Reunião Anual da

ASSOCIAÇÃO PORTUGUESA DE CLASSIFICAÇÃO

E ANÁLISE DE DADOS (CLAD)

Programa e Livro de Resumos

AS JOCLAD 2017 TIVERAM O APOIO INSTITUCIONAL DE:

CLAD
Associação Portuguesa de
Classificação e Análise de Dados

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Ficha Técnica

Presidente das Jornadas

José Gonçalves Dias (Presidente da CLAD)

Secretário das Jornadas

Isabel Silva Magalhães (Universidade do Porto)

Comissão Organizadora

Conceição Rocha (LIAAD - INESC TEC)

Fernanda Sousa (Universidade do Porto)

Isabel Silva Magalhães (Universidade do Porto)

Luís Miguel Grilo (Instituto Politécnico de Tomar)

Presidente da Comissão Científica

Adelaide Figueiredo (Universidade do Porto)

Título: XXIV Jornadas de Classificação e Análise de Dados (JOCLAD 2017). Livro de Resumos.

Produzido: Instituto Nacional de Estatística

Editores: Adelaide Figueiredo, Isabel Silva Magalhães, José Gonçalves Dias, Luís Miguel Grilo, Conceição Rocha, Fernanda Sousa

ISBN: 978-989-98955-3-9

Prefácio

No Campus da Asprela, a Faculdade de Engenharia da Universidade do Porto (FEUP) acolhe este ano as XXIV Jornadas de Classificação e Análise de Dados (JOCLAD 2017), no edifício sede do INESC-TEC (Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência).

A FEUP recebeu a sua denominação atual em 1926 e desde então tem sido o vivo exemplo de uma instituição de criação, transmissão e difusão do conhecimento, da tecnologia e da cultura na área da engenharia. A formação dos jovens para o exercício da profissão de engenheiro é sustentada em Investigação e Desenvolvimento de excelência, contemplando as vertentes científica, técnica, ética e cultural.

É com grande gosto que a FEUP recebe, pela primeira vez, as Jornadas Anuais da Associação Portuguesa de Classificação e Análise de Dados (CLAD), fundada em 1994. Estas Jornadas Científicas, que têm ocorrido anualmente desde então, sem qualquer interrupção, têm como principais objetivos fomentar e desenvolver a investigação em *data science*, apresentando ferramentas que facilitem a interpretação e validação da crescente quantidade de dados que surgem em cada instante nas mais variadas áreas do conhecimento.

O habitual caráter multidisciplinar das Jornadas reflete-se no Programa das JOCLAD 2017, que engloba de forma equilibrada a apresentação de trabalhos teóricos e aplicados, onde as diversas temáticas da Análise de Dados em domínios transversais à sociedade são destacadas. Começamos por agradecer aos oradores das sessões plenárias, os Professores Pedro Larrañaga (Universidad Politécnica de Madrid, Espanha), Rosanna Verde (Università della Campania "Luigi Vanvitelli", Itália) e Victor Lobo (ISEGI / UNL e Escola Naval).

Agradecemos também a todos os autores dos trabalhos de proposta livre (em formato oral e *poster*) e moderadores de sessões, aos membros da Comissão Científica, bem como aos participantes convidados e aos colegas que procederam à revisão dos trabalhos que constam neste livro. Um agradecimento particular aos Professores Maria do Rosário Oliveira (Instituto Superior Técnico) e Pedro Larrañaga (Universidad Politécnica de Madrid, Espanha) que lecionam os mini-cursos, bem como à Doutora Sónia Cardoso Pintassilgo (em conjunto com a Direção da Associação Portuguesa de Demografia), Doutora Filipa Lima (Banco de Portugal) e Dr. Carlos Marcelo (Instituto Nacional de Estatística) que se disponibilizaram para organizar as Sessões Temáticas que constam no Programa.

Com o objetivo de estimular a atividade de estudo e investigação científica nas áreas científicas da CLAD entre os jovens que trabalham nas áreas de Classificação e Análise de Dados, foram este ano instituídas as Bolsas CLAD. Cada bolsa destina-se a custear a participação nas Jornadas. Numa sessão especial serão atribuídas duas Bolsas CLAD 2017 e os premiados apresentarão os seus trabalhos.

Por último, agradecemos também a todas as entidades que direta ou indiretamente apoiaram ou patrocinaram estas Jornadas.

O nosso muito obrigado a todos!

Pel'A Comissão Científica das JOCLAD 2017	Pel'A Comissão Organizadora das JOCLAD 2017	Pel'A Direção da CLAD
Adelaide Figueiredo	Isabel Magalhães	José Gonçalves Dias

Porto, abril de 2017

ORGANIZAÇÃO

Presidente das Jornadas

José Gonçalves Dias (Presidente da CLAD)

Secretário das Jornadas

Isabel Silva Magalhães (FEUP - Universidade do Porto)

Comissão Organizadora

Conceição Rocha (LIADD - INESC TEC)
Fernanda Sousa (FEUP - Universidade do Porto)
Isabel Silva Magalhães (FEUP - Universidade do Porto)
Luís Miguel Grilo (Instituto Politécnico de Tomar)

Presidente da Comissão Científica

Adelaide Figueiredo (FEP - Universidade do Porto)

Comissão Científica

A. Manuela Gonçalves (Universidade do Minho)
Ana Lorga da Silva (Universidade Lusófona)
Ana Sousa Ferreira (Universidade de Lisboa)
Anabela Afonso (Universidade de Évora)
Carlos Ferreira (Universidade de Aveiro)
Carlos Soares (Universidade do Porto)
Catarina Marques (Instituto Universitário de Lisboa)
Conceição Amado (Universidade de Lisboa)
Conceição Rocha (INESC-TEC e Universidade do Porto)
Fátima Salgueiro (Instituto Universitário de Lisboa)
Fernanda Otília Figueiredo (Universidade do Porto)
Fernanda Sousa (Universidade do Porto)
Fernando Nicolau (Universidade Nova de Lisboa)
Gilda Soromenho (Universidade de Lisboa)
Helena Bacelar-Nicolau (Universidade de Lisboa)
Irene Oliveira (Universidade de Trás-os-Montes e Alto Douro)
Isabel Silva Magalhães (Universidade do Porto)
Jorge Cadima (Universidade de Lisboa)
José Gonçalves Dias (Instituto Universitário de Lisboa)

Luís Miguel Grilo (Instituto Politécnico de Tomar)

Manuela Neves (Universidade de Lisboa)

Margarida Cardoso (Instituto Universitário de Lisboa)

Paula Brito (Universidade do Porto)

Paula Vicente (Instituto Universitário de Lisboa)

Paulo Infante (Universidade de Évora)

Pedro Campos (Universidade do Porto)

Pedro Duarte Silva (Universidade Católica Portuguesa)

Rosário Oliveira (Universidade de Lisboa)

Susana Faria (Universidade do Minho)

Victor Lobo (Universidade Nova de Lisboa)

APOIOS


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO


INESCTEC
TECNOLOGIA E CIÊNCIA
LABORATÓRIO ASSOCIADO


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


BANCO DE PORTUGAL
EUROSISTEMA


PSE
PRODUTOS E SERVIÇOS DE ESTATÍSTICA


ipt
Instituto
Politécnico
de Tomar


1834
ASSOCIAÇÃO
COMERCIAL
DO PORTO


Visit Porto.

PROGRAMA

QUINTA-FEIRA, 20 DE ABRIL

9:00 Registo e entrega de documentação

9:30 **Mini-curso** – Auditório A

Análise em Componentes Principais, Robustez e Deteção de *Outliers*: Uma Introdução em R
Maria do Rosário Oliveira, p. 3.

Moderador: Fernanda Otília Figueiredo

10:45 **Pausa para café**

11:00 **Mini-curso** (*cont.*)

12:30 **Almoço**

13:30 **Mini-curso** – Auditório A

Supervised classification methods

Pedro Larrañaga, p. 5.

Moderador: Conceição Amado

15:00 **Pausa para café**

15:15 **Mini-curso** (*cont.*)

16:30 **Sessão de Abertura das Jornadas** – Auditório B

17:00 **Sessão Plenária I** – Auditório B
Análise de Dados Geoespaciais utilizando Mapas Auto-Organizados
Victor Lobo, p. 9.

Moderador: Fernanda Sousa

18:00 Sessão Paralela I

	Auditório B ***Aplicações em Ciências da Saúde I*** Moderador: Manuela Neves	Auditório A ***Classificação I*** Moderador: Luís Grilo
18:00	**Análise cinemática da marcha como ferramenta complementar de decisão na doença de Parkinson Vascular** Flora Ferreira, Miguel Gago, Nafiseh Mollaei, Lurdes Rodrigues, Nuno Sousa, Pedro Pereira Rodrigues, Estela Bicho, p. 47.	**Regression based Approaches for Predicting Age at Onset** Maria Pedroto, Alípio M. Jorge, João Mendes-Moreira, Teresa Coelho, p. 53.
18:20	**Melhorar o diagnóstico da Apneia Obstrutiva do Sono com dados clínicos: uma abordagem Bayesiana** Daniela Ferreira-Santos, Pedro Pereira Rodrigues, p. 49.	**A likelihood-ratio control chart for monitoring the parameters of a logistic process** Fernanda Otília Figueiredo, Maria Ivette Gomes, Amitava Mukherjee, p. 55.
18:40	**Multimorbidade de pacientes com enfarte agudo do miocárdio descrita usando abstrações temporais e redes bayesianas** Carla Silva, Mariana Lobo, Pedro Pereira Rodrigues, p. 51.	**O modelo logístico multinomial na saúde ocupacional** Luís M. Grilo, Helena L. Grilo, p. 57.

SEXTA-FEIRA, 21 DE ABRIL

8:30 Registo e entrega de documentação

9:00 **Sessão Plenária II** – Auditório B

Bayesian networks in neuroscience

Pedro Larrañaga, p. 11.

Moderador: Isabel Silva

10:00 Sessão Paralela II

	Auditório A ***Modelos Estocásticos e Modelos Espaciais*** Moderador: Paulo Infante	Auditório B ***Data Mining*** Moderador: Conceição Rocha
10:00	**O efeito do delineamento de amostragem em modelos de análise de sobrevivência** Ana Laura Carreiras, Paulo Infante, Anabela Afonso, p. 59.	**Disaggregation and classification of loads** Inês Soares, David Rua, João Gama, p. 67.
10:20	**Delineamentos experimentais para ensaios iniciais de programas de melhoramento de plantas: aplicação à selecção de castas antigas de videira** Elsa Gonçalves, Antero Martins, p. 61.	**Inflammatory Bowel Disease: evidence, classification and prediction** Claudia Camila Dias, Pedro Pereira Rodrigues, Altamiro da Costa-Pereira, Guilherme Macedo, Fernando Magro, p. 69.
10:40	**Deteção de distribuições atípicas em sequências genómicas** Ana Helena Tavares, Vera Afreixo, Paula Brito, p. 63.	**Detection of Centrality and Community Patterns in Multilayer Networks – An Application in Medicine** Patrícia C. T. Gonçalves, Pedro Campos, p. 71.
11:00	**l1-norm posterior probability Support Vector Machines** A. Pedro Duarte Silva, p. 65.	**Exploratory Data Analysis in Pavement Engineering Using Python** Pedro Marcelino, Maria de Lurdes Antunes, Eduardo Fortunato, Marta Castilho Gomes, p. 73.

11:20 Pausa para café + Sessão de *Posters* I

Avaliação de métodos de estimação da variância em amostras complexas
Eládio Muianga, Anabela Afonso, p. 105.

A comparison of team cooperation levels in different sports
Bruno Martins, Ana Cristina Costa, p. 107.

Mapeamento das ONGD em Portugal . Um estudo sobre a comunicação estratégica
Cláudia Silvestre, César Neto, Mafalda Eiró Gomes, p. 109.

Comparing data from CAPI and CAWI surveys
Paula Vicente, Elizabeth Reis, p. 111.

Mapas dos valores europeus
Margarida G. M. S. Cardoso, Luís Chambel, p. 113.

A utilização da Internet para viagens: Comparação da Geração Y com os seus antecessores
Carla Henriques, Suzanne Amaro, Paulo Duarte, p. 115.

12:00 **Sessão Temática I – Banco de Portugal**

	Auditório B ***Economia e Finanças*** Moderador: Filipa Lima
12:00	**How to use International Banking Statistics?** João Falcão Silva, Vânia Lopes, p. 17.
12:20	**To be or not to be... a general government unit: borderline between the financial corporations and general government sectors** Sérgio Branco, Filipe Morais, Sónia Mota, p. 19.
12:40	**Looking into R&D intensity in Portugal in the last years** Carla Ferreira, Cloé Magalhães, Luís Pinto, Mário Lourenço, p. 21.

13:00 **Almoço**

14:15 **Sessão Temática II – Instituto Nacional de Estatística**

	Auditório B ***Desafios nas Estatísticas Oficiais VI*** Moderador: Carlos Marcelo
14:15	**É possível substituir os Censos por informação administrativa?** Anabela Delgado, Paula Paulino, Sandra Lagarto, p. 23.
14:35	**O uso do *Web Scraping* nas estatísticas oficiais** Maria José Fernandes, p. 25.
14:55	**Padrão das disponibilidades alimentares para o período 2012/16** Carlos Carvalho, Sofia Duarte, p. 27.
15:15	**Mismatch between jobs and skills in the EU** João Lopes, Marco Moura, Sónia Quaresma, p. 29.

15:35 **Sessão BOLSAS CLAD 2017**

	Auditório B Presidente: Adelaide Figueiredo
15:40	**Deteção de *Outliers* para Melhoria do Diagnóstico de Perdas em Sistemas de Abastecimento de Água** Maria José de Almeida e Silva, p. 41.
16:00	**Comparação estatística de medidas de qualidade de imagem: da visualização à quantificação** Patrícia Manuel Branco das Neves, p. 43.

16:20 **Pausa para café + Sessão de *Posters* II**

Anova com dois fatores não paramétrica
Anabela Afonso, Dulce Pereira, p. 127.

Investimento em Redes Rodoviárias Europeias: um Estudo Comparativo
Pedro Marcelino, Marta Castilho Gomes, p. 129.

Avaliação do desempenho dos alunos Portugueses em Matemática – PISA 2015
Susana Faria, Manuel João Castigo, p. 131.

O Papel Mediador da Avaliação Cognitiva na Relação entre a Mútua Interferência Trabalho-Família e o *Burnout*
Jéssica Rodrigues, A. Manuela Gonçalves, Susana Faria, Clara Simães, A. Rui Gomes, p. 133.

17:00 Sessão Paralela III

	Auditório B ***Classificação II*** Moderador: José Dias	Auditório A ***Séries Temporais e Dados Espaciais*** Moderador: Pedro Duarte Silva
17:00	**Efeitos da gestão do fundo de maneio na rentabilidade das empresas do setor mobiliário português** Ruben Silva, Paulo Peixoto, Susana Silva, Cristina Lopes, p. 75.	**Spatial Econometrics Modeling of the *Saúde24* line calls** Paula Simões, M. Lucília Carvalho, Sandra Aleixo, Sérgio Gomes, Isabel Natário, p. 83.
17:20	**Clustering co-authorship networks across scientific fields using parameter distributions** Conceição Rocha, Paula Brito, Fernando Silva, Alípio M. Jorge, p. 77.	**MIBEL Electricity Prices Forecasting** Eliana Costa e Silva, Ana Borges, p. 85.
17:40	**Segmentação com dados espaciais Gaussianos. Uma comparação de algoritmos** Luís F. Domingues, José G. Dias, p. 79.	**Forecasting Tourism Demand with SVARMA models** Cristina Lopes, Eliana Costa e Silva, Filomena Soares, p. 87.
18:00	**Hidden Markov models applied to technical analysis data** José G. Dias, p. 81.	**Previsão das dormidas mensais na região Norte de Portugal em anos recentes: um estudo preliminar** Joaquim Silva, Hugo Alonso, Isabel Silva, p. 89.

18:00 Reunião da Assembleia Geral da CLAD – Auditório A

20:30 **Programa Social e Jantar das Jornadas**

SÁBADO, 22 DE ABRIL

9:00 Registo e entrega de documentação

9:30 **Sessão Paralela IV**

	Auditório A ***Modelos com Variáveis Latentes*** Moderador: Maria Fátima Salgueiro	Auditório B ***Aplicações em Ciências da Saúde II*** Moderador: Anabela Afonso
9:30	**Modelling wage trajectories in Brazil using data from RAIS Identificada** Maria de Fátima Salgueiro, Marcel D.T. Vieira, Ricardo Freguglia, p. 91.	**Detectar anomalias a fim de facilitar o diagnóstico médico por meio de agrupamento de dados e abstrações temporais** Giovana Jaskulski Gelatt, André P.L.F. de Carvalho, Pedro Pereira Rodrigues, p. 97.
9:50	**O efeito da escolha do método de estimação em modelos com variável latente contínua: Um estudo de simulação** Paula C.R. Vicente, Maria de Fátima Salgueiro, p. 93.	**Lúpus: Comparação de duas medidas da atividade da doença** Ana Cristina Matos, Carla Henriques, Diogo Jesus, p. 99.
10:10	**Relationship between sustainability and competitiveness of nations: a macro level analysis** Nikolai Witulski, José G. Dias, Catarina Roseta Palma, p. 95.	**Avaliação do risco de morte em doentes operados com pancreatite aguda: É possível ir além dos *scores* clínicos?** Carla Henriques, Ana Cristina Matos, Jorge Pereira, Carlos Daniel, Tercio de Campos, p. 101.

10:30 **Sessão Temática III – Associação Portuguesa de Demografia**

	Auditório B ***Demografia*** Moderador: Pedro Campos
10:30	**Migrações de Substituição e Idade Prospetiva: projetar as necessidades laborais em Portugal (2015-2060)** Daniela Craveiro, João Peixoto, Isabel T. Oliveira, Jorge Malheiros, Eduarda M. Costa, Diogo Abreu, Vítor Escária, Paula C. Albuquerque, Maria Teresa M. Garcia, Cristina S. Gomes, Maria João G. Moreira, José Alves, p. 31.
10:50	**Modelação Estatística da Fecundidade Realizada** Teresa Lopes, Maria Filomena Mendes, Paulo Infante, p. 35.
11:10	**Uma projeção coerente dos padrões de mobilidade dos estudantes do ensino superior em Portugal** Filipe Ribeiro, Lídia Patrícia Tomé, Maria Filomena Mendes, Rita Freitas, p. 37.

11:30 **Pausa para café**

12:00 **Sessão Plenária III** – Auditório B
Distributional Data Clustering and Visualization for Official Statistics
Rossana Verde, p. 13.

Moderador: Paula Brito

13:00 Sessão de Encerramento das Jornadas

RESUMOS

ÍNDICE

Sessão Temática II – Instituto Nacional de Estatística

Sessão Temática III – Associação Portuguesa de Demografia

SESSÃO BOLSAS CLAD 2017

SESSÕES PARALELAS

Aplicações em Ciências da Saúde I

Classificação I

Modelos Estocásticos e Modelos Espaciais

Data Mining

Classificação II

Séries Temporais e Dados Espaciais

Modelos com Variáveis Latentes

Aplicações em Ciências da Saúde II

POSTERS

Sessão I

MINI-CURSO

5ª Feira, 20 de Abril – Auditório A (9:30 – 12:30)

Análise em Componentes Principais, Robustez e Deteção de *Outliers*: Uma Introdução em R

M. Rosário Oliveira
CEMAT e Dep. Matemática, Instituto Superior Técnico, Universidade de Lisboa, rosario.oliveira@tecnico.ulisboa.pt

A Análise de Componentes Principais (ACP) é uma das metodologias estatísticas mais usadas na resolução de problemas reais. No entanto, a existência de observações atípicas pode conduzir a conclusões enganadoras. A identificação destas observações é, em geral, uma tarefa difícil e a sua remoção pode traduzir-se em perdas inaceitáveis de informação. A solução natural reside no uso de procedimentos robustos. Estes têm a capacidade de detetar e acomodar as observações atípicas, propondo estimadores que, não sendo ideais no caso de se verificarem as hipóteses teóricas associadas ao modelo em estudo, dão melhores resultados quando ocorrem pequenos desvios destas mesmas hipóteses.

No estudo de problemas reais, a existência de *outliers* nos dados é muitas vezes encarada como um obstáculo, que se converte em efeitos perturbadores na análise. Mas por vezes o interesse principal do estudo reside na deteção dessas observações atípicas: a descoberta dos buracos na camada de ozono derivou da identificação de valores irregulares e inesperados nas medições efetuadas; um ataque numa rede de computadores pode ser evitado se o tráfego anómalo associado for identificado e bloqueado a tempo. A ACP robusta é um método de deteção de *outliers* que utiliza as potencialidades da ACP como método de redução da dimensionalidade em benefício da identificação de observações atípicas.

Neste curso introduzimos ACP e discutimos as limitações e vantagens dos estimadores clássicos e robustos. A deteção de *outliers* baseada na distância de Mahalanobis e em ACP é discutida. A formulação teórica e as propriedades dos métodos são exploradas, mas é dada igual importância à análise de dados reais usando o *software* R. Traga o seu portátil para explorarmos, com um conjunto de dados, como estas metodologias funcionam.

Palavras-chave: Análise de Componentes Principais, Detecção de *Outliers*, Robustez, R.

5ª Feira, 20 de Abril – Auditório A (13:30 – 16:30)

Supervised classification methods

Pedro Larrañaga[1]
[1] *Universidad Politécnica de Madrid, pedro.larranaga@fi.upm.es*

The tutorial will cover some of the most popular supervised classification methods: non-probabilistic, such as K-nearest neighbours and classification trees; probabilistic, such as logistic regression and Bayesian classifiers and also metaclassifiers. For each method the general algorithm as well as some of its variants will be explained.

Honest validation methods --repeated training and test, k-fold cross-validation and bootstrap-- for estimating the performance of a classifier will be discussed. Guidelines about their use will be presented. Several performance metrics as the accuracy, sensitivity, specificity, F1 measure, and the area under the ROC curve will be shown.

Feature subset selection based on filter (univariate and multivariate) and wrapper approaches for selecting relevant and non-redundant variables will be discussed.

The attendees of the tutorial will do some practical exercises with WEKA software applied to a real microarray data set.

SESSÕES PLENÁRIAS

Sessão Plenária I – 5ª Feira, 20 de Abril – Auditório B (17:00)

Análise de Dados Geoespaciais utilizando Mapas Auto-Organizados

Victor Lobo[1]
[1] *ISEGI / UNL e Escola Naval; vlobo@isegi.unl.pt*

A evolução das tecnologias utlizadas para georreferenciação, sensores, e sistemas de informação tem levado a uma revolução na quantidade e variedade de dados geoespaciais disponíveis, e nas análises que é possível fazer. Embora seja uma área que é estudada há bastante tempo, por exemplo pela comunidade de Engenharia de Minas, as aplicações mais recentes têm estado a usar uma panóplia de novas técnicas, usando modernos Sistemas de Informação Geográfica (SIG), Inteligência Artificial (IA), e sistemas de visualização de dados.

Entre as técnicas utilizadas para análise deste tipo de dados estão os Mapas Auto-Organizados (SOM), também conhecidos como Redes Neuronais de Kohonen. No entanto, os SOM são muitas vezes usados apenas como alternativas à técnica de K-médias. Nesta palestra será explicado o funcionamento dos SOM, discutida a sua utilização na análise de dados geoespaciais, e revistos alguns avanços recentes e aplicações desta técnica.

Sessão Plenária II – 6ª Feira, 21 de Abril – Auditório B (9:00)

Bayesian networks in neuroscience

Pedro Larrañaga[1]
[1] *Universidad Politécnica de Madrid, pedro.larranaga@fi.upm.es*

Neuroscience is a scientific domain with many challenging problems to be solved. The development of new recording technologies has caused an increase in size and complexity of neuroscience datasets. Data driven methods allowing the discovery of complex interactions between the brain and behavior and the development of tools for the diagnosis and prognosis of neurological disorders and psychiatric diseases are becoming an important issue in computational neuroscience.

Bayesian networks are computational models that represent the conditional independence relationships among the variables in a given domain. They allow for any kind of inference, from predictive or diagnostic to intercausal, bidirectional and also total and partial abduction. In addition, there exist some methods for learning Bayesian networks from data. This type of probabilistic graphical model has been applied in a large number of neuroscience problems due to its transparency and computational properties (Bielza and Larrañaga, 2014, Frontiers in Computational Neuroscience). During this talk, some Bayesian network applications developed inside the Human Brain Project, whose aim is to decipher the most highly organized piece of excitable matter in the known universe --the human brain-- will be presented.

Bayesian classifiers will be used for solving a supervised classification problem with imprecise labels for discriminating among different types of interneurons. The problem of classifying and naming neurons has been a topic of debate since the original findings of Santiago Ramón y Cajal, more than 100 years ago. The dataset was obtained by means of a web-based interactive system that collected data about the terminological choices for a set of 320 cortical interneurons by 42 experts in the field (DeFelipe et al. 2013, Nature Reviews Neuroscience).

Morphology of dendritic spines appears to be critical from the functional point of view. Although many different classifications of spines have been proposed, the most widely used categorization is still that of Peters and Kaiserman-Abramof, which traditionally groups spines into four types (thin, mushroom, stubby and filopodium). We will introduce a quantitative and objective method to clustering dendritic spines based on finite mixture models where each component of the mixture factorizes according to a Gaussian Bayesian network, that incorporates linear and directional variables. More than 7000 individually 3D reconstructed dendritic spines from human cortical pyramidal neurons to cluster spine morphologies have been used in the experiments (Luengo et al, 2017, in preparation).

Neuron morphology is crucial for neuronal connectivity and brain information processing. Computational models are important tools for studying dendritic morphology and its role in brain function. During the talk, an approach based on Bayesian networks will be used for modeling and simulation of more than 350 layer III pyramidal dendritic trees from three different regions of the neocortex of the mouse (Lopez-Cruz et al. 2011, Neuroinformatics).

Multilabel classification based on multidimensional Bayesian network classifiers applied to a real-world data set containing 488 Parkinson's disease patients will be presented in the last part of the talk. The aim is to predict the European Quality of Life-5Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39) (Borchani et al. 2012, Journal of Biomedical Informatics).

Sessão Plenária III – Sábado, 22 de Abril – Auditório B (12:00)

Distributional Data Clustering and Visualization for Official Statistics

Rosanna Verde[1]
[1] *Dipartimento di Matematica e Fisica - Università della Campania "Luigi Vanvitelli", Caserta, Italy, rosanna.verde2@gmail.com*

Distributional Data Analysis (DDA) is a new field of research related to Symbolic Data Analysis (Bock and Diday, 2000). The statistical units, or objects, are described by empirical distribution of values observed for numerical variables. In some cases, the distributions are synthesis or aggregated data, in order to preserve the confidentiality of the information. Many Official Statistics are in form of "histogram data", like the "Financial Characteristics for Housing Units With a Mortgage" data of the ACS American Community Survey 2015, of the UScensus Bureau. In such a case, an interval on proportions is furnished as a sort of "confidence interval".

