{k..}} \right)}{\sum_{k=1}^{p} \left| w_k \frac{EBLUP_{k_j}}{\overline{y}_{k..}} \right|} \tag{2}$$

where I_j is the index value for the genotype j; $EBLUP_{k_j}$ is the EBLUP of genotypic effect of the genotype j for the trait k; $\bar{y}_{k..}$ is the phenotypic mean of the studied population for the trait k; w_k is the economic weight for the trait k; and p is the number of traits.
The proposed index uses the EBLUP of the genotypic effect instead of the broad-sense heritability and phenotypic value, which enables its application with more complex models. Additionally, for a given trait, the EBLUP of the genotypic effect is divided by the overall phenotypic mean, which makes this index resistant to differences between the units in which the traits are expressed. Furthermore, since the denominator is the absolute value of the numerator, the value of the index belongs to the interval $[-1, 1]$, resulting the values 1 and -1 from genotypes with positive predictors and negative predictors, respectively, for all traits. Therefore, this index allows not only to rank the genotypes, but also to assess a genotype by itself. The genotype is better, the closer its index value is to 1. All these features distinguish this index from the existing ones.
Using *nlme* package of R software, linear mixed models were fitted to yield and grape quality traits data obtained in selection field trials of several grapevine varieties. To perform polyclonal selection (selection of a group of genotypes) based on those traits, an Excel macro was constructed. The genotypes were ranked according to the value of each index (Smith-Hazel index and the new proposed index) and the superior 20 genotypes were selected. For each group of selected genotypes, the predicted genetic gain for each trait (the mean of the EBLUPs of the genotypic effects of the 20 selected genotypes) was computed. The ultimate goal is to select a group of genotypes revealing genetic gains for all the important traits. The results confirmed the above mentioned disadvantages of the Smith-Hazel index. Polyclonal selection performed with the new proposed index revealed genetic gains in all the traits used for selection.
Acknowledgements Sónia Surgy thanks the national funding by Fundação para a Ciência e Tecnologia (FCT), Portugal through the Ph.D. grant 2020.07338.BD. LEAF and CEAUL have been funded by FCT UIDP/04129/2020 and UID/MAT/00006/2019, respectively.

References

[1] P. Cotterill and C. Dean. *Successful tree breeding with index selection.* CSIRO, Australia, 1990.

10 December, 10:00 - 10:20

Waste management on subsurface ships: a case study

M. Filomena Teodoro[1,2], **Suzana Lampreia**[2], **Tomás Mendes**[2]
[1] CEMAT, IST, Lisbon University,Portugal, mteodoro64@gmail.com,
[2]CINAV, Naval Academy, Portuguese Navy, Portugal

The objective of this investigation is to analyze and evaluate the waste management system in Portuguese Ships "Tridente" and "Arpão", considering the different types of waste on board and the used treatment systems, taking into account the legal and doctrinal framework in the Navy, national and international legislation and the state of the art of the scientific community. Military representatives of both garrisons of the Trident class submarines were interviewed, the staff who were interviewed did not answer a questionnaire. The surveyed participants did perform management functions or waste treatment.

Keywords: waste management, environment, marine pollution, submarines, EFA

The management of waste [4] on board ships can take complex proportions due to the number of people in the crew, the maintenance operations that may occur while sailing, the type and time of mission or just a less careful management. When a naval organization uses submarines in its missions, waste management must comply [1, 2] with a management model implemented and followed by all elements of the garrison. It was intended to demonstrate the importance of waste management in the Portuguese Navy's submarines, carrying out a study case. The Portuguese Navy is a branch of the Armed Forces that uses the sea and complies with legal regulations [3] that aim to reduce its ecological footprint. It was carried out a research work using interviews with elements of both crews of Trident class submarines, who perform functions related to the management and treatment of waste on board. Also, a questionnaire, built and validated in [5], was implemented considering the submarine garrisons, whose analysis and processing of data obtained from respondents' answers are subject to statistical treatment: first a descriptive analysis followed by an Exploratory Factorial Analysis (EFA), similarly to the study find in [6]. We were able to identify important queries to the latent variables that take into account waste management skills. Analysis of variance techniques (both univariate and multivariate cases) were used to obtain important independent variables that helped to explain the chosen EFA factors. The initial part of questionnaire concerns the socio-demographic information about each participant. The second part consists in questions of open or closed response, with the possibility of choosing more than one answer in each question and also some questions in the form of Likert scale. This second part aimed at evaluating participants' knowledge, attitudes and practices regarding waste management, comprising questions about knowledge issues, other about attitudes and some questions that consider practice details. It is

concluded that the existing training actions about the legal regulations on waste management on board are not directly based on instructing the Tridente class submarine crews. However, on board waste management is based on a perspective of applying good practices and common sense.
In a preliminary data analysis of questionnaires, and taking into account the non-quantitative nature of the involved variables, were calculated measures of association, nonparametric Spearman correlation coefficient, nonparametric test of Friedman for paired samples, etc. The application of EFA, when we restrict the variance of each factor greater than one, conduced to the following factors: treatment of oily waste, treatment of urban solid waste and garrison behavior, treatment of special waste, knowledge of environmental rules, suitability of equipment and processes, improvement of on-board equipment, responsibility and formation of the garrison.

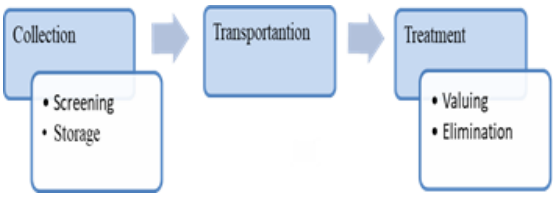

Figure 1: Waste management system scheme

Acknowledgements This work was supported by Portuguese funds FCT, through the CEMAT, University of Lisbon, Portugal, project UID/Multi/04621/2019, and CINAV, Portuguese Naval Academy.

References

[1] L. Bilgili. Environmental and economic analysis of waste management scenarios for a warship in life cycle perspective. *Journal of Material Cycles and Waste Management*, 5:116–130, 2020.

[2] L. Brioschi, L.C. Simonetti Gonçalves, and A. Sant'Anna Pedra. Dever internacional de reciclagem de resíduos plásticos pelos navios. *Revista Científica Foz*, 2(2):71–88, 2019.

[3] Associação Portuguesa de Certificação. Np en ISO 14001:2015 – sistema de gestão ambiental. `https://apcergroup.com/pt/certificacao/pesquisa-de-normas/169/iso-14001`, accessed at 2/9/2021, 2019.

[4] Agência Portuguesa do Ambiente. Resíduos. `https://apambiente.pt/index.php?ref=16&subref=84`, accessed at 2/9/2021, 2021.

[5] J.B. Rebelo. *Impacto Ambiental da Marinha Portuguesa. Análise e resolução da Gestão de Resíduos no mar*. Master thesis, Escola Naval, 2019.

[6] M.F. Teodoro, J.B. Rebelo, and S. Lampreia. Waste management and embarked staff. In O. Gervasi and et al, editors, *Computational Science and Its Applications – ICCSA 2020, Lecture Notes in Computer Science*, volume 12251, pages 402–503. Springer, 2020.

11 December, 9:00 - 9:20

Sparse Divisive Feature Clustering

Ndèye Niang[1], **Mory Ouattara**[2], **Gilbert Saporta**[3]
[1] CEDRIC-CNAM, ndeye.niang_keita@cnam.fr
[2] SFA-UNA, ouattaramory.sfa@univ-na.ci
[3] CEDRIC-CNAM, gilbert.saporta@cnam.fr

We propose an approach based on a divisive algorithm for clustering variables in order to identify in a large data table underlying dimensions that are not necessarily orthogonal. The number of clusters does not have to be defined in advance. The clusters, which are as unidimensional as possible, are then represented in a parsimonious way by a small number of variables or components.

Keywords: Feature Clustering, Simple structure, Sparse principal component

Let us consider the context of unsupervised analysis of a large data set where variables are assumed to be structured in homogeneous blocks. Two situations may occur: either the features are divided into blocks defined beforehand (expert knowledge or "natural" clusters such as answers to questionnaires according to different themes) and multiblock methods such as STATIS or RGCCA are well adapted. We are interested here in the alternative case where the blocks are constructed from the data. Finding blocks of variables is linked to the objective of reducing the dimension of the features space in order to facilitate the interpretation. But also, more fundamentally, it is related to the objective of discovering simple structures as in factor analysis in the sense that each factor would be correlated with a small number of features and each feature would be correlated with few factors. It is then natural to look for clusters that are as unidimensional as possible. Each cluster will be summarized by a prototype or by a parsimonious linear combination of features.

Feature clustering solves problems that PCA cannot address because of dual orthogonality constraints on factors and components. The orthogonality constraints lead to optimal projections for units but usually not to simple structures for the components. Hence the use of orthogonal or oblique rotations or sparse PCA [4]. Sparse PCA facilitates the interpretation because each sparse component is related to few features, but the degree of sparsity depends on a parameter whose tuning remains problematic.
Compared to the clustering of individuals, variable clustering received much less attention in the literature. Proposed methods often simply copy usual methods of clustering of units which is intellectually unsatisfactory. Among specific methods for classifying features, let us mention the latent variables based CLV method [6], ClustOfVar package [1], the interpretable principal components based on [2] and the methods based on the likelihood of the link [3]. Previous methods are hierarchical or K-means like and suffer from well-known shortcomings: hierarchical methods are not adapted to the case of a large number of objects, K-means like methods assume that the number of clusters is fixed in advance.

The VARCLUS procedure of SAS software, which has never been the subject of scientific articles, presents several interests. It is a top-down hierarchical method that separates iteratively the features into two sub-clusters until there is only one eigenvalue larger than 1 in the PCA of each cluster. The condition on the eigenvalues gives VARCLUS two important advantages: the number of clusters is naturally obtained and the resulting clusters which are associated with a first large eigenvalue are unidimensional to some extent. Clustering features into unidimensional blocks is a simple and efficient way to search for so-called "oblique" factors. Each block can then be represented by a single component combining only its features and then necessarily sparse relatively to the number of variables. When the cluster size is still too large, a further simplification is needed.

We therefore propose a multi-step strategy. Firstly, VARCLUS is performed. This provides the optimal number of clusters and the associated partition is used as an initialization for CLV, which avoids the computational cost of the hierarchical agglomerative clustering or of the several random runs of an initial partition as proposed to get the number of clusters in CLV method. Secondly CLV is performed and noise (or isolated) variables are discarded in order to keep only relevant clusters in the spirit of [5]. For the last step of prototype determination: the prototype can be the first sparse principal component, its closest feature in terms of maximal correlation or the ≪ medoid ≫ feature.
All these different strategies are evaluated on simulated data and illustrated on real data.

References

[1] M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco. Clustofvar: An R package for the clustering of variables. *Journal of Statistical Software*, 50(13):1–16, 2012.

[2] D.G. Enki, N.T. Trendafilov, and I.T. Jolliffe. A clustering approach to interpretable principal components. *Journal of Applied Statistics*, 40(3):583–599, 2013.

[3] F. Nicolau and H. Bacelar-Nicolau. Some trends in the classification of variables. In *Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan*, pages 89–98. Springer, 1998.

[4] N. T. Trendafilov. From simple structure to sparse components: a review. *Computational Statistics*, 29(3):431–454, 2014.

[5] E. Vigneau and M. Chen. Dimensionality reduction by clustering of variables while setting aside atypical variables. *Electronic Journal of Applied Statistical Analysis*, 9(01):134–153, 2016.

[6] E. Vigneau and E.M. Qannari. Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, 32(4):1131–1150, 2003.

11 December, 9:20 - 9:40

Clustering domestic water consumption profiles: a Portuguese case study

Elisa Araújo[1], **Flora Ferreira**[2], **Duarte Silva**[3], **Estela Bicho**[4], **Wolfram Erlhagen**[1]

[1] Department of Mathematics, University of Minho, pg40157@alunos.uminho.pt

[2] Centre of Mathematics, University of Minho, fjferreira@math.uminho.pt, wolfram.erlhagen@math.uminho.pt

[3] Águas do Norte, duarte.silva@adp.pt

[4] Algoritmi Center, University of Minho, estela.bicho@dei.uminho.pt

Regional water systems serve consumers of different consumption profiles. For an efficient planning and management of water demand it is crucial to understand this heterogeneity. Having this challenge in mind, this study aims to detect domestic water consumption profiles of a group with 483 individual customers from the north of Portugal. A cluster analysis (based on the K-means algorithm) was used to identify subgroups of households taking into account their average hourly, daily, and monthly water consumption. We find that 85% of the households can be clustered into six groups. The hourly, daily, and monthly water consumption curves of each cluster provide a summary view of consumption profiles which can be used to obtain estimates of the spatio-temporal distribution of demand, thus allowing better monitoring and management of the water demand.

Keywords: clusters analysis, household water consumption, K-means

Water companies are concerned with efficient management of the water demand for increasing numbers of households in order to detect and decrease water loss. One of the challenges in the development of intelligent models to monitoring and planning water demand is the heterogeneity of the consumers served by the regional water system [1]. With the emergence of smart meters, a better understanding of the consumption profiles is now accessible. Recently, the Portuguese company, Águas do Norte SA, has been installing smart water meters in their network system with the goal to improve the management of water supply and network operation. In this study, we take the first dataset provided by these smart meters to identify distinct groups of customers according to average hourly, daily, and monthly water consumption. The dataset contains water consumption measurements in liters per hour (L/h) gathered between December 2020 and June 2021 (7 months) from 483 households located in the north of Portugal. Cluster analysis was performed with the K-means clustering algorithm separately for each temporal scale: hourly, daily, and monthly consumption basis. In each case, principal component analysis (PCA) was

applied to reduce the data dimension prior to K-means clustering. The Elbow Method was used to find the optimal number of clusters. Then, the results obtained in the three cases were organized in a triple table and the customers with similar consumption behavior at all temporal scales were clustered, and water consumption curves were calculated for each cluster.

We found 6 main clusters (clusters with more than 12 households) containing around 85% of the households. A small monthly variability in the average water consumption was observed in all these 6 clusters. The main difference between them is the amount of consumed water with cluster 1 having the lowest and cluster 6 having the highest average consumption. However, different daily and hourly water consumption profiles were also observed (Figure 1). The 130 households assigned in cluster 1 show low water consumption (about 28 liters per day) with small daily and hourly variability. Cluster 2, 3, and 6, with 98, 84, and 22 households, respectively, show a similar daily consumption profile with a peak in the morning, a decline in usage during the afternoon and a high peak again in the evening. A slightly higher consumption on Friday compared to the other days of the week was also observed in these three clusters. Clusters 4 and 5, with 48 and 30 households, respectively, show a similar daily consumption profile with the highest peak of consumption in the evening, followed by another considerably lower peak at the end of the morning. Both show higher consumption on Friday and Saturday.

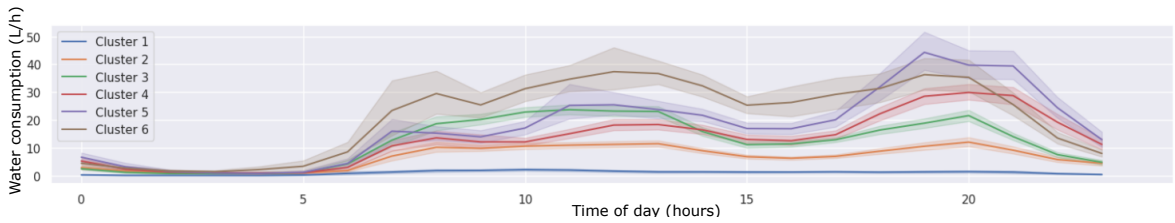

Figure 1: Hourly water consumption curves of the main 6 clusters.

These results provide for the company a summary view of the temporal water consumption profiles of their customers, which can be used to obtain estimates of the spatio-temporal distribution of water demand. The next steps of the ongoing investigation are the upscaling of the work to all company customers and the development of a water monitoring model for each of the groups formed by the clustering algorithm capable of detecting possible abnormalities in water consumption.

Acknowledgements Supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] Noa Avni, Barak Fishbain, and Uri Shamir. Water consumption patterns as a basis for water demand modeling. *Water Resources Research*, 51(10):8165–8181, 2015.

11 December, 9:40 - 10:00

Clustering times series of electricity consumption

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

---

Modelling the behaviour of electricity load time series has many applications in power systems. The goal of the present work is to obtain clusters of electricity consumption time series, data referring to the Portuguese Transmission System Operator (TSO). A semi-hierarchical clustering procedure is adopted which reveals some consumption patterns that could be anticipated, but also new (more detailed) patterns, invisible to the experts' eyes.

Keywords: Clustering time series, distance measures,electricity load data

---

Understanding the behaviour of electricity load time series enables better investment decisions and contributes to the forecasting of electricity load, which allows the fine tuning of the daily operations of the power system, such as the scheduling of power plants, or the reconfiguration of electrical network topology.
We conduct a clustering analysis of electricity consumption time series. The data were obtained through the Portuguese TSO website with a discretization of 15-minutes.
The K-Medoids algorithm is used to cluster the daily load time series data. As result, well separated groups of days are obtained, such that days within the same group have similar load profiles and dissimilar days are allocated to different groups.
We capitalize on the K-Medoids capacity of dealing with several distance measures to provide different clustering alternatives: 1) Euclidean distance is used to capture scale differences; 2) a Pearson based distance identifies the similarity of trends; 3) a Periodogram based distance identifies cyclical behaviour; 4) an autocorrelation based measure is able to distinguish between time series' autocorrelation structures. Finally, we resort to the use COMB distance, of a convex combination of the four referred distance measures, [1], which captures differences between time series from diverse perspectives and is able to provide meaningful results. The number of clusters is determined using the Silhouette index [2].

In the preliminary data analysis we deal with some missings in times series raw data, resorting to imputation: time series have one missing hour (Daylight Saving Time), that was imputed by the average of the two nearest hours data, and a redundant hour data that was removed.
The first clustering results obtained, based on individual distances 1) to 4), provide some interesting insights but also very limited ones. The first K-Medoids analysis based on COMB

indicates the existence of two clusters essentially dividing working and non-working days. To obtain a richer solution from a substantive point of view, we conduct a semi- hierarchical clustering procedure by replicating the same K-Medoids procedure within the clusters constituted. In the second clustering level four clusters are obtained which are able to separate the seasons of the year. An additional level of clustering (within the previous level clusters) reveals relevant patterns of electricity consumption which could not be identified with an expert's eye only. E.g. within a first level cluster that mainly distinguishes non-working days, and within a second level cluster that mostly reunites Autumn and Winter days, we discover one third level cluster that also includes five "working days" which are Christmas Eves and New Year's Eves and another that includes every Christmas days. We then conclude the analysis with nine (third level clusters).
Future research should focus on the weights of COMB distance, eventually relying on measures of agreement between the clusterings derived from individual distance measures and the COMB based clusters.

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

References

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

[2] P.J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, pages 197 – 212, 1987.

11 December, 10:00 - 10:20

Clustering of EFCs in European Economies

Eliana Costa e Silva[1], **Aldina Correia**[1], **Ana Borges**[1]
[1] CIICESI, ESTG, Politécnico do Porto, Felgueiras, Portugal, {eos,aic,aib}@estg.ipp.pt

The Global Entrepreneurship Monitor (GEM) project annually assess the national level of entrepreneurial activity. Here, the 12 indicators of the entrepreneurial ecosystem are studied. Clustering is used to infer how European experts' perceptions have changed between 2000 and 2019. For each year, internal validation measures and four different algorithms were used. The optimal validation measures were obtained for two clusters and hierarchical methods. Significant differences between the clusters obtained over the years were found.

Keywords: Clustering, entrepreneurship, European economies

Entrepreneurship is explained as an individual's ability to place ideas into practice, articulating project planning, management, ability to take calculated risks, innovation and creativity to achieve previously defined goals [2]. Part of GEM consists on gathering experts' opinions concerning components of the entrepreneurship ecosystem using a National Expert Survey (NES) [3]. The present study focus on the 12 Entrepreneurial Framework Conditions (EFCs) compiled by the NES survey. To study the European countries, based on the experts' perceptions on EFCs over the years, cluster analysis was used to group the countries in homogeneous groups. For each year, in order to determine the best number of clusters, Hierarchical, Partitioning Around Medoids, K-means and Fuzzy ANalYsis methods were used. Furthermore, Dunn Index, Silhouette Width and Connectivity internal validation measures were computed. For the majority of the years, the optimal values of the validation measures were found for two clusters and hierarchical methods. Thus, $k = 2$ clusters and hierarchical agglomeration method was considered. It was possible to observe significant differences between the clusters obtained over the years. Furthermore, the distribution of the countries in each cluster varies considerably (Fig. 1). The European economies in Cluster 1 present an average below, while the ones on Cluster 2 have an average above, the global average of the 12 EFCs (for details see [1]). E.g., in 2019 the overall average was 2.85, while in Clusters 1 and 2 it was 2.61 and 3.10, respectively. Thus, the perceptions of entrepreneurship are higher in Cluster 1 economies, namely Germany (DE), Ireland (IE), Latvia (LV), Luxembourg (LU), Netherlands (NL), Norway (NO), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), United Kingdom (GB).
To understand the pattern and the differences in the cluster agglomeration over the years, the allocations of the European economies with the best three and the worst three values of the Total Early-Stage Entrepreneurial Activity[1] (TEA) values, were considered. Italy

[1]TEA indicator represents the percentage of the population with ages between 18 and 64 years old who are either a nascent entrepreneur or owner-manager of a new business.

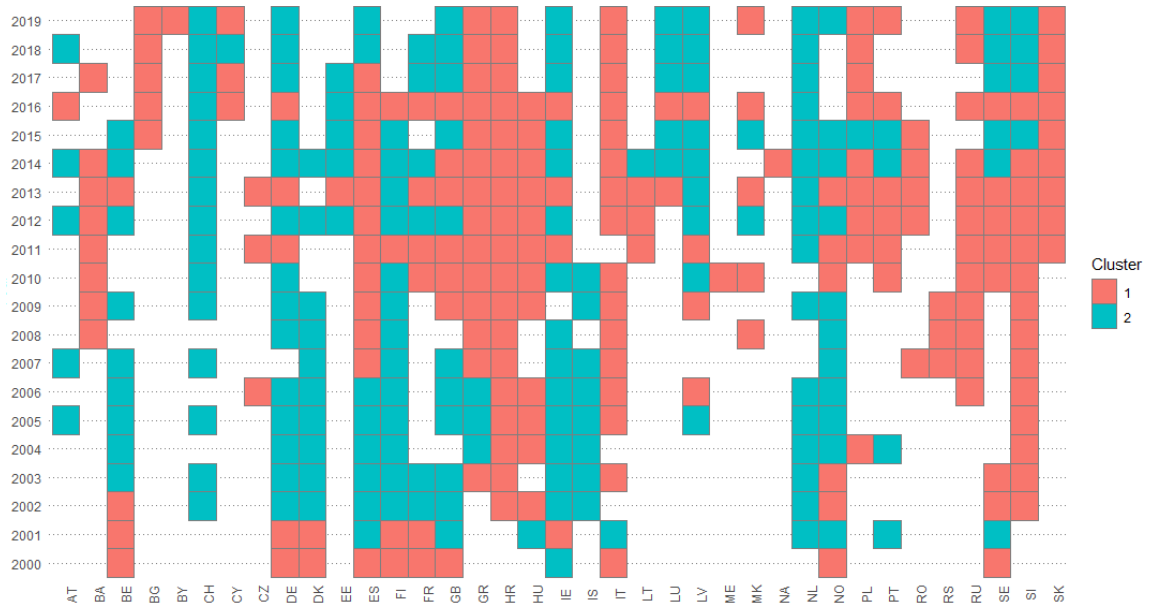

Figure 1: European economies' cluster membership.

(IT), Poland (PL) and Belarus (BY) are the three countries with lower TEA values (2.79, 5.39 and 3.78, respectively), and for all the years, except for Poland 2015, they are allocated to Cluster 1. For 2019, the three economies with the highest TEA are Latvia (LV), Slovakia (SK) and Portugal (PT). Latvia (TEA=15.43) and Portugal (TEA=12.89) change its allocation between Cluster 1 and Cluster 2. Contrary to what was expected, Slovakia (TEA=13.33) maintains its allocation to Cluster 1. The results warn for the need to consider annual and intra-country dynamics. Most studies (e.g.[2, 1]) perform cross-sectional studies combining information from GEM to group economies. However, neglecting to consider a longitudinal dynamic, may result in biased results. In future, a longitudinal clustering approach will be performed.

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

References

[1] E. Costa e Silva, A. Correia, and A. Borges. Unveiling the dynamics of the european entrepreneurial framework conditions over the last two decades: A cluster analysis. *Axioms*, 10(3), 2021.

[2] L. Farinha, J. Lopes, S. Bagchi-Sen, J.R. Sebastião, and J. Oliveira. Entrepreneurial dynamics and government policies to boost entrepreneurship performance. *Socio-Economic Planning Sciences*, 72:100950, 2020.

[3] S. Singer, M. Herrington, and E. Menipaz. Global Entrepreneurship Monitor: Global Report 2017/18. Technical report, 2018.

11 December, 11:00 - 11:20

Visually supported curriculum development

Rogério Duarte[1], **Ângela Lacerda Nobre**[2], **Fernando Pimentel**[1], **Marc Jacquinet**[3],
[1] ESTSetúbal, CINEA, Instituto Politécnico de Setúbal
[2] ESCE, Instituto Politécnico de Setúbal
[3] DCSG, CEMRI, Universidade Aberta

Curriculum development processes are of paramount importance for the preparation of future professionals. An impediment to these processes lays in stakeholders' communication difficulties. This study considers the combination of information and data science techniques namely, the use of classification, natural language processing and network graphs, to make curriculum representations self-explanatory, bridging communication gaps and helping stakeholders to articulate their expert knowledge.

Keywords: Network graphs, Natural language processing, Classification, Curriculum analytics, Education

---

Accreditation bodies require "scientific" curriculum development processes reliant on outcomes based education and on constructive alignment principles. To implement these principles *Learning Outcomes Mappings* use courses as building blocks and the visual representation—typically a chart, table or map—depicting vertical (from year to year) and horizontal (within a year) relations between courses conveys a sense to the study program structure. From the analysis of each course Learning Outcomes (LO) it is possible to infer further connections between courses and gain access to the "mechanics" behind a program plan. However, course LO presume tacit understanding of concepts specific to disciplinary and scientific sub-areas and this renders LO-statements seldom clear and unequivocal.

This study reports the combination of information and data science techniques including classification of LO-statements, natural language processing, the use of network graphs, to communicate—visually and quantitatively—information included in the study program that is typically difficult to access and to articulate.

Figure 1, obtained as described in detail in [1], is the end product of a data visualization method that considers the following steps: (1) classification of course learning objective statements into *broader terms*; (2) use of Natural Language Processing (NLP) to convert *broader terms* into quantitative frequencies of key program concepts; (3) visualization of the links between courses based on common key concepts.