Starting from previous techniques, proposed for analyzing DDA (e.g., clustering, principal component analysis), we introduce a further information in the data, given by the intervals of proportions. That leads to consider that each statistical unit is described by a set of distributions for each variable, expressed by the convex combination of the interval values on proportions.

A comparison between objects is then performed using a suitable measure between distributions: the L2 Wasserstein distance (Rüschendorf, 2001). This metric is particularly appropriate in DDA, and some of its properties have been demonstrated, especially, the decomposition of the L2 Wasserstein distance in three components related to the position, size and shape characteristics of the distributions.

A hierarchical clustering analysis allows to discover classes of objects according to the characteristics of the distribution of the values, taking also into account their variability, expressed by a set of distributions, carried out as convex combination of the bounded distributions.

A new visualization tool of the characteristics of the distributions, in a reduced subspace, has been furnished by a Principal Component Analysis (PCA) technique for distributional data (Verde et al., 2016).

An extension of PCA to this kind of Official Statistics allows a visualization of the variability of the data related to the several components of DDA.

Finally, DDA presents strong potentiality in the analysis of Big Data, and all that concerns the Volume of values and the Variability.

Bibliography

Bock H.-H., Diday E. (2000) *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organization. Springer-Verlag, Berlin.

Dias, S., Brito, P. (2015) Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8(2), 75-113.

Irpino A., Verde R. (2006) A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. IN: BATANJELI, V., BOCK, H.H., FERLIGOJ, A. & ZIBERNA, A. (Eds.) *Data science and classification, IFCS 2006*. Springer, Berlin, 185–192.

Irpino A., Verde R. (2008a) Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recogn Lett*, 29, 1648–1658.

Irpino A., Verde R. (2008) Comparing histogram data using a Mahalanobis-Wasserstein distance. IN: BRITO, P. (Ed.) COMPSTAT 2008. Physica-Verlag, Heidelberg, 77–89.

Kim J., Billard L. (2013) Dissimilarity measures for histogram-valued observations. *Communications in Statistics: Theory and Methods*, 42, 283–303.

Rüschendorf L. (2001) Wasserstein metric. IN: HAZEWINKEL, M. (Ed.) *Encyclopedia of mathematics.* Springer, New York.

Verde R., Irpino A., Balzanella A. (2016). Dimension Reduction Techniques for Distributional Symbolic Data. *IEEE Transactions on Cybernetics*, 46, 344-355.

SESSÕES TEMÁTICAS

Sessão Temática I – Banco de Portugal

Sessão Temática II – Instituto Nacional de Estatística

Sessão Temática III – Associação Portuguesa de Demografia

ST I – Banco de Portugal – 6ª Feira, 21 de Abril, Auditório B (12h00)

How to use International Banking Statistics?

João Falcão Silva[1], Vânia Lopes[2]
[1] *Banco de Portugal, jmfsilva@bportugal.pt;*
[2] *Banco de Portugal, vilopes@bportugal.pt*

Abstract: International Banking Statistics were created to monitor the activity of banking systems, providing internationally comparable measures of their exposure to several risks. In this article we focus on the consolidated banking statistics analysis and show that the Portuguese banking system country risk has changed over the years.

Key words: International banking statistics, consolidated banking statistics

Introduction

Globalization has deepened financial international interlinkages which have become more and more complex. Financial markets and the economies became more intertwined both domestically and abroad. In this respect, International Banking Statistics play a fundamental role to provide international comparable measures of the banks' cross border activities. These data comprises two different data sets: the locational banking statistics (LBS) and the consolidated banking statistics (CBS). The purpose of this article is to present International Banking Statistics and its main components. In addition, a descriptive analysis of the CBS is presented for Portugal since 2004.

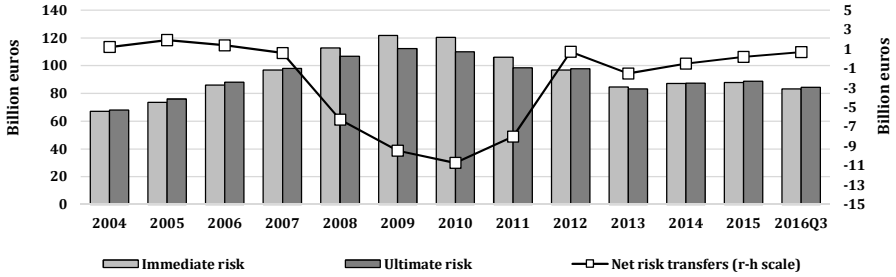
LBS measure banks' international holdings/ liabilities based on the concept of residency. It uses unconsolidated information of financial claims and liabilities in the balance sheet of banks resident in reporting countries. The CBS address national banking exposures to country risk. It is based on the nationality of the reporting bank, i.e., the country where the bank's head office is located (excluding inter-office positions).

CBS offers two different perspectives: the immediate risk basis and the ultimate risk basis. The immediate risk basis covers claims allocated to the country of residence of the counterpart which signed the agreement with the bank, even when the claim is guaranteed by a third party. On the contrary, the ultimate risk basis corresponds to foreign financial claims against the counterpart that is ultimately responsible for the compliance with the agreement. At a country level, the difference between the ultimate risk basis and the immediate risk basis corresponds to the risk transfers across countries. Under the current statistical methodology, a risk transfer is recorded with a negative sign when the ultimate risk of a foreign claim held by a bank of the reporting country is addressed by an entity resident in the reporting country. On the other hand, when the ultimate risk of a domestic asset held by a bank of the reporting country is addressed by a non-resident entity, the risk transfer is recorded with a positive sign.

Data Analysis: Immediate risk basis, ultimate risk basis and net risk transfers

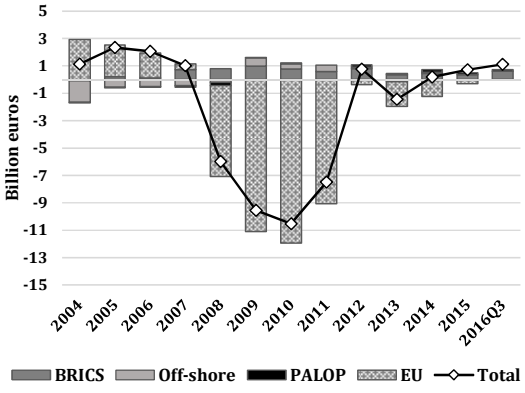
Chart 1 shows the international financial assets of Portuguese banks (CBS), split between the immediate risk basis, the ultimate risk basis and the correspondent risk transfers. Chart 2 shows the net risk transfers broken down by relevant geographic aggregates (European Union – EU; BRICS; Off-shore financial centers and Portuguese Speaking African countries – PALOP).

Chart 1: International financial assets of Portuguese banks (2004-2016Q3)



Between 2004 and 2007 (Chart 1) the net risk transfers were systematically positive, reflecting a period where foreign financial claims of Portuguese banking groups were more exposed to other countries on an ultimate risk than on an immediate risk. This situation was reversed between 2008 and 2011 where exposure to ultimate risk fell below that on an immediate risk basis. This implies that part of international immediate risk was ultimately borne by residents in Portugal. Analysing the same period by geographical counterparty (Chart 2), European Union exhibits a negative net risk transfer. This means that at an ultimate risk basis, a major part of the assets held by Portuguese banks vis-à-vis entities resident in the European Union are guaranteed by entities resident in other geographical areas.

Chart 2: Net risk transfers, by geographical aggregates



For the most recent years, the size of net risk transfers has decreased markedly. This suggests that the value of foreign claims on an immediate risk basis is very similar to that of claims on an ultimate risk basis.

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

Banco de Portugal (2015). Statistics on foreign financial claims of Portuguese banks on a consolidated basis. *Statistical Press Release, 13.*

ST I – Banco de Portugal – 6ª Feira, 21 de Abril, Auditório B (12h20)

To be or not to be... a general government unit: borderline between the financial corporations and general government sectors

Sérgio Branco[1], Filipe Morais[2], Sónia Mota[3]

[1] *Banco de Portugal, jscbranco@bportugal.pt;*

[2] *Banco de Portugal, fmmorais@bportugal.pt;*

[3] *Banco de Portugal, scmota@bportugal.pt*

Abstract: Sector classification issues related with the borderline between the financial corporations and general government sectors have been relevant in the context of European government interventions in financial institutions since 2008. This paper present different views concerning the sector classification of entities. We argue the whole statistical community should be involved in order to perform a review of the rules on sector classification of entities. Besides ensuring a correct sector classification, statisticians should also guarantee that the classification is unique.

Key words: Classification, Financial corporations, General government.

The classification of entities in the European framework relies on the European system of national and regional accounts in the European Union (ESA2010). In the specific case of general government sector the Manual on Government Deficit and Debt – 2016 edition (MGDD) is also of relevance.

The practical application of the theoretical rules on the classification of entities established in the international statistical manuals could be complex divergent and not straightforward. Moreover, the existing rules on the delimitation of institutional sectors do not always offer a clear guidance for the classification of entities, especially when they are engaged in financial activities and, at the same time, government units have a strong influence in their activities. In those cases there could be valid arguments for the classification of the units in the financial corporations sector or in the general government sector. The deep analysis of the rules on the classification of entities presented in the various statistical manuals can lead to different interpretations by different specialists that can conduct to some incoherence in specific cases and, therefore, to inconsistent statistical results, namely concerning captive financial institutions, defeasance structures and protection funds.

In the case of **captive financial institutions** the classification controversy arises from the incoherence between ESA2010 and MGDD in relation to the status (or not) of the captive financial institutions as separate institutional units. In fact, ESA §2.20 refers that only the captive financial institutions with no independence of action are to be classified in general government. Notwithstanding, according to §47 and §48 of Part I.6 of MGDD, if certain conditions are met, other than that of independence, all government controlled captives financial institutions would be classified in the government sector. The case:

Hungarian Eximbank is a government controlled Hungarian Export-Import Bank. According to Eurostat, Eximbank is a captive financial institution and should, therefore, be classified in general government sector. The Hungarian authorities consider that Eximbank is an entity belonging to the financial corporations sector, as it acts on its own risk.

A tenuous borderline exists between **defeasance structures** and financial intermediary under resolution. The relevant question is which body assumes the majority of the risk: the general government that controls the defeasance structure or the financial intermediary under resolution. As defined by the §2.57 of ESA2010 «*the financial intermediation process channels funds between third parties with a surplus and those with a lack of funds. A financial intermediary does not only act as an agent for other institutional units, but places itself at risk by acquiring financial assets and incurring liabilities on its own account.*» However, the features of the entities created under the resolution schemes often raise doubts about their sector classification, since they often have residual activities which can still be considered as financial intermediation. The case: Portuguese Banif - Banco Internacional do Funchal, S.A. is a residual entity that arose from the resolution process of that entity. It is an institutional unit that still produces on a limited scale financial intermediation services and with a low probability of receiving additional government support. However, Banif was reclassified from deposit-taking corporations except the central bank sector to the general government sector, considering Banif as a defeasance structure that holds problematic assets.

In the case of **protection funds** it is relevant to define the degree of control by the State and the autonomy of decision. As stated in §13 of Part I.5 of MGDD, if the protection fund has a full autonomy of decision, it should be classified as a financial auxiliary, otherwise the unit should be classified in the general government sector. However, this rule does not always have a straightforward application. The case: The Hellenic deposit and Investment Guarantee Fund (TEKE) comprises three different activities: a resolution fund, a deposit guarantee fund and an investment fund. According to Eurostat, TEKE fulfils the criteria to be classified in the general government sector since its activity is controlled by the Greek government and it is financed by compulsory payments. The Bank of Greece considers that: TEKE is not controlled by the Greek government; it has autonomy of decision; and it is a market producer. Hence, it should be included in the financial sector. A CMFB consultation pointed to the classification of TEKE in general government sector.

To conclude, we argue that it is highly relevant to achieve further harmonization across countries, statistical domains and systems regarding the statistical classification of units, ensuring the clarification of rules, a correct and unique sector classification of units, especially of entities that are involved in the type of activities presented above.

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.

ST I – Banco de Portugal – 6ª Feira, 21 de Abril, Auditório B (12h40)

Looking into R&D intensity in Portugal in the last years

Carla Ferreira[1], Cloé Magalhães[2], Luís Pinto[3], Mário Lourenço[4]

[1] Banco de Portugal, ciferreira@bportugal.pt;

[2] Banco de Portugal, clmagalhaes@bportugal.pt;

[3] Banco de Portugal, lmpinto@bportugal.pt;

[4] Banco de Portugal, mfllourenco@bportugal.pt;

Abstract: This paper presents some results regarding the economic and financial behavior of manufacturing companies based on their technological intensity, using OECD's classification of these activities. Results show that profitability levels are higher among high Research and Development (R&D) intensity sectors of activity, which present a greater link with the export sector. These sectors' performance regarding bank loans and foreign direct investment on these activities are also discussed.

Key words: Corporations, Indebtedness, Investment, Profitability, R&D.

Evaluating industries' technological intensity has proven to be a challenging task for those who investigate this feature. The OECD has historically proposed a taxonomy where manufacturing activities are labelled according to different technological levels. Galindo-Rueda *et. al.* (2016) discussed this taxonomy, proposing the use of the expression "R&D intensity" instead of "technological intensity". The authors suggest four levels of R&D intensity (high, medium-high, medium-low and low R&D intensity) based on the International Standard Industrial Classification (ISIC) and provide its correspondence with NACE Rev. 2 codes. This classification was used in this study.

Banco de Portugal's Central Balance Sheet Data (CBSD) show that low and medium-low R&D intensity sectors prevail in the Portuguese manufacturing sector concerning a large number of indicators. In 2015, low R&D intensity sectors gathered 59% of the companies, 44% of the turnover and 57% of the number of employees of manufacturing companies as a whole. High and medium-high R&D intensity sectors, which stood for only 10% of the companies, represented around 26% of manufacturing sector's turnover and 17% of its number of employees.

Foreign direct investment (FDI) positions in Portuguese manufacturing represented 8% of total inward equity direct investment in 2016; it was mainly linked to low and medium-low R&D intensity sectors, which had 57% of that investment. Nevertheless, FDI in equity of medium-high R&D intensity sectors was also significant (39%).

Data from Banco de Portugal's Central Credit Register (CCR) reveal that loans granted to the manufacturing sector by the resident financial sector at the end of 2016 were also mostly directed to low R&D intensity sectors (52%), followed by medium-low (29%), medium-high (16%) and high (3%) R&D intensity sectors.

Results also point to differences between levels of R&D intensity through the 2011-15 period, with high and medium-high R&D intensity sectors performing better in several economic and financial indicators. In 2015, the average turnover of a company in the high R&D intensity sectors was 3.6 times higher than the standard for manufacturing companies, while the low R&D intensity sector's average turnover stood for only 75% of the average turnover of manufacturing companies in Portugal.

Export companies were more relevant among high R&D intensity sectors, standing for 27% of these sectors' enterprises and 88% of its turnover.

Profitability, measured according to the return on equity ratio, was generally higher among high and medium-high R&D intensity sectors (13% and 15%, respectively in 2015), although it is worth mentioning profitability developments within low R&D intensity sectors in recent years (7 p.p. increase from 2011 to 2015).

The highest non-performing loans ratios, according to CCR data, were registered among medium-low and low R&D intensity sectors (14.9% and 11.1%, respectively, at the end of 2016) while high and medium-high R&D intensity sectors exhibited significantly lower non-performing loans ratios (1.7% and 6.3%, respectively, at the same period).

In spite of the better economic and financial performance of high and medium-high R&D intensity sectors, it is worth mentioning that the 1% increase in the manufacturing sector's turnover from 2011 to 2015 was sustained by low R&D intensity sectors, while the remaining segments registered turnover decreases. Low R&D intensity sectors were also the major contributors to the 30% increase of manufacturing sector's EBITDA in the same period.

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

Banco de Portugal (2013), *Statistics on non-financial corporations of the Central Balance Sheet Database – Metodological notes*, Supplement to the Statistical Bulletin 2 | 2013.

Banco de Portugal (2015b), *Central de Responsabilidades de Crédito*, Banco de Portugal Booklets, No 5, April 2015 (Portuguese version only).

Banco de Portugal (2015c), *Estatísticas da Balança de Pagamentos e da Posição de Investimento Internacional – notas metodológicas*, Supplement to the Statistical Bulletin 2 | 2015 (Portuguese version only).

Galindo-Rueda, F. and Verger, F. (2016), *OECD Taxonomy of Economic Activities Based on R&D Intensity*, OECD Science, Technology and Industry Working Papers, 2016/04, OECD Publishing, Paris.

ST II – Instituto Nacional de Estatística – 6ª Feira, 21 de Abril, Auditório B (14h15)

É possível substituir os Censos por informação administrativa?

Anabela Delgado[1], Paula Paulino[2], Sandra Lagarto[3]
[1] Instituto Nacional de Estatística, anabela.delgado@ine.pt;
[2] Instituto Nacional de Estatística, paula.paulino@ine.pt;
[3] Instituto Nacional de Estatística, sandra.lagarto@ine.pt

Sumário: Apresenta-se uma síntese dos trabalhos desenvolvidos pelo Instituto Nacional de Estatística, no âmbito da utilização de informação administrativa para a construção de uma Base de População Residente para Portugal: as fontes dos dados, a metodologia utilizada e os resultados obtidos.

A construção da Base de População Residente é o primeiro passo de uma estratégia de médio e longo prazo que visa a mudança para um Censo administrativo, mais eficiente e que permita a divulgação anual de dados de cariz censitário.

Palavras-chave: Censos, Estimação da população, Registos administrativos.

Os organismos internacionais como o Eurostat e as Nações Unidas têm vindo a incentivar a adoção de práticas que permitam melhorar a eficiência das operações censitárias, nomeadamente através da redução dos custos, da diminuição da carga estatística sobre os cidadãos e de uma maior frequência na divulgação de informação de cariz censitário. Neste contexto, e alinhado com as tendências internacionais, o Instituto Nacional de Estatística (INE) tem em curso um estudo de viabilidade que visa a análise do contributo da informação administrativa para fins censitários.

Em Portugal, ao contrário de outros países que já evoluíram para um modelo censitário baseado em registos administrativos, não existe um registo central da população residente, não existe um número único de identificação de cidadão e o País não possui o enquadramento legal necessário para aceder aos nomes e moradas dos indivíduos que constam nos registos administrativos.

A construção de uma Base de População Residente (BPR) a partir de informação administrativa é o primeiro passo de uma estratégia de mudança para um Censo administrativo. A BPR resulta da interligação de dez ficheiros administrativos, disponibilizados por diferentes entidades da administração pública.

O ficheiro de partida para a criação da BPR foi a Base de Dados de Identificação Civil (BDIC) do Instituto de Registos e Notariado. A BDIC apresenta, contudo, fortes limitações para a contagem da população residente em Portugal, na medida em que é uma base de registo civil de cidadãos portugueses com morada legal de contacto em Portugal, o que não é sinónimo de residência efetiva no país, e como tal não está de acordo com o conceito censitário. Por um lado, não inclui os cidadãos de nacionalidade estrangeira residentes em Portugal; por outro lado, sobrestima a população residente no país em cerca de 10% face à população recenseada nos Censos 2011 (mais de 1,1 milhões de indivíduos).

Para construção da BPR foi desenvolvida uma metodologia baseada em indícios de residência, ou 'signs of life', dada pela presença de um indivíduo nos diferentes ficheiros administrativos.

A interligação entre os vários ficheiros para a criação de um registo único relativo ao indivíduo (e identificação da presença nas diferentes fontes administrativas) foi um dos principais obstáculos identificados, e decorre da não existência de um identificador comum aos vários ficheiros administrativos.

A figura 1 ilustra os resultados obtidos no exercício da BPR 2015, enquadrados pelos valores das Estimativas da População e da BDIC para o mesmo ano. A população estimada através da BPR, 10 434 161 indivíduos, apresenta um desvio de 0,9% relativamente às Estimativas da População (cerca de 93 mil indivíduos), com estrutura etária e distribuição por sexo semelhantes.

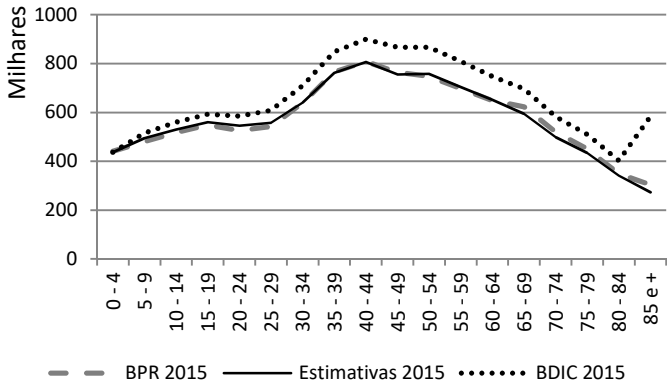


Figura 1: População residente por grupo etário, 2015

Os resultados obtidos são encorajadores, sendo no entanto necessário efetuar mais estudos que possibilitem garantir elevados níveis de qualidade na estimação da população, nomeadamente para níveis geográficos detalhados. A construção da BPR é um trabalho pioneiro no INE que está a dar os primeiros passos. Muito trabalho deverá ser ainda realizado para que a BPR possa ser explorada para fins estatísticos e designadamente para a produção de estatísticas censitárias.

O INE irá continuar o desenvolvimento destes estudos tendo em vista o potencial de utilização que a BPR oferece, designadamente a divulgação anual de estatísticas de cariz censitário.

A modernização dos Censos 2021 passa pela adoção de um novo modelo. Perspetiva-se a realização de um Censo digital, em linha com as expectativas da sociedade e com as oportunidades que o uso de tecnologias oferece.

ST II – Instituto Nacional de Estatística – 6ª Feira, 21 de Abril, Auditório B (14h35)

O uso do *Web Scraping* nas estatísticas oficiais

Maria José Fernandes
Instituto Nacional de Estatística, mjose.fernandes@ine.pt

Sumário: No contexto de modernização dos processos de recolha o INE tem vindo a explorar o potencial da internet como fonte de informação complementar ou alternativa às fontes convencionais.

O objetivo desta apresentação é partilhar e divulgar o trabalho efetuado no desenvolvimento de um modelo de recolha baseado em técnicas de *Web Scraping*. A infraestrutura criada – extração, armazenamento e análise - usa o Python como linguagem base de programação e é totalmente assente na utilização de ferramentas "free" e "open source".

Palavras-chave: Índice de preços no consumidor, Ipython notebook, Mongodb, Python, Web scraping.

A quantidade de informação disponibilizada na internet tem crescido exponencialmente o que configura um desafio quando pensamos na sua recolha. A recolha manual é enfadonha, lenta, falível, cara e impraticável quando dirigida a grandes volumes de informação, como por exemplo a um site de retalho com 30.000 produtos.

Torna-se por isso crucial simular a interação humana com a internet, ou seja, automatizar a web. E é isso mesmo que faz o web scraping, automatiza o processo de recolha através de ferramentas ("scrapers", "spiders", "crawlers", "internet robots", etc.) que navegam e extraem a informação não estruturada de páginas web.

Não obstante as diversas soluções oferecidas pelo mercado para fazer web scraping - quer em forma de interfaces "point and click", de plugins de browsers ou softwares dedicados - que não exigem conhecimentos técnicos avançados, o INE optou por desenvolver o seu próprio modelo. Esta opção implica um elevado nível de conhecimentos técnicos mas proporciona um total controlo do processo. É ainda a abordagem que se considera mais importante porque potencia o uso de novas linguagens/ferramentas estimulando a investigação por parte dos intervenientes e incentivando o debate sobre os processos tecnológicos em uso.

A infraestrutura de captura, armazenamento e análise foi construída tendo como base a linguagem de programação Python: os web scrapers são desenvolvidos em Python, a informação é armazenada numa base de dados não relacional Mongodb e o pré-processamento e análise são efetuados com o Ipython Notebook – todas estas ferramentas são "free" e "open source".

O projeto encontra-se ainda em fase de desenvolvimento mas já com duas aplicações em produção: na recolha de preços e características dos produtos do site do IKEA no

âmbito do Índice de Preços do Consumidor e na atualização da amostra do Inquérito ao Transporte Rodoviário de Mercadorias.

A aplicação deste método permite a automatização de processos, reduzindo a intervenção manual com eventuais erros de leitura e registo, promovendo a melhoria da qualidade dos resultados.

Este projeto de I&D é cofinanciado pelo Eurostat no âmbito da modernização das estatísticas de preços no consumidor.

Referências

Cavallo, R., Rigobon R. (2016) The Billion Prices Project: Using Online Prices for Inflation Measurement and Research. *Journal of Economic Perspectives,* 30(2), 151-78.

McKinney, Wes (2012) *Python for Data Analysis. Data Wrangling with Pandas, Numpy, and Ipython*. O'Reilly Media

Mitchell, Ryan (2015) *Web Scraping with Python. Collecting Data from the Modern Web*. O'Reilly Media

ST II – Instituto Nacional de Estatística – 6ª Feira, 21 de Abril, Auditório B (14h55)

Padrão das disponibilidades alimentares para o período 2012/16

Carlos Carvalho[1], Sofia Duarte[2]
[1]*Instituto Nacional de Estatística, carlos.carvalho@ine.pt;*
[2]*Instituto Nacional de Estatística, sofia.duarte@ine.pt*

Sumário: A Balança Alimentar Portuguesa (BAP) é um instrumento analítico de natureza estatística que disponibiliza um quadro de informação exaustivo relativo ao padrão de abastecimento alimentar de um dado país para um período de referência específico. Serão apresentados os resultados para o período 2012/2016.

Palavras-chave: Calorias, Consumo, Disponibilidades alimentares, Estatísticas, Nutrientes.

A primeira referência a esta operação estatística data de 1948, com a sua aprovação pelo então Ministro da Presidência e atribuição ao INE da sua execução. Foi então publicada em 1951 a "Balança Alimentar do Continente Português" para o período de referência 1948-1949. Em 1977 foi reeditada a "Balança Alimentar do Continente Português" para o período de referência 1963-1975. Este projeto só voltaria a ser retomado em 1994, ano em que se publicou, pela primeira vez, dados a nível nacional: "Balança Alimentar Portuguesa 1980-1992".

O campo de observação da BAP integra todos os produtos da agricultura, pescas e indústria alimentar, cuja principal aptidão seja a alimentação humana, sistematizados na Classificação para efeitos de Balança Alimentar Portuguesa (Produtos da balança alimentar portuguesa, Bebidas da balança alimentar portuguesa); a formulação é a seguinte:

$$R_h = \sum_{h=1}^{k}[Prod_h + I_h - (Pd_h - E_h)],$$

onde

$Prod = Produção\ alimentar\ utilizável$ $\quad$ $I = Importações$

$E = Exportações$ $\quad$ $Pd = Perdas$

$R = Recursos\ alimentares\ disponíveis$

Para o cálculo dos consumos de cada um dos grupos de produtos alimentares e bebidas estabelecem-se equilíbrios entre recursos e empregos a nível tão desagregado quanto a informação disponível o permite.

Do lado da utilização, as disponibilidades são desagregadas entre consumo humano bruto, alimentação animal (consumos de matérias-primas na produção de alimentos para animais), utilização industrial (quantidades de produtos que são utilizados como matérias-primas para a produção de produtos não alimentares, incluindo bebidas) e

sementeira/incubação (quantidades de produto primário do ano n destinado à sementeira da campanha agrícola do ano n+1 e ao número de ovos destinados à incubação no ano n+1); a fórmula geral é dada por

$$C_h = \sum_{h=1}^{k} R_h + Ve_h \ - (Ut_h + A_h + S_h + Inc_h)],$$

onde

$C = Consumo\ Humano\ Bruto$ $\quad$ $Ve = Variação\ de\ Existências$

$Ut = Utilização\ Industrial$ $\quad$ $A = Alimentação\ animal$

$S = Sementeiras$ $\quad$ $Inc = Incubação$

A oferta *per capita* de cada alimento disponível para consumo humano é então obtida dividindo a respetiva quantidade pela população residente em território nacional. A partir da tabela de composição dos alimentos, aplicando os coeficientes de conversão da porção edível para cada produto, obtém-se a Capitação Edível anual que de forma expedita pode ser convertida em capitação edível diária. Uma vez qua a tabela de composição dos alimentos fornece igualmente a composição dos géneros alimentícios em nutrientes e calorias, é possível calcular a capitação edível expressa em unidades de energia e por tipo de nutrientes.

Concetualmente, os balanços alimentares medem o abastecimento alimentar da população. Este resultado fica habitualmente aquém das expectativas dos utilizadores. De facto a BAP mede o consumo de alimentos do ponto de vista da oferta de alimentos, mas não dá qualquer indicação das diferenças que podem existir na dieta consumida pelos diferentes grupos populacionais: grupo etário, perfil socioeconómico, nível geográfico, entre outras. Também não fornecem informações sobre as variações sazonais no fornecimento total de alimentos.

Na ausência destes inquéritos à dieta alimentar, a BAP representa a única fonte harmonizada de informação que permite comparações internacionais ao longo do tempo.

A 7 de abril passado, dia mundial da saúde, o INE divulgou a informação da BAP para o período 2012-2016. Neste estudo é efetuada uma análise das disponibilidades alimentares e a sua relação com a roda dos alimentos, o efeito da crise económico-financeira e a caracterização das disponibilidades expressa em calorias e nutrientes. Salienta-se que pela primeira vez foi efetuada uma análise às disponibilidades de micronutrientes (vitaminas e sais minerais).

ST II – Instituto Nacional de Estatística – 6ª Feira, 21 de Abril, Auditório B (15h15)

Mismatch between jobs and skills in the EU

João Lopes[1], Marco Moura[2], Sónia Quaresma[3]
[1] *Instituto Nacional de Estatística, joao.lopes@ine.pt*
[2] *Instituto Nacional de Estatística, marco.moura@ine.pt*
[3] *Instituto Nacional de Estatística, sonia.quaresma@ine.pt*

Abstract: We analysed the impact of Labour Market Attractiveness on pressing policy questions, namely, the mismatch between jobs and skills at regional level, and labour market mobility. For this purpose, we used clustering techniques, multivariate regression analyses and weighted network correlation analyses. Our approach provides a proof-of-concept of a tool available for common citizens and enterprises seeking for labour market information, and for policy makers aiming to reduce the skills mismatch.

Key words: Labour Market Attractiveness, Labour Market Mobility, Skills Demand, Skills Supply.

The European Big Data Hackathon took place in Brussels from 13 to 15 March 2017, in parallel with the New Techniques and Technologies for Statistics (NTTS) 2017. This event, organised by the European Commission (Eurostat), gathered 22 teams from 21 European countries to compete for the best data product combining official statistics and big data to support policy makers in one pressing policy question Europe is facing. The question, revealed the day prior to the start of the competition, was related to European Commission (EC)'s priority "Jobs Growth and Investment", namely, "How would you support the design of policies for reducing mismatch between jobs and skills at regional level in the EU through the use of data". The data product proposed was required to be supported by relevant data, statistical analysis and visualization. Teams were also invited to use some datasets provided (including European Employment Services (EURES) data from the EC on jobseekers and on job vacancies), and additional publicly available data sources with international applicability (e.g. Eurostat online database).

The development of our data product was focused on three main ideas: 1) combine official statistics data from Eurostat at NUTS2 level (i.e. solid data sets known for being well-structured, clean and accurate, but also characterized by a morose collect and release process) with real-time unstructured "big data"; 2) explore the notion of Labour Market Attractiveness as an important factor in the mismatch between skills demand and supply, as well as in the potential to reduce this skills mismatch, in labour market mobility, and in ordinary emigration; 3) create a flexible, interactive and user-friendly product that allows for customization of the answers in order to be used either by policy makers or by both citizens searching for help on jobseeking and enterprises looking for particular labour market characteristics.

The definition of Labour Market Attractiveness has to be considered carefully, thus, our approach should be seen has a first-step towards a more mature definition. We considered 17 variables from 6 Eurostat data sets, namely "reg_demo" for demographics data, "earn" for earnings structure data, "edtr" for data on education and training, "ilc" for life conditions information, "employ" for employment/unemployment data, and "na10" for national accounts data. These variables were broken by several categorical levels (e.g. "age groups", "level of education", "qualifications", "occupations") originating more than 70 variables. Several data mining techniques were then considered to analyse the data. Using the Labour Market Attractiveness set, we calculated the distance between regions and visualize those using networks, we further clustered the regions using Partition Around Medoids (PAM) method on those distances creating a categorical variable with grouping information. This variable, along with variables on skills mismatch, labour market mobility, and emigration were used separately as dependent variables on multivariate linear and non-linear regression analyses using the Labour Market Attractiveness set as independent variables. We further constructed eigen variables from the considered set and performed Weighted Correlation Network Analysis (WCNA) on the dependent variables. Analyses were performed at country-level and at NUTS1- and NUTS2-level.

In our work we assumed two major simplifications in the construction of the skills mismatch indicator, however, these simplifications do not affect our product in terms of proof-of-concept and can be dropped in later developments. The first one was to use previously cleaned and treated data on job vacancies and education attainment from the Eurostat's "labour" and "edtr" data sets, respectively. Instead, a better approach would be to use the freshly collected EURES data provided, but the use of this data would have two caveats: a) the cleaning of the data requires a considerable expertise on the subject; b) the normalizing of the data, using for example marginal calibration techniques, requires several demographic data at the required regional level in order to successfully capture the populations considered. The second simplification was to use an *ad hoc* mapping between qualifications (classified using ISCED-F 13) and the cross between occupations (defined using ISCO-08) and economic activity (defined using NACE Rev. 2). Nonetheless, a formal mapping will be released in mid-2017 by European Skills, Qualifications and Occupations (ESCO) from the EC.

We used exclusively Microsoft tools to store the data (SQL Server 2008 Express), to load and parse the data and to establish communications with a cloud service (SQL Server Management Studio 2016 and Azure Cloud), and to develop dashboards and reports available in Desktops and mobile devices (Power BI). All statistical analyses were performed in R using libraries "cluster", "glmulti", "Hmisc", "MASS", "nnet", "sna" and "WGCNA".

ST III – Associação Portuguesa de Demografia – Sábado, 22 de Abril, Auditório B (10h30)

Migrações de Substituição e Idade Prospetiva: projetar as necessidades laborais em Portugal (2015-2060)

Daniela Craveiro[1], João Peixoto[2], Isabel T. Oliveira[3], Jorge Malheiros[4], Eduarda M. Costa[5], Diogo Abreu[6], Vítor Escária[7], Paula C. Albuquerque[8], Maria Teresa M. Garcia[9], Cristina S. Gomes[10], Maria João G. Moreira[11], José Alves[12]

[1] SOCIUS/CSG, Instituto Superior de Economia e Gestão da Universidade de Lisboa daniela.craveiro@gmail.com;
[2] SOCIUS/CSG, Instituto Superior de Economia e Gestão da Universidade de Lisboa, jpeixoto@iseg.ulisboa.pt;
[3] CIES, ISCTE — Instituto Universitário de Lisboa, isabel.tiago.oliveira@gmail.com;
[4] IGOT, Universidade de Lisboa, jmalheiros@campus.ul.pt;
[5] IGOT, Universidade de Lisboa, eduarda.costa@campus.ul.pt;
[6] IGOT, Universidade de Lisboa, diogo.abreu@campus.ul.pt;
[7] CIRIUS, Instituto Superior de Economia e Gestão da Universidade de Lisboa, vescaria@iseg.ulisboa.pt;
[8] SOCIUS/CSG, Instituto Superior de Economia e Gestão da Universidade de Lisboa, pcma@iseg.utl.pt;
[9] UECE, Instituto Superior de Economia e Gestão da Universidade de Lisboa, mtgarcia@iseg.ulisboa.pt;
[10] Unidade de Investigação Governança Competitividade e Políticas Públicas, Universidade de Aveiro, mcgomes@ua.pt;
[11] CEPESE, Instituto Politécnico de Castelo Branco, mjgmoreira@ipcb.pt;
[12] UECE, Instituto Superior de Economia e Gestão da Universidade de Lisboa, jose.borges.alves@gmail.com

Sumário: Simplificadamente, as migrações de substituição correspondem ao volume de migrantes necessário para compensar o declínio e o envelhecimento de uma população. Nesta comunicação, introduz-se a noção de idade prospectiva no cálculo de migrações de substituição, ajustando os limites etários da população em idade activa aos ganhos de esperança média de vida. Considerando isto, realiza-se um exercício para Portugal, que consiste na estimativa dos saldos migratórios necessários para estabilizar a dimensão da população activa e o índice de sustentabilidade potencial até 2060.

Palavras-chave: Envelhecimento populacional, Idade prospectiva, Migrações de substituição, População activa.

O envelhecimento populacional descreve uma transição demográfica global e transformadora das nossas sociedades. Os ganhos ao nível da esperança média de vida associados à tendência prologada no decréscimo das taxas de natalidade das sociedades contemporâneas têm resultado em estruturas etárias inéditas, aumentando o peso percentual da população das faixas etárias mais altas. No ano 2000, a Organização das Nações Unidas (ONU, 2000) publicou um estudo de grande impacto projectando o volume de migrações necessário para compensar as transformações populacionais associadas ao

envelhecimento populacional em vários países – *Replacement migration: is it a solution to declining and ageing population?* O relatório descreve como, por exemplo, para assegurar a manutenção da dimensão da população activa igual à de 1995, os países da União Europeia teriam de quase duplicar o saldo migratório anual durante o período até 2050. Por outro lado, seria necessário aumentar o fluxo anual de migrantes em quase 15 vezes para manter o índice de sustentabilidade potencial (rácio entre população em idade activa e população idosa) igual ao registado em 1999. Este exercício tornou evidente a extensão da irreversibilidade das tendências demográficas previstas, atendendo ao grande volume de população imigrante que teria de ser integrada para manter constantes: o volume total da população, o volume da população em idade activa e o rácio entre idosos e activos. A ideia de migrações de substituição foi criticada, entre outras razões, pelo enorme volume de migrações que seria necessário para alcançar os objectivos propostos.

O presente estudo pretende contribuir para o desenvolvimento deste corpo empírico propondo a integração do conceito de idade prospectiva. Esta ideia permite rever os limites etários clássicos usados para definir a população idosa (convencionalmente definida pela população com 65 anos ou mais). De acordo com Sanderson e Sherbov (2005, 2007, 2008), os indicadores referentes ao envelhecimento populacional não devem ter em conta apenas os anos decorridos desde o nascimento ("idade retrospetiva") mas o número de anos que um individuo pode esperar viver (esperança de vida remanescente, ou "idade prospectiva"/"idade remanescente"). Segundo esta perspectiva, o limite etário que define a população idosa é a idade a partir da qual se pode esperar viver 15 anos ou menos.

Nesta comunicação comparam-se os volumes de migrações necessários para responder às questões colocadas pelas Nações Unidas em 2000, tendo como limite definidor da população idosa o limite fixo de 65 anos e um limite dinâmico que corresponde à idade cuja esperança de vida remanescente é de 15 anos. Tomando Portugal como estudo de caso, comparam-se as migrações de substituição necessárias para assegurar a manutenção da dimensão da população em idade activa e o rácio entre a população em idade activa e a idosa (índice de sustentabilidade potencial). Os cálculos baseiam-se nas hipóteses centrais do Instituto Nacional de Estatística (INE) relativamente à fecundidade e à mortalidade até 2060 e foram realizados com base no método de componentes por coortes.

Os resultados mostram que as necessidades de migrantes para compensar a perda da população em idade activa ou mesmo a diminuição dos índices de sustentabilidade potencial diminuem significativamente quando estimadas pelos critérios prospectivos. Para manter o tamanho da população activa, os saldos migratórios teriam de tomar valores médios anuais na ordem dos 75 mil por ano considerandos os limites clássicos; e de 54 mil por ano na versão que considera a idade prospectiva. Para manter constante o índice de sustentabilidade potencial, seriam necessários saldos anuais médios de 589 mil e de 219 mil, de acordo com a perspetiva adoptada. Dado que, desde o início, a ONU considerou inviável a manutenção do índice de sustentabilidade potencial, fará sentido centrar a discussão dos resultados e suas implicações no exercício relativo à manutenção

da população em idade activa. Na verdade, esta estabilizaria, mas tornar-se-ia progressivamente mais diversificada (em termos étnicos e de nacionalidades) e tendencialmente mais envelhecida.

Agradecimentos: A comunicação integra-se no trabalho desenvolvido no âmbito do projecto de investigação mais amplo, financiado pela Fundação Francisco Manuela dos Santos.

Referências

Sanderson, W. & Scherbov, S. (2008). Rethinking age and aging. *Population Bulletin* 63 (4). 3-16.

Sanderson, W. & Scherbov, S (2007). A new perspective on population aging. *Demographic Research*, 16(2), 27-58.

Sanderson, W. & Scherbov. S. (2005). Average remaining lifetimes can increase as human populations age. *Nature* 435(7043), 811-813.

Organização Nações Unidas (ONU) (2000). Replacement Migration: Is It a Solution to Declining and Ageing Populations?, New York: United Nations

ST III – Associação Portuguesa de Demografia – Sábado, 22 de Abril, Auditório B (10h50)

Modelação Estatística da Fecundidade Realizada

Teresa Lopes[1], Maria Filomena Mendes[2], Paulo Infante[3]

[1] CIDEHUS e Mestrado em Modelação Estatística e Análise de Dados, teresa_lopes_1993@hotmail.com;

[2] CIDEHUS e DSOC/ECS, Universidade de Évora, mmendes@uevora.pt;

[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt

Sumário: A fecundidade é um tema bastante abordado nos últimos anos em Portugal e no Mundo devido ao seu decréscimo gradual. Neste trabalho, tomando como base os dados do Inquérito à Fecundidade de 2013, realiza-se um estudo comparativo entre alguns modelos estatísticos possíveis para modelar a fecundidade realizada (número de filhos tidos). Foram considerados os modelos de Poisson, zeros inflacionados (ZIP), Hurdle e Bayesiano. Conclui-se que não só a significância das variáveis é afetada pelo modelo como também o seu efeito na variável resposta.

Palavras-chave: Análise de dados, Fecundidade, Modelação estatística**,** Número de filhos.

O declínio da fecundidade é um problema emergente em Portugal. Nunca se registaram tão poucos nascimentos como atualmente, tendo o nosso país um dos mais baixos níveis de fecundidade da Europa e do mundo (INE & FFMS, 2014).

Os padrões de baixa fecundidade na população portuguesa devem-se aos efeitos conjuntos da diminuição do número de filhos tidos (*quantum*) e do avanço da idade média em que as mulheres têm esses filhos (*tempo*); a tendência de recuperação dos nascimentos adiados (fecundidade tardia) não permitiu compensar as perdas; as mulheres portuguesas continuam a adiar o nascimento do primeiro filho e têm, em média, apenas esse filho. A diminuição da transição para o segundo filho é o que marca a fecundidade em Portugal (Mendes *et al.*, 2016).

Este trabalho tem como principal objetivo modelar estatisticamente a fecundidade realizada (número de filhos tidos) com base nos dados do Inquérito à Fecundidade de 2013, elaborando vários modelos de regressão, procurando não só avaliar a abordagem mais adequada para os dados em estudo, mas também avaliar o impacto que a escolha do modelo tem nas conclusões a que o investigador chega.

Com o decréscimo do número de filhos, instalou-se em Portugal a tendência do filho único, verificando-se que, em 2013, segundo o INE & FFMS (2014), 33% dos casais tinham apenas um filho. Perante isto, decidimos utilizar uma nova abordagem para o modelo ZIP (Poisson de Zeros-Inflacionados) e para o modelo Hurdle, transformando a variável resposta Y=X-1, de forma a podermos analisar o excesso de filhos únicos.

A abordagem estatística Bayesiana considera que toda a informação de que dispomos é útil e deve ser aproveitada, enquanto que a abordagem clássica se restringe apenas às informações obtidas com a observação dos dados amostrais.

Identificámos os fatores potenciadores de um maior número de filhos em cada modelo e observámos que alguns fatores são comuns em todos os modelos e outros fatores aparecem num só modelo. Os fatores comuns a todos os modelos são a idade, o número de indivíduos no agregado familiar, a escolaridade do próprio, opinião sobre se é preferível ter menos filhos com mais oportunidades e menos restrições e a existência de enteados.

O modelo Poisson demonstrou ser o modelo mais parcimonioso de entre os modelos analisados, onde o sexo, a idade, o número de indivíduos no agregado, escolaridade, idade em que deixou o agregado parental de origem, a compensação, a existência de enteados e a existência de primeira conjugalidade são fatores determinantes da fecundidade.

O modelo Bayesiano mostrou ser o menos parcimonioso e mais complexo, uma vez que se obtiveram muitas interações significativas, dificultando muito a interpretação e comunicação dos resultados. Contudo, foi o modelo que apresentou o maior valor do R^2 de Nagelkerke. Por outro lado, este modelo identifica alguns fatores que não se mostraram significativos nos restantes modelos considerados: opinião se é prejudicial para a criança até à idade escolar que a mãe trabalhe fora de casa, realização pessoal e escalão de rendimento.

As variáveis opinião sobre a conciliação materna do trabalho com a vida familiar e adiamento da maternidade apenas se mostraram significativas no modelo Hurdle.

Agradecimentos: Agradecemos à Fundação Francisco Manuel dos Santos que permite a utilização dos dados do Inquérito Português à Fecundidade pela equipa do Estudo "Determinantes da Fecundidade em Portugal" em artigos e comunicações de índole científica.

Referências

Mendes, M., Infante, P., Afonso, A., Maciel, A., Ribeiro, F., Tomé, L.P., & Freitas, R.B. (2016) *Determinantes da Fecundidade em Portugal,* Fundação Francisco Manuel dos Santos.

INE, & FFMS (2014) *Relatório do Inquérito à Fecundidade de 2013,* Lisboa: Instituto Nacional de Estatística.

Cameron, A. C., & Travedi, P. K. (2013). *Regression Analysis of Count Data.* 2nd Edition, Cambridge: Cambridge University Press.

Tang, W., He, H., & Tu, X. M. (2012). *Applied Categorical and Count Data Analysis.* Nova Iorque: CRC Press.

ST III – Associação Portuguesa de Demografia – Sábado, 22 de Abril, Auditório B (11h10)

Uma projeção coerente dos padrões de mobilidade dos estudantes do ensino superior em Portugal

Filipe Ribeiro[1], Lídia Patrícia Tomé[2], Maria Filomena Mendes[3], Rita Freitas[4]

[1] CIDEHUS.UE, Universidade de Évora, fribeiro@uevora.pt;

[2] CIDEHUS.UE, Universidade de Évora, lidiatome@uevora.pt;

[3] CIDEHUS.UE, Universidade de Évora, mmendes@uevora.pt;

[4] CIDEHUS.UE, Universidade de Évora, rfreitas@uevora.pt;

Sumário: A problemática do envelhecimento populacional agravada pelo declínio acentuado da mortalidade e da fecundidade a nível global, acaba por influenciar diretamente as Instituições de Ensino Superior. Perante uma diminuição no número de jovens, aquelas vêem-se obrigadas a uma reorganização de acordo com o número de candidatos que conseguem captar. Neste estudo, recorremos à abordagem composicional do método de Lee-Carter (Oeppen, 2008) para estimar os futuros fluxos de candidatos ao ensino superior. Resultados preliminares estimam que Instituições de Ensino Superior situadas em grandes cidades se revelem ainda mais atrativas no futuro, bem como ocorra uma diminuição no número global de candidaturas.

Palavras-chave: Ensino Superior; Portugal; Projeções Demográficas.

Na área científica das ciências sociais, muito raramente se chega a um consenso no respeitante a uma determinada metodologia a aplicar. No caso específico da demografia, a aplicação de um modelo de projeção de população por coortes e componentes é utilizada sem qualquer contestação a nível mundial (Rowland, 2003; Preston *et al*. 2001).

Com o intuito de estimarmos os fluxos de candidatos ao ensino superior, recorremos ao método de Lee-Carter, amplamente utilizado na estimativa de taxas de mortalidade por idade, período, coorte, expresso numa abordagem composicional (Oeppen, 2008). Esta abordagem resulta na elaboração de estimativas coerentes uma vez que constrange a sua evolução ao longo dos anos a valores viáveis.

Em termos globais, e apesar da manutenção da tendência de envelhecimento populacional, o impacto de uma fecundidade extremamente baixa e de saldos migratórios negativos resultará num declínio da população residente. Claramente, são as idades mais jovens as que serão mais afetadas.

Na análise da evolução populacional identifica-se um aumento entre 1940 e 2011, tendo-se identificado que posteriormente a este período a população apresenta uma continuada tendência de diminuição. Em 2011, ano do mais recente recenseamento geral da população portuguesa, a população residente em Portugal era de 10 562 178 habitantes, sendo expectável que em 2021, ignorando os movimentos migratórios possíveis, esse valor possa vir a ser de 10 064 219 ($IC_{95\%}$: 9 991 349; 10 137 292. Este valor tenderá a diminuir de forma ainda mais acentuada quando considerada na análise a

componente das migrações, podendo esperar-se que venha a diminuir para 10 043 008, apresentando consequentemente um maior intervalo de variação ($IC_{95\%}$: 9 702 069; 10 393 018).

Por outro lado, e se nos focarmos na população de interesse deste estudo, aquela com 18 anos de idade, i. e., identificando como grupo alvo os residentes nas idades em que os jovens se candidatam em maior número ao Ensino Superior, concluímos que, apesar de alguma inconstância ao longo dos anos, esta subpopulação começou a diminuir ainda antes de 2011. Em 1997 registava já um valor perto dos 150 mil habitantes (151 999). Tomando uma vez mais o ano do último recenseamento como referência, verificou-se que, em 2011, a população portuguesa com 18 anos era igual a 114 291, projetando-se para 2021 que venha a atingir 103 547 ($IC_{95\%}$: 103 522; 103 569). Esta última estimativa é o resultado de um exercício de projeção que não inclui a componente das migrações que, por sua vez, virá acentuar a tendência de diminuição. Na verdade, a projeção considerando aquela componente, estima em 106 029, o número de residentes jovens com 18 anos de idade, para 2021 ($IC_{95\%}$: 101 196; 111 008).

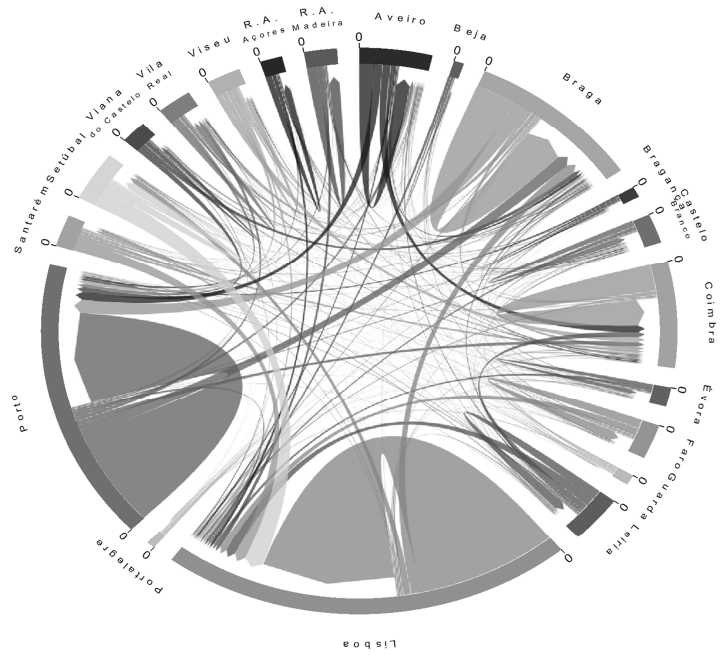


Figura 1: Projeção dos padrões de mobilidade dos estudantes do ensino superior para 2021.

Ao nível das escolhas dos candidatos ao ensino superior, é expectável que em 2021 sejam as Instituições de Ensino Superior localizadas nas grandes cidades, como Lisboa e Porto, que continuarão a ser as mais atrativas (como mostra a figura 1). Contudo, serão seguidas de perto por Braga e Coimbra.

Referências

Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: a test using Japanese cause-of-death data. In *European Population Conference 2008*. European Association for Population Studies, July.

Preston, S. H., Heuveline, P., & Guillot, M. (2001). Demography. Measuring and Modeling Population Processes. Oxford, England: Blackwell Publishing.

Rowland DT (2003). Demographic Methods and Concepts. Oxford University Press, New York.

Sessão BOLSAS CLAD 2017

Sessão BOLSAS CLAD 2017 – 6ª Feira, 21 de Abril, Auditório B (15h40)

Deteção de *Outliers* para Melhoria do Diagnóstico de Perdas em Sistemas de Abastecimento de Água

Maria Silva[1], Conceição Amado[2], Dália Loureiro[3]

[1] Instituto Superior Técnico e Laboratório Nacional de Engenharia Civil, mjsilva@lnec.pt;

[2] CEMAT – Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, conceicao.amado@tecnico.ulisboa.pt;

[3] Laboratório Nacional de Engenharia Civil, dloureiro@lnec.pt.