In Figure 1, `MATH`, `PHY`, `STAT`, etc. are acronyms for courses included in the study program. The relative position of the courses as well as the number of connecting lines provides both

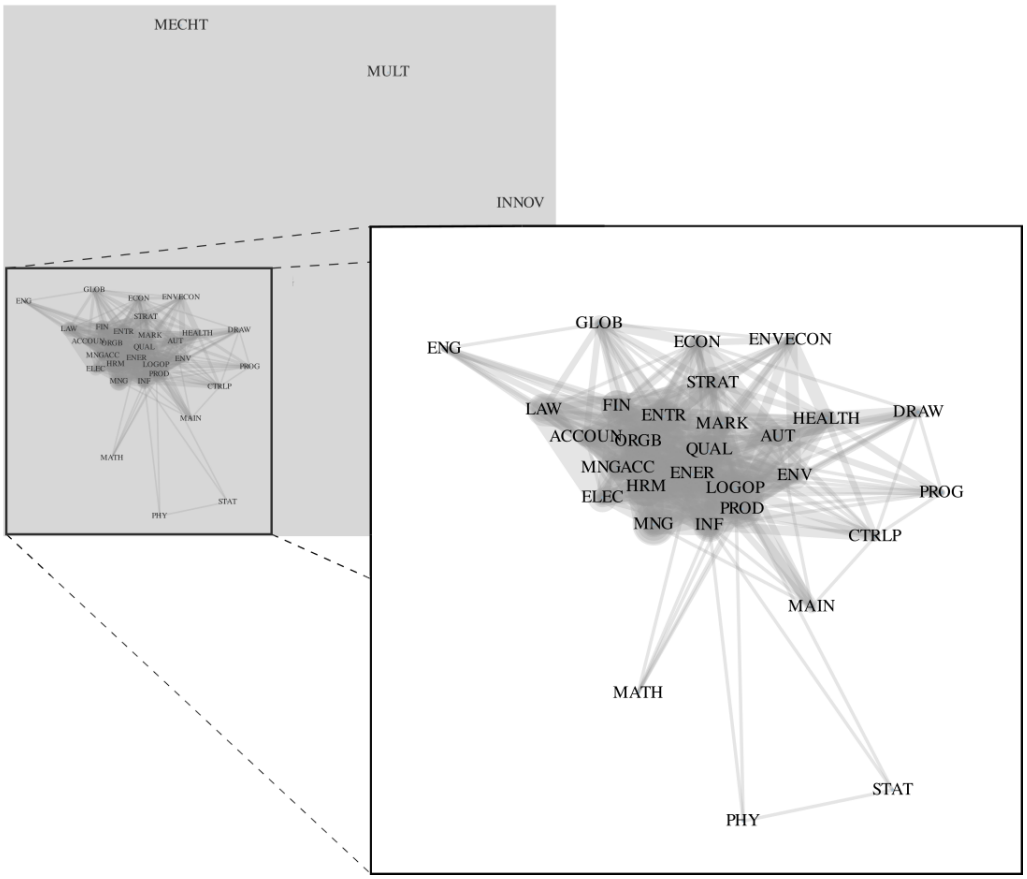

Figure 1: Visualization of links between courses of a bachelor degree in Technology and Industrial Management [1]. The forward (white) plane presents an enlarged detail with 30 of the 33 courses present in the background (gray) plane.

visual and quantitative measure of courses connectivity. Figure 1 highlights the lack of connections to three courses depicted in the background (gray) plane, `MECHT`, `MULT`, `INNOV`. It becomes obvious (inspecting the forward white plane) that fundamental courses such as mathematics, physics, statistics (`MATH`, `PHY`, `STAT`) are detached from the figure's mass center, where the majority of the courses lay.

References

[1] Rogério Duarte, Ângela Lacerda Nobre, Fernando Pimentel, and Marc Jacquinet. Broader terms curriculum mapping: Using natural language processing and visual-supported communication to create representative program planning experiences. *CoRR*, abs/2102.04811 (https://arxiv.org/abs/2102.04811), 2021.

11 December, 11:20 - 11:40

How to analyze online behavior as a source for political information in the Portuguese 2019 European Parliament election?

Cláudia Silvestre[1], **Rodrigo Pinheiro**[2], **Filipe Montargil**[3],
[1] ESCS – Escola Superior de Comunicação Social, csilvestre@escs.ipl.pt
[2] ISCTE – Instituto Universitário de Lisboa, rodrigopinheiro2110@gmail.com
[3] ESCS – Escola Superior de Comunicação Social and LLMCP – Living Lab on Media Content and Platforms, fmontargil@escs.ipl.pt

---

This study explores the use of online media as a source of political information in the 2019 European elections in Portugal. For this purpose we resort to a panel developed and maintained by Netquest, that collects web navigation actions, on computer and mobile devices. We analyze user's online behavior during a period of two month around the election, looking for patterns and trends in the use of political information in online media. However, we will begin this data mining journey with the challenges of data preparation.

Keywords: European elections, online behavior, political news,CRISP-DM, data preparation

---

The internet plays an important role in our society, namely in the circulation of political ideas [2] knowing that, political actors have been using web's potential to invigorate their campaigns [4]. Obama's 2008 presidential campaign is a well-known example [1].
In this study, which is part of a larger one about European elections, we intend to examine Portuguese' use of online media in relation to political involvement in the 2019 European Parliament election. This project, which is being developed in partnership with Netquest (an opinion and market research company), uses a database of web navigation actions (WNA) from its Internet user panel in Portugal. This data set includes navigation actions on computer and mobile devices, for a sample of 1,288 users. Our data were collected between April 26 and June 26, 2019 (a period of two months, around the elections, held on May 26, in Portugal), and contains 20,137,355 WNA.
In order to analyze this data set we applied the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [3]. After business and data understanding, we are in phase 3, the most time-consuming task - data preparation. In this phase a binary variable was added to identify if a WNA refers to an online media or not. This classification was based on the list of media provided by the Entidade Reguladora para a Comunicação Social (Portuguese Regulatory Authority for the Media).
The next step is to identify which of these WNA of Portuguese media are about politics. First, we select subdomains or tabs of the WNA URL address that contains the words:

"politica" (politics), "eleicao" or "eleicoes" (election). Other options will be apply text mining to news titles in the WNA url address or use HTML scraping and text mining algorithms to analyze online news content.
This study is focused on challenges we faced during data preparation. To make sure that the database is consistent and does not include duplicate or redundant information it was necessary to understand what each variable actually represented. Then we recoded the data to numerical values. And some variables were grouped, such as the region, the level of education and the area of study. We have also turned date of birth into age and standardized standardized the time spent online. In addition, mobile and desktop WNA information have been tuned to be expressed in the same way.
Finally, we present the preliminary results of the identification of WNA related to policy issues and an exploratory analysis of the information will be carried out.

References

[1] E. Bomberg and B. Super. The 2008 US presidential election: Obama and the environment. *Environmental Politics*, 18:424–430, 2009.

[2] M. Bonchek. *From broadcast to netcast: The Internet and the flow of political information.* PhD thesis, Harvard University, 1997.

[3] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. H. Orallo, M. Kull, N. Lachiche, M. J. Quintana, and P. Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33:3046–3061, 2019.

[4] L. Parisi and R. Rega. Disintermediation in political communication: chance or missed opportunity? *Leadership and New Trends in Political Communication*, pages 123–148, 2011.

11 December, 11:40 - 12:00

The role of the pre-university life path in the performance of students who access higher education: the case study of the Master's Degree in Civil Engineering at FEUP

Fernanda Campos de Sousa[1], **Isabel Martins Ribeiro**[2]
[1] Faculty of Engineering, University of Porto, Porto, Portugal, fcsousa@fe.up.pt
[2] Faculty of Engineering, University of Porto, Porto, Portugal, iribeiro@fe.up.pt

---

This study aims to characterize the profile academical and social of the students enrolled to higher education, suggesting measures that lead to the improvement of their academic performance. It is based on the case study which covers the pre-university and first year students enrolled in the Integrated Master in Civil Engineering (MIEC), at Engineering Faculty of Porto University (FEUP) between 2015 and 2020.

Keywords: Data analysis, non-parametric tests, principal component analysis, higher education, academic performance

---

In Portugal, admission to public higher education is conducted through a national competition, when takes into consideration both secondary education grades and specific exams for the intended course. Each course has a pre-determined limited number of vacancies, which is called numerus clausus.

The first-year university student faces a series of challenges and new experiences, often intertwined with various factors and milestones: reaching adulthood, leaving parental home, moving to a new residence, learning different teaching methods, begin enrolled in the desired course, etc. However, in order to facilitate the integration of the new students, the Integrated Master's degree in Civil Engineering has developed and launched, in the academic year 2015/2016, a Peer Mentoring Programme, called CIVIL'in [1]. It is widely known that performance in the first year of the course, namely the number of curricular units completed, is a determining factor for success in the following years, in particular for the number of years required to complete the course. This explains the various measures implemented to support the integration of students in higher education and the growing attention with the first year students [1]. This work focuses on students enrolled in the Integrated Master in Civil Engineering (MIEC) at Engineering Faculty of Porto University (FEUP), between 2015 and 2020.

The area of civil construction is strongly connected to national economic cycles, having suffered a severe crisis in the past decade, which had repercussions on the ability to attract students to Civil Engineering courses. MIEC always filled the number of vacancies available, but the average application grades tended to decrease.

The objectives of this study are to characterize students enrolled to MIEC, to find pre-university curricular and/or social factors that influence their academic performance in the first year of the course, and to understand if there is any relevant trend during the 6-year period under analysis.
The information used in this work was provided by the course board. The variables under study are the student's entry score, the grade obtained in the high school, the rank position of MIEC option on the higher education application, the type of school attended in high school (namely public or private) and its location, the access regime, the entry stage in MIEC, the student status at FEUP, the number of course units completed in the first year and, when this number is not less than 5, the respective average classification, the number of completed curricular units in the scientific area of Mathematics (Mathematical Analysis 1 and 2 and Algebra) and average classification obtained, when the student completes at least 2 of these.
A univariate and bivariate descriptive analysis of the data set, for each of the years under study, considering the typology of the different variables, revealed that some modality and/or binary variables had relevant information content, namely to interpret the dependency relationships with the quantitative variables. The variables i) type of school attended in high school, public or private, and ii) location of usual place of residence, translated into permanence in the usual place of residence or need to reside abroad, depending on whether or not to the location belongs to the Greater Porto area, proved to be variables to be explored in more detail. Using non-parametric Wilcoxon Mann-Whitney tests, it was possible to verify that these variables are significantly influencing the students' performance in the first year of the course, measured by the number of course units completed, in general subjects and in mathematics, and by the respective average scores.
Interestingly, the variables student's entry score and grade obtained in the high school did not prove to be systematically determinants of the first year performance. In order to further explore the possible relationships between the quantitative variables in the study, a Principal Component Analysis was conducted for these variables and for each year. For the 6 years under study, the first 3 main axes were found to explain between 85% and 90% of the inertia of the point cloud. The first axis can be called "continuity of good performance", to which students who been enrolled with high grades and maintain the level in the first year contribute. The second axis that we call "in-adaptation/change of course", is associated with students who, been enrolled with high grades, completed a reduced number of course units. The third axis has different interpretations throughout the 6 years.

References

[1] Ribeiro I.M., Henriques A., Carvalho B., Guimarães A., and Sousa V. The civil'in programme - a peer mentoring programme for first-year students of civil engineering. *Engaging Engineering Education - SEFI 48TH ANNUAL CONFERENCE*, pages 416–431, 2020.

11 December, 12:00 - 12:20

Data Analysis to Advance Immersive Technologies: examples from Augmented Reality-based assembly procedures

Carlos Ferreira[1], **Bernardo Marques**[2], **João Alves**[2], **Paulo Dias**[2], **Beatriz Sousa Santos**[2],
[1] IEETA/DEGEIT, University of Aveiro, carlosf@ua.pt
[2] IEETA/DETI, University of Aveiro, bernardo.marques@ua.pt, jbga@ua.pt, paulo.dias@ua.pt, bss@ua.pt

---

Immersive technologies have been evolving in the latest years and have currently many applications, such as training, medicine, and entertainment. Augmented Reality (AR) enables users to interact with their augmented physical environment through the overlay of digital information, yet it still has many challenges concerning how to show and to let users interact with digital information. Adequate experimental evaluation is paramount to overcome these challenges, namely user studies, which generally entail quantitative and qualitative data collected from relatively small groups of participants. Examples of data analysis used to develop and compare interaction and visualization methods for industrial AR-based assembly procedures are presented.

Keywords: Data analysis, Augmented Reality, User studies, Assembly procedures

---

Immersive technologies have been evolving in the latest years and have currently many applications, namely in training, medicine, and entertainment to name just a few. Augmented Reality (AR) is such a technology that enables users to interact with their augmented physical environment through the overlay of digital information. While its adoption has been made easier due to the wider availability of Hardware such as smartphones, tablets and head mounted displays (HMD), and the Pokémon Go phenomenon has demonstrated AR's potential to be adopted by mainstream culture, there are still a lot of challenges to a more generalized adoption. Many of these challenges are related to the usability and acceptability of AR-based applications [2] , namely concerning how to show digital information to the users and how to let them interact with it. Using adequate experimental evaluation is paramount to overcome these hurdles, propose new interaction and visualization methods as well as develop new useful and usable systems. User studies are the "work horse" of evaluation in this field, generally entailing quantitative and qualitative data collected from relatively small groups of participants making more difficult their analysis. This communication presents examples of data analysis used to develop and compare interaction and visualization methods for industrial AR-based systems for assembly procedures [1] [3].

Figure 1: User study comparing interaction conditions based on: C1 - Mobile device surface gestures (Left); C2 - Moving the mobile device (Middle); C3 - Controllers and HMD [3].

In both cases a user study and controlled experiment was performed to assess usability and acceptance of different interaction and visualization methods for AR-based assembly scenarios. The collected qualitative and quantitative data were analysed using EDA, ANOVA, nonparametric tests and cluster analysis allowing establish the preferred methods, highlight their strengths and limitations, leading to potential advantages in specific use cases, and suggesting their integration in collaborative contexts.

Acknowledgements We thank everyone involved in the user studies for their time and expertise. This work was supported by IEETA - Institute of Electronics and Informatics Engineering of Aveiro, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UID/CEC/00127/2019.

References

[1] J. Alves, B. Marques, C. Ferreira, P. Dias, and B. Sousa Santos. Comparing augmented reality visualization methods for assembly procedures. *Virtual Reality*, 2021.