Sumário: As perdas de água que ocorrem nos sistemas de abastecimento de água constituem um grave problema económico e ambiental, que se reflete, tipicamente, em picos de consumo das séries temporais de caudal. Com o objetivo de estimar com maior fiabilidade as perdas reais, aplicaram-se vários métodos de deteção de *outliers,* resultantes de inovações com base no teste do desvio studentizado extremo generalizado, no método de Tukey e no SAX. Este último revelou-se o melhor para a maioria das séries.

Palavras-chave: Deteção de *outliers,* Sistemas de abastecimento de água, SAX, Séries temporais de caudal.

As perdas de água que ocorrem nos sistemas de abastecimento de água, devidas a fugas em condutas e ramais, podem ser estimadas através da deteção de eventos anómalos, também designados por *outliers,* nas séries temporais de caudal. Neste trabalho foram estudados vários métodos que permitem efetuar a deteção destes eventos.

Iniciou-se o estudo pela aplicação do modelo TBATS (De Livera et al., 2011) a três séries temporais de caudal, construindo-se intervalos de previsão, de confiança e intervalos combinados para a deteção de *outliers*. Contudo, os resultados não foram satisfatórios devido ao elevado número de falsos alarmes produzido, para além do elevado peso computacional deste método.

Com base num conjunto de dados de vinte e oito séries temporais de caudal, e dado que estas apresentavam comportamentos e sazonalidades diversificados, optou-se por começar por realizar uma análise de *clusters* destas séries. Essa análise permitiu estudar o comportamento de cada método de deteção de *outliers* em cada um dos *clusters* obtidos e, desta forma, entender qual o melhor método que se deverá aplicar em cada *cluster*.

Para a deteção de *outliers,* vários procedimentos presentes na literatura foram considerados. Para além destes, foram desenvolvidos neste trabalho outros métodos que surgem ou de modificações dos existentes, ou de abordagens distintas ao problema. Os métodos utilizados foram: o teste do desvio studentizado extremo, o método de Tukey e o SAX (*Symbolic Aggregation approXimation*). Relativamente ao primeiro, este foi aplicado aos resíduos da decomposição da série temporal, onde a tendência foi substituída por um estimador

de localização, tal como descrito em Silva (2016). Neste caso, a mediana e o estimador de localização de Huber foram os utilizados, juntamente com os estimadores de dispersão MAD e *Qn.* O método clássico de Tukey foi modificado para que se pudesse ter em consideração a sazonalidade destas séries, para o qual foi necessário previamente efetuar uma análise de *clusters* das séries temporais mensais. Este agrupamento é também usado no método baseado no SAX (Lin et al., 2007), onde a série temporal é transformada numa sequência de símbolos (designada por alfabeto). Neste caso, com base num dia da semana de meses semelhantes, é construído um padrão para o dia que se está a testar. Assim, são consideradas como *outliers* as observações do dia em teste que não forem representadas pela mesma "letra" que o padrão naquele instante de tempo. Para diminuir o número de falsos alarmes produzido pela classificação de *outlier* em observações com valores muito próximos, mas que eram representadas por letras distintas, é proposto o uso de um parâmetro de afinação, δ, relacionado com uma vizinhança associada a cada ponto.

Na aplicação prática de todos estes métodos, foram utilizados os valores das séries originais, os valores dos seus logaritmos e ainda os valores após a transformação de Box-Cox. Previamente, houve a necessidade de proceder à reconstrução das séries, uma vez que estas apresentavam alguns valores omissos. O procedimento usado encontra-se descrito em Silva (2016).

Em resultado da análise de *clusters* das séries temporais de caudal, obtiveram-se três *clusters*. No primeiro, encontram-se as séries que apresentam maior sazonalidade anual, caracterizada por maiores consumos nos meses quentes. Já no segundo *cluster,* estão as séries sem sazonalidade anual. O último é constituído por três séries que não apresentam um padrão definido e para as quais não foi possível encontrar um bom método de deteção de *outliers.*

Os vários métodos de deteção de *outliers* foram aplicados às séries temporais de cada um dos três *clusters*. No caso dos métodos baseados no teste do desvio studentizado extremo generalizado e no de Tukey, os resultados não foram satisfatórios em nenhum dos *clusters,* devido essencialmente ao elevado número de falsos alarmes. Já o método baseado no SAX, após a padronização das séries e o uso do parâmetro δ, permitiu obter resultados satisfatórios para as séries dos *clusters* 1 e 2. No caso do primeiro *cluster,* os melhores resultados foram obtidos quando é utilizada a transformação de Box-Cox, enquanto, no segundo *cluster,* não foi necessário aplicar nenhuma transformação aos dados.

Referências

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

Lin, J., Keogh, E., Wei, L. & Lonardi, S. (2007) Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery,* 15(2), 107-144.

Silva, M. (2016) Modelação da incerteza e deteção de *outliers* para melhoria do diagnóstico de perdas em sistemas de abastecimento de água. Tese de mestrado, Instituto Superior Técnico, Lisboa.

Sessão BOLSAS CLAD 2017 – 6ª Feira, 21 de Abril, Auditório B (16h00)

Comparação estatística de medidas de qualidade de imagem: da visualização à quantificação

Patrícia Neves[1], Conceição Amado[2], Eunice Carrasquinha[3]
[1] Instituto Superior Técnico, patricia.neves@tecnico.ulisboa.pt
[2] CEMAT - Instituto Superior Técnico, conceicao.amado@tecnico.ulisboa.pt
[3] IDMEC - Instituto Superior Técnico, patricia.neves@tecnico.ulisboa.pt

Sumário: Existem diferentes medidas de qualidade de imagem, sendo assim necessário estudar quais são mais sensíveis e mais eficazes a detetar variações de qualidade nas imagens. Para cumprir esse objetivo, foram usados métodos estatísticos, em particular a Análise de agrupamentos e a Análise de variância. Escolheram-se algumas medidas de qualidade para avaliar imagens de diferentes tipos de representação: a preto e branco e a cores quando sujeitas a ruído e/ou distorção. Foi realizada uma avaliação estatística das medidas de qualidade de imagem de acordo com o seu desempenho.
Palavras-chave: Análise de variância, *Clustering***,** Medidas de qualidade, Processamento de imagem, Qualidade de imagem**.**

A imagem (do latim *imago*, significa representação, forma, imitação, aparência) começou a ser um objeto de estudo na antiguidade grega, apresentando assim diferentes significados e interpretações. A nível conceptual, Platão (428a.C-348a.C) e Aristóteles (384a.C-322a.C) defendiam que a imagem não era mais que uma projeção da mente. Mais tarde, do conceito passou-se à prática, e a imagem começou a exprimir uma ideia, sendo representada em desenhos, gravuras, pinturas ou outras formas de arte. No entanto, foi a partir de 1826, ano em que foi conseguida a primeira fotografia pelo físico francês Joseph Nicéphore Niépce (1765-1833), e, posteriormente, a partir da segunda metade do século XX, com o avanço dos meios técnicos de reprodução de imagem, que esta passou a ter um impacto mais significativo na sociedade. O desenvolvimento da imagem faz parte da história da humanidade, sendo um meio de comunicação com o exterior, rápido e sofisticado. Mais tarde, em 1957, Russel Kirsch foi o responsável pela produção em computador da primeira imagem digital.

Com o avanço tecnológico surgiram novas áreas de investigação, nomeadamente o processamento de imagem com aplicação num vasto conjunto de áreas do conhecimento. Os "dados" de estudo são imagens, tais como fotografias, vídeos, exames médicos, imagens de satélite, entre outros (Trigueirão, 2015, p. 81).

A qualidade de uma imagem digital é afetada a partir do momento em que é capturada até ao momento em que chega ao observador. Em processamento de imagem, a avaliação da qualidade de uma imagem é uma área com um papel bastante importante. Medir a qualidade de uma imagem é um processo difícil, uma vez que a opinião de um observador depende dos seus parâmetros físicos e psicológicos. Portanto, têm surgido

muitas técnicas de avaliação objetiva que contornam as desvantagens de uma avaliação subjetiva.

Realizou-se uma abordagem estatística, baseada no trabalho de Sheikh et al. (2006), de acordo com o desempenho das medidas de qualidade, em alternativa às abordagens que se baseiam unicamente nas características e propriedades das medidas ou que se baseiam numa avaliação subjetiva da qualidade da imagem. Apesar de a ideia do estudo aqui realizado ter por base o artigo referido, as metodologias utilizadas e a forma como o estudo se desenvolveu apresenta moldes diferentes, aumentando-se o rigor matemático do processo, os métodos utilizados e a possível futura aplicabilidade.

Foram escolhidas dezanove medidas de qualidade para avaliar imagens de diferentes tipos de representação: a preto e branco e a cores. Às imagens em estudo foram aplicados dois tipos de degradação: ruído e distorção. A imagem (original) foi comparada com a respetiva imagem com ruído ou distorção. Para o cálculo das medidas, que indicam a diferença das imagens, utilizou-se uma aplicação do *software* MATLAB, *Full-Reference Image Quality Assessment Tool*. Com os valores das medidas calculados, apresenta-se, posteriormente, uma abordagem estatística no âmbito da avaliação das medidas de qualidade de imagem, em particular recorreu-se à Análise de Agrupamentos e à Análise de Variância. Os valores obtidos pelas diferentes medidas foram submetidos a testes estatísticos, com recurso ao software estatístico R (R Core Team, 2015), que permitiram agrupar as medidas consoante o seu desempenho. Posteriormente, os agrupamentos obtidos foram avaliados com o intuito de escolher as medidas mais eficazes na deteção do nível de degradação de uma imagem.

Foi assim possível selecionar, *a priori*, a/as medida/as mais adequada/as para um determinado problema de quantificação de qualidade de imagem. Os resultados vão permitir que em trabalhos futuros o tempo computacional de avaliação de qualidade de uma imagem diminua, da mesma forma que possibilitará que o processo de reconstrução de imagem seja mais eficiente.

Agradecimentos: Um obrigado à minha orientadora Professora Conceição Amado e à minha co-orientadora Eunice Carrasquinha pelas valiosas sugestões e pela ajuda constante ao longo da elaboração deste trabalho.

Referências

Sheikh, H. R., Sabir, M. F. & Bovik, A. C. (2006) A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15.11, 3449-3451.

Trigueirão, E. (2015) Principal component analysis and circulant matrices. Tese de doutoramento. Universidade Técnica de Lisboa – Instituto Superior Técnico.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

SESSÕES PARALELAS

Aplicações em Ciências da Saúde I – 5ª Feira, 20 de Abril, Auditório B (18h00)

Análise cinemática da marcha como ferramenta complementar de decisão na doença de Parkinson Vascular

Flora Ferreira[1,5], Miguel Gago[2,3], Nafiseh Mollaei[3], Lurdes Rodrigues[2], Nuno Sousa[3], Pedro Pereira Rodrigues[4], Estela Bicho[1]

[1] Centro Algoritmi, Escola de Engenharia, Universidade do Minho, Portugal
[2] Serviço de Neurologia, Hospital da Senhora da Oliveira, Portugal
[3] ICVS, Escola de Medicina, Universidade do Minho, Portugal
[4] CINTESIS, Faculdade de Medicina, Universidade do Porto, Portugal
[5] CIICESI, ESTG, Politécnico do Porto, Portugal

Sumário: O objetivo deste estudo é avaliar o papel da análise quantitativa e objetiva da marcha na decisão clínica da dose do fármaco L-Dopa na doença de Parkinson Vascular (DPV). Foram analisados 13 pacientes com DPV em dois momentos: estado "On" (sem L-dopa) e estado "Off" (com L-dopa), sendo quantificados vários parâmetros cinemáticos da marcha através de sensores portáteis colocados nos pés. Esta análise permitiu detetar e quantificar a resposta clínica à terapêutica farmacológica, constituindo-se como uma potencial ferramenta complementar à decisão clínica personalizada na DPV.

Palavras-chave: Doença de Parkinson, Avaliação da marcha, Testes não paramétricos.

Introdução

A análise cinemática da marcha com dispositivos portáteis foi prefigurada nos últimos anos como uma nova ferramenta de diagnóstico complementar na avaliação de doenças neurodegenerativas, com a capacidade, de diagnosticar desvios à normalidade, tais como os parâmetros de variabilidade da marcha e simetria (Paterson et al., 2009). A doença de Parkinson Vascular (DPV) é uma síndrome Parkinsónico raro com critérios clínicos não consensuais (Zijlmans et al., 2015). Esta doença é caracterizada por rigidez e bradicinesia, essencialmente dos membros inferiores, com extrema incapacidade motora manifesta em marcha lenta com instabilidade postural e risco de queda. As opções terapêuticas são limitadas, essencialmente por estimulação do sistema dopaminérgico através do fármaco L-Dopa. Esta opção farmacológica é dúbia, pela raridade de ensaios clínicos, mas também pela escassez de ferramentas de avaliação, que persistem essencialmente qualitativas (escalas clínicas). Apesar da crescente utilização e inovação de sensores na medicina, em particular na avaliação da marcha, a utilização desta tecnologia na DPV ainda persiste numa fase muito preliminar.

Com o objetivo de avaliar o papel da análise objetiva da marcha na decisão da dose de L-Dopa nos DPV observou-se 13 pacientes com DPV, de uma amostra de conveniência de base Hospitalar, com idade média de 80 anos [76,90], em dois momentos diferentes: estado "Off" (24h sem L-dopa) e estado "On" (com 150% da dose diária de L-dopa). Nos dois momentos foi pedido ao paciente para caminhar em terreno plano, 60 metros, com velocidade de marcha livre. Os valores de 24 parâmetros cinemáticos da marcha foram extraídos a partir de sensores colocados nos pés de acordo com as instruções do

fabricante GaitUp® das quais foram analisadas: a *velocidade*, a *variabilidade da marcha*, a *simetria* e o *comprimento da passada* (parâmetros mais sensíveis a risco de queda). Os pacientes também foram avaliados nos dois momentos pelo clínico de acordo com a escala MDS-UPDRS, Parte III. As medianas dos 4 parâmetros da marcha e da pontuação de MDS-UPDRS-III foram comparadas estatisticamente a partir do teste não paramétrico de Wilcoxon, com nível de significância de 5%. Para completar a estatística inferencial calculou-se o tamanho do efeito (r) e a sua interpretação segue as indicações de Cohen (1988).

Resultados

De acordo com os resultados da Tabela 1, pode-se observar diferenças estatisticamente significativas ($p < 0.05$) não só na pontuação da MDS-UPDRS-III como nas variáveis *velocidade* da marcha e *comprimento da passada*, com tamanho do efeito médio ($0.50 \leq |r| \leq 0.79$).

Tabela 1: Resultados da análise comparativa das medianas de parâmetros cinemáticos da marcha e da pontuação MDS-UPDRS-III em dois momentos (estado "Off" e "On").

Parâmetros		Estado "Off"	Estado "On"	P D	DN	*Z*	*p*	*r*
		Me [Min, Max]	Me [Min, Max]					
Velocidade (m/s)		0.62 [0.20,1.05]	0.66 [0.21, 1.04]	10	1	-2.80	0.005	-0.55
Variabilidade da marcha (%)		10.53 [5.52, 20.53]	8.55 [4.87, 24.46]	6	7	-0.03	0.972	-0.01
Simetria (s)		1.05 [1.00, 1.13]	1.05 [1.01, 1.23]	5	8	-0.82	0.413	-0.16
Comprimento da Passada(m)	PE	0.68 [0.19, 1.11]	0.76 [0.21, 1.10]	11	2	-2.70	0.007	-0.53
	PD	0.67 [0.20, 1.11]	0.76 [0.21, 1.11]	12	1	-2.97	0.003	-0.58
MDS-UPDRS-III		44 [22, 70]	40 [17, 67]	13	0	-3.20	0.001	-0.63

Notas: PE = Pé esquerdo, PD = Pé direito, Me= Mediana, Min=Mínimo, Max=Máximo, DP= Número total de diferenças positivas, DN= Número total de diferenças negativas, *Z*= teste de Wilcoxon, *p*= significância (bilateral), *r*= tamanho do efeito ($r = z/\sqrt{26}$, (Field, 2009)). A *variabilidade da marcha* para cada indivíduo corresponde ao coeficiente de variação (CV) do tempo do ciclo da marcha (CV=desvio padrão/média ×100).

Conclusões

Diferentes circuitos neuronais de controlo da marcha (locomoção e estabilidade postural), com envolvimento dopaminérgico ou não dopaminérgico, poderão estar representados em diferentes parâmetros cinemáticos. A intervenção terapêutica com L-Dopa teve um benefício significativo apenas na *velocidade* e *comprimento da passada*, representativos de circuitos dopaminérgicos, mantendo-se a *variabilidade* e a *simetria* da marcha como refratárias. Ao permitir a quantificação da resposta à L-Dopa e identificação de parâmetros refratários, a análise cinemática da marcha surge como uma potencial ferramenta clínica de personalização terapêutica na DPV.

Agradecimentos: Este trabalho foi parcialmente suportado pelos projetos NORTE-01-0145-FEDER-000026 (DeM-Deus Ex Machina) e NORTE-01-0145-FEDER-000016 (NanoSTIMA) financiados pelo Programa Operacional Regional do Norte (NORTE2020), através do Acordo de Parceria PORTUGAL2020 e do Fundo Europeu de Desenvolvimento Regional (FEDER),

Referências

Cohen, J. (1988). Statistical power analysis for the behavioral sciences Lawrence Earlbaum Associates. *Hillsdale, NJ*, 20-26.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Zijlmans, J., Daniel, S. E., Hughes, A. J., Révész, T., & Lees, A. J. (2004). Clinicopathological investigation of vascular parkinsonism, including clinical criteria for diagnosis. *Movement Disorders*, *19*(6), 630-640.

Paterson, K. L., Lythgo, N. D., & Hill, K. D. (2009). Gait variability in younger and older adult women is altered by overground walking protocol. *Age and ageing*, *38*(6), 745-748.

Aplicações em Ciências da Saúde I – 5ª Feira, 20 de Abril, Auditório B (18h20)

Melhorar o diagnóstico da Apneia Obstrutiva do Sono com dados clínicos: uma abordagem Bayesiana

Daniela Ferreira-Santos[1], **Pedro Pereira Rodrigues**[1,2]

[1] *Center for Health Technology and Services Research – CINTESIS, ferreiradossantos.daniela@gmail.com*

[2] *MEDCIDS-FMUP, Faculty of Medicine of the University of Porto, pprodrigues@med.up.pt*

Sumário: A apneia obstrutiva do sono, pela sua prevalência e consequências clínicas, é atualmente considerada um problema de saúde pública. As consequências médicas, sociais e económicas substanciais do não tratamento; o número esmagador de pacientes que escaparam à deteção clínica; e a probabilidade de tratamento bem-sucedido justificam fortemente o rastreio para esta doença. Um modelo de uso diário e sua implementação nos cuidados de saúde primários poderia reduzir o número de polissonografias com resultado normal, otimizando recursos hospitalares e priorizando pacientes.

Palavras-chave: Apneia obstrutiva do sono, Diagnóstico, Fatores de risco, Modelo clínico, Rede bayesiana

Na Apneia Obstrutiva do Sono (AOS), o esforço respiratório é mantido, mas há diminuição ou ausência da ventilação devido á oclusão parcial ou total da via aérea superior. Afeta 4% dos homens e 2% das mulheres na população mundial. A gravidade da AOS é avaliada com o índice apneia-hipopneia (IAH), considerado como o número de apneias e hipopneias por hora de sono. A patologia está presente quando o IAH é maior ou igual a 5, apresentando 3 níveis: ligeiro (IAH: 5 a 14); moderado (IAH: 15 a 29) e severo (IAH: maior ou igual a 30). Os fatores de risco descritos para AOS incluem: obesidade, idade avançada, sexo masculino, roncopatia, aumento da circunferência do pescoço (NC), aumento da circunferência abdominal (AC), hipertensão arterial (HTA), hipertensão pulmonar (HTP), fibrilação auricular (FA), insuficiência cardíaca congestiva (ICC), acidente vascular cerebral (AVC), enfarte do miocárdio (EM), refluxo gastro esofágico (RGE), hipotiroidismo, diabetes, ansiedade, consumo de álcool, histórico de tabagismo, uso de sedativos, condutores de longas distâncias, anomalias craniofaciais e da via área superior (CFA), etnia e histórico familiar. O nosso objetivo foi definir um método auxiliar de diagnóstico que suporta a decisão de realizar polissonografia (PSG), baseado em sinais e sintomas, priorizando pacientes. A nossa amostra é constituída por pacientes com mais de 18 anos com suspeita de AOS referenciados para realizar polissonografia no Laboratório de Sono do Centro Hospitalar de Vila Nova de Gaia/Espinho, entre janeiro e maio de 2015. Foi realizada uma revisão da literatura para definir as variáveis mais relevantes a serem recolhidas, num total de 39, divididas em demográficas, exame físico, história clínica e co-morbilidades. Toda a informação clínica (diários e diagnósticos) foi extraída do SAM e diretamente do Laboratório do Sono. Dois classificadores de redes Bayesianas foram utilizados: Naïve Bayes (NB) e Tree

Augmented Naïve Bayes (TAN). Estes foram usados para construção dos modelos com 39 variáveis ou com variáveis selecionadas (considerando um nível de significância de 5% ou 20% se pelo menos 10 pacientes foram observados). Curvas ROC, sensibilidade, especificidade, valor preditivo positivo e negativo e validação cruzada foram analisadas. O software utilizado foi o R e o SamIam. 241 pacientes foram estudados, mas apenas 194 preenchiam os critérios de inclusão, com 123 (63%) homens e uma idade média de 58 anos. 66 (34%) pacientes apresentavam um resultado normal e 128 (66%) o diagnóstico de AOS. Um total de quatro modelos foram avaliados: NB39, TAN 39, NB13 e TAN13. A sensibilidade e especificidade de cada modelo foi a seguinte: NB39 – 75% e 57%, TAN39 – 82% e 55%, NB13 – 76% e 55% e TAN13 – 81% e 48%, sendo o melhor modelo o TAN39 com uma curva ROC de 79% e uma precisão de 73%. Obtivemos um total de 66 pacientes (34%) com resultado normal, revelando um elevado número de exames dentro da normalidade (IAH <5) que são realizados diariamente. Estes resultados continuam a demonstrar a necessidade de um método diagnóstico de apoio á decisão ao nível dos cuidados de saúde primários, pois só assim poderemos otimizar a necessidade de realizar PSG. É dando conhecimento e novas ferramentas aos médicos de clínica geral que iremos colmatar a desnecessária realização de PSG nos cuidados secundários. Portugal não dispõe de um método validado para a triagem ou encaminhamento de doentes com suspeita de AOS nos cuidados primários passível de implementação, pelo que pensamos que o nosso modelo (TAN13) consiste num método válido. Um modelo de uso diário e sua implementação nos cuidados primários poderia reduzir o número de exames com resultado normal, otimizando os recursos hospitalares disponíveis e priorizando os pacientes.

Agradecimentos: Project "NORTE-01-0145-FEDER-000016" (NanoSTIMA) is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

Referências

Chung, S., Jairam, S., Hussain, M. R. G., & Shapiro, C. M. (2002) How, what, and why of sleep apnea. Perspectives for primary care physicians. *Canadian Family Physician*, 48, 1073–1080.

Lam, J. C. M., Sharma, S. K., & Lam, B. (2010) Obstructive sleep apnoea: Definitions, epidemiology & natural history. *Indian Journal of Medical Research*, 131, 165–170.

Epstein, L. J., Kristo, D., Strollo, P. J., Friedman, N., Malhotra, A., Patil, S. P., Weinstein, M. D. (2009) Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *Journal of Clinical Sleep Medicine*, 5(3), 263–276.

Darwiche, A. (2010) Bayesian Networks. *Communications of the ACM*, 53(12), 80–90.

Aplicações em Ciências da Saúde I – 5ª Feira, 20 de Abril, Auditório B (18h40)

Multimorbidade de pacientes com enfarte agudo do miocárdio descrita usando abstrações temporais e redes bayesianas

Carla Silva[1], Mariana Lobo[2], Pedro Pereira Rodrigues[3]

[1] CINTESIS - Center for Health Technology and Services Research, Centro de Investigação Médica, Faculdade de Medicina da Universidade do Porto, carla.maps@gmail.com;
[2] CINTESIS - Center for Health Technology and Services Research, Centro de Investigação Médica, Faculdade de Medicina da Universidade do Porto, nanalobo@gmail.com;
[3] CINTESIS - Center for Health Technology and Services Research, Centro de Investigação Médica, Faculdade de Medicina da Universidade do Porto e MEDCIDS - Departamento de Medicina da Comunidade, Informação e Decisão em Saúde, Faculdade de Medicina da Universidade do Porto, pprodrigues@med.up.pt

Sumário: A pesquisa tem como objetivo explicar a complexidade da multimorbidade (dois ou mais problemas de saúde) em pacientes com enfarte agudo do miocárdio através de abstrações de redes bayesianas baseadas no tempo. Foram conduzidas análises univariada e multivariada através de redes bayesianas dinâmicas e temporais, considerando a monitorização dos pacientes durante um período de tempo (5 follow-up). A estrutura de uma rede pretende resumir graficamente as metáforas bayesianas para a multimorbidade.

Palavras-chave: Abstrações temporais, Enfarte agudo do miocárdio, Multimorbidade, Redes bayesianas dinâmicas, Redes bayesianas temporais

A modelação bayesiana aplicada à multimorbidade decorre da necessidade de extrair conhecimento avançado sobre a modelação de doenças, podendo assim acelerar o desenvolvimento de estratégias profícuas identificando fatores de multimorbidade. Dado que neste tipo de problema coexistem fatores de risco dependentes do tempo e uma vez que estes pacientes são susceptíveis de experimentar múltiplas comorbidades e desenvolver novas condições, é necessário um desenho de estudo com componente analítica temporal. Realizamos uma análise retrospectiva utilizando dados obtidos a partir de registos de saúde electrónicos (RSE). A análise teve como base uma amostra aleatória de 500 pacientes admitidos com enfarte agudo do miocárdio (EAM) em hospitais portugueses, onde foram observados 28 diferentes tipos de problemas de saúde, registados do ano 2011 a 2015. Começamos com uma análise univariada, foram calculadas as frequências para as variáveis categóricas. Realizamos o teste não paramétrico de Kolmogorov-Smirnov para observar se as variáveis contínuas, dias de internamento e idade nos anos de admissão, seguiam uma distribuição Normal. Em seguida, conduzimos uma análise onde usamos o teste exato de Fisher (com a correção de Bonferroni) para analisar as proporções e procuramos encontrar associações fazendo combinações *pairwise* na forma de comparações múltiplas (Figura 1). Por fim, realizamos uma análise multivariada, através de redes bayesianas, sendo que, uma rede bayesiana representa uma distribuição conjunta de um conjunto de variáveis, estabelecendo a

suposição de independência entre todas, representando a interdependência por um grafo acíclico direcionado.

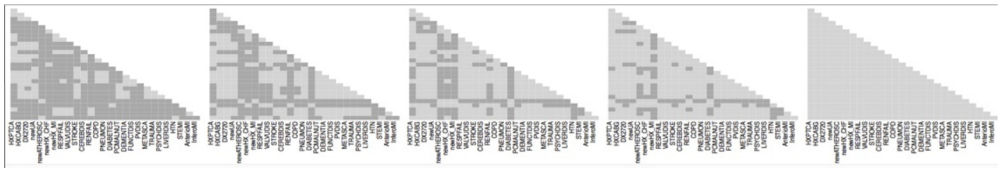


Figura 1: *5 Follow-up.* Associações de comorbidades evoluindo no tempo por eventos {0-1}, {1-2}, {2- 3}, {3-4} e {4-5}. Teste exato de Fisher (com a correção de Bonferroni), cor escura significa associações estatisticamente significativas.

Aprendemos a estrutura da rede bayesiana em cada abstração baseada em tempo usando uma pesquisa *tabu*, um algoritmo *hill-climbing* modificado, utilizando o critério de informação bayesiano.

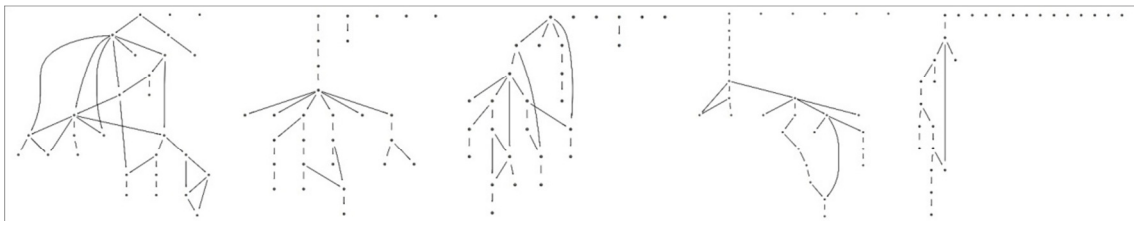


Figura 2: *5 Follow-up.* Esboço Bayesiano representando as abstrações das redes de multimorbidade evoluindo no tempo por eventos {0-1}, {1-2}, {2- 3}, {3-4} e {4-5}.

O problema de pesquisa é assim dividido em duas direções analíticas baseadas em tempo: redes bayesianas dinâmicas (modelação por eventos, Figura 2) e redes bayesianas temporais (modelação por intervalos), de modo que cada representação de tempo fosse estendida ao longo de um intervalo temporal em vez de um único ponto de tempo como em redes discretas. A análise preliminar mostrou que houve diminuição das associações estatisticamente significativas entre comorbidades no tipo de abstração baseada em eventos e um número constante de associações na abstração baseada em intervalos. Portanto, observamos diferentes comportamentos na avaliação da multimorbidade, ao assumir diferentes abstrações temporais, o que pode levar a caminhos de pesquisa mais precisos na área.

Agradecimentos: Projeto "NORTE-01-0145-FEDER-000016" (NanoSTIMA) é financiado pelo Programa Operacional Regional do Norte (NORTE2020), através do Acordo de Parceria PORTUGAL2020 e do Fundo Europeu de Desenvolvimento Regional (FEDER).

Referências

Lappenschaar, M., Hommersom, A., Lucas, P.J., Lagro, J., Visscher, S., Korevaar, J.C. & Schellevis, F.G. (2013) Multilevel temporal bayesian networks can model longitudinal change in multimorbidity. *Journal of Clinical Epidemiology,* 66, 1405-1416.

Orphanou, K., Stassopoulou, A. & Keravnou, E. (2016) Dbn-extended: A dynamic bayesian network model extended with temporal abstractions for coronary heart disease prognosis. *IEEE Journal of Biomedical and Health Informatics,* 20(3), 944-952.

Orphanou, K., Stassopoulou, A. & Keravnou, E. (2014) Temporal abstraction and temporal bayesian networks in clinical domains: a survey. *Artificial intelligence in medicine,* 60(3), 133-49.

Classificação I – 5ª Feira, 20 de Abril, Auditório A (18h00)

Regression based Approaches for Predicting Age at Onset

Maria Pedroto[1], Alípio M. Jorge[2], João Mendes-Moreira[3], Teresa Coelho[4]
[1] *INESC TEC, Faculdade de Engenharia, Universidade do Porto,* maria.j.pedroto@inesctec.pt*;*
[2] *INESC TEC, Faculdade de Ciências, Universidade do Porto,* amjorge@fc.up.pt*;*
[3] *INESC TEC, Faculdade de Engenharia, Universidade do Porto,* jmoreira@fe.up.pt*;*
[4] *Unidade Corino de Andrade, Centro Hospitalar do Porto,* tcoelho@netcabo.pt*.*

Summary: This work describes the usage of Classical Regression based approaches to predict the Age at Onset of Patients diagnosed with TTR-FAP, a genetic disease first diagnosed in Portugal. In our case, we used traditional Machine Learning algorithms, namely: Decision Trees, Support Vector Machines, and Random Forests. Our results showed that Random Forests achieved the best results with an average mean absolute error of 3.1 years.

Keywords: Feature Engineering, Genetic Diseases, Supervised Learning

Introduction

Transthyretin Familial Amyloid Polyneuropathy (TTR-FAP) is a genetic disease that was first diagnosed in Porto, in 1952, by a Portuguese neurologist named Mário Corino de Andrade. This is a severe, progressive, and life-threatening disease characterized by sensory, motor and autonomic neuropathies (Parman et all). Age at Onset is a medical term referring to the age on which a patient first acquires, develops, or experiments the first symptoms of a disease. The age at onset of patients with TTR-FAP has historically taken values between 19 and 82 years old. After the first symptoms, a patient with TTR-FAP has, on average, 10 years' survival rate. Nowadays, TTR-FAP medical approaches are only able to delay its progression. These rely on organ transplant (liver, heart, or kidney transplants) and medical prescription treatments (Tafamidis). A factor that can greatly impact the evolution of the disease is related with a "as soon as possible" effective diagnosis of the disease.

Methodology

For our work, we considered three levels of information related with the patients, namely: i) patient related information; ii) patient's close family related information; iii) and full length pedigree information. For each of these classes we took into consideration specific features related with the patient and its familiar structure.

In Figure 1 we have the representation of the prediction pipeline defined for this work. Its main characteristics are related with the usage of a 10-fold Cross Validation method to validate the accuracy of the models and, in the Support Vector Machines (SVM) experiments, the preprocessing of the dataset to scale the final features to a common order of greatness. All the implementation was done in Python version 3, with "scikit-learn" and "Jupyter Notebooks" as main developer tools and prediction APIs.

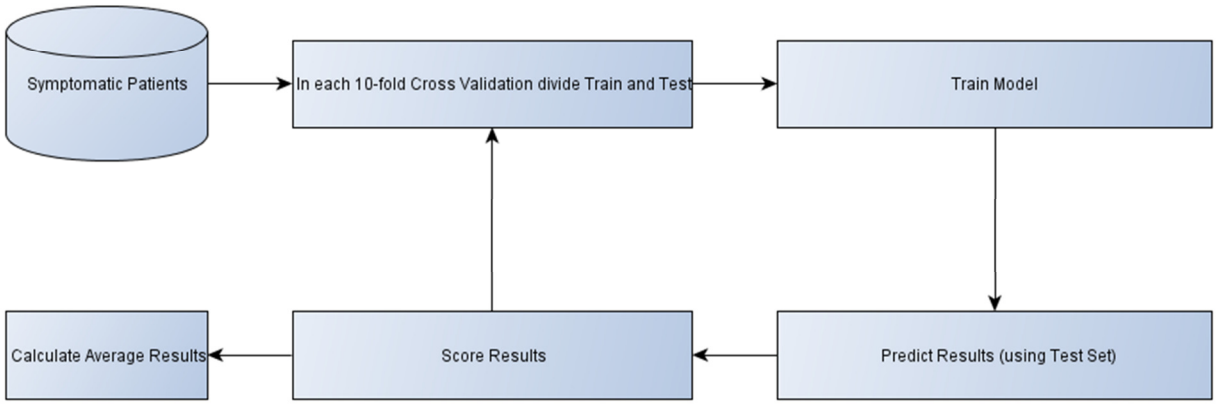


Figure 1: Process Modeling

Results and Conclusions

We used three different traditional algorithms: Decision Trees, SVM and Random Forests. Our results can be observed in Table 1, where it is clear the better performance of Random Forests with an average mean absolute error of 3.1 years (approximately). Further work needs to be performed, namely related with feature engineering and parameter tuning tasks. Domain experts clinically working with patients with TTR-FAP consider these interesting baseline results.

Table 1: Evaluation results for Regression Approaches for Predicting Age of Onset

Algorithm	Metric	Regression
Decision Tree	MAE	3.5 (Avg) ± 0.26 (Std)
	MSE	26.38 (Avg) ± 7.29 (Std)
	RMSE	5.09 (Avg) ± 0.71 (Std)
Random Forest	MAE	3.07 (Avg) ± 0.28 (Std)
	MSE	20.55 (Avg) ± 7.52 (Std)
	RMSE	4.48 (Avg) ± 0.75 (Std)
SVM	MAE	3.88 (Avg) ± 0.35 (Std)
	MSE	27.19 (Avg) ± 9.46 (Std)
	RMSE	5.09 (Avg) ± 1.18 (Std)

MAE – Mean Absolute Error, **MSE** – Mean Square Error, **RMSE** – Root Mean Square Error.

Acknowledgements: We would like to thank Centro Hospitalar do Porto EPE for supporting our work and all the staff of Unidade Corino de Andrade (UCA) at Hospital de Santo António do Porto.

References

Parman, Y., Adams, D., Obici, L. (2016) *Sixty years of transthyretin familial amyloid polyneuropathy (TTR-FAP) in Europe: where are we now? A European network approach to defining the epidemiology and management patterns for TTR-FAP,* Current Opinion in Neurology 29 (Suppl 1).

Classificação I – 5ª Feira, 20 de Abril, Auditório A (18h20)

A likelihood-ratio control chart for monitoring the parameters of a logistic process

Fernanda Otília Figueiredo[1], Maria Ivette Gomes[2], Amitava Mukherjee[3]
[1] FEP, University of Porto and CEAUL, otilia@fep.up.pt;
[2] FCUL, DEIO and CEAUL, University of Lisbon, ivette.gomes@fc.ul.pt;
[3] Xavier School of Management, India, amitmukh2@gmail.com.

Abstract: A likelihood-ratio control chart for monitoring the parameters of a logistic distribution is developed and its performance is analyzed. A practical application concerning the monitoring of the diameters and lengths of stopper-corks is considered. We illustrate the adequacy of the logistic distribution to model such type of data and the performance of the developed control chart.

Key words: Average Run-Length**,** Logistic Distribution, Stopper-Cork.

This work focuses on the development of a likelihood-ratio control chart for the logistic distribution, here called L–chart, and evaluation of its performance to monitor a cork stopper's process production. The logistic model, LD(α,β), with probability density function (pdf) given by

$$f(x;\propto,\beta)=\frac{1}{\beta}e^{-\frac{x-\propto}{\beta}}\left(1+e^{-\frac{x-\propto}{\beta}}\right)^{-2}, \qquad -\infty<x<+\infty,$$

and the cumulative distribution function (cdf) given by

$$F(x;\propto,\beta)=\left(1+e^{-\frac{x-\propto}{\beta}}\right)^{-1}, \qquad -\infty<x<+\infty,$$

is similar to the normal in shape, but with slightly heavier/longer tails (higher kurtosis) than the normal, as it also happens with the student–t distribution. This model can be appealing for use in applications whenever there is a potential deviation from the normal assumption, with the advantage of having simpler expressions than the student-t for both the pdf and cdf. For details on this distribution see, for instance, Johnson *et al.* (1995).

In the process production to be monitored, the corks must have some specific physical characteristics in terms of length and diameter. First, from an available sample of corks' lengths and diameters, collected from a stable in-control process, we have analyzed the adequacy of the logistic distribution to model such type of data. Then, we have determined the process nominal values, α_0 and β_0.

Let us consider the L–chart to monitor the corks diameter of the process production, and let us assume that the diameters follow a LD(α, β). When the process is in-control (IN) state, $\alpha = \alpha_0 = 24$ and $\beta = \beta_0 = 0.0724$ (these nominal values will be considered as known). Otherwise, i.e., if $\alpha \neq \alpha_0$ and/or $\beta \neq \beta_0$, the process is out-of control (OOC). The control statistic of the L–chart is defined by

$$\Lambda = \left(\frac{\hat{\beta}}{\beta_0}\right)^{-n} exp\left(-x_i\left(\frac{1}{\hat{\beta}} - \frac{1}{\beta_0}\right) + \frac{\hat{\alpha}}{\hat{\beta}} - \frac{\alpha_0}{\beta_0}\right)\prod_{i=1}^{n}\left(\frac{1 + exp\left(-\frac{x - \hat{\alpha}}{\hat{\beta}}\right)}{1 + exp\left(-\frac{x - \alpha_0}{\beta_0}\right)}\right)^{-2},$$

with $(\hat{\alpha}, \hat{\beta})$ denoting the maximum likelihood estimators of (α, β), and (α_0, β_0) the process nominal values. The upper control limit (UCL) of the chart is a quantile of the estimated distribution of the control statistic Λ, determined in order to have a fixed in-control average run-length, ARL_0 (IC ARL). The L–chart triggers a signal whenever Λ > UCL, and the process is then considered in an OOC state.

Note that the RL (run-length) has a geometric distribution with parameter $p(\alpha, \beta) = P(\Lambda > UCL \mid \alpha, \beta)$. Consequently, the $ARL_0 = 1/p(\alpha_0, \beta_0)$ and the control limit UCL is the quantile of probability $1 - 1/ARL_0$ of the estimated distribution of Λ.

In the case study, we have implemented the L–chart in order to have an IC ARL=500. We have computed the OOC ARL of the chart by Monte Carlo simulations, using 100000 runs, for samples of size 5, 10 and 15, and different magnitudes of shifts in one or both parameters of the distribution. More precisely, when the diameters follow:

- $LD(\alpha_0 + \Delta\alpha, \beta_0)$, for $\Delta\alpha > 0$. In this case $\mu = 24 + \Delta\alpha$ and $\sigma = 0.1313 = \sigma_0$;
- $LD(\alpha_0, \beta_0 \Delta\beta)$, for $\Delta\beta \neq 0$. In this case $\mu = 24 = \mu_0$ and $\sigma = 0.1313\ \Delta\beta$;
- $LD(\alpha_0 + \Delta\alpha, \beta_0 \Delta\beta)$, for $\Delta\alpha > 0$ and $\Delta\beta \neq 0$. In this case we have changes in the process mean and standard deviation, i.e., $\mu = 24 + \Delta\alpha$ and $\sigma = 0.1313\ \Delta\beta$.

If we consider $\Delta\alpha < 0$ instead of $\Delta\alpha > 0$ the results are similar.

From this study we can draw the following conclusions:

- To detect shifts in the location of the logistic distribution, i.e., in the parameter α, or equivalently in μ, the L–chart is an ARL-unbiased chart with an interesting performance, even to detect small shifts. If we increase the sample size, the efficiency of the chart increases substantially, as expected.
- To detect shifts in the scale of the logistic distribution, i.e., in the parameter β or equivalently, in σ, the L–chart remains an ARL-unbiased chart. The chart is more efficient to detect increases in β (or equivalently, in σ) than decreases, but in both cases it has a good performance. As before, the efficiency of the L–chart to detect shifts in β, or in σ, increases substantially with n.
- To detect shifts in both of the logistic parameters, or equivalently, simultaneously in μ and σ, the L–chart continues to exhibit a very good performance, being very fast to trigger a signal even in the case of shifts of very small magnitude. Again, increasing the value of n enables to get a control chart with a better performance.

Acknowledgements: Research partially supported by National Funds through FCT- Fundação para a Ciência e a Tecnologia, project UID/MAT/00006/2013 (CEA/UL). We also thank the referee for the comments.

References

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995) *Continuous Univariate Distributions*, 2nd edition, John Wiley & Sons, Inc. New York.

Classificação I – 5ª Feira, 20 de Abril, Auditório A (18h40)

O modelo logístico multinomial na saúde ocupacional

Luís M. Grilo[1], Helena L. Grilo[2]

[1] Instituto Politécnico de Tomar, Unidade Departamental de Matemática e Física, e Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Portugal, lgrilo@ipt.pt;
[2] Instituto Politécnico de Tomar, Centro de Sondagens e Estudos Estatísticos, Portugal, helenagrilo56@gmail.com.

Sumário: De acordo com a Agência Europeia para a Segurança e a Saúde no Trabalho os riscos psicossociais e o *stress* relacionado com este estão entre as questões mais desafiadoras da segurança e saúde ocupacional. Tendo por base um questionário com variáveis em escala tipo Likert com 5 classes, foi obtida uma amostra representativa de trabalhadores de um setor de serviços português. A regressão multinomial foi usada para estimar a probabilidade de cada uma das classes da variável dependente "qualidade geral da saúde", onde entre outras aparecem como estatisticamente significativas as variáveis independentes "*stress"* e "*burnout"*.

Palavras-chave: Regressão**,** Riscos psicossociais, Variáveis categóricas.

A percentagem de trabalhadores na União Europeia exposta a fatores de *stress* psicossocial no local de trabalho tem vindo a crescer em muitos setores de atividade, com consequências desastrosas para funcionários, empresas e sociedade em geral (Bakker *et al*., 2014, Kristensen *et al*., 2005, Neto *et al*., 2014). O Copenhagen Psychosocial Questionnaire (conhecido por COPSOQ) é utilizado internacionalmente para avaliar riscos psicossociais e o seu impacto na saúde e bem-estar dos trabalhadores, sendo que inclui a maioria das dimensões relevantes de acordo com várias teorias sobre fatores psicossociais. O COPSOQ na versão curta foi aplicado em 2015 a trabalhadores de um importante sector de serviços em Portugal (n = 5020). Recorremos à regressão logística multinomial para estimar a probabilidade de cada uma das 5 classes da variável dependente "qualidade da saúde geral" (codificadas do modo seguinte: "1 - deficiente", "2 - razoável", "3 - boa", "4 - muito boa" e "5 - excelente") em função de 6 variáveis independentes que representam dimensões psicossociais com evidência epidemiológica para a saúde. Foram consideradas como variáveis independentes quantitativas o "*stress*", o "*burnout*", os "conflitos trabalho-família" e os "comportamentos ofensivos", e, ainda, as variáveis qualitativas "problemas de sono" e "sintomas depressivos" (ambas com as classes: "1 - nada/quase nada", "2 - pouco", "3 - moderadamente", "4 - muito" e "5 - extremamente").

Método

Seja Y a variável dependente com K = 5 classes possíveis, o modelo multinomial é constituído por um sistema de 5 modelos logísticos que, depois de normalizado relativamente a uma classe de referência de Y (neste estudo de caso consideramos a primeira classe: "1 - deficiente"), permite calcular a probabilidade de Y tomar o valor de

qualquer uma das K = 5 classes. Na notação matricial, com **X** matriz de variáveis independentes e **β** o vetor dos coeficientes, temos (Hosmer & Lemeshow, 2000):

$$P(Y=1\mid\mathbf{X})=\frac{1}{1+\sum_{k=2}^{5}e^{\mathbf{X}\boldsymbol{\beta}_k}},\quad P(Y=2\mid\mathbf{X})=\frac{e^{\mathbf{X}\boldsymbol{\beta}_2}}{1+\sum_{k=2}^{5}e^{\mathbf{X}\boldsymbol{\beta}_k}},\ldots,\quad P(Y=5\mid\mathbf{X})=\frac{e^{\mathbf{X}\boldsymbol{\beta}_5}}{1+\sum_{k=2}^{5}e^{\mathbf{X}\boldsymbol{\beta}_k}}.$$

As $K-1$ = 4 chances de ocorrer uma das classes de *Y*, relativamente à classe de referência, são dadas por,

$$\frac{P(Y=2\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=e^{\mathbf{X}\boldsymbol{\beta}_2},\quad \frac{P(Y=3\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=e^{\mathbf{X}\boldsymbol{\beta}_3},\quad \frac{P(Y=4\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=e^{\mathbf{X}\boldsymbol{\beta}_4},\quad \frac{P(Y=5\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=e^{\mathbf{X}\boldsymbol{\beta}_5}$$

e, logaritmizando os dois membros da igualdade obtemos as 4 chances, relativas à classe de referência de *Y*, que constituem o modelo multinomial,

$$\ln\frac{P(Y=2\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=\mathbf{X}\boldsymbol{\beta}_2,\quad \ln\frac{P(Y=3\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=\mathbf{X}\boldsymbol{\beta}_3,\quad \ln\frac{P(Y=4\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=\mathbf{X}\boldsymbol{\beta}_4,\quad \ln\frac{P(Y=5\mid\mathbf{X})}{P(Y=1\mid\mathbf{X})}=\mathbf{X}\boldsymbol{\beta}_5.$$

O modelo é ajustado iterativamente com o método de máxima verosimilhança.

Resultados

O modelo ajustado é estatisticamente significativo, dado que o resultado da estatística de qui-quadrado, obtido da diferença entre do teste do rácio de verosimilhanças para o modelo nulo (só com a constante) e para o modelo final completo (com todas as variáveis independentes), é $\chi^2_{(48)}$ = 1421,94 com *p-value* < 0,001. Concluímos, assim, que existe pelo menos uma variável independente que influência significativamente a "qualidade da saúde geral" dos trabalhadores. Em consonância com esta conclusão estão os valores AIC e BIC, que são inferiores no caso do modelo final (AIC final = 9185,611; BIC final = 9524,349).

Os resultados dos testes do rácio de verosimilhanças para cada uma das variáveis independentes indicam que a variável "comportamento ofensivo" não é estatisticamente significativa (*p-value* = 0,307 > 0,05), mas as variáveis "*burnout*", "*stress*", "problemas de sono", "sintomas depressivos" e "conflitos trabalho-família" tem um efeito estatisticamente significativo (todos os *p-values* são inferiores a 1%) sob o *logit* da probabilidade de ter uma "qualidade da saúde geral" pelo menos razoável, relativamente à classe de referência.

Referências

Bakker, A., Demerouti, E. & Sanz-Vergel, A. I. (2014) Burnout and Work Engagement: The JD–R Approach, *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1), 140114155134003.

Hosmer, D. W. & Lemeshow, S. (2000) *Applied Logistic Regression* (*2nd Edition*), John Wiley & Sons, New York.

Kristensen, T. S., Borritz, M., Villadsen, E. & Christensen, K. B. (2005) The Copenhagen Burnout Inventory: A new tool for the assessment of burnout, *Work & Stress* 19, 192–207.

Neto, H. V., Areosa, J. & Arezes, P. (2014) *Manual sobre Riscos Psicossociais no Trabalho*, Civeri Publishing (coleção RICOT).

Modelos Estocásticos e Modelos Espaciais – 6ª Feira, 21 de Abril, Auditório A (10h00)

O efeito do delineamento de amostragem em modelos de análise de sobrevivência

Ana Laura Carreiras[1], Paulo Infante[2], Anabela Afonso[3]

[1] MMEAD - Universidade de Évora, analaura.carreiras@gmail.com

[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt

[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, aafonso@uevora.pt

Sumário: Quando se usam dados provenientes de amostras complexas usualmente são disponibilizadas variáveis de pesos e outras informações adicionais do delineamento de amostragem. Neste trabalho apresentamos as principais diferenças entre vários tipos de pesos, avaliando-se, em particular, a importância de considerar os pesos normalizados corrigidos pelo efeito do delineamento (*deff*) e o impacto que uma incorreta estimação do *deff* pode ter, na significância das variáveis, em modelos de análise de sobrevivência.

Palavras-chave: Amostras complexas, Pesos, Efeito do delineamento, Análise de Sobrevivência.

As amostras complexas resultam da combinação de vários métodos de amostragem para a seleção de uma amostra representativa da população possuindo, deste modo, pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais e ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008, 113-115). Por isso, ao contrário do que acontece numa amostra aleatória simples, numa amostra complexa nem todos os indivíduos da amostra representam o mesmo número de indivíduos da população, sendo essa informação dada por pesos. Existem vários tipos de pesos, como o peso do delineamento, o peso da pós-estratificação e o peso da população, entre outros.

Na modelação estatística e análise dos dados é preciso incluir os pesos, sendo o investigador alertado para isso aquando da cedência dos dados. Mas nem sempre o investigador está familiarizado com os pesos disponibilizados, o que dificulta a identificação dos pesos a utilizar na análise que pretende realizar, bem como se estes devem ou não ser alvo de alguma transformação. Por outro lado, alguns investigadores optam por não considerar os pesos, abordagem que pode produzir incorreções, tanto para as estimativas, como para as respetivas variâncias, enviesando os resultados e as conclusões (Osborne, 2011).

Neste trabalho para além de tecermos algumas considerações sobre o que representam os vários tipos de pesos, que transformações se podem ou devem realizar e que pesos utilizar, focamos essencialmente a estratégia que consiste em utilizar as técnicas disponíveis baseadas na amostragem aleatória simples, mas ponderando os dados usando os pesos normalizados corrigidos pelo efeito do delineamento (*deff*). Este efeito é definido pelo quociente entre a variância do estimador de um parâmetro sob um plano de amostragem complexa e a variância do estimador do mesmo parâmetro sob o

pressuposto de um plano de amostragem aleatório simples com uma amostra do mesmo tamanho. Como o valor do *deff* nem sempre é fornecido e pode variar consoante as variáveis em estudo (Sturgis, 2004), mostramos como se pode estimar o *deff* (algo que muitas vezes é complexo) e analisamos o impacto que uma estimativa incorreta tem na significância das variáveis.

São considerados dados obtidos no "National Survey on Drug Use and Health" (NSDUH) sobre a temática do consumo de álcool e cocaína, tendo-se selecionado um conjunto de variáveis sociodemográficas para o ajuste de modelos de análise de sobrevivência (Hosmer et al., 2008), tomando-se como variável resposta o tempo até à primeira vez que o indivíduo consumiu a substância e como variável *status* se alguma vez a consumiu.

Consideram-se duas abordagens utilizando os seguintes valores de *deff* 0,5; 1; 1,5; 2,5 e 3:

1) Comparar os modelos obtidos com o uso de um conjunto de valores do *deff* pré definidos, o que permite avaliar o impacto de diferentes valores do *deff* no modelo ajustado;

2) Considerar o modelo obtido com a informação correta do delineamento de amostragem e ajustar novamente esse modelo considerando diferentes valores do *deff*, o que permite avaliar o efeito do *deff* nas estimativas e na sua significância.