[2] M. Billinghurst. Grand challenges for augmented reality. *Frontiers in Virtual Reality*, 2, 2021.

[3] B. Marques, J. Alves, M. Neves, R. Maio, I. Justo, A. Santos, R. Rainho, D. Costa, C. Ferreira, P. Dias, and B. Sousa Santos. Interaction with virtual content using augmented reality:a user study in assembly procedures. *Proceedings ACM on Human-Computer Interaction*, 4, 2020.

Poster Sessions

10 December, 10:20 - 10:40

Statistical Modeling of User Influx to the HESE's Emergency Service

Loide Ascenso[1], **Hugo Quintino**[2], **Paulo Infante**[3], **Gonçalo Jacinto**[4],
[1] ISLA Santarém, raquel.ascenso@islasantarem.pt
[2] Hospital do Espírito Santo EPE, hquintino@hevora.min-saude.pt
[3] DMAT/ECT e CIMAA/IIFA, Universidade de Évora, pinfante@uevora.pt
[4] DMAT/ECT e CIMA/IIFA, Universidade de Évora, gjcj@uevora.pt

---

The growing demand for emergency services in hospitals is a global problem affecting health professionals and users. This overcrowding is associated with several factors, including reduced access to other medical emergency or primary care services, and has led to delays in the care of urgent patients. Hospital do Espírito Santo de Évora, EPE (HESE) is the largest and the main hospital unit in Alentejo, offering greater differentiation and, in addition to receiving patients from Central Alentejo, treating an ever-increasing number of patients from Alto Alentejo, Baixo Alentejo and Alentejo Litoral. In this work, based on generalized linear models and control charts, we analyze the inflow of users to the emergency service, seeking to provide support for management decision making by the service management team.

Keywords: control charts; exploratory data analysis; emergency department; generalized linear models; hospital user

---

Public hospitals are an essential part of the National Health Service (NHS), and one of the main and most complex areas is its Emergency Department (ER). Medical care provided by this service is focused on urgent cases. However, there has been an increase in emergency department visits for episodes classified as less or non-urgent, according to data from the Reassessment Committee of the National Emergency/Urgent Care Network in 2010. The influx of non-urgent patients to the ER may lead to overcrowding of a hospital's emergency department, resulting in excessive waiting times, deterioration of the clinical response and patient dissatisfaction. This international problematic of the ER has received special attention from political authorities and the press [2] and is referred to as a worldwide problem [1], [3] and [4]. Characterizing a frequent user of the Emergency Department of Hospital Espírito Santo in order to understand the factors that lead him to make an excessive demand for this type of services can be very useful for decision makers. Based on statistical quality control, we will monitor the inflow to the ER according to the Manchester Protocol, allowing us to identify some factors that influence inflow by color and estimate the process's ability to meet the Protocol's specifications.

The data we will analyze corresponds to 148.120 emergency episodes in 2018 and 2019. Average user's age per episode is about 42 years, with a standard deviation of 29 years. About 53% of episodes were recorded by women. The season of the year with the highest influx / inflow was the fall, in the months of October and December. Yellow episodes were about 40% in the two years of data collection, constituting the highest occurrence, followed by 26% Orange and 24% Green. The affluence in the ER was also analyzed considering days before and after a holiday, i.e., holiday's eve, day before a holiday, day after a holiday, with 37% being registered in the days after a holiday. In relation to daily affluence, as can be seen in Figure 1, it increases from 6 a.m. on, with the highest peak occurring between 9 and 11 a.m., and another peak being registered at 3 p.m.. Monday is the day of the week with the highest influx.

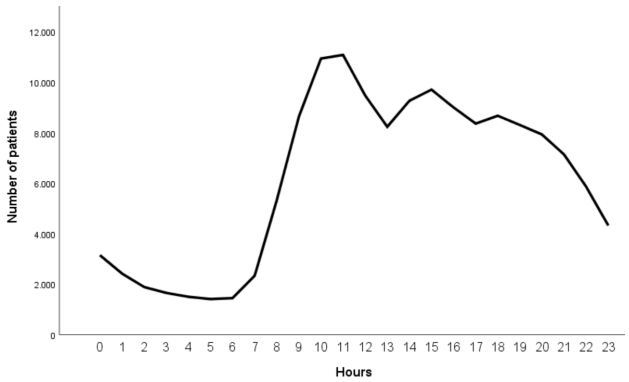

Figure 1: ER hourly influx.

Acknowledgements The Centro de Investigação em Matemática e Aplicações is supported by the Fundação para a Ciência e a Tecnologia, project UID/04674/2020.

References

[1] J. Boyle, M. Jessup, J. Crilly, D. Green, J. Lind, M. Wallis, and G. Fitzgerald. Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5):358–365, 2010.

[2] N. R. Hoot and D. Aronsky. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136, 2008.

[3] N. R. Hoot, S. K. Epstein, T. L. Allen, S. S. Jones, K. M. Baumlin, N. Chawla, and D. Aronsky. Forecasting emergency department crowding: An external, multicenter evaluation. *Annals of Emergency Medicine*, 54(4):514–522, 2009.

[4] J. L. Wiler, R. T. Griffey, and T. Olsen. Review of modeling approaches for emergency department patient flow and crowding research: Academic emergency medicine. 18(12):1371–1379, 2011.

10 December, 10:20 - 10:40

Quality control techniques for tidal data in near real time

Dora Carinhas[1], **Margarida Alves**[2], **Paulo Infante**[3], **António Martinho**[4]
[1] Instituto Hidrográfico; IIFA/Universidade de Évora, dora.carinhas@hidrografico.pt
[2] Instituto Hidrográfico, margarida.alves@hidrografico.pt
[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt
[4] Marinha Portuguesa, santos.martinho@marinha.pt

The classical question of metrology related to the quality of the tide gauge measurements has become more important this last decade or so as new technologies have emerged and tide gauge networks are modernized. This work allowed to evaluate not only the performance of tide gauges in the Portuguese tidal network, but also to present quality control (QC) techniques for tidal data in near real time. The validation of sea level in situ observations will be based on the QC procedures and flagging system suggested by GLOSS. To examine and evaluate the measurement performance of tide gauges, the Van de Casteele test is revisited. The Global Sea Level Observing System (GLOSS) target of 1 cm accuracy in the individual sea level measurement; for the tide gauge station under study, Viana do Castelo tide gauge station, we obtained a 4mm difference between observations of two adjacent tide gauges, this result complies with GLOSS guidelines.

Keywords: accuracy, tide, tide gauge, time series

The exchange of in situ sea level observations within the European countries has grown significantly during the last decade. This is caused by an increased interest in getting access to real time information of the variations in the sea level on a regional scale [2], [3]. Sea level observations are used in applications with many different purposes and on different time scales such as tsunami warning issues, now casting and forecasting the sea level variations inoperational oceanography and long term estimates for climate change studies.

To estimate the accuracy of sea level measurements, laboratory or field experiments can be undertaken in which the tide gauge is compared with an independent higher quality standard or reference gauge. The Van de Casteele test involves taking simultaneous sea level heights with both a tide gauge (the gauge to be checked) and a standard (the reference gauge) over a full tidal cycle [4]. The data obtained are then used to construct a simple diagram in which the sea level elevation (y axis) is plotted against the gauge error (x axis). The gauge error (DH) is determined as the difference in sea level height measured by the reference tide gauge (H) and the sea level height measured by the tide gauge we are

checking (H'). In the case of a perfect gauge the diagram results in a vertical line centered at zero.
Near-real time quality control of sea level data is recommended for the main applications related with operational oceanography [1]. This implies the need of implementing automatic software of error detection and flagging. Data are considered to arrive in near-real time for latencies normally between 1 hour and several weeks, and this is normally the situation for storm surge forecasting or altimetry data calibration. Quality control consists basically of detection of strange characters, wrong assignment of date and hour, spikes, outliers and computation of residuals (Figure 1).

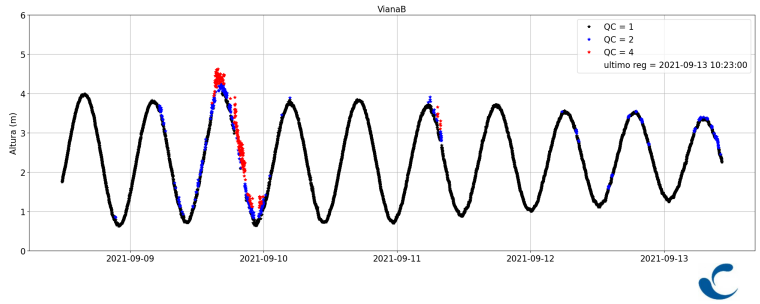

Figure 1: Error detection and flagging for tide gauge of Viana do Castelo.

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.

10 December, 10:20 - 10:40

Time series synchronization: An application to COVID-19 data

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

This research analyzes the dynamics of COVID-19 times series to understand how countries can be clustered based on their similarities. We apply an extended hidden Markov model. Results show that the best model (BIC) contains three regimes and three clusters. Thus, COVID-19 time series can be described by three regimes or states and grouped into three clusters of countries with distinct dynamics within.

Keywords: COVID data, clustering, hidden Markov models, dynamic models

Since its onset at the end of 2019, the COVID-19 pandemic has created a health and socioeconomic crises across the whole world. The current estimate is that global GDP per capita declined by 5.3% in 2020 with negative growth in real GDP per capita in 2020 [1]. In most countries, the monitoring of the severe acute coronavirus 2 (SARS-CoV-2) infection and COVID-19 was covered by existing epidemiological surveillance of respiratory viral agents and death certificate information on the natural cause of death has not been standardized worldwide. Despite these general concerns, these daily updated figures with obvious limitations have been used by media coverage and in scientific debate. Under notification and under-reporting have been the most studied outcomes of the epidemiological surveillance system and reported in most countries of the world. For instance, Kupek estimates under-reporting of COVID-19 deaths in 22.62% in Brazil [4]. Under-reporting has previously shown to occur due to low rate of laboratory testing for SARS-CoV-2, reporting delay, inadequate access to medical care, and its poor quality, leading to the low sensitivity of epidemiological surveillance and poor outcomes. Additionally, in many situations there is a lack of laboratory confirmation of the cause of death. Most of the approaches to the quality of the reported data have been based on comparing the same weeks/months in pre-COVID-19 and COVID-19 months after controlling some factors. This approach does not take into account the dynamics of the reporting of different countries as it takes the cross-section of each week/month at the time.

This work applies a panel hidden Markov model that is an extension of the hidden Markov model [2, 3]. Apart from a dynamic latent variable that models transitions between regimes, the panel version contains a cross-sectional dimension that corresponds to a mixture model (statistic latent variable). Given the dimension of the latent space, the maximum likelihood estimation of the parameters of the model uses an adaptation of the EM algorithm, the

Baum-Welch algorithm [2]. Model selection of the number of regimes and clusters can be based on the Bayesian information criterion (BIC).
The data set used in this research comes from the European Centre for Disease Prevention and Control. Each time series has 214 daily observations of cumulative number for 14 days of COVID-19 cases per 100000 from May 15th to December 14th 2020. The panel data corresponds to time series of log-variations for 194 countries.
BIC shows that the best model contains three regimes for the dynamic latent variable and three clusters for the static latent variable. Thus, the position of each time series is well described by the transitions between three regimes, each one representing a different type of behavior: zero-inflated regime, standard regime, and outlier regime. On the other hand, the cross-country heterogeneity between countries can be described by three clusters, each one with similar dynamics within.
In summary, the application of the panel hidden Markov model allows the clustering of time series with heterogeneous regime switching dynamics within. In particular, it detects cross-sectional unobserved heterogeneity, in which countries are grouped into clusters with different propensity to switch between regimes.

Acknowledgements Funding from Fundação para a Ciência e Tecnologia (Portugal), UID/GES/00315/2020.

References

[1] World Bank. *Global Economic Prospects*. Washington, DC: World Bank, 2021.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[3] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[4] E. Kupek. How many more? Under-reporting of the COVID-19 deaths in Brazil in 2020. *Tropical Medicine & International Health*, 26(9):1019–1028, 2021.

10 December, 10:20 - 10:40

A Case Study of Pavement Texture Performance using Linear Mixed Models

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, CBMA – Centro de Biologia Molecular e Ambiental, Guimarães, Portugal
[3] Universidade do Minho, Departamento de Engenharia Civil, ISISE – Instituto para a Sustentibilidade e Inovação em Estruturas de Engenharia, Guimarães, Portugal

---

Linear Mixed Models are an extension of classic statistical procedures that provide analysis flexibility in correlated longitudinal data and allow researchers to model the covariance structures that represent its random effects. These 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. In this work, we develop a linear mixed model to describe the pavement texture performance throughout the time and identify whether the traffic, climate conditions, pavement structure, and geometric characteristics of the highway may influence that performance.