Para ambas as abordagens serão criados modelos focados essencialmente na estratégia da modelação de uma amostra aleatória simples, mas ponderando os dados usando os pesos finais normalizados corrigidos pelos valores de *deff* acima referidos. Os pesos finais utilizados em todos os modelos obtidos foram fornecidos pela base de dados NSDUH. Através dos modelos criados, para as respetivas abordagens, pretende-se comparar a significância das variáveis e das suas interações de forma a verificar o impacto de um valor de *deff* mal estimado.

Concluímos que a estimação incorreta do *deff* afeta a significância das variáveis e das interações entre as variáveis. Tal pode condicionar a discussão de resultados feita pelo investigador e a sua teorização sobre o fenómeno em estudo. Além disso, com a primeira abordagem concluímos que quanto menor o valor do *deff* maior o número de variáveis significativas. Com a segunda abordagem, concluímos que as estimativas dos coeficientes não são afetadas pelo valor do *deff* mas que para valores do *deff* mais reduzidos tende a existir uma subestimação do desvio-padrão.

Referências

Hosmer, D.W., Lemeshow, S. & May, S. (2008) *Applied Survival Analysis.* Wiley Online Library.

Lavrakas, P. J. (2008) *Encyclopedia of survey research methods*. SAGE Publications.

Osborne, J.W. (2011) Best Practices in Using Large, Complex Samples: The Importance of Using Appropriate Weights and Design Effect Compensation. *Practical Assessment, Research & Evaluation*16, 1-7.

Sturgis, P. (2004) Analysing Complex Survey Data: Clustering, Stratification and Weights. *Social Research Update* 43.

Modelos Estocásticos e Modelos Espaciais – 6ª Feira, 21 de Abril, Auditório A (10h20)

Delineamentos experimentais para ensaios iniciais de programas de melhoramento de plantas: aplicação à selecção de castas antigas de videira

Elsa Gonçalves[1], Antero Martins[2]

[1] Universidade de Lisboa, Instituto Superior de Agronomia/ Departamento de Ciências e Engenharia de Biossistemas, Secção de Matemática, LEAF, Linking Landscape, Environment, Agriculture and Food, Tapada da Ajuda 1349-017 Lisboa, Portugal, elsagoncalves@isa.ulisboa.pt.

[2] Universidade de Lisboa, Instituto Superior de Agronomia, LEAF, Tapada da Ajuda 1349-017 Lisboa, Portugal, anteromart@isa.ulisboa.pt.

Sumário: Neste trabalho comparam-se delineamentos experimentais destinados a ensaios de campo para conservação, quantificação da variabilidade genética e selecção de um grupo de genótipos superiores em castas antigas de videira. Compara-se o delineamento em blocos completos casualizados com o delineamento linha-coluna, utilizando dados reais de rendimento obtidos em vários ensaios. A maior eficiência da selecção (traduzida num valor de heritabilidade mais elevado) consegue-se com o delineamento linha-coluna e é tanto maior quanto menor for o número de falhas de plantas na parcela.

Palavras-chave: Delineamento Experimental, Ensaios de Campo, Melhoramento de Plantas.

Os delineamentos em blocos completos casualizados desenvolvidos por *Fisher* são dos métodos mais simples para o controlo da variação espacial em ensaios agronómicos. As aproximações posteriores, correspondentes à classe dos delineamentos em blocos incompletos, visaram os mesmos objectivos, contudo, foram direccionadas para o controlo da variação espacial em ensaios com maior número de tratamentos, tipicamente o que acontece frequentemente em melhoramento de plantas.

De facto, quando o número de tratamentos (v) é grande, o controlo da heterogeneidade espacial poderá ser conseguido constituindo pequenos conjuntos homogéneos de unidades experimentais, isto é, uma repetição de tamanho v pode ser dividida em s blocos incompletos, cada um de tamanho k ($k < v$), sendo $v = ks$. Um delineamento com r repetições (blocos completos) deste tipo é designado por delineamento experimental em blocos incompletos resolúvel. Os delineamentos em blocos incompletos equilibrados originais de *Yates* requerem que v seja o quadrado de k. Devido a esta restrição, surgiu uma classe de delineamentos em blocos incompletos resolúveis, particularmente adequada para ensaios com um grande número de tratamentos, os delineamentos alfa (Patterson e Williams, 1976). Neste caso, apenas se exige que v/k seja um número inteiro, embora se alcance maior eficiência quando o

tamanho do bloco, k, for menor que a raiz quadrada do número de tratamentos, v, e então, menor que s (Giesbrecht e Gumpertz, 2004). Outra classe de delineamentos engloba os designados linha-coluna (Williams e John, 1989). Estes são usados quando se suspeita da existência de variação espacial em duas direcções no campo, sendo as linhas e colunas utilizadas conjuntamente para introduzir dois factores de controlo da heterogeneidade espacial. Neste caso, têm-se blocos incompletos simultaneamente nas linhas e nas colunas e o delineamento é resolúvel se, em cada repetição, o número de tratamentos v for igual ao número de linhas l, multiplicado pelo número de colunas c, isto é, $v = lc$.

Os delineamentos alfa e linha-coluna são correntemente usados em ensaios agrícolas e florestais, estando desenvolvidos muitos algoritmos para a sua construção, assim como instrumentos para avaliar a sua eficiência (Giesbrecht e Gumpertz, 2004).

No que respeita ao delineamento de ensaios de selecção da videira, têm sido conduzidos estudos com base em dados simulados com o objectivo de identificar os delineamentos experimentais mais eficientes na quantificação da variabilidade genética intravarietal e na selecção de um grupo de genótipos superiores (Gonçalves *et al*., 2010). O presente trabalho procura agora avaliar essa eficiência mas com base em dados de rendimento obtidos em ensaios de campo reais. Nomeadamente, procura-se identificar o que se perde na eficiência da selecção: (1) quando existem falhas de plantas no campo e, consequentemente, não se consegue avaliar a totalidade das plantas inicialmente previstas em cada parcela; (2) quando se opta por um delineamento em blocos completos casualizados (RCBD) em vez de delineamentos linha-coluna (RCD). Essa perda de eficiência na selecção é traduzida pela redução do valor da heritabilidade, um parâmetro de genética quantitativa largamente utilizado no melhoramento de plantas, e que corresponde a um índice do afastamento entre os verdadeiros efeitos genotípicos e os efeitos genotípicos previstos segundo determinado modelo. Os resultados mostram que quando não se consegue avaliar a totalidade das plantas inicialmente previstas em cada parcela do ensaio pode perder-se até cerca de 40% de eficiência na selecção, quando apenas se avalia uma planta por parcela. De entre os ensaios analisados, quando se compara o RCD com o RCBD, o primeiro delineamento revela-se mais eficiente, perdendo-se entre 1% a 6% de eficiência na selecção genética quando se opta por um RCBD.

Referências

Giesbrecht, F. G. & Gumpertz, M.L. (2004) *Planning, construction and statistical analysis of comparative experiments*, New York: Wiley series in probability and statistics, John Wiley & Sons.

Gonçalves, E., St.Aubyn, A. & Martins, A. (2010) Experimental designs for evaluation of genetic variability and selection of ancient grapevine varieties: a simulation study. *Heredity,* 104, 552-562.

Patterson, H. D. & Williams, E. R. (1976) A new class of resolvable incomplete block designs. *Biometrika*, 63, 83-92.

Wiliams, E. R. & John, J. A. (1989) Construction of row and column designs with contiguous replicates. *Applied Statistics*, 38**,** 149-154.

Modelos Estocásticos e Modelos Espaciais – 6ª Feira, 21 de Abril, Auditório A (10h40)

Deteção de distribuições atípicas em sequências genómicas

Ana Helena Tavares[1], Vera Afreixo[2], Paula Brito[3]
[1] Departamento de Matemática & CIDMA, Universidade de Aveiro, ahtavares@ua.pt;
[2] Departamento de Matemática & CIDMA & iBiMED, Universidade de Aveiro; vera@ua.pt;
[3] FEP & LIAAD-INESC TEC, Universidade do Porto, mpbrito@fep.up.pt

Sumário: Neste trabalho abordamos o problema de deteção de distribuições atípicas no contexto das distâncias entre palavras genómicas, seguindo uma abordagem funcional, tendo em conta não apenas a existência de outliers de magnitude, mas também a possível existência de outliers de forma. Aplica-se um procedimento baseado na medida *directional outlyingness*, que permite detetar *outliers* de magnitude e *outliers* de forma.

Palavras-chave: *Directional outlyingness*, Distribuição atípica, Palavra genómica.

O DNA pode ser descrito por uma longa sequência de A, C, G e T's. Apesar de utilizar um alfabeto de apenas 4 letras, é nesse texto que se encontra codificada a informação necessária à vida dos organismos. A deteção de palavras genómicas atípicas, ou *outliers*, é um assunto de muito interesse no estudo do DNA. Neste trabalho utilizamos as distâncias entre palavras como descritor da distribuição de uma palavra ao longo do genoma.

A distância entre-palavras é definida como a diferença entre a última letra de duas ocorrências sucessivas da mesma palavra. Por exemplo, na sequência ACTCGACAC as distâncias entre os AC são 5 e 2. Olhando para um conjunto de distribuições de distâncias como um conjunto funcional, podemos utilizar medidas e técnicas de análise de dados funcionais, para a deteção de *outliers*. Neste contexto, uma função *outlier* não é apenas aquela que contém valores atipicamente baixos ou elevados (*outliers* de magnitude), mas também aquela que apresente um padrão diferente do da maioria (*outliers* de forma).

Recentemente, Rousseeuw *et. al.* (2016) propuseram uma nova abordagem, baseada na medida *directional outlyingness* (DO), que atribui um valor robusto de atipicidade a cada ponto do domínio de uma função. Baseados na medida de *outlyingness* de Stahel-Donoho (Stahel, 1981; Donoho, 1982) definem a medida DO de um ponto y, em relação a uma amostra univariada $Y = \{y_1, \cdots, y_m\}$, como

$$DO(y;Y) = \begin{cases} \dfrac{y - Med(Y)}{S_a(Y)}, & se\ y \geq Med(Y) \\ \dfrac{y - Med(Y)}{S_b(Y)}, & se\ y \leq Med(Y) \end{cases}$$

onde S_a e S_b são estimadores de escala robustos para os subconjuntos de pontos acima e abaixo da mediana, respetivamente. Considere-se agora uma função f, um conjunto de funções $F = \{f_1, \dots, f_m\}$, todas de domínio univariado, e $\{t_1, \dots, t_T\}$ um conjunto discreto do domínio onde o valor das funções é observado. Para cada ponto t_i pode calcular-se o DO

de $f(t_i)$ com respeito ao conjunto de valores tomados pelas outras funções nesse mesmo ponto. Os valores de DO obtidos para cada função podem ser reduzidos a um par ordenado, por via de uma medida de localização, *fDO*, (uma média ponderada dos valores DO) e uma medida de variabilidade, *vDO*. Representando os pares ordenados (*fDO*, *vDO*) num gráfico de dispersão obtém-se o *functional outlyingness map* (FOM), uma ferramenta gráfica que permite identificar possíveis candidatos a *outliers*.

Neste trabalho segue-se essa abordagem na deteção de palavras *outliers* no genoma humano. A cada conjunto funcional, F_k, formado pelas 4^k distribuições de distâncias entre palavras associadas às palavras de tamanho k, aplica-se o procedimento baseado na medida DO e na representação gráfica FOM ($k = 1, \dots, 5$). Numa primeira abordagem considera-se que as distâncias contribuem todas com o mesmo peso na definição do *fDO* (distribuição uniforme).

O FOM associado ao conjunto de distribuições F_2 evidencia a distribuição de distâncias entre-CG como *outlier*, pois esta apresenta valores elevados de *fDO* e de *vDO*. De facto, a distribuição de distâncias entre-CG distingue-se das restantes por ser mais "achatada". À medida que o valor de k aumenta, mais inesperado é o padrão de algumas curvas, apresentando um ou mais picos que podem ocorrer tanto em distâncias curtas como em distâncias mais longas. No caso de $k = 5$, o FOM resultante aponta 17 casos como *outliers*. O método permite capturar distribuições com um padrão que se diferencia do padrão da maioria, assim como distribuições com picos em sub-intervalos onde as outras distribuições não os apresentam. São efetuadas experiências utilizando diferentes função peso e os resultados são comparados.

Os resultados iniciais indicam que o procedimento baseado na medida DO é promissor, colocando em evidência distribuições atípicas que eram mascaradas pelas restantes curvas.

Agradecimentos: Este trabalho é parcialmente financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) e FCT (Fundação Portuguesa para a Ciência e Tecnologia) através dos projetos UID/MAT/04106/2013 do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações), UID/BIM/04501/2013 do Instituto de Biomedicina (iBiMED), UID/EEA/50014/2013 do INESC-TEC (Instituto de Engenharia de Sistemas e Computadores do Porto), POCI-01-0145-FEDER-006961 do COMPETE 2020 (Programa Operacional Competitividade e Internacionalização) e bolsa de doutoramento PD/BD/105729/2014 de AT.

Referências

Rousseeuw, Peter J., Raymaekers, J. & Hubert, M. (2016). A measure of directional outlyingness with applications to image data and video. *Preprint arXiv*:1608.05012.

Stahel, Werner A. (1981), Breakdown of Covariance Estimators, *Research Report*, vol. 31, Fachgruppe für Statistik, E.T.H. Zürich, Switzerland.

Donoho, David L. (1982), Breakdown properties of multivariate location estimators. Ph.D. diss., Harvard University.

Modelos Estocásticos e Modelos Espaciais – 6ª Feira, 21 de Abril, Auditório A (11h00)

l1-norm posterior probability Support Vector Machines

A. Pedro Duarte Silva
Católica Porto Business School e CEGE, Univ. Católica Portuguesa (Porto)
psilva@porto.ucp.pt

Abstract: It is proposed a novel support vector machine approach to the binary regression problem. This proposal relies on a sequence of linear programming models optimizing functionals on Reproducing Kernel Hilbert Spaces, with l1-norm regularization penalties. The choice of l1-norm penalties allows for the handling of big data problems. The proposed method is compared with classical binary regression models under different data conditions.

Key words: Support vector machines, Binary regression**,** Machine learning**,** Big data

Vapnik's principle of never solving a problem more general than you actually need to solve lead to development of Support Vector Machines (SVMs), which have quickly established themselves as one of the most important paradigms in statistical machine learning. Originally proposed for two group classification, SVMs have been successfully extended to tackle difficult problems in nonparametric regression, general multi-group classification, and density estimation, with both classical, complex and big data (see *e.g.* Cristianini and Shawe-Taylor 2000). However, somehow surprisingly, SVMs extensions to address categorical regression problems do not seem to exist. In this presentation, one SVM extension specially designed for binary regression will be proposed.

Assume that a set of given objects partitioned into two well defined groups can be described by the random pairs $\mathbf{z} = (\mathbf{x}, y)$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, where $\mathcal{X}$ is some attribute domain, and $\mathcal{Y} = \{-1,+1\}$ is a set of group labels. Then, the theoretical Bayes rule maximizing the probability of correct classifications is given by $\hat{y} = \phi_B(\mathbf{x}) = sign(P(y=+1 \mid \mathbf{x}) - 1/2)$. Furthermore, given a training sample of l independent examples, $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l$ drawn from some unknown but common distribution, and a Reproducing Kernel Hilbert Space with kernel $K(.,.)$, $\mathcal{H}_K$, rich enough to approximate $\phi_B(\mathbf{x})$ to any desired accuracy level, it follows (Lin 2002) that, for any positive value of the trade-off parameter λ , $\phi_B(\mathbf{x})$ can be consistently (as $l \to \infty$) estimated by the standard two-group classification SVM rule: $\hat{y} = \phi_{SVM}(\mathbf{x}) = sign(h(\mathbf{x}) + b)$, where the constant $b \in \Re$, and function $h(\mathbf{x}) \in \mathcal{H}_K$, minimize $\frac{1}{l}\sum_{i=1}^{l} \varepsilon_i + \lambda \| h \|_{H_k}^2$ under the constraints $y_i(h(\mathbf{x}) + b) \geq 1 - \varepsilon_i$; $\varepsilon_i \geq 0$ $(i = 1,\ldots,l)$.

However, if the training sample examples are drawn separately from the two classes with proportions, p^+ and p^- , that do neither coincide, nor consistently estimate, the true *a*

priori probabilities, π^+ and π^-, then $\phi_{SVM}(\mathbf{x})$ does not approximate $\phi_B(\mathbf{x})$ neither can be modified to do so by a simple shift in the value of the bias parameter, *b* (Lin, Lee and Wahba 2002). Furthermore, if the misclassification costs differ across groups, with $c^+ \neq c^-$ being respectively the misclassification costs for a false positive and a false negative, $\phi_{SVM}(\mathbf{x})$ has no clear relation with the minimal expected cost Bayes rule $\hat{y} = \phi_{ECB}(\mathbf{x}) = sign\left(P(y=+1 \mid \mathbf{x}) - \frac{c^+}{c^+ + c^-}\right)$. In fact, Lin, Lee and Wahba (2002) show that standard SVMs do not carry any information about the *a posterior* membership probabilities, P(y=+1|**x**) and P(Y=-1|**x**), other than the sign of P(y=+1|**x**)–½. Nevertheless, non-standard SVMs that minimize instead $\frac{1}{l}\sum_{i=1}^{l} L(y_i)\,\varepsilon_i + \lambda \| h \|^2_{H_k}$ with L(+1) := $c^+ \pi^+ p^-$ and L(-1) := $c^- \pi^- p^+$ are consistent estimators of $\phi_{ECB}(\mathbf{x})$.

Based on these insights, here it will be proposed a *posterior probability* regression model, obtained by solving a succession of non-standard SVMs with varying L(+1) and

L(-1) specifications. However, these models replace Lin, Lee and Wehba (2002) regularization term $\| h \|^2_{H_k}$, by a *l1-norm* penalty regularization in the spirit of similar proposals of Mangasarian and Thompson (2008) for standard two group classification SVMs. It is known (Cristianini and Shawe-Taylor 2000) that *l1-norm* based SVMs can lead to linear rather than quadratic optimization problems, which on the one hand leads to a reduction of the number of support vectors employed potentially improving their generalization ability, and other the other hand are able to handle efficiently much larger data sets than *l2-norm* based SVMs.

The proposed method will be compared with classical regression models for binary data, under several relevant data conditions.

References

Cristianini, N. & Shawe-Taylor, J. (2000) *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press.

Lin, Y. (2002) Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery* 6(3), 259-275.

Lin, Y., Lee, Y. & Wahba, G. (2002) Support vector machines for classification in nonstandard situations. *Machine Learning* 46(1-3), 191-202.

Mangasarian, O.L. & Thompson, M.E. (2008) Chunking for massive nonlinear kernel classification. *Optimisation Methods and Software* 23(3), 365-374.

***Data Mining* – 6ª Feira, 21 de Abril, Auditório B (10h00)**

Disaggregation and classification of loads

Inês Soares[1], David Rua[2], João Gama[3]
[1] INESC TEC, aisoares@inesctec.pt; [2]INESC TEC, drua@inesctec.pt; [3]FEP, INESC TEC, jgama@fep.up.pt

Summary: Nowadays, driven by the growing awareness on energy efficiency, consumers want to know more about their household consumption. The possibility of knowing the consumption of each devices that they have at home is a potential tool in allowing them to become more efficient in the way that they use the devices and to take advantage of more advantageous pricing schemes. The intrusiveness of monitoring may be an obstacle so to address this issue, we tested three algorithms for non-intrusive load monitoring.

Keywords: Non-intrusive load monitoring, Machine Learning,

Information and communication technology plays an important role in approaching energy issues throughout the world. The advent of smart meters introduced a paradigm shift in allowing the remote metering of the energy use. These devices can also provide information in real-time for families about their electricity consumption and thus help them save energy. However, smart meters still cannot provide detailed information about which device is activated and how much it consumed. There are two ways to obtain this information: using devices that measure the consumption of each device individually (through intrusive monitoring which is inconvenient and costly), or by inferring this information from the total electricity consumption, measured by a smart meter or another central metering device (non-intrusive monitoring methods).

In our work, the aim was to study how non-intrusive load monitoring (NILM) algorithms behave in detecting when a given appliance was turned on or off. Therefore, three non-intrusive load monitoring (NILM) algorithms were described and tested. The first algorithm tested was proposed by Baranski and Voss (2004b), where unsupervised methods are used. The second one, created by Weiss et al. (2012), uses supervised methods, which means that pre-classified data is required to realise an experiment. The last algorithm employs a semi-supervised method and it was proposed by Parson et al. (2012).

For the realisation of the experiments we chosen an existing framework available on-line, in Matlab, NILM-Eval, that implements the aforementioned algorithms and two datasets were selected: one European (ECO) and one American (REDD). Both datasets have six houses each, data of individual appliances and the aggregated consumption of each house.

The Parson's algorithm was selected because it disaggregates the appliances one by one given that different patterns emerge when devices are successfully disaggregated. Being semi-supervised, it is necessary to define several parameters of the device that will be predicted (e.g., mean and variance of the on and off states power, probability of transition between states, etc.).

To compute the performance of the algorithm with REDD dataset was necessary to make some changes in the pre-defined parameters and in the data. A new value for the two means (on/off) were chosen and different training hours were used. We decided to change the number of

training hours since the number of consecutive days in the dataset to perform the test was very small (on average, five consecutive days for training and another five for test) and an increasing of the number of hours on training enhances the accuracy of the inference. Whatever, we conclude that given a certain number of training hours the results started to be worst (overfitting). In general, the algorithm detected the performance of the device with high precision.

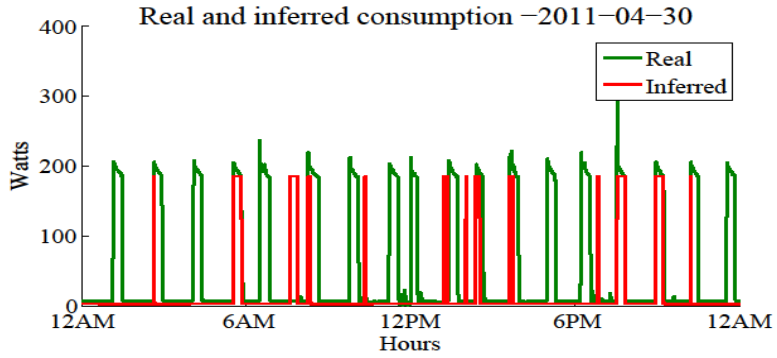


Fig. 4 - Three hours of training

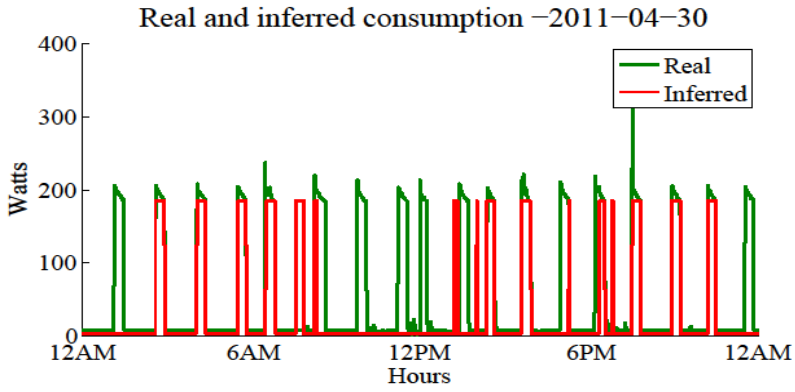


Fig. 5 - Five hours of training

With the experiments carried out it was concluded that data pre-processing techniques are fundamental to guarantee a better quality of the analysis and estimation of the algorithms of disaggregation. It was also determined that is very important to have, at least, one week of consecutive days to train, in order to obtain consistent results.

Acknowledgment: This work is financed by the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project SmartGP/0003/2015 under the framework of ERA-Net Smart Grids Plus.

References

Beckel, C., et al.(2014). The eco data set and the performance of non-intrusive load monitoring algorithms. In Proceedings of the 1st ACM International Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys 2014). Memphis, TN, USA, pp. 80–89. ACM

Magoules, E. and Zhao, H. (2016). Data Mining and Machine Learning in Building Energy Analysis: Towards High Performance Computing. Wiley-ISTE

Parson, O. (2014). Unsupervised training methods for non-intrusive appliance load monitoring from smart meter data. Ph. D. thesis, University of Southampton.

***Data Mining* – 6ª Feira, 21 de Abril, Auditório B (10h20)**

Inflammatory Bowel Disease: evidence, classification and prediction

Claudia Camila Dias[1], **Pedro Pereira Rodrigues**[2], **Altamiro da Costa-Pereira**[3], **Guilherme Macedo**[4], **Fernando Magro**[5]

[1] *Center for Health Technology and Services Research & MEDCIDS, Faculty of Medicine of the University of Porto, camila@med.up.pt;*

[2] *Center for Health Technology and Services Research & MEDCIDS, Faculty of Medicine of the University of Porto, pprodrigues@med.up.pt;*

[3] *Center for Health Technology and Services Research & MEDCIDS, Faculty of Medicine of the University of Porto, altamiro@med.up.pt;*

[4] *Center for Health Technology and Services Research & Centro Hospitalar São João,* gmacedo@med.up.pt

[5] *Biomedicine Department & Centro Hospitalar São João, fm@med.up.pt*

Abstract: Inflammatory bowel disease (IBD) is a chronic disease with unknown aetiology. This work aims to summarize the evidence regarding IBD outcomes and risk factors, identify and assess risk factors, and develop and validate prognostic models. Clinical and demographical factors should be used more frequently in the prognosis of IBD as they are easier to collect than serologic or genetic ones. The prognostic models seem easily applicable as bedside clinical tools that can help physicians during therapeutic decisions in early disease management.

Keywords: Inflammatory bowel disease, prognostic models, Bayesian networks, decision tree

Inflammatory bowel disease (IBD) is a chronic disease with unknown aetiology. The two most prevalent types of IBD are Crohn's disease and ulcerative colitis which, given their similarities, are often difficult to distinguish. Their chronic characteristics, along with a high risk of complications (including relapses, surgeries and disabling, among others) make it essential to develop accurate prediction models.

This work has three objectives: summarize the evidence regarding IBD outcomes and risk factors, identify and assess risk factors for identified outcomes, and develop and validate prognostic models for those outcomes, based on the identified clinical and demographic factors.

To identify factors associated with the prognosis, systematic reviews and meta-analysis for Crohn's disease and ulcerative colitis were developed. Factors such as young age at diagnosis, perianal disease, initial use of steroids for the first flare and ileocolic disease location were identified as independent factors for disabling disease in Crohn's patients. Concerning ulcerative colitis, male gender, non-smoking habits, extensive disease, need for corticosteroids and hospitalization were associated with colectomy. In order to identify and assess risk factors for outcomes of Crohn's disease, an observational study was conducted. Early surgery or immunosuppression seem to not prevent global

disabling disease, but an early start of immunosuppression by itself is associated with fewer surgeries and should be considered in daily practice as a preventive strategy. The third objective of this work is to develop and validate prognostic models for the identified outcomes. Two main paths of work have been followed for modelling predictive classifiers, one based on Bayesian networks and the other on decision trees. Bayesian network models achieved high area under the curve (AUC) for disabling disease and reoperation, and were included in an online tool, allowing the application of the classifier at bedside. Risk matrices - based on age at diagnosis, perianal disease, disease aggressiveness and early therapeutic decisions - exhibited also good performance for the most important prognostic criteria: high positive post-test odds for disabling disease and low negative post-test odds for reoperation. The risk matrices seem also easily applicable as bedside clinical tools that can help physicians during therapeutic decisions in early disease management. In the second path, decision trees were able to predict disabling, surgery and reoperation with high AUC, and were shown to be a valid and useful approach to depict outcome risks. The defined cut-off risk levels expressed high odds for a positive test for disabling, while excluding surgery and reoperation with low odds for a negative test.

In conclusion, clinical and demographical factors should be used more frequently in the prognosis of IBD as they are easier to collect than serologic or genetic ones. The predictive models for CD prognosis could enhance the initial approach and, therefore, improve the clinical outcome.

Acknowledgment: This work has been developed under the scope of GEDII (Grupo de Estudo da Doença Inflamatória Intestinal), which gave an invaluable institutional support for this thesis, and project NanoSTIMA (NORTE-01-0145- FEDER-000016) which is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

Louis, E., Belaiche, J., and Reenaers, C. (2010) Do clinical factors help to predict disease course in inflammatory bowel disease? World Journal of Gastroentereology, 16 (21): 2600-3

Dias, C.C., Rodrigues, P.P., et al (2016) Development and validation of risk matrices for Crohn's disease outcomes in patients submitted to early therapeutic intervention. Journal of Crohn's and Colitis jjw171. doi: 10.1093/ecco-jcc/jjw171

Dias, C.C., Rodrigues, P.P, et al (2017) The risk osf disabling, surgery and reoperation in Crohn's disease: a decision tree-based approach to prognosis. PloSOne (accepted)

***Data Mining* – 6ª Feira, 21 de Abril, Auditório B (10h40)**

Detection of Centrality and Community Patterns in Multilayer Networks – An Application in Medicine

Patrícia C. T. Gonçalves[1], Pedro Campos[2]
[1] *LIAAD INESC-TEC, pctg@inesctec.pt;*
[2] *LIAAD INESC-TEC and FEP, pcampos@fep.up.pt*

Summary: We can visualize relationships between a disease and different degrees of kinship, for example, through networks of individuals, because they constitute a practical and effective way of referencing patients and the links between them. Relating patient data through social networks makes it possible to, for example, detect "hidden" links between illnesses, or identify patterns of geographic dispersion. Multilayer networks allow us to represent and analyse complex data in a multivariate form.

Keywords: Data Visualization, Health Data, Multilayer Networks, Pattern Identification.

Many diseases are related to several aspects, such as heredity, geography, demography, social aspects, etc. It is possible to visualize relationships in the form of networks, in which patients are represented by network nodes and links between them are represented by edges. The networks enable the analysis of complex data of this type – through the analysis of social networks –, and the level of complexity can increase even more when several variables are included simultaneously, such as geographic location, kinship or type of disease – through the analysis of multilayer networks. Strictly speaking, according to the extensive study of Kivelä et al. (2014), the multilayer networks we use are *layer-disjoint networks*, since each node is present only in a single layer.

There are several types of patterns that can be extracted from social networks. Network centrality, for example, measures the level of importance of the nodes in the network. We use random walks applied to multilayer networks: the more likely it is that a random walker can be found at a node-layer, the more important that node-layer is, and thus more central (Meng, 2015). Community measures (or modularity) evaluate the network partition into modules, or groups, thus identifying individuals with high degrees of homogeneity (Jackson, 2008). Detecting regularities in the patterns of the movement of a random walker is a way of revealing communities in a network (Rosvall, 2009).

The possibility to perform these mappings in multilayer networks and extract information is a new topic of great importance. In particular for the health sciences, as this allows (i) to relate patient data in order to detect possible "hidden" links between diseases that apparently would not have any connection between them, (ii) to detect links between relatives, or (iii) to identify geographical patterns of concentration of a disease.

The objective of this study lies in the mapping of this multi-layered information and information extraction that allows to find these hidden links among diseases, geographical concentration, or kinship by obtaining centrality and community measures

in multilayer networks. The application of this methodology in real context is therefore a valuable asset, if it can be done in real databases of patients. Figure 1 shows an example of what can be done in terms of multilayer networks between cancer patients, providing a valuable visualization of data. (The network contains fictitious random data, where each layer represents the age groups of patients, and each node represents a patient.)

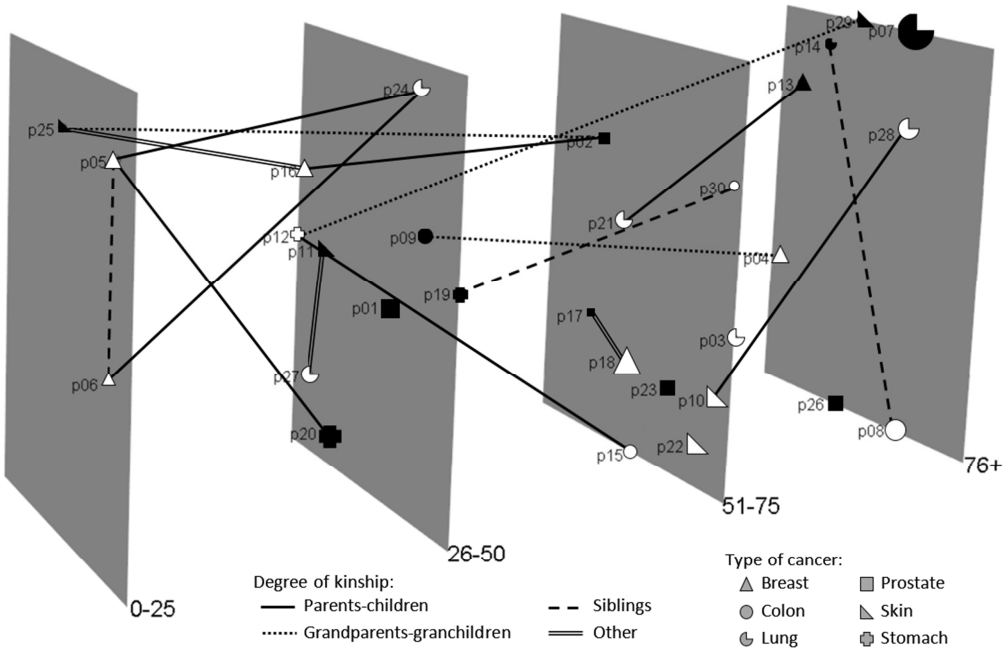


Figure 1: **Family ties between cancer patients**. The four layers represent age ranges. The nodes depict the patients: men in black, women in white. The size of each node represents the stage of the disease, being a node as bigger as the more advanced the disease is. The links between the nodes represent the degree of kinship between patients. (Fictitious data randomly generated.)

Acknowledgments: This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The authors would like to thank project NanoSTIMA – Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics (NORTE-01-0145-FEDER-000016).

References

Jackson, M. O. (2008) Observing and Measuring Social Interaction - Community Structures, Block Models, and Latent Spaces. In M. O. Jackson, *Social and Economic Networks* (pp. 573-590). Princeton University.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014) Multilayer Networks. *Journal of Complex Networks, 2*(3), 203-271.

Meng, X. (2015) *Centrality Measures in Multilayer Networks*, Extended Essay. University of Oxford.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009) The Map Equation. *The European Physical Journal Special Topics*, 178, 13-23.

***Data Mining* – 6ª Feira, 21 de Abril, Auditório B (11h00)**

Exploratory Data Analysis in Pavement Engineering Using Python

Pedro Marcelino[1], Maria de Lurdes Antunes[2], Eduardo Fortunato[3], Marta Castilho Gomes[4]

[1] Laboratório Nacional de Engenharia Civil, pmarcelino@lnec.pt;

[2] Laboratório Nacional de Engenharia Civil, mlantunes@lnec.pt;

[3] Laboratório Nacional de Engenharia Civil, efortunato@lnec.pt;

[4] CERIS, CESUR, Instituto Superior Técnico, marta.gomes@tecnico.ulisboa.pt

Abstract: This paper presents an exploratory data analysis in pavement engineering problems using Python. A case study with data from the U.S. Long-Term Pavement Performance (LTPP) database illustrates Python applications for data analysis and visualization. This work demonstrated that Python can play an important role in the analysis of pavement engineering data, such as inventory, climate, traffic and pavement monitoring data. Further extensions of this research can lead to the development of more complex analyses, like the development of prediction models for pavement deterioration.

Keywords: Data analysis, Data science, Data visualization, Pavement engineering, Python.

This paper addresses the use of Python to analyse pavement engineering data. Python is an open source programming language whose popularity increased in recent years. One of the key features of Python is that it supports packages. These packages extend its capabilities, providing tools that simplify data computation, analysis and visualization.

Python has been successfully applied in different areas of science and engineering. However, the number of studies on Python applications for pavement engineering in the literature is scarce. Nowadays, many decisions on pavement maintenance and rehabilitation treatments are supported by models derived from the analysis of extensive pavement performance data, in an effort to improve infrastructure management. In this context, Python can play an important role because it offers a set of solutions that simplify data analysis and visualization, besides being a free alternative to commercial software.

In this paper, the use of Python to perform several data analyses and visualization tasks is illustrated for a case study that uses data from the Long-Term Pavement Performance (LTPP, 2017) database. For pavement engineering researchers, the LTPP database is an important source of information. Data from more than 2500 test sections of in-service highways all over the United States and Canada is available in the LTPP database. This database is subjected to rigorous control checks in order to ensure data quality.

This work also identifies and discusses the relevance of some Python packages for pavement engineering studies. There are more than 90,000 Python packages available from the official online repository. Since many of these packages are designed for scientific research, in particular for data analysis and visualization, those that can be most useful for the pavement engineering community have to be selected. Considering the usual needs of a pavement engineer and the capabilities of each package, a set of suitable packages is presented and their application to the case study described.

The current study has shown that Python can be used to import, visualize and analyse several broad classes of data, such as inventory, climate, traffic and pavement monitoring data. This establishes the baseline for more complex studies, such as the development of prediction models for pavement deterioration. Moreover, Python is a free and easy to learn programming language, supported by a strong online community. As a result, it fosters collaboration and knowledge sharing, contrasting with proprietary software packages, which can be expensive and not so cooperation-friendly. The experience reported in this study is expected to encourage pavement engineers to use Python in their work.

References

McKinney, W. (2012) *Python for Data Analysis,* USA: O'Reilly.

Hair, J., Black, W., Babin, B. & Anderson, R. (2014) *Multivariate Data Analysis,* USA: Pearson.

Ayer, V., Miguez, S. & Toby, B. (2014) Why scientists should learn to program in Python, *Powder Diffraction*, 29(S2), S48–S64.

Long-Term Pavement Performance (LTPP) Database. (2017) *LTPP InfoPave*, https://infopave.fhwa.dot.gov (Accessed Feb. 2017).

Classificação II – 6ª Feira, 21 de Abril, Auditório B (17h00)

Efeitos da gestão do fundo de maneio na rentabilidade das empresas do setor mobiliário português

Ruben Silva[1], Paulo Peixoto[2], Susana Silva[3], Cristina Lopes[4]

[1] *ISCAP – Polytechnic of Porto, rrs@eu.ipp.pt*

[2] *ISCAP – Polytechnic of Porto, paulojuniogavinapeixoto@hotmail.com*

[3] *ISCAP – Polytechnic of Porto, inessilva.2010.2011@gmail.com*

[4] *LEMA & CEOS.PP, ISCAP – Polytechnic of Porto, cristinalopes@iscap.ipp.pt*

Resumo: O objetivo do presente estudo longitudinal é averiguar se uma eficiente gestão de fundo de maneio está relacionada com o aumento da rentabilidade dos ativos das empresas portuguesas do setor mobiliário. A amostra é composta por 626 empresas que foram analisadas durante o período de 2011 a 2015. Os resultados obtidos através de modelos de efeitos fixos sugerem que a rentabilidade de uma empresa aumenta quando se diminui o tempo médio de pagamentos, tempo médio de recebimentos e tempo médio de inventários. Do mesmo modo, uma diminuição do ciclo de tesouraria sugere um aumento da rentabilidade.

Palavras Chave: Fundo de Maneio, Rentabilidade, Setor Mobiliário, Dados em painel.

O objetivo do presente estudo é testar o impacto da gestão do fundo de maneio na rentabilidade das empresas do sector mobiliário Português. A literatura das Finanças Empresariais está, sobretudo, focada no estudo de investimentos, estrutura dos capitais e/ou dividendos, dando mais importância às decisões de longo prazo (García-Teruel & Martínez-Solano, 2007). A realidade é que nos dias de hoje as empresas investem muito mais em curto prazo em detrimento do longo prazo e a maioria dos seus recursos são de maturidade inferior a 1 ano. Sendo a maior parte das empresas analisadas PME's, detêm a maior parte dos seus ativos e passivos no curto prazo e portanto necessitam de, constantemente, tomar decisões de curto prazo. Com a crise económico-financeira que paira sobre Portugal, e a dificuldade de acesso ao crédito bancário, uma gestão eficiente do fundo de maneio é cada vez mais preponderante nos dias de hoje.

Os dados usados neste estudo foram obtidos na base de dados SABI e constituem uma amostra de dados em painel de 626 empresas do setor mobiliário português (CAE 31091) no período de 2011 a 2015. A fim de introduzir o efeito do ciclo económico sobre os níveis investidos no fundo de maneio, foram obtidas informações na PORDATA sobre o crescimento anual do PIB em Portugal. Todos os cálculos foram efetuados no R Studio. A rentabilidade operacional dos ativos (ROA), analisada antes dos juros e impostos, foi usada como variável dependente em modelos de regressão linear para dados em painel. Como variáveis independentes foram utilizados o tempo médio de pagamentos (TMP), tempo médio de inventários (TMI), tempo médio de recebimentos (TMR) e o ciclo de tesouraria (CT), como sugerido em Shin & Soenen (1988). O tempo médio de inventários

indica, em média, o tempo de posse dos ativos pela empresa; um maior valor nesta variável representa um maior investimento nos inventários. Somando o tempo médio de recebimentos com o tempo médio de inventários e subtraindo o tempo médio de pagamentos obtemos o ciclo de tesouraria. Como variáveis de controlo, usou-se a dimensão da empresa (que corresponde ao logaritmo dos ativos), o grau de alavancagem (que é o quociente Passivo/Capital próprio), o crescimento das vendas e o crescimento do PIB. As empresas foram ainda caracterizadas pela sua Forma Jurídica (variável binária: empresas por quotas ou outra) e pela Região (variável qualitativa com 7 níveis).

O efeito da gestão de fundo de maneio na rentabilidade das empresas do setor mobiliário português foi testado através da metodologia de dados em painel. Inicialmente considerou-se o modelo de regressão linear (OLS), que não separa a tendência geral da especificidade de cada empresa deste estudo. Por isso, ajustamos modelos de efeitos fixos com o objetivo de comparar com os modelos OLS e escolhermos o mais adequado. O teste F permitiu-nos concluir que o modelo de efeitos fixos era preferível ao modelo OLS. Subsequentemente estimamos modelos de efeitos aleatórios individuais e para o tempo, com o intuito de analisar se estes se traduziriam em alguma melhoria em relação ao modelo de efeitos fixos. Efetuado o teste de Hausman para os modelos em análise (com e sem as variáveis constantes no tempo) a hipótese nula foi rejeitada, assumindo-se assim que existe uma correlação entre os efeitos individuais e as variáveis explicativas (Murteira *et al*, 2016), devendo optar-se por um modelo com efeitos fixos.

Com este estudo chegou-se à conclusão da existência de uma relação negativa entre a variável dependente (ROA) e as variáveis independentes (TMP, TMR, TMI e CT). De um modo geral, a diminuição do ciclo de tesouraria das empresas, tempo médio de recebimentos e tempo médio de inventários potencia o aumento da rentabilidade dos ativos, o que vai de encontro aos resultados esperados. Uma vez que o tempo médio de pagamentos representa o número de dias, em média, que a empresa demora a solver os seus compromissos aos fornecedores, seria expectável que um aumento nesta variável implicasse também um aumento na rentabilidade dos ativos, mas os resultados obtidos nos modelos evidenciam que a rentabilidade dos ativos aumenta com uma diminuição do tempo médio de pagamentos. Relativamente às variáveis de controlo, são todas estatisticamente significativas. Os resultados obtidos demonstram que o crescimento da rentabilidade dos ativos é potenciado por uma maior dimensão da empresa, por um crescimento das vendas e nos períodos de crescimento económico do PIB.

References

Murteira, J.; Castro, V.; Martins, R. (2016) *Introdução à Econometria*. Coimbra: Almedina.

Shin, H.H. and Soenen, L. (1998), Efficiency of working capital and corporate profitability, *Financial Practice and Education* , Vol. 8, pp. 37-45.

García-Teruel,P., Martínez-Solano;P. (2007) Effects of working capital management on SME profitability *International Journal of Managerial Finance*, 3 (2), pp.164-177.

Classificação II – 6ª Feira, 21 de Abril, Auditório B (17h20)

Clustering co-authorship networks across scientific fields using parameter distributions

Conceição Rocha[1], **Paula Brito**[2], **Fernando Silva**[3], **Alípio M. Jorge**[4]

[1] *LIAAD-INESC TEC, Universidade do Porto, Porto, Portugal, conceicao.n.rocha@inesctec.pt*

[2] *Fac. Economia & LIAAD-INESC TEC, Univ. Porto, Porto, Portugal; mpbrito@fep.up.pt*

[3] *Fac. Ciências & CRACS-INESC TEC, Univ. Porto, Porto, Portugal, fds@fc.up.pt*

[4] *Fac. Ciências & LIAAD-INESC TEC, Univ. Porto, Porto, Portugal, amjorge@fc.up.pt*

Abstract: We analyze co-authorship networks, of different scientific fields, and in different time periods. Each network is represented by the distributions of node or edge statistical indicators, in the form of quantile vectors. Hierarchical clustering is then applied, using an appropriate distance and the Ward criterion. This leads to a typology of scientific fields in each time period, on the one hand, and to the identification of stable periods and changing points in each scientific field, on the other hand.

Keywords: Clustering networks, Co-authorship networks, Distributional data, Quantile representation.

It is common knowledge that research collaboration is different across scientific fields. In this work, we focus on the comparison and clustering of different scientific fields based on their research collaboration networks. To this purpose, we use co-authorship networks, where nodes are researchers and the edges indicate the existence of at least one common publication. The data covers 22 scientific fields, and publications ranging from 1972 to 2016, organized in different time periods: before and after 2000, and per year after 2000.

Following previous work by Giordano and Brito (2014), we represent each network by the distributions of statistical indicators, such as degree, eigenvector centrality, closeness, betweeness, clustering coefficient. However, since data cannot be considered uniform by intervals, the representation of distributions by histogram-valued variables is not appropriate. We then use, alternatively, a representation by quantile vectors (Ichino, 2008).

Distributions are compared using the discrete version of the Mallows distance between quantile functions (Levine, Bickel, 2001). However, network parameters cannot be considered independent, their correlation structure should be taken into account, so that the (corresponding discrete version of the) Mahalanobis-Wassertein distance (Verde, Irpino, 2008) is used.

Hierarchical clustering is then applied, using the Ward aggregation index, leading to a typology of scientific fields for each time period.

A separate analysis of co-authorship networks of a given field in different years allows for a better insight of the evolution of research collaboration along time, identifying stable periods and moments of change in the collaborations structure.

Comparing networks by the distributions of network indicators, the proposed approach avoids the direct computation of graph distances, such as the graph edit distance, a problem know to be NP-complete. It should be noticed that in some fields networks easily reach thousands of nodes, so that high complexity algorithms are not easily applicable. Furthermore, the method allows for the choice of the indicators to be used in the clustering process, as well as for the evaluation of their respective importance in a final partition.

Acknowledgments: This work is partially funded by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF) within project "TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020"

References

Giordano,G., Brito, P. (2014). Social Networks as Symbolic Data. In: VICARI, D., OKADA, A., RAGOZINI, G., AND WEIHS, C. (Eds.) *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*. Springer.

Ichino, M. (2008). Symbolic PCA for histogram-valued data. *Proceedings IASC* Conf.

Levina, E., Bickel, P. (2001). The earth mover's distance is the Mallows distance: Some insights from statistics. In: *Proc. Eighth IEEE International Conference on Computer Vision*, Vol. 2, 251-256. IEEE.

Verde, R., Irpino, A. (2008). Comparing histogram data using a Mahalanobis–Wasserstein distance. In: BRITO, P. (Ed.), *Proc. COMPSTAT 2008,* 77-89. Physica-Verlag HD.

Classificação II – 6ª Feira, 21 de Abril, Auditório B (17h40)

Segmentação com dados espaciais Gaussianos. Uma comparação de algoritmos

Luís F. Domingues[1], José G. Dias[2]

[1] *Instituto Politécnico de Beja, Lisboa, Portugal, luis.domingues@ipbeja.pt*

[2] *Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Lisboa, Portugal, jose.dias@iscte.pt*

Sumário: Este estudo compara cinco algoritmos frequentemente utilizados na estimação de campos de Markov com variáveis latentes aplicados a dados simulados no contexto da segmentação de imagem com dados espaciais Gaussianos. O desempenho dos algoritmos é comparado em situações extremas numa perspectiva de imagem, mas de interesse do ponto de vista genérico da segmentação, isto é, quando a componente espacial que os distingue é definida pelas variâncias das distribuições das Gaussianas e não pelos seus valores médios.

Palavras-chave: Algoritmo EM, Dados espaciais, Hidden Markov random fields, Segmentação.

As áreas do conhecimento mais activas na detecção de *clusters* com base em informação espacial são as ciências da computação, em particular no contexto da segmentação de imagens, reconhecimento de padrões e processamento de sinal. Desta forma, não é de estranhar que a maior parte das técnicas e algoritmos mais utilizados nestes contextos tenham a sua origem nestas áreas do conhecimento. Em geral, assume-se que a deturpação ou erro do sinal de recepção conduz a que as cores observadas apresentem o comportamento de variáveis aleatórias Gaussianas centradas na cor original. A correcção da cor de um pixel é feita no pressuposto de a informação dos vizinhos (pixéis mais próximos) contribui para a determinação da probabilidade de pertença do próprio pixel a uma cor, o que induz a existência de um campo de variáveis latentes cuja propriedade Markoviana fica definida pela dependência dos vizinhos mais próximos.

Num caso mais geral, num processo de segmentação de dados espaciais Gaussianos, o objectivo não se restringe à melhor identificação das médias das distribuições das variáveis observadas como no contexto do processamento de imagens. O verdadeiro objectivo de um processo de segmentação de dados espaciais é alocar os dados observados nos segmentos mais prováveis bem como obter as *melhores* estimativas dos parâmetros das distribuições subjacentes. Neste contexto, o objectivo deste trabalho é comparar o desempenho de vários algoritmos de classificação quando a componente espacial que os distingue é fundamentalmente definida pela variância das distribuições Gaussianas e não pela respectiva média. Uma vez que os dados são simulados, a verdadeira classificação e os parâmetros das distribuições são conhecidos, pelo que o desempenho dos algoritmos pode ser avaliado pela percentagem de valores observados correctamente classificados e pela diferença entre as estimativas obtidas e os verdadeiros

valores dos parâmetros das distribuições. A Figura 1 representa o padrão espacial utilizado e na Figura 2 são apresentados alguns dos histogramas das correspondentes funções densidade simuladas utilizados como teste neste estudo.

Figura 1: Padrão utilizado

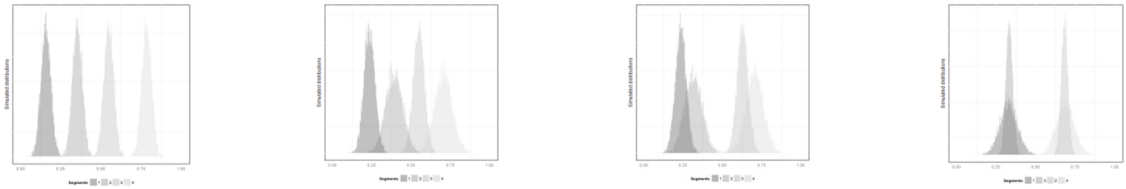


Figura 2: Histogramas das densidades dos segmentos simulados

O estudo compara os algoritmos mais utilizados em processamento de imagem, nomeadamente o algoritmo de classificação *K-means*, o modelo de mistura finita clássico (FMM) que não possuem dependência espacial, e dois algoritmos com interacção espacial: uma adaptação do algoritmo *Iterative Conditional Modes* (ICM) de Besag (1986) proposta por Celeux *et al.* (2003) e uma versão recente do algoritmo *Spatially Variant Finite Mixtures* (SVFM) de Gopel *et al.* (1998) proposta por Blekas *et al.* (2005).

Os resultados mostram que o algoritmo SVFM apresenta as melhores percentagens de valores observados correctamente classificados, seguindo-se o algoritmo ICM e por último os algoritmos sem componente espacial. No que respeita às estimativas dos parâmetros das distribuições, o modelo de mistura clássico apresenta as estimativas mais próximas dos verdadeiros parâmetros das densidades, seguindo-se os algoritmos SVFM, ICM e por último o algoritmo *K-means*.

Referências

Besag, J. (1986) *On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society, Series B-Methodological, 3(48) , 259-302*.

Blekas, K. et al (2005) *A Spatially Constrained Mixture Model for Image Segmentation. IEEE Transactions on Neural Networks, 2(16), 494-498*.

Celeux, G. et al (2003) *EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation. Pattern Recognition, 36, 131-144.*

Sanjay-Gopel, S. and Hebert, T. J. (1989) Bayesian Pixel Classification using Spatially Variant Finite Mixtures and the Generalized EM Algorithm. IEEE Transactions on Image Processing, 7(7), 1014-1028.

Classificação II – 6ª Feira, 21 de Abril, Auditório B (18h00)

Hidden Markov models applied to technical analysis data

José G. Dias
Instituto Universitário de Lisboa ISCTE IUL, BRU-IUL, Lisboa, Portugal, jose.dias@iscte.pt

Abstract: This paper introduces hidden Markov models (HMMs) in the context of technical analysis (candlestick data structures). Model specification, selection, and estimation are discussed. SP500 time series index is used to illustrate the application. Results show that HMMs can reveal the underlying structure in candlestick time series data.

Key-words: Data mining, Hidden Markov model, Technical analysis, Stock markets.