Keywords: Linear mixed models, Longitudinal data, Performance model, Texture

---

Modeling pavement performance is an essential activity of a Pavement Management System (PMS). The models play a crucial role in several aspects of the PMS, in particular, to predict when maintenance will be required for individual road sections and how to prioritize competing maintenance requirements.
The methods used to model pavement performance indicators are many. They include stochastic methods such as the Markov chain, linear or nonlinear mixed effects models and artificial intelligence modeling techniques.
In order to describe the pavement texture performance, we develop a linear mixed effect model by monitoring and analyzing the conditions of road pavements in a period of eight years. Linear mixed models are particularly appropriate for modeling the dependencies in the data that were measured on the same statistical unit at various levels over time (for further discussion, see [1], [2] and [3]).
This study is based on a real database obtained over 8 years for the Ascendi Concession network. It covers six different districts of Portugal, in a total of 7204 pavement sections of 1km. All analyses were performed using the R statistical software using the *nlme* package. The *texture* is used as the dependent variable. Pavement structure (type of surface course); traffic (accumulated or annual average daily, light and heavy vehicles, and day or night);

climate conditions (temperature, precipitation, and relative humidity of the air); geometric characteristics of the vertical and horizontal alignments (plan, profile), and the lane and hypsometry are used as covariates. Some of these variables were identified in the state-of-the-art review; others were defined according to the objectives of the work.
The structure of the random effects is examined by testing whether the random effects specified in this model should be included. The final model includes random effects for the intercept and for the slope of the *time* variable.
Since the observations are taken longitudinally on the same statistical unit (pavement section), the within-group (i.e. within pavement section) errors are probably autocorrelated and an AR(1) model was chosen as correlation structure.
Likelihood ratio tests are applied for choosing between two nested models, to select the final model. Akaike information criterion (AIC) are also used to compare several alternative models.
From the analysis of the final model, we can see that texture decreases with time and average number of days with maximum temperature above 25°C. It increases with accumulated annual daily average traffic, the minimum air temperature value and average of total precipitation. It was demonstrated that texture increases on the right lane and the slow lane, in comparison with the left lane. Regarding the type of surface, texture increases in the porous asphalt and decreases in gap-graded asphalt concrete with a high percentage of Rubber Modified Binder when compared with the gap-graded asphalt concrete surface course. Also, texture decreases in the Low Altitude sections when compared with Medium Altitude sections.
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 programme UID/BIA/04050/2019 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] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer - Verlag, New York, 2000.

10 December, 10:20 - 10:40

Comparing Time Series Models for Forecasting Meteorological Data

A. Manuela Gonçalves[1], **Marco Costa**[2], **Cláudia Costa**[3]
[1] Department of Mathematics, Centre of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[2] Águeda School of Technology and Management, Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt
[3] Department of Mathematics, University of Minho, Portugal, claudiacostafm@hotmail.pt

---

The main objective of project TO CHAIR - Optimum Challenges in Irrigation is to study and understand management irrigation problems in order to more efficiently plan the use of water in irrigation systems. This requires identifying the most suitable forecasting models for meteorological time series that have an impact on the evapotranspiration process and on soil humidity. Therefore, the main goal of this study is to estimate and forecast meteorological variables (wind speed, minimum and maximum air temperature and precipitation) in real time (daily) for a given location: a farm. In particular, it is presented a comparison of two forecasting methods, the TBATS models and the linear regression models with correlated errors.

Keywords: time series, meteorological variables, forecasting, TBATS, regression with correlated errors

---

In a world where climate change and increasing social conflicts are a reality, a proper management of the existing scarce resources is vital. This study is carried out in the context of project "TO CHAIR - Optimum Challenges in Irrigation" funded by the European Regional Development Fund (ERDF), the Competitiveness and Internationalization Operational Program (COMPETE 2020) and the Foundation for Science and Technology (FCT) and its main objective is to study and to understand management irrigation problems in order to more efficiently plan the use of water in irrigation systems.
The main goal of this study is to identify the most suitable forecasting models for modeling meteorological time series that have an impact on the evapotranspiration process and soil humidity. For that, it is necessary to forecast meteorological variables like wind speed, minimum and maximum air temperature and precipitation at a location (in this case, at a farm in Carrazeda de Ansiães, in the district of Bragança in the North of Portugal) where there are historical observations but current measurements are not available, including various steps for forecasting (i.e., 7 days). The data under study consist of daily records observed from January 1, 2010 to April 23, 2019. This study presents a comparison of two forecasting methods, the TBATS models (Box-Cox transformation, ARMA errors, trend

and trigonometric seasonal components), [3], [2], and the linear regression models with correlated errors [1]. These models were selected due to their ability to model seasonal fluctuations strongly present in meteorological data, in particular when dealing with time series with complex seasonal patterns. The collected data were divided into two sets: training data (in-sample data) and testing data (out-of-sample data) in order to test the accuracy of the suggested forecasting models.
These models have successfully been applied to modeling and foresight scenarios of environmental and meteorological variables. In this study, we have shown that both TBATS and linear regression models with correlated errors (for forecasting time series with complex seasonal patterns) can efficiently capture the behaviour of the meteorological time series in the studied site. The obtained results show that the application of TBATS and linear regression models with correlated errors to these time series provides valuable insights into the studied data structures and their components, thus being a good basis for accurate estimations and forecasts.

Acknowledgements A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM (Centre of Mathematics). This work has also received funding from FEDER/COMPETE/NORTE2020/POCI/FCT through grants PTDC-EEI-AUT-2933-2014116858-TOCCATA and from To CHAIR - POCI-01-0145-FEDER-028247 Financial support from the Portuguese FCT in the framework of the Strategic Financing UIDIFIS/04650/2013. The authors are grateful for access to meteorological data provided by the "Direção Regional de Agricultura e Pescas do Norte de Portugal". Marco Costa was partially financed by the Centre for Research and Development in Mathematics and Applications (CIDMA) through FCT, within project UIDB/04106/2020 and UIDP/04106/2020.

References

[1] T. Alpuim and A. El-Shaarawi. On the effciency of regression analysis with ar(p) errors. *Journal of Applied Statistics*, 35(7):717–737, 2008.

[2] R.J. Hyndman, A.B. Koehler, R.D. Snyder and S. Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–459, 2002.

[3] A.M. De Livera, R.J. Hyndman and R.D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

10 December, 10:20 - 10:40

Determinants for the existence of victims in road accidents in the district of Setúbal

Paulo Infante[1], **Gonçalo Jacinto**[1], **Anabela Afonso**[1], **Rodrigo Cesar**[2], **Pedro Nogueira**[3], **Marcelo Silva**[3], **Vitor Nogueira**[3], **José Saias**[3], **Paulo Quaresma**[3], **Patrícia Gois**[4], **Paulo Rebelo Manuel**[5]
[1] DMat/ECT, CIMA/IIFA, Universidade de Évora,
pinfante@uevora.pt,gjcj@uevora.pt,aafonso@uevora.pt
[2] DMat/ECT, Universidade de Évora, rcfs@uevora.pt
[5] DInf/ECT, NOVALINCS/IIFA, Universidade de Évora,
pmn@uevora.pt,marcelogs@uevora.pt,vbn@uevora.pt,jsaias@uevora.pt,pq@uevora.pt
[4] DAVD/EA, Universidade de Évora, pafg@uevora.pt
[5] CIMA/IIFA, Universidade de Évora, pjsrm@uevora.pt

Between 2016 and 2019 there were in Portugal 136654 accidents with victims, resulting in 2466 deaths. In this period, Setúbal was one of the districts with highest number of accidents. Using data collected by GNR and joining information about road conditions and meteorological information and starting from an identification of municipalities with identical profiles, logistic regression and machine learning models were obtained to identify some determinants for the existence of victims in traffic accidents that occurred in Setúbal district.

Keywords: Getis Ord-Gi*, Local Moran's I, logistic regression, machine learning, road accidents

---

According to the National Road Safety Authority (ANSR), the number of accidents with victims in Portugal has increased since 2011. Between 2016 and 2019 there were 136654 accidents with victims, resulting in 2466 deaths [1]. The district of Setúbal is one of the districts with the highest number of road accidents. In this 4-year period, 28103 accidents were recorded in the area under the jurisdiction of the GNR Territorial Command of Setúbal, which resulted in 8260 victims, with 510 serious injuries and 167 fatal injuries. This work analyzes the data collected with the Statistical Bulletin of Road Accidents, with an update of the ANSR for victims at 30 days, and complemented with meteorological information provided by IPMA. Initially, a spatial analysis of the accidents was carried out, using the Getis Ord-Gi* statistic to identify hotspots and the Local Moran's I statistic for spatial autocorrelation, which allowed the identification of municipalities with identical profiles for fatalities and serious injuries. Subsequently, a logistic regression model was used to identify some determinants for the existence of victims in traffic accidents that occurred in the district of Setúbal. We found that the determinants for the occurrence of accidents with victims are: geographical factors (county and area where the accident

occurs), temporal factors (month, day of the week and time of day when the accident occurs), environmental factors (wind at the time of the accident), driver and vehicle factors (gender, age and number of drivers involved in the accident, age of vehicles, occurrence of escape) and associations between the type of road and the type of accident. The results obtained were compared with *machine learning* models, and an agreement was observed in the conclusions obtained.

Acknowledgements This work is a contribution to the MOPREVIS project "FCT DSAIPA/DS/0090/2018" funded by the FCT-Foundation for Science and Technology, under the National Initiative on Digital Skills e.2030, Portugal INCoDe.2030. We would also like to thank the project's partner entities.

References

[1] Relatório Anual de Segurança Rodoviária: 2019. Autoridade nacional de segurança rodoviária. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Documents/2020, 2020.

10 December, 10:20 - 10:40

Likelihood function approximation through the delta method in mixed SDE models

Nelson T. Jamba[1,4], **Patrícia A. Filipe**[1,3], **Gonçalo Jacinto**[1,2], **Carlos A. Braumann**[1,2]

[1] Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora

[2] Departamento de Matemática, Escola de Ciência e Tecnologia, Universidade de Évora

[3] Iscte - Instituto Universitário de Lisboa, Iscte Business School, Departamento de Métodos Quantitativos para Gestão e Economia

[4] Liceu nº 918 do município dos Gambos, Angola

d39830@alunos.uevora.pt, patricia.filipe@iscte-iul.pt gjcj@uevora.pt, braumann@uevora.pt

---

Individual growth is modeled through mixed stochastic differential equations models and maximum likelihood estimation method is used to estimate the models parameters. Obtaining a closed form expression for the likelihood function is not always possible due to the complexity of the expressions. The delta method is applied to solve the integrals involved in the likelihood function. An application to cattle weight data is shown and comparison between methodologies is presented.

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

---

We use a class of stochastic differential equations (SDE) to model the evolution of cattle weight, taking the form $dY_i(t) = \beta(\alpha - Y_i(t))dt + \sigma dW_i(t)$, $Y_i(t_0) = y_{i,0}$, $i = 1, \ldots, M$, where $Y_i(t) = h(X_i(t))$ is a transformed weight obtained by applying a strictly increasing C^1 function h to the actual weight $X_i(t)$ of the i^{th} animal at age t, α is the average transformed weight at maturity. β is a growth parameter, σ measures the intensity of environmental fluctuations and $W_i(t)$ $(i = 1, \ldots, M)$ are independent standard Wiener processes. Depending on the function h chosen, we obtain stochastic versions of the most commonly used deterministic animal growth models. We have seen that, for our type of data, one of the most adequate choices of the function h was the logarithm of the weight, which corresponds to the stochastic Gompertz model. For this reason, we have presented the results for this model.

Since model parameters may vary from animal to animal, we have extended the study to SDE mixed models where the variation among animals of the parameters α, β or both is assumed to be random. Here we present the results for the case where both parameters are Gaussian uncorrelated random variables, $\alpha_i \sim N(\mu, \theta^2)$ and $\beta_i \sim N(\lambda, \omega^2)$, $(i = 1, \ldots, M)$.

The maximum likelihood estimation method was applied to estimate the parameters μ, θ, λ, ω and σ. Approximation methods can be particularly useful when a closed-form expression for the likelihhod function cannot be obtained, as it happens in our case. The delta method is one of the methodologies that can be applied to solve the integrals involved in the likelihood function [1].

To illustrate the performance of this method, we compare the estimated parameters obtained from a simulated dataset with the estimates from existing packages [2]. We have simulated data weights from 500 animals with the same time vector of 49 observation ages, from birth till 4 years of age, making a total of 24500 observations. The results show a very good performance of the delta method, outperforming the existing methods (Table1). The proposed method also has the advantage of being applicable in real situations where the weight observations of the different animals are taken at different and/or unevenly spaced ages. The existing packages, as far as we know, require the same time vector of observation ages for all animals. We illustrate the results using real data on the weight from a large and heterogeneous sample of 16029 Mertolengo cattle males, where each animal has several observations with a minimum of 3 and a maximum of 33 weights at ages varying between birth and a maximum age that ranges from 0.2 until 16 years, totalling 96204. The delta method also have the advantage of always resulting in simpler and closed-form expressions for the likelihood function.

Table 1: True parameters, maximum likelihood estimates and corresponding 95% asymptotic confidence intervals obtained using the delta method, for both simulated and real data, and estimates obtained from the *MsdeParEst* R package, for the simulated data.

	Simulated data			Real data
	True	**delta method**	**MsdeParEst**	**delta method**
$\mathrm{h}^{-1}(\mu)$	632.70	634.05 ± 29.86	646.12	629.70 ± 6.30
θ	0.15	0.0970 ± 0.0377	0.1401	0.0909 ± 0.0083
λ	1.43	1.4231 ± 0.0703	1.2671	1.4261 ± 0.0122
ω	0.30	0.1795 ± 0.0471	$2.7e{-}05$	0.1998 ± 0.0050
σ	0.33	0.3438 ± 0.0098	0.4975	0.3273 ± 0.0016

Acknowledgements The Centro de Investigação em Matemática e Aplicações is supported by the FCT, project UID/04674/2020. This work was developed within the Operational Group PDR2020-1.0.1-FEADER-031130 - Go BovMais, funded by PDR 2020.

References

[1] G. Casella and R. L Berger. *Statistical Inference.* Duxbury Press, 2 edition, 2001.

[2] M. Delattre and C. Dion. *MsdeParEst: Parametric Estimation in Mixed-Effects Stochastic Differential Equations*, 2017. R package version 1.7.

10 December, 10:20 - 10:40

The impact of e-government on sustainable development: a logit model

Cristina Lopes[1], **Conceição Castro**[1]
[1] CEOS.PP, ISCAP, Politécnico do Porto, cristinalopes@iscap.ipp.pt

The impact of e-government on sustainable development is analysed using a logit model with a sample of 103 countries in the period 2003–2018. Sustainable development is proxied by adjusted net savings, a variable that embraces a country's economic, social and environmental development. The results show that countries with higher e-government development are more likely to attain sustainable development, particularly in developing and transition economies.

Keywords: Logistic regression, Sustainable development, E-government, Adjusted net savings

It is recognized that the quality of institutions affects sustainable development. Better government may contribute to proper resource allocation fostering sustainable development [3]. The OECD defines e-government as "the use of information and communication technologies (ICTs), and particularly the Internet, to achieve better government".
In this work, the empirical investigation on the impact of e-government and other variables in sustainable development is examined for 103 countries in the period 2003 to 2018 [1], using the e-Government Development Index (EGOV) from the United Nations, which ranges from zero to one, where higher scores denote better e-government development.
Sustainable Development can be defined as non-declining wealth ($\frac{dW_t}{dt} \geq 0$) [3], where wealth (W_t) includes manufactured capital, human capital and natural capital. Sustainable development will be proxied by Adjusted net savings in percentage of Gross National Income (ANS), an indicator provided by the World Bank that is commonly adopted as a broad indicator of sustainability over the long run. If the Adjusted net savings of a country are positive, it suggests that the present value of welfare is increasing. On the contrary, persistently negative ANS are indicating that the economy is in an unsustainable path. Therefore, a binary variable was defined – $ANSbin_i$ – to flag when a country i has $ANS_i \geq 0$, and a logistic regression model (1) was developed to relate the probability of having non-negative ANS with several key factors for sustainable development.

$$ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 EGOV_i + \beta_2 AgeDep_i + \beta_3 EconGrowth_i + \beta_4 NatRents_i + \beta_5 GNI_i \quad (1)$$

where $\mu_i = P(ANSbin_i = 1) = P(ANS_i \geq 0)$ is the probability of a country i having non-negative ANS and $\mu_i/(1-\mu_i)$ is the odds ratio in favour of having non-negative ANS,

that is, the ratio of the probability of a country having sustainable development to the probability that ANS will be negative.

The choice of the control variables was guided by previous empirical studies on the determinants of sustainable development, namely the economic growth, gross national income per capita (GNI), age dependency, and natural resource rents. It is expected that economic growth indicts an increase in investment resources, contributing to the accumulation of a productive base, and so increasing Adjusted net savings [3]. Nevertheless, economic growth can affect negatively sustainable development due to increases in environmental pollution [2]. Natural resources rents measure the degree to which an economy depends on natural resources to generate income.

The logit model in (1) was estimated for the whole sample, and separately for the subsample of developing and transition economies. The likelihood ratio tests, significant at a 1% level, show the overall goodness of fit. Although the values obtained from pseudo R^2 are moderate, the models correctly classify the outcome for 88.6% and 84.9% of the cases. The areas under the ROC (Receiver Operating Characteristic) curves, 0.815 and 0.802, reveal a good discriminating capacity, demonstrating the usefulness of these logit models for classifying new observations.

The results suggest that e-government is an important determinant of the odds of having sustainable development. With an increase of 0.1 in e-Government Development Index, and maintaining the remaining variables constant, it is expected that the odds of having sustainable development increase about 26.52% ($(e^{\beta_i 0.1} - 1) \times 100$). In the subsample of developing and transition economies, the odds of having sustainable development increase by about 36.43%, which is higher than in the whole sample.

This evidence highlights the importance for developing and transition economies to invest in the use of ICTs by governments, as a part of an overall public policy strategy to achieve sustainable development. The results also suggest that economic growth and GNI per capita are significant positive influences, and that increases in age dependency and natural resource rents may reduce the likelihood of a country having sustainable development. The negative impact of natural resource rents is consistent with the resource curse hypothesis.

Acknowledgements This work is financed by portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] C. Castro and C. Lopes. Digital government and sustainable development. *Journal of the Knowledge Economy*, 2021.

[2] T. Güney. Governance and sustainable development: How effective is governance? *The Journal of International Trade and Economic Development*, 23(6):316–335, 2017.

[3] M. Sato, S. Samreth, and K. Sasaki. The impact of institutional factors on the performance of genuine savings. *International Journal of Sustainable Development and World Ecology*, 25(1):56–68, 2018.

11 December, 10:20 - 10:40

A hurdle-gamma regression model for the average number of undeveloped pine nuts per cone

Anabela Afonso[1], **Dulce G. Pereira**[2], **Ana Cristina Gonçalves**[3]
[1] Departamento de Matemática/ECT, Centro de Investigação em Matemática e Aplicações/IIFA, Universidade de Évora, aafonso@uevora.pt
[2] Departamento de Matemática/ECT, Centro de Investigação em Matemática e Aplicações/IIFA, Universidade de Évora, dgsp@uevora.pt
[3] Departamento de Engenharia Rural/ECT, Mediterranean Institute for Agriculture, Environment and Development/IIFA, Universidade de Évora, acag@uevora.pt

---

Pinus pinea fruit production plays an important role to the local development, in Portugal, especially of Setubal district. A hurdle-gamma model was used to analyze the average number of undeveloped pine nuts per cone and per tree in relation to cone, tree and stand structures. Stem height, diameter at breast height, cone weight, number of pine nuts, seed efficiency, year and plot contribute to explain the existence or the average number of undeveloped pine nuts.

Keywords: hurdle-gamma regression, pine nut and kernel efficiency, stem and crown diameter, weight

---

Pinus pinea stands are characteristic of the Mediterranean region. Its fruit yield is of importance for local development, due to its highly nutritional kernels and its economic value.

One major concern with fruit production, both for regeneration and for human consumption, is their quantity and quality. In conifers in general and in *Pinus* spp, in particular, the number of cones per tree, the number of seeds per cone and the seed efficiency (defined as the percent of the number of fully developed seeds per cone) are determinant for the seed availability for both uses.

The main goal of this study is to understand the relations of stand type, tree' dimensions and cone weight on the pine nut efficiency with a large data set.

The data was collected in four plots located in Alcácer do Sal, Portugal. In 120 trees per plot cone were harvested, 30 manually and 90 mechanically, during three years. Trees were allocated to manual and mechanical harvest through a random stratified sampling, with strata defined by 0.1 m diameter at breast height classes (for details see [2]). In each harvest, 3 cones were selected randomly per tree [1]. Data set is composed of pine nuts and kernels of 3313 cones.

The analysis of the differences in the average number of undeveloped pine nuts per cone between trees was done with a hurdle-gamma regression [3], due to the high occurrence of

zeros in the data set. The explanatory variables considered were year, plot, average of cone fresh weight per tree, average of the cone moisture content per tree (quotient between the difference of fresh and dry cone weight and cone fresh weight), average number of pine nuts fully developed per cone and tree, average seed efficiency, tree characteristics (diameter at breast height, total height, stem height, crown length, height of the beginning of the live crown, crown diameter).
Total height, crown length, height of the beginning of the live crown, crown diameter, and the average of the cone moisture content per tree did not contribute to explain either the existence or the average number of undeveloped pine nuts per cone and tree. The odds of a cone having undeveloped pine nuts were lower for trees in plots managed for timber and higher for trees in plots managed for fruit production. The odds were highest in 2005 in all plots. For all plots and years, the odds were lower for cones with high number of pine nuts fully developed, however these odds increased with the stem height of the tree. Among the cones with undeveloped pine nuts, the average number of undeveloped pine nuts per cone and tree decreased with the increase in the diameter at breast height, stem height and seed efficiency. In all plots, this average was smaller in year of 2004 and higher in 2005. There is a significant interaction between the average of the fresh cone weight per tree and the plots.

Acknowledgements The work was financed by PROGRAMA AGRO 200 (Project AGRO/200/2001: "Colheita mecânica da pinha (Pinus pinea L.)"). This work is funded by National Funds through FCT - Foundation for Science and Technology under the Project UIDB/05183/2020 (MED) and UIDB/04674/2020 (CIMA).

References

[1] Ana Cristina Gonçalves, Anabela Afonso, Dulce G Pereira, and Anacleto Pinheiro. Influence of umbrella pine (pinus pinea l.) stand type and tree characteristics on cone production. *Agroforestry systems*, 91(6):1019–1030, 2017.

[2] Ana Cristina Gonçalves, António Bento Dias, Anabela Afonso, Dulce G Pereira, Anacleto Pinheiro, and José Oliveira Peça. Mechanical versus manual harvest of pinus pinea cones. *Biosystems Engineering*, 143:50–60, 2016.

[3] Alain F Zuur and Elena N Ieno. *Beginner's guide to zero-inflated models with R*. Highland Statistics Limited United Kingdom, 2016.

11 December, 10:20 - 10:40

How the indicators for EU countries in the period 2010-2019 are approaching targets of Europe 2020 agenda?

Adelaide Figueiredo[1], **Fernanda Otilia Figueiredo**[2]
[1] Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC, adelaide@fep.up.pt
[2] Faculdade de Economia da Universidade do Porto and CEAUL, otilia@fep.up.pt

The Europe 2020 strategy is the European Union agenda for growth and jobs for the current decade. According to the European Commission, this strategy is a way to overcome the structural weaknesses in Europe's economy, improve its competitiveness and productivity. In this study, we carry out a multivariate analysis of some indicators of Europe 2020 strategy for the European Union countries during the period 2010-2019. Our aim is to analyze the evolution of the indicators along this period of time to verify whether they are converging to the defined targets in Europe 2020 strategy and also to find out which countries are closer to achieve those targets.

Keywords: Europe 2020, European countries, PCA, Statis methodology

The Europe 2020 strategy includes five areas: Employment, Education, Research and Development, Poverty and Social Exclusion, and Climate Change and Energy. The indicators considered in these areas are employment rate; early leavers from education and training rate; population, aged 30 to 34, with higher education; expenditure on R&D as % of GDP; population at risk of poverty; greenhouse gas emissions; renewable energy consumption; energy efficiency – primary energy consumption and final energy consumption.
We use data from Pordata of the previous indicators for the European Union countries for several years of the period 2010-2019. Thus, we obtain several data tables of countries described by quantitative indicators, evaluated in different years. For a conjoint analysis of the data tables, we apply the STATIS (*Structuration de Tableaux à Trois Indices de la Statistique*) methodology: STATIS and dual STATIS. This methodology of multivariate data analysis was developed in Lavit [2] and Lavit *et al.* [3]. It allows us to analyze simultaneously several data tables of individuals (in our case, European countries) described by quantitative variables (here, the several indicators), collected at different moments in time or circumstances (in our case, different years of the period 2010-2019). In this work we can apply both methods: the STATIS and the dual STATIS. The STATIS method enables us to compare the structure of the countries along the data tables and the dual STATIS enables us to compare the relations between the variables along the data tables. Both methods

include three important steps: the interstructure, the intrastructure and the analysis of the evolution of each individual or variable.
In the STATIS method, we begin to determine, for each year considered in the study, a matrix of the scalar products between the individuals, i.e. the representative object of each data table. In the first step, the interstructure, we compare globally the data tables. To obtain the distances between two objects corresponding to different years we compute the scalar product of Hilbert-Schmidt, which is equal to the vector correlation coefficient RV proposed by Escoufier [1]. A PCA based on the matrix of RV coefficients gives us the Euclidean image of the series of data tables. This graphical representation of the interstructure allows us to visualize the proximities and differences between the years. In the second step, the intrastructure, we determine a single table representative of the common structure of the individuals in all data tables, the compromise object, which is defined by the weighted mean of the normed objects. A PCA based on the compromise object enables us to obtain the Euclidean image of the compromise. The correlations of the variables with the principal components of the compromise enable us to interpret the compromise axes and the compromise positions of the individuals. Finally, in the third step, we identify which individuals contribute the most (or least) to the observed differences between the data tables. We also represent the different positions of the individuals on the compromise Euclidean image, i.e., their trajectories.
The dual STATIS method, analogous to the STATIS method, focuses on the relations between variables instead of the distances between individuals. The representative object of each data table corresponds either to the covariance matrix or correlation matrix (in case of standardized data) and describes the structure of the variables in the data table. The three steps of dual STATIS method are similar to those of the STATIS method. Based on the trajectories of the variables, we also find out which indicators are unchangeable along the years or instead are converging (or possibly diverging) to targets of Europe 2020.

Acknowledgements This project was financed by the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, through national funds, within the project UIDB/00006/2020 (CEAUL), and co-funded by the FEDER, where applicable.

References

[1] Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, (29):751–760, 1973.

[2] C. Lavit. *Analyse Conjointe de Tableaux Quantitatives*. Masson, 1988.

[3] C. Lavit, Y. Escoufier, R. Sabatier, and P. Traissac. The ACT (STATIS method). *Computational Statistics and Data Analysis*, (18):87–119, 1994.

11 December, 10:20 - 10:40

A state space framework for daily temperature forecasting

A. Manuela Gonçalves[1], **F. Catarina Pereira**[2], **Marco Costa**[3]
[1] Department of Mathematics, Centre of Mathematics, University of Minho, Portugal, mneves@math.uminho.pt
[2] Department of Mathematics, Centre of Mathematics, University of Minho, Portugal, up202010700@edu.fe.up.pt
[3] Águeda School of Technology and Management, Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal, marco@ua.pt

---

In the context of the project "TO CHAIR - Optimum Challenges in Irrigation", a statistical modeling process is established to improve the accuracy of temperature forecasting obtained from the weatherstack.com website with a dataset of real observations in a farm in Portugal. The proposed model establishes a stochastic linear relationship between the observed temperature and the h-step-ahead forecast produced by the website. This relation is modeled in a state space framework associated to the Kalman filter predictors. An alternative Generalized Method of Moments estimators was considered in the models' parameters estimation.

Keywords: state space model, Kalman filter, generalized method of moments, forecasting time series, maximum temperature

---

In the context of the TO CHAIR project, it is necessary to improve short-term forecasts of meteorological variables. Indeed, more accurate forecasts of these variables can improve the results of the optimization routines to obtain a more efficient use of water in irrigation systems [3]. In this project, the main goal of statistical modeling is to improve the accuracy of the forecast of meteorological variables obtained from the weatherstack.com website for the location under analysis, a farm in Portugal. However, agricultural researchers that investigate in this area know that forecasts have significant errors compared with observations obtained locally by a portable weather station.
This work aims to establish a state space framework that combines forecasts with the observations in order to correct or "calibrate" a forecast by comparing it with knowledge from the past, namely through an estimated model based on few data observations.
The statistical analysis was performed using a dataset that includes forecasts (obtained from the weatherstack.com website) of daily maximum temperature (in Celsius degrees) for the location of the farm Senhora da Ribeira in Portugal, between February 20 and October 11, 2019. Additionally, we use observations of daily temperatures obtained by a portable weather station installed in the farm during that period of 234 days.

The optimization of the log-likelihood is done by numerical procedures via the Newton-Raphson method or, more often, by the EM algorithm [4]. However, previous modeling has shown that the normality is rejected in the residuals analysis. Thus, alternative methods are needed. In this context, we propose to adapt the distribution-free estimators initially proposed in [1] and subsequently generalize them for multivariate models in [2].
The results show that this approach allows reducing the RMSE of the uncorrected forecasts in 16.90% considering the 6-step-ahead forecasts and in 60.45% considering the 1-step-ahead forecasts, compared with the initial RMSE. Additionally, empirical confidence intervals at the 95% level have a coverage rate similar to this confidence level. Thus, this approach has proven suitable for this type of forecasts correction since it considers a stochastic calibration factor in order to model time correlation of this type of variable.
The state space approach shows that it can be considered for improving weather variables forecasts obtained from some accessible sources, even if those sources produce data with significant errors, as long as more accurate data is available to estimate the parameters of the models.

Acknowledgements A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/20 20 and UIDP/00013/2020 of CMAT-UM (Centre of Mathematics). Marco Costa was partially financed by the Centre for Research and Development in Mathematics and Applications (CIDMA) through FCT, within project UIDB/04106/2020 and UIDP/04106/2020. This work has also received funding from FEDER/COMPETE/NORTE2020/POCI/FCT through grants PTDC-EEI-AUT-2933-2014116858-TOCCATA and from To CHAIR - POCI-01-0145-FEDER-028247 Financial support from the Portuguese FCT in the framework of the Strategic Financing UIDIFIS/04650/2013. The authors are grateful for access to meteorological data provided by the "Direção Regional de Agricultura e Pescas do Norte de Portugal".

References

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

[2] A.M. Gonçalves and M. Costa. Predicting seasonal and hydro-meteorological impact in environmental variables modelling via kalman filtering. *Stoch. Environ. Res. Risk Assess*, 27:1021–1038, 2013.

[3] C. Costa, A.M. Gonçalves, M. Costa, and S. Lopes. Forecasting temperature time series for irrigation planning problems. *In: Proceedings of the 34th IWSM International Workshop on Statistical Modelling*, 5:116–130, 2019.

[4] R.H. Shumway and D.F. Stoffer. *Time series analysis and its applications: with R examples.* Springer, New York, 2011.

11 December, 10:20 - 10:40

A sociological portrait of the Portuguese, based on their religiosity and values: A cross-time comparison with Europe

Maria Paula Lousão[1], **Cláudia Vasconcelos Silvestre**[2], **José Luís Casanova**[3],
[1] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, mlousao@escs.ipl.pt
[2] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt
[3] Instituto Universitário de Lisboa (ISCTE-IUL)-Centro de Investigação e Estudos de Sociologia (CIES-IUL), jose.casanova@iscte-iul.pt

This communication is part of an ongoing investigation that studies the religious phenomenon and human values in the European Union. In this study we will present evolution of human values between 2002 and 2018 in Portugal and in European Union. Further, we will single out the main differences between religious and non-religious people, and will describe their sociological portrait. This study is based on data from European Social Survey rounds 1 to 9 and statistical analysis was performed using IBM-SPSS, version 27.0.

Keywords: Multiple Correspondence Analysis, Eta coefficient, religiosity, human values

Roccas and Schwartz [4], and Schwartz and Huismans [5] showed that there is a correlation between people's religiosity and the human values defined by SH Schwartz: conformity and tradition values positively correlate with religiosity, while hedonism, self-determination, achievement and power have negative correlations. The remaining human values are not significantly related, however safety and benevolence are positively correlated while universalism and stimulation are negatively correlated. The purpose of this communication is to present the evolution of human values taking into account the religiosity of people in Portugal, from a perspective of European comparison. This communication also intends to trace the sociological profile of people who claim to belong to a religion (called religious) and people who do not belong to any religion (called non-religious) in relation to human values. To achieve this objective, several sociographic variables are used to characterize the individuals and also two indices, created from European Social Survey (ESS) variables, the Index of Relation to Religion (IRR) and the Index of Religious Practice (IPR).
As can be seen in Figure 1, throughout all rounds, in Portugal, Tradition, Hedonism and Stimulation stand out as the most differentiating values between individuals who practice a religion and those who do not, the most prominent differentiator for Tradition in the last rounds observed.

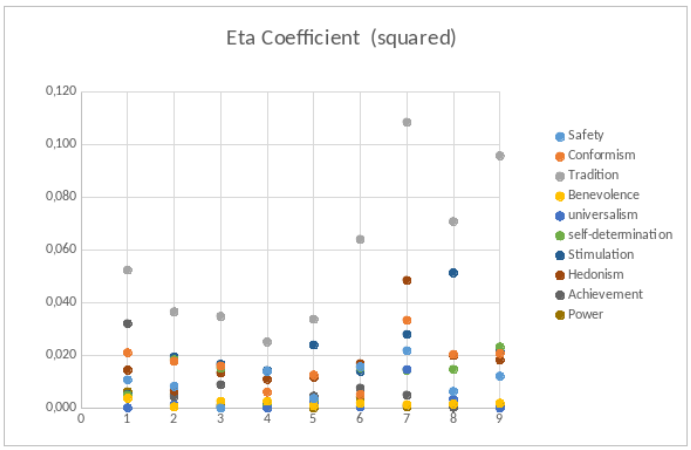

Figure 1: Evolution of the comparison between religious vs non-religious people (Portugal).

Data processing and analysis focused on the nine rounds of the European Social Survey, between 2002 and 2018. After preparing the data for analysis, statistical techniques used, such as Multiple Correspondence Analysis, and non-parametric tests of mean difference were obtained using IBM SPSS version 27.0 software.
By comparing the averages of human values, it is concluded that Portugal is closer to European values, from round 6 (2012). Multiple Correspondence Analysis allowed us to conclude that, as expected, people who belong to a religion have a formal relationship with it. While those who do not belong to any religion are associated with a residual or null relationship, but also with informal relationships, and sometimes medium or intense. Which means that, even though they do not belong to a religion, they may have some kind of relationship with it.

References

[1] A. Carneiro, M. A. P. Dinis, A. Leite, et al. Human Values and Religion: Evidence from the European Social Survey. *Social Sciences*, 10:75–92, 2021.

[2] B. Le Roux and H. Rouanet. *Multivariate Correspondence Analysis.* Sage Publications, London, 2010.

[3] L. Lebart and M. Morineau, A.and Piron. *Statistique Exploratoire Multidimensionalelle.* Dunod, Ed., Paris, 1995.

[4] S. Roccas and S.H. Schwartz. Church-state relations and the association of religion with values: A study of catholics in six countries. *Cross-cultural Research*, 31:356–375, 1997.

[5] S.H. Schwartz and S. Huismans. Value priorities and religiosity in four Western religions. *Social Psychology Quaterly*, pages 88–107, 1995.

11 December, 10:20 - 10:40

A longitudinal analysis of the severity of road accidents in the district of Setúbal between 2016 and 2019

Paulo Infante[1], Anabela Afonso[2], Gonçalo Jacinto[3], Leonor Rego[4], Pedro Nogueira[5], Marcelo Silva[6], Vitor Nogueira[7], José Saias[8], Paulo Quaresma[9], Patrícia Gois[10], Paulo Rebelo Manuel[11]

[1] DMat/ECT, CIMA/IIFA, Universidade de Évora, pinfante@uevora.pt
[2] DMat/ECT, CIMA/IIFA, Universidade de Évora, aafonso@uevora.pt
[3] DMat/ECT, CIMA/IIFA, Universidade de Évora, gjcj@uevora.pt
[4] DMat/ECT, Universidade de Évora, lrego@uevora.pt
[5] DInf/ECT, NOVALINCS/IIFA, Universidade de Évora, pmn@uevora.pt
[6] DInf/ECT, NOVALINCS/IIFA, Universidade de Évora, marcelogs@uevora.pt
[7] DInf/ECT, NOVALINCS/IIFA,Universidade de Évora, vbn@uevora.pt
[8] DInf/ECT, NOVALINCS/IIFA, Universidade de Évora, jsaias@uevora.pt
[9] DInf/ECT, NOVALINCS/IIFA, Universidade de Évora, pq@uevora.pt
[10] DAVD/EA, Universidade de Évora, pafg@uevora.pt
[11] CIMA/IIFA, Universidade de Évora, pjsrm@uevora.pt

Between 2016 and 2019, the district of Setúbal was one of the districts in Portugal with the most accidents with deaths or serious injuries. The work developed focuses on accident's characterization by severity over these four years. We used data from the Traffic Accidents Statistical Bulletin. Using parametric and non-parametric tests and control charts we identified periods of the day, weekdays and months of the year with the highest and lowest severity accidents.

Keywords: control charts, gravity, non-parametric tests, parametric tests, road accident

---

Road accidents are one of the greatest social problems of modern societies, not only because of the high number of victims, but also because of the high associated costs. In 2016 road traffic injuries were the 8th leading cause of death and are predicted to become the 7th leading cause of death by 2030 [3]. Moreover, road traffic costs represent about 1-3% of gross domestic product (GNP) worldwide [3].

In 2019, Portugal recorded the 6th highest rate of road fatalities among the 27 members of the European Union (EU), with 16 more fatalities per million inhabitants than the EU as a whole [1]. Beyond the impact caused by fatalities, road accidents had an economic and social impact equivalent to 1.2% of GDP, i.e. EUR 2.3 billion [2].

Between 2016 and 2019, the district of Setúbal was one of the districts in Portugal with the most accidents resulting in fatalities or serious injuries. These results motivated a

partnership between the GNR District Command of Setúbal and the University of Évora to reduce the injury severity of road accidents in this district. In this paper, we present a characterization of the road accidents in Setúbal's district by their severity over this period of 4 years. Daily data from the Statistical Bulletin of Road Accidents were used, which were collected and validated by the Territorial Command of the GNR of Setúbal. The data were updated by the National Road Safety Authority (ANSR) with the severity of the victims to 30 days. Parametric and non-parametric tests, and U control charts, were used to identify periods of the day, weekdays and months of the year with higher and lower severe accidents in this district.
Our results show that the severity of the accidents varies among hours and weekdays. Also, some types of accidents with victims are more likely to occur on Fridays and others on Sundays. Fatalities are related to the occurrence of the accident during working hours.

Acknowledgements This work is a contribution to the MOPREVIS project "FCT DSAIPA/DS/0090/2018" funded by the FCT-Foundation for Science and Technology, under the National Initiative on Digital Skills e.2030, Portugal INCoDe.2030. We would also like to thank the project's partner entities.

References

[1] Eurostat. Road accidents: number of fatalities continues falling. `https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/ddn-20210624-1`, 2021. Accessed on 25 Jun. 2021.

[2] Lusa. Sinistralidade rodoviária tem impacto económico e social negativo de 1,2% do pib - governo. `https://www.rtp.pt/noticias/pais/sinistralidade-rodoviaria/tem-impacto-economico-e-social-negativo-de-12-do-pib-governo_n1112193`, 2018. Accessed on 18 Nov. 2018.

[3] World Health Organization. Global status report on road safety 2018. `https://apps.who.int/iris/rest/bitstreams/1164010/retrieve`, 2018. Accessed on 8 Jun. 2021.

11 December, 10:20 - 10:40

Data, Technology and Journalism

Cláudia Silvestre[1], **Pedro Frazão**[2],
[1] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt
[2] NetCF org, pfrazao@netcf.org

Despite of journalism is still based on ethical values, such as truth, balance, accuracy, and objectivity, it has been changing in many ways. Technological development has had a major impact in the way of thinking and practicing "the" journalism. How does it affect teaching and learning activities? In this work, we present a first approach to Python programming language, with an emphasis on data preparation and exploration, to analyze media coverage of COVID-19, based on 2025 news headlines.

Keywords: Data analysis, journalism students, Python, teaching

Technological development has had a major impact on all levels of our societies [3], namely in the way of thinking and practicing journalism [5]. Consequently new job types like data journalist or infographic designers have been evolved in newsrooms. This reveals that journalism students should be trained to learn the contemporary technologies. "Interestingly, although transformation of journalism practices after the artificial intelligence and algorithms is on the agenda of recent communication studies, the reflection of this technology in journalism education is still a neglected field of study." ([4]:170). This technological change also include the way we teach data analysis, since data are really used as a primary source for telling a story [6].

Thinking about that, we carried out a small project using Python programming language. Communication students usually do not have any contact with a programming language, but it is a valued skill [1]. The goal of this project was to perform a statistical analysis of the Portuguese media coverage of the pandemic, focused on news headlines. We have chosen three Portuguese large-circulation newspapers, two daily newspapers: *Diário de Notícias* and *Público*, and a weekly, *Expresso*. The data set have 2025 headlines related to pandemic and were collected between between January and March, 2021.

Data cleaning was the first step. Incomplete data and duplicate news were removed, after that we got 2020 headlines. Then we enriched the dataset by adding more variables, such as, news is on the first page (yes/no), references to numbers in the headline (yes/no) and the tone of the news (positive/neutral/negative). With this analysis we can conclude that most (80%) news had a neutral tone, and numerical information was rarely used (10%), but when it was used it often highlighted positive or negative information. The most common words used in headlines were *pandemia* (pandemic), *vacinas* (vaccines) and

COVID. And headlines with numbers generally referred to the pandemic evolution, like number of infected, dead or recovered, and to COVID-19 crisis economic impact.
Although this project is quite simple, we find it useful to introduce a programming language, namely Python [1]. It also helps to increase technological skills knowledge which is "one of the requirements of the modern era in additional of traditional journalism courses in journalism education" ([2]:7471).

References

[1] J. Brunner and J. Kim. Teaching data science. *Procedia Computer Science*, 80:1947–1956, 2016.

[2] Özen C. New Technologies Challenging the Practice of Journalism and The Impact of Education: Case of Northern Cyprus. *EURASIA Journal of Mathematics, Science and Technology Education*, 13:7463–7472, 2017.

[3] J. Lasser, D. Manik, A. Silbersdorff, B. Säfken, and T. Kneib. Introductory data science across disciplines, using Python, case studies, and industry consulting projects. *Teaching Statistics*, 43:190–200, 2021.

[4] B. Narin. Teaching high tech storytelling: Reorganizing journalism education for programmer journalists and data journalists. *Centro de Publicaciones PUCE*, pages 169–201, 2018.

[5] V. Somayyeh. The impact of technology on journalism. *International Journal of Advance Engineering and Research Development*, 5:550–555, 2018.

[6] D. Spiegelhalter. *The Art of Statistics : How to Learn from Data*. Basic Books, United States, 2021.

11 December, 10:20 - 10:40

Field test validation for predicting VO2max in the Portuguese Military Academy

Rui Lucena[1], **Lucas Nogueira**[2], **Nuno Almeida**[3], **Cristiano Almeida**[4], **Paula Simões**[5]
[1] CINAMIL, Academia Militar, Instituto Universitário Militar and Military Readiness Lab, Portugal, rui.lucena@academiamilitar.pt
[2] CINAMIL, Academia Militar, Instituto Universitário Militar, Portugal nogueira.las@exercito.pt
[3] CIPER – Faculdade de Motricidade Humana, Universidade de Lisboa and Military Readiness Lab, Portugal, nrcalmeida@gmail.com
[4]CINAMIL, Academia Militar, Instituto Universitário Militar and Military Readiness Lab, Portugal, almeida.cf2@exercito.pt
[5] CINAMIL, Academia Militar, Instituto Universitário Militar and Military Readiness Lab and CMA– Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal, paula.simoes@academiamilitar.pt

---

The aim of this study was to determine, between three commonly used physical field (indirect) tests, which test can be used as a suitable measure to reproduce indirect maximum oxygen uptake (VO2max) laboratory prediction provided by the Ellestad test, in the case of Military Academy cadets. The YO-YO test reveals to be the most suitable field test, allowing to enhance our understanding of indirect VO2max prediction and what type of scenario should be implemented in the military environment.

Keywords: VO2max prediction, Military Academy, Hypothesis Testing

---

The training of the Portuguese military Academy Cadets is comprised of a mixture of high intensity exercises ranging from obstacle courses to long, loaded marches. Aerobic power, ie maximum oxygen uptake (VO2max), is an indicator of aerobic capacity, essential for characterising and understanding cadets' responsiveness to military training. The direct process of VO2max determination, through measurement of inspired and expired gas exchange during a maximal test, is not an option for application, so the use of validated indirect tests are the best tools for this purpose. The Cooper test is an indirect protocol used by the Military Academy to predict VO2max. However, reliable alternative field tests have been developed to better predict this parameter, such as the multistage 20-m shuttle run test and the YO-YO Intermittent Recovery Test Level 1, among others. In order to validate the field test which better predicts the VO2 max in the Portuguese Military Academy, the Ellestad test was performed. Despite being non-invasive, this protocol has been validated [3], thus providing the necessary conditions to fulfill this study.

The purpose of this study was to determine, between three commonly used physical field (indirect) tests, which test can be used as a suitable measure to reproduce the indirect maximum oxygen uptake (VO2max) laboratory prediction provided by the Ellestad test, in the case of Military Academy cadets. The indirect VO2max calculation associated with the referred tests (field and laboratory) were performed considering the equations for VO2max scientifically validated and commonly cited (see Table 1).

Table 1: Equation for VO2máx (ml/kg/min)associated with the considered test

Cooper	$22,351D$–$11,288$	D - distance(km)
VV20	$31.025 + 3.238X - 3.248A + 0.1536AX$	A - age (years) X - level speed (km/h)
YO-YO	$0,0084D + 36,4$	D - distance (m)
Ellestad	$3,933T + 4,46$	T -time

Thirty-Three healthy cadets from Portuguese Military Academy participated in the study (age: 22 ± 0.9 years; body mass: 73.2 ± 7.8 kg; height: 175 ± 0.1 cm; body mass index: 23.6 ± 1.6). All the participants were free from any injury or pain that would prevent maximal effort during performance testing. After receiving a thorough explanation of the protocols, all gave their written informed consent to the study. It was only possible to apply the Ellestad test to 12 of the 33 cadets (due to pandemic constraints). Data was analysed using IBM SPSS statistical software [2].

The study is performed in two main phases. First, considering the enunciated physical field tests, it was analysed if it can be considered that them provide different results of VO2max. Here, by resorting to parametric statistical methods for hypothesis testing development, considering the One Way Analysis of Variance (ANOVA) with multiple pairwise comparisons. Second, taking into account that significant differences were detected, we are interested in conduct a analyses which enables the selection of the most appropriate test for depicting the values of VO2max obtained through the Ellestad laboratory test. In this stage, based on non-parametric statistical methods, using the Kruskal-Wallis test [1].

The results revelled significant (mean) differences between the considered field physical tests, and that the most suitable field test for reproducing the indirect VO2max given by the considered laboratory test were the YO-YO Intermittent Recovery Test Level 1, allowing to enhance our understanding of indirect VO2max prediction and what type of scenario should be implemented in the military environment.

References

[1] G. Casella and R. Berger. Statistical inference. duxbury. *Pacific Grove, CA*, 2002.

[2] J. Marôco. *Análise Estatística com o SPSS Statistics.: 7ª edição*. ReportNumber, Lda, 2018.

[3] M. Pollock, R. Bohannon, K. Cooper, J. Ayres, and A. Linnerud. A comparative analysis of four protocols for maximal treadmill stress testing. *American Heart Journal*, 92(1):39–46, 1976.

11 December, 10:20 - 10:40

Mechanical Behavior of Skin: an ANOVA Approach

M. Filomena Teodoro[1,2], **Teresa Oliveira**[3,4]
[1] CEMAT, IST, Lisbon University, Portugal mteodoro64@gmail.com,
[2] CINAV, Naval Academy, Portuguese Navy, Portugal,
[3] CEAUL, Faculdade de Ciências, Lisbon University, Portugal,
[4] Sciences and Technology Department, Open University, Lisboa, Portugal.

The skin and adjacent soft tissues of each individual assume a mechanical behavior when subjected to external forces. Knowing that the majority of studies from literature related with the present theme from Mechanical Engineering are traditionally applied numerical methods, this work aims to verify whether it is possible to analyze the mechanical behavior of the skin by performing a statistical analysis of the experimental data. The present study investigates the differences between genders in the perception of pain and soft tissue deformation, also complementing, with the study of differences in BMI and adipose fold. Eighty individuals (40 men and 40 women) were tested in a single anatomical region, the forearm, applying an indenter. The participants filled a pre and a post questionnaires and underwent several measurements and the indentation test. It was concluded that there are significant differences between genders, both in the measured variables and in the variables related to the indentation test. In view of these differences (biological and physiological differences), the variable maximum strength, which presents $20N$ of average difference between the maximum values of each gender, stands out. These differences were evidenced using an ANOVA approach.

Keywords: gender differences, indentation method, pain perception, pain threshold, body mass index, fat fold, ANOVA, nonparametric ANOVA

Rehabilitation equipment has a high rate of rejection / discontinuity of use (over 30% [1]). A solution to overcome this problem is the use of simulation to develop rehabilitation equipment suitable for the intended function, but which also takes into account the comfort of the end user [4]. For this it is essential that the contact between the user and the equipment is accounted for. However, to numerically simulate the contact between equipment and the skin, it is necessary to have equations that satisfactorily reproduce the mechanical behavior of the skin and there is also the need to know the limits of load application for the user's safety and comfort. Taking into account the variability of the mechanical behavior of the skin, it is desirable that the determination of the coefficients of these same equations and the limits of load application are based on experimental results. The mechanical behavior of the skin depends on the place where the contact occurs, on the gender and

age of the individual, among other factors, as well as on the test parameters used, which makes it difficult to obtain these coefficients and limits. We verify that it is possible to analyze the mechanical behavior of the skin performing an analysis of variance to explore the experimental data, whether the results are consistent for a group of individuals and to verify whether the limits of load application in safety and comfort remain stable for that group of individuals.
We analyze the differences between men and women in the measured variables (body mass index (BMI) and fat fold) and in the variables obtained through the indentation test (Strengths, Deformations and absorbed Energy at pain threshold[1]). In the present work we made a preliminary evaluation of the psychological behavior and physical condition of individuals through questionnaires, before and after the application of the indentation test. We also investigate if there is some association between the considered variables through using an ANOVA approach [3, 2] in a first approach. Some of the hypothesis of using ANOVA conduced to $p - values$ between 0.05 and 010. Considering this issue, we have applied nonparametric ANOVA with better results.

Acknowledgements This work was supported by Portuguese funds FCT, through the CEMAT, University of Lisbon, Portugal, project UID/Multi/04621/2019, and CINAV, Portuguese Naval Academy.

References

[1] W.C. Mann. *Smart Technology For Aging, Disability and Independence: The State Of Art.* Wiley, New York, 2005.

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

[3] M.L. Morgado, M.F. Teodoro, and T. Perdicoulis. *Métodos Estatísticos em Ciências Biomédicas. Série Didáctica, Vol. 394.* Universidade de Trás-os-Montes e Alto Douro (UTAD), Vila Real, 2010.

[4] P. Silva. *Computational Modelling of a Wearable Ankle-Foot Orthosis For Locomotion Analysis and Comfort Evaluation.* PhD thesis, Instituto Superior Técnico, Universidade de Lisboa, 2012.

[5] A.O. Vitor, E.L. Ponte, P. Soares, M.E. De Sousa Rodrigues, R. Lima, K.M. Carvalho, M. Patrocínio, and S. Vasconcelos. Psychophysiology of pain: a literature review. *Reciis*, 2(1):85–94, 2008.

[1]Perception of Pain: "Unpleasant emotional and sensory experience associated with tissue damage" [5]

11 December, 10:20 - 10:40

The Relationship of the Human Capital Index with the Level of Education and the Adult Survival Rate

Alexandra Marques[1], **Ana Pinheiro**[1], **Maria Carolina Matos**[1], **Cristina Torres**[2], **Cristina Lopes**[2], **Isabel Vieira**[2]
[1] ISCAP, Polytechnic of Porto
[2] CEOS.PP, ISCAP, Polytechnic of Porto, ctorres@iscap.ipp.pt

The relationship between the level of education and the adult survival rate with the human capital index and its components is studied using cluster analysis, applied to a sample of 174 countries covering all continents. The results allow the grouping of countries into three clusters that are consistent with three different levels of development. It was also observed that the lower the values of survival rate and level of education, the lower the human capital index of a country is.

Keywords: Clusters Analysis, Human Capital Index, Expected Years of School, Adult Survival Rate

This study aims to relate the human capital index in 174 countries around the world with the adult survival rate and the number of expected years of school, using data updated to september 2020, retrieved from The World Bank (https://data.worldbank.org/) database. The mission of The World Bank is to reduce poverty and build shared prosperity in the international development and economic sustainability of countries by fighting inequalities [1]. In the database used the human capital index calculates the contributions of health and education to worker productivity. The index values are between 0 and 1 and measure the productivity as a future worker of a person born in the year in question in relation to the full health and complete education. These variables were chosen taking into account their impact on society and their interconnection with socio-economic advances. It is suposed that a country whose adult survival rate is higher, will make greater advances in the health sector. On the other hand, it can be assumed that a higher adult survival rate indicates that the population will work up to an older age, so they have more years to invest in education, becoming a vicious cycle, since high educational parameters, presuppose growth in the economy, better living conditions, more access to health, higher survival rates, and closing the cycle, incentives to education. This analysis is an asset in achieving The World Bank's mission, as it provides data and facts that support theoretical bases that can be applied in society and fulfill its objective. In this study, multivariate analysis (cluster analysis) was used in order to assess the position of each country with regard to education and health (represented by the adult survival rate), and what is its relationship with the human capital index. Combining the different variables through the various agglomerative clustering

methods, it was concluded that the Furthest Neighbour Method was the one that obtained a dendrogram that better distinguished the formed clusters. The quadratic Euclidean measure and the z-score standardization were used. Three clusters were generated and, according to the characteristics of the profile of the countries that make up each cluster, they were designated as: Underdeveloped Countries, Developed Countries and Developing Countries. It was found that underdeveloped countries have a lower adult survival rate, and that the inhabitants attend the educational system for fewer years, which concurs with the low index of human capital. In contrast, Developed Countries, having a higher adult survival rate, have a greater margin of years spent on education and consequently a higher human capital index. The results for Developing Countries are found between the values of the other two groups, sometimes approaching those of Developed Countries, mainly in the variable expected years of school. It can be assumed that this approximation is the result of an effort on the part of Developing Countries to increase schooling and enhance economic-technological development.

Acknowledgements This work is financed by portuguese national funds through FCT − Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

References

[1] *World Development Report 2021: Data for Better Lives.* The World Bank, 2021.

[2] João Marôco. *Análise Estatística com o SPSS Statistics.* Report Number, 6.ª edition, 2014.

Author Index

SPONSORS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


UNIVERSIDADE
BEIRA INTERIOR


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


BANCO DE PORTUGAL
EUROSISTEMA


COVILHÃ
MUNICÍPIO
A TECER O FUTURO


Quinta dos Termos
A EXCELÊNCIA DA NATUREZA


QUEIJOS BRAZ