Technical analysis has been a key instrument among traders and financial analysts for many years (Person, 2004). Also known as “charting", it has been a part of financial practice for many decades, because computing power was not available at the time. It uses charts in an attempt to find shapes and patterns that can anticipate trends in financial time series. One of the main difficulties of the technical analysis has been its highly subjective nature as many times the presence of shapes in the historical data is very dependent on the analysts' insights and subjective believes. There are many approaches in technical analysis (e.g., candlestick charting, Dow Theory, and Elliott wave theory), but most analysts combine different techniques between these two extremes: some analysts use subjective judgement to decide patterns and trends; whereas others tend to be more objective using system techniques in the identification of patterns. One of the main instruments of technical analysis is the candlestick chart. It is said to be developed by Munehisa in the beginning of 18th century and it is today a pivotal tool in technical analysis (Azzopardi, 2010). The candlestick fills a box between the open and close prices to emphasize the difference between two consecutive days. Colors may change: usually red and black represent days when close price is lower than open price; white, green, or grey colors are used otherwise. The whiskers represent the maximum and minimum of the day.

This approach has been received with criticism and scepticism in academia, given its descriptive and intuitive visual analysis in reading the market mood. Academic research has taken mostly the fundamental analysis of economic factors that influence the way investor's price financial products (e.g., earnings, market conditions, contagion, dividends) as they do not believe that market's prices reflect all relevant information. In some circles, technical analysis is known as “voodoo finance", and for some researchers the difference between fundamental and technical analysis is the same as between astronomy and astrology. However, despite its jargon and methods, intraday variation contains useful information for market prices.

This paper explores the richness of this type of data by extending hidden Markov models to candlestick data. Thus, it complements the visual inspection of charts with a statistical model that can identify cycles and trends in time series data.

The data set contains daily prices from the Standard & Poor's 500 (SP500) from 2 January 1962 to 21 February 2017 drawn from Yahoo!Finance. The series is denominated in US dollars. For each day t we observed four prices: open, high, low, and close. These prices are arranged in a vector $\boldsymbol{P_t} = (O_t, H_t, L_t, C_t)$ that corresponds to the opening price, the maximum price of the day, the minimum price of the day, and the closing price (corrected for dividends). The daily rates of return are defined as the percentage log-return by $y_t = 100 \times log\ (C_t/C_{t-1}) = 100 \times log\ (P_t/O_t)$. Additionally, we use the range or intraday variation that is given by $r_t = 100 \times log\ (H_t/L_t)$.

Our model is based on an extension of the hidden Markov model, which has been used successfully in financial time series modelling (Rossi & Gallo, 2006). We extended this model by taking a GARCH structure, a skew-t distribution for innovations, and a time-varying structure for transition probabilities between states. Additionally, to the traditional autoregressive structure of the log-returns in means and variances, we consider the intraday variation on means and transition probabilities. Model selection of the number of regimes is based on the Bayesian Information Criterion (BIC).

The HMM with two regimes succeeds in picking the different market structure for bull and bear phases, showing distinct volatility and asymmetry. Additionally, it shows that the intraday range impacts the meaning structure and transition process between regimes.

References

Azzopardi, P. V. (2010). *Behavioural Technical Analysis: An introduction to Behavioural Finance and its Role in Technical Analysis*. Petersfield: Harriman House.

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

Person, J. L. (2004). *A Complete Guide to Technical Trading Tactics: How to Profit using Pivot Points, Candlesticks & other Indicators*. Hoboken, N.J.: Wiley.

Rossi, A. & Gallo, G. M. (2006). Volatility estimation via hidden Markov models. *Journal of Empirical Finance*, 13(2), 203-230.

Séries Temporais e Dados Espaciais – 6ª Feira, 21 de Abril, Auditório A (17h00)

Spatial Econometrics Modeling of the *Saúde24* line calls

Paula Simões[1], M. Lucília Carvalho[2], Sandra Aleixo[3], Sérgio Gomes[4], Isabel Natário[5]

[1] Centro de Matemática e Aplicações (CMA), FCT, UNL and Área Departamental de Matemática, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal, paulasimoes@adm.isel.pt

[2] Centro de Estatística e Aplicações (CEAUL), Faculdade de Ciências, Universidade de Lisboa, Portugal, mlucilia.carvalho@gmail.com

[3] Centro de Estatística e Aplicações (CEAUL), Faculdade de Ciências, Universidade de Lisboa and Área Departamental de Matemática, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal, sandra.aleixo@adm.isel.pt

[4] Direção Geral de Saúde (DGS), Portugal, sergiogomes@dgs.pt

[5] Centro de Matemática e Aplicações (CMA), FCT, UNL and Departamento de Matemática, FCT, UNL, Portugal, icn@fct.unl.pt

Abstract: Saúde24 line (LS24), a Portuguese national health line, was created in 2007, by the Health Ministry in order to better respond to citizens' health needs, by expanding and improving accessibility to existing services, as well as rationalizing the use of existing resources by directing users to the most appropriate institutions of the public national health services. For LS24 data, the location attribute is an important source of information to describe its use. This study analyzes the calls received, at the municipal level, under a spatial econometrics approach, aiming to describe and evaluate the use of LS24, for future development of decision support indicators.

Keywords: Bayesian Analysis, Spatial Econometrics, Hierarchical models, Autoregressive models, Poisson.

An initiative to improve accessibility and to rationalize the use of existing resources was carried out by the Portuguese Health Ministry through the creation of a national health line, LS24, in April of 2007. These objectives are achieved by directing users to the most appropriate institutions of the public national health service or by counseling self-care home measures. Aiming a future development of decision support indicators in a hospital context, based on the economic impact of the use of LS24, rather than resorting to horpital urgency services, the present study analyses the number of calls to LS24, in 2014, at the municipal level, since for LS24 data the location attribute is an important source of information to describe its use. The methodology proposed consisting of analyzing the received calls classified as Triage, counseling and routing (TCR) in disease situations, through spatial models, describing and evaluating the use of LS24 using both a hierarchical and an autoregressive perspective.

Using hierarchical Poisson bayesian models, the spatial structure assumed for the risk of what is being counted is included, in the first level of the hierarchy, through a prior

distribution of spatially structured random effects. Additionally, some covariates might be considered for their relevance, as well as some non-spatially structured random effects, to account for risk variation not yet explained. These models are better considered under the bayesian paradigm and their inference needs to be based on simulation methods, namely the Markov Chain Monte Carlo (MCMC) (Banerjee et al. 2004). Within the scope of spatial autoregressive models, there are alternatives for modeling counts, such as the ones we explore here: an expanded standard spatial lag model, recently developed within an Integrated Nested Laplace Approximations (INLA) framework, by Gómez-Rubio, Bivand and Rue in 2015; a spatial autoregressive lag model of counts, developed by Lambert, Brown and Florax in 2010, under a classical perspective; additionally, a spatial lag autoregressive component is incorporated in the model for counts, using INLA methodology, combining the two abovementioned models, resulting in a spatial lag Poisson model. This is implemented in R-INLA (Blangiardo et al. 2015).

In our study, firstly standard spatial econometrics techniques are used to look for spatial dependence in the number of calls to LS24 in each municipality by considering a neighborhood contiguity structure, as well as in the residuals of a baseline log-Poisson model with covariates. The covariates are selected under different scenarios, and only two of the most relevant ones are analyzed. The number of calls is then further analyzed by using the hierarchical and autoregressive methods described above, resulting in similar conclusions for both perspectives. Due to their relevance in explaining the use of LS24, the average number of years of schooling stands out, for scenario 1 of the analysis, and the average monthly income stands out, for scenario 2. Additionally, the spatial component for both cases was quite relevant, as confirmed by the high estimates obtained for the spatial autocorrelation parameter. We consider that in possession of the results of this study, and in cooperation with the Portuguese Directorate-General of Health, it will be possible, in a near future, to analyze, test and implement different management government policies at the hospital level, or even to predict future behaviours, under distinct scenarios. The savings from the properly use of LS24, by avoiding urgent healthcare in a hospital, can then be channeled for other needed areas.

Acknowledgments: This work is financed by national funds through FCT - Foundation for Science and Technology - under the projects UID/MAT/00297/2013 and UID/MAT/00006/2013.

References

Banerjee, S., Carlin, B. & Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data,* Chapman and Hall/CRC.

Blangiardo, M. & Cameletti, M. (2015) *Spatial and Spatial-temporal Bayesian Models with R-INLA* Wiley.

Goméz-Rubio, V., Bivand, R. & Rue, H. (2015) A new latent class to fit spatial econometrics models with Integrated Nested Laplace Approximations. *Spatial Statistics: Emerging Patterns-Part2,* 27, 116-118.

Lambert, D., Brown, J. & Florax, R. (2010) A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application. *Regional Science and Urban Economics,* 40, 241-252.

Séries Temporais e Dados Espaciais – 6ª Feira, 21 de Abril, Auditório A (17h20)

MIBEL Electricity Prices Forecasting

Eliana Costa e Silva[1], Ana Borges[2]
[1] CIICESI, ESTG Polytechnic of Porto & Centro Algoritmi, University of Minho, eos@estg.ipp.pt;
[2] CIICESI, ESTG Polytechnic of Porto, aib@estg.ipp.pt

Summary: At the 119th European Study Group with Industry, the Company EDP proposed a challenge concerning electricity prices simulation with the main purpose of short-term Electricity Price Forecasting (EPF). In this work a vector autoregressive model (VAR) with exogenous variables is proposed for short-term EPF. The results show that the VAR model, with the season of the year and the type of day as exogenous variables, explains the intra-day and intra-hour dynamics of the hourly prices, and replicates the daily pattern.

Keywords: Electricity Price Forecasting, Exogenous Variables, Multivariate Time Series, Vector Autoregressive Models.

The study of times series data is considered one of the 10 challenges in data mining (Yang & Wu, 2006). In particular, motif discovery, i.e. the discovery of interesting patterns, is non-trivial and has become one of the most important tasks in data mining. A large variate of methods has been proposed to deal with Electricity Price Forecasting (EPF). Weron (2014) reduces the existing methods into five major categories: (i) multi-agent models, (ii) fundamental models, (iii) reduced-form models, (iv) statistical models and (v) computational intelligence models. However, the frontier between these categories is not always clear. For instance, several statistical methods are classified as computational intelligent models. According to Weron (2014) statistical methods present some of the most promising results. The statistical EPF models, for short term price forecasting, that have been widely used are: (i) univariate AutoRegressive models (AR); (ii) AutoRegressive Moving Average models (ARMA); (iii) AutoRegressive Integrated Moving Average (ARIMA); (iv) Seasonal ARIMA models (SARIMA).

EPF literature has mainly concerned on models that use information at daily level. In the present analysis the interest focus on forecasting intra-day prices using hourly data (disaggregated data), therefore, VAR models (that explore the complex dependence structure of the multivariate price series) are here proposed. To capture the seasonal pattern of the process mean, the exogenous variables: (i) day type (working day vs. weekend) and (ii) the annual seasons defined in terms of meteorological conditions, were added yielding a VARX model.

A day-ahead market consists in a system where agents submit their bids and offers for the delivery of electricity during each hour of the next day before a certain market closing time (Weron, 2014). The electricity prices, (available online at www.mibel.com) is given as a strip of prices, all simultaneously observed once at a given time of each day. In the present analysis the hourly prices from March 10th 2014 to June 26th 2016, in a total of

20205 observations, were used (see Costa e Silva *et al* (2017) for more details). The data analysis was performed using R (version 3.3.0). After the log transformation, for augmented Dickey-Fuller Test, one may assume that the time series is stationary. The existence of linear dynamic dependence in the data is supported by the multivariate Ljung–Box test.

A VARX(7,0) was estimated using hourly prices $Y_{kt}, k = 1, \dots, 24$, from March 10th 2014 to May 29th 2016, in a total of 19461 observations: $\boldsymbol{Y}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{7} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-1} + \boldsymbol{\beta}_1 \boldsymbol{X}_{1t} + \boldsymbol{\beta}_2 \boldsymbol{X}_{2t} + \boldsymbol{a}_t$, where $\boldsymbol{\phi}_0$ denotes the constant vector, $\boldsymbol{\Phi}_i$ are 24×24 matrices of autoregressive parameters, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the exogenous variables and $\boldsymbol{a}_t$ the residuals. A total of 4104 parameters were estimated. The autoregressive coefficients of the VARX(7,0) explain the existence (or not) of dependence within the hourly prices. The results show that there are several non-diagonal elements of $\boldsymbol{\Phi}_i$ that are nonzero, therefore there are several hourly prices that are dynamically correlated to prices of other hours in previous days.

The hourly prices from March 30th to June 28th 2016 were used for comparison between the real (log) hourly prices and the forecast of the VARX(7,0) model. For evaluating the point forecasts the root mean square errors (RMSE) was used. Although the forecast does not exactly replicate the real price they are quite similar, showing approximately the same pattern: decrease in prices after 10PM until 3AM, an almost constant price until to 5AM, increasing prices until 10AM, constant prices until 6PM, increasing prices until 10PM. The RMSE is below 34. For instance, the introduction into the model of exogenous variables that better explain the meteorological conditions could improve the model estimates. Moreover, further analysis of the orders of the VARX should be performed.

Acknowledgment: This work was partially support by ESGI119 - an initiative supported by COST Action TD1409, Mathematics for Industry Network (MI-NET), COST is supported by the EU Framework Programme Horizon 2020. The authors were supported by Center for Research and Innovation in Business Sciences and Information Systems (CIICESI), ESTG - P.Porto. We would like to thank M.F. Teodoro and M.A.P. Andrade for the discussions during ESGI119 and R. Covas for bringing this challenge to ESGI119 (www.estg.ipp.pt/esgi).

References

Costa e Silva, E., Borges, A., Teodoro, M.F., Andrade, M.A.P. & Covas, R. (2017) Time Series Data Mining for Energy Prices Forecasting: An Application to Real Data. IN A.M. Madureira et al. (Eds.), Intelligent Systems Design and Applications, Advances in Intelligent Systems and Computing 557.

Weron, R. (2014) Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4), 1030-1081.

Yang, Q. & Wu, X. (2006) 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04), 597-604.

Séries Temporais e Dados Espaciais – 6ª Feira, 21 de Abril, Auditório A (17h40)

Forecasting Tourism Demand with SVARMA models

Cristina Lopes[1], Eliana Costa e Silva[2], Filomena Soares[3]
[1] LEMA & CEOS.PP, ISCAP – Polytechnic of Porto, cristinalopes@iscap.ipp.pt
[2] CIICESI, ESTG – Polytechnic of Porto, eos@estg.ipp.pt
[3] ESHT- Polytechnic of Porto, filomenasoares@estg.ipp.pt

Abstract: We present a multivariate time series approach to forecast the number of overnight stays in tourist accommodations (hotels, guesthouses, and apartments) in the portuguese region of Algarve. We adjust a seasonal vector autoregressive and moving averages model (SVARMA) to ten year monthly data and compared the forecast values with the actual values of the overnight stays in 2016. These forecast values can be used by a hotel manager to predict their occupancy and to generate the best pricing policy.

Keywords: Hotel Occupancy, Seasonality, Time Series Forecasting, Seasonal Vector Autoregressive and Moving Averages Model.

Tourism has been a key factor for the growth of the Economy and for the development of some regions in Portugal. Foreign tourists' expenditures have allowed Portugal to maintain a smaller deficit in its balance of payments, a higher economic growth and sustainable development in some regions popular with tourists (Turismo de Portugal, 2007). The Algarve region is particularly focused on attracting foreign tourists and has built over the years a large offer of diversified hotel units, from the most basic hostels and guest houses to luxury hotels and resorts. Nowadays, a hotel room rate is rarely fixed; hotel managers are interested in using revenue management strategies to establish the best pricing policy, which should be adjusted to the demand.

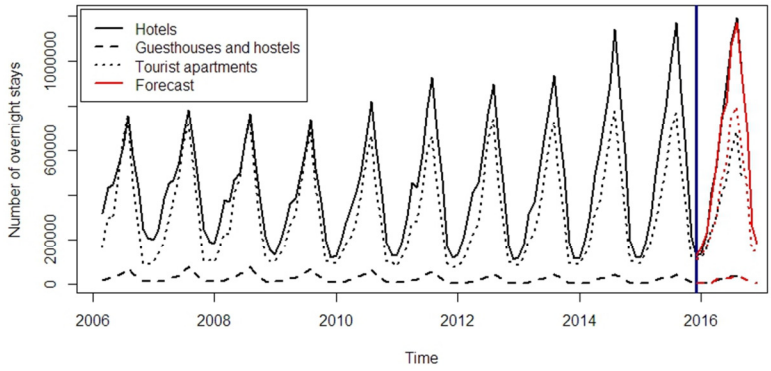
The data set considered in this work consisted of three variables concerning the number of overnight stays in: (i) hotels, H_t, (ii) guesthouses or hostels, G_t, and (iii) tourist apartments, A_t, all of them located in Algarve. Monthly data from March 2006 until September 2016 (127 observations) was collected from INE. All computations were performed with R Studio.

The three time series showed seasonality and trend, therefore seasonal models were required (Brockwell & Davis, 1996). The number of overnight stays in hotels showed a linear trend and seasonal effects depending on the level, as well as the number of stays in tourist apartments (Fig. 1). The guesthouses series showed seasonality and no trend, with a considerable lower level than the two other series. We first analysed the three series individually and adjusted a seasonal autoregressive integrated and moving averages model (ARIMA) to each one, using the order of the integration filters as decided by the KPSS and OCSB tests (Osborn *et al*, 1988). The best model obtained for the guesthouses series G_t, was an ARIMA$(3,1,0)(0,1,0)_{12}$; for the Apartment series, A_t was an ARIMA$(3,1,1)(2,1,0)_{12}$; for the Hotel series, H_t, was an ARIMA$(2,1,0)(1,1,1)_{12}$. The Ljung

Box test showed a significant autocorrelation of the residuals, meaning that there was still some information in the data not being captured by the model.

We found that there were significant cross-correlations between the three series, therefore we decided to analyse these three time series, not independently, but as one multivariate time series, using a Seasonal Vector Autoregressive Integrated Moving Averages model (SVARMA) (Tsay, 2014). Several SVARMA models were considered. The model which obtained the best AIC (55.9827) was a SVARMA(3,1,0)(1,1,1)$_{12}$ model. This model was further refined by removing the non significant coefficients in the model. This led to a model with smaller AIC= 55.5036 and BIC= 55.9585. Having validated the SVARMA model with the Shapiro-Wilk normality test and the multivariate Ljung-Box Q statistical test, it was applied iteratively to compute the forecast vectors $\hat{Z}_t$ of the hotel occupancy for the months between January 2016 and December 2016. Figure 1 presents the observed time series and the forecasted values.

Figure 1: The multivariate series of overnight stays at Hotels, Guesthouses and Tourist Apartments and its forecast values computed with the SVARMA(3,1,0)(1,1,1)$_{12}$ model



The forecast values were compared with the data that became available meanwhile and the mean absolute percent error (MAPE) for the number of stays was below 5% for Hotels, below 15% for Apartments, and a not as good MAPE=27% for Guesthouses. These forecasts could be used by a hotel manager to generate the best pricing policy and increase their profits even when the demand is not at its peak.

References

Turismo de Portugal (2007) *Plano estratégico nacional do turismo (PENT) para o desenvolvimento do turismo em portugal,*. Lisboa: Turismo de Portugal. Retrieved from: http://www.turismodeportugal.pt/Portugu%C3%AAs/turismodeportugal/publicacoes/Documents/PENT%202007.pdf.

Brockwell, P.J., Davis, R.A. (1996) *Introduction to Time Series and Forecasting*. New York: Springer-Verlag.

Osborn,D., Chui, A., Smith, J. & Birchenhall, C. (1988) Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics,* 50, 361–377.

Tsay, R. (2014) *Multivariate Time Series Analysis with R and Financial Applications.* New Jersey: Wiley.

Séries Temporais e Dados Espaciais – 6ª Feira, 21 de Abril, Auditório A (18h00)

Previsão das dormidas mensais na região Norte de Portugal em anos recentes: um estudo preliminar

Joaquim Silva[1], Hugo Alonso[2], Isabel Silva[3]

[1] Universidade Lusófona do Porto, joaquimdasilva2014@outlook.pt;

[2] Universidade Lusófona do Porto e Universidade de Aveiro, hugo.alonso@ua.pt;

[3] Faculdade de Engenharia, Universidade do Porto, ims@fe.up.pt

Sumário: O setor do Turismo é um dos mais importantes para a economia da região Norte de Portugal e tem crescido de ano para ano. Este trabalho considera dados recentes das dormidas mensais nos alojamentos turísticos da região, analisa-os e compara os resultados de previsões obtidas por meio da aplicação de redes neuronais artificiais e da análise espectral singular.

Palavras-chave: Análise espectral singular, Dormidas, Previsão, Redes neuronais artificiais, Região Norte de Portugal.

O setor do Turismo tem uma importância estratégica para a região Norte de Portugal, ajudando na criação de emprego e na captação de divisas para diversos setores da economia, e está em franco crescimento, nomeadamente no que respeita ao número de dormidas, número de hóspedes e proveitos globais (Silva, 2017).

Um problema atual e relevante para a gestão dos alojamentos turísticos é o da previsão das dormidas mensais. Resolvê-lo é importante para se ter uma visão antecipada da procura turística e com isso ser mais fácil gerir os alojamentos e criar condições mais favoráveis à boa receção dos visitantes. Verifica-se que são vários os estudos na literatura que procuram dar-lhe resposta para a região Norte de Portugal, de acordo com a revisão bibliográfica em Silva (2017). Tais estudos propõem metodologias baseadas em modelos de regressão lineares, modelos integrados mistos auto-regressivos e médias móveis sazonais e redes neuronais artificiais e apresentam bons resultados. No entanto, os dados considerados não vão além de dezembro de 2010. Sucede que o facto de o Porto ter sido considerado o melhor destino europeu nos anos de 2012 e 2014 pode ter ajudado a atrair mais turistas, não só para a cidade, mas também para a região Norte de Portugal. Assim, este e outros fatores podem ter contribuído para uma mudança no comportamento da série das dormidas mensais na região em anos recentes. Neste contexto, importa perceber se os bons resultados das previsões se mantêm.

O presente trabalho considera dados atuais das dormidas mensais na região Norte de Portugal, mais propriamente dados até dezembro de 2015. A série das dormidas é analisada e compara-se a qualidade das previsões realizadas por meio da aplicação de redes neuronais artificiais (Haykin, 2009) e da análise espectral singular (Golyandina *et al.*, 2001). Os resultados das previsões das redes neuronais artificiais já foram apresentados em Silva (2017).

As redes neuronais artificiais são sistemas não lineares, formados por um conjunto intrincado de subsistemas elementares, representativos de neurónios, que têm a capacidade de aprender a partir de dados conhecidos, sem assumirem *a priori* hipóteses sobre a distribuição desses dados, e de generalizarem quando são considerados dados novos. Além disso, as redes são robustas na presença de perturbações e podem ser facilmente adaptadas para lidarem com mudanças num ambiente não estacionário. Mais ainda, têm uma estrutura distribuída massivamente paralela que as faz exibir um elevado poder computacional e capturar um comportamento verdadeiramente complexo de uma forma altamente hierarquizada. Apesar de complicados, estes sistemas são aqui aplicados de forma relativamente simples, depois de estarem definidos, bastando-lhes indicar o ano e o mês para os quais se pretende obter uma previsão das dormidas.

A análise espectral singular (SSA) é uma técnica não paramétrica que permite decompor a série original num pequeno número de componentes independentes e interpretáveis, que podem ser consideradas como componente de tendência, oscilatórias e de ruído. Uma das vantagens da SSA é que não é necessário que a série em análise seja estacionária. A previsão de valores futuros da série temporal é calculada através dos coeficientes de uma fórmula de recorrência linear.

Agradecimentos: Este trabalho foi parcialmente financiado pela Fundação Portuguesa para a Ciência e Tecnologia (FCT), através do CIDMA - Centro de Investigação e Desenvolvimento em Matemática e Aplicações, no âmbito do projeto UID / MAT / 04106/2013.

Referências

Golyandina, N., Nekrutkin, V. & Zhigljavsky, A. (2001) *Analysis of Time Series Structure: SSA and Related Techniques*. New York, USA: Chapman & Hall.

Haykin, S. (2009) *Neural networks and learning machines* (3rd Edition). New Jersey: Pearson Education, Inc.

Silva, J. (2017) *Previsão das dormidas mensais nos alojamentos turísticos da região Norte de Portugal* (Dissertação de Mestrado). Universidade Lusófona do Porto, Porto, Portugal.

Modelos com Variáveis Latentes – Sábado, 22 de Abril, Auditório A (9h30)

Modelling wage trajectories in Brazil using data from RAIS Identificada

Maria de Fátima Salgueiro[1], Marcel D.T. Vieira[2], Ricardo Freguglia[3]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, fatima.salgueiro@iscte.pt;

[2] Universidade Federal de Juiz de Fora, Brasil, marcel.vieira@ice.ufjf.br;

[3] Universidade Federal de Juiz de Fora, Brasil, ricardo.freguglia@ufjf.edu.br

Abstract: This paper discusses the use of the latent growth curve modelling framework to model wage trajectories of Brazilian formal workers, using data from RAIS Identificada, for the period 2006-2013. Growth curve models with different types of restrictions are presented. The advantages of the proposed approach over the classical econometric models (with fixed and random effects) are discussed.

Key words: Fixed-effects model, Latent growth curve model, RAIS Identificada, Random-effects model.

Several studies have been conducted in various countries aiming at understanding wage trajectories for individuals with similar characteristics and jobs. This has been an interesting and longstanding issue in labour economics. Indeed, the true identification of the nature of the observed wage trajectories is important on the grounds of policy, research and individual welfare. There are some well-established facts in the economics literature on the analysis of individual wage determination, namely that education level, the type of occupation and the experience of the worker play a key role. Additionally, there is evidence that the type of industry, the size of the company and the geographical region also need to be taken into account.

Two classical approaches to modelling longitudinal data are often considered by econometricians: fixed-effects models (FE) and random-effects models (RE). Such models provide a way to control for all the time-invariant unmeasured variables (whether known or unknown) that influence the dependent variable. However, these models are not flexible enough, as pointed out by Bollen & Brand (2010). They also have some limitations: in FE models it is not possible to estimate the effect of observed time-invariant variables; estimated coefficients are fixed over time; no effects from the lagged dependent variables are allowed; residual error variances are equal across all waves; regarding unobserved heterogeneity, latent time-invariant variables either freely correlate with all time-varying covariates, as in FE models, or they must be uncorrelated with all covariates, as in RE models.

There is a vast literature in econometrics suggesting models and estimators that, separately, try to address some of these limitations. The current paper proposes using the flexibility of the latent growth curve modelling framework (Bollen & Curran, 2006) to model wage trajectories. Brazilian data from the Ministry of Labor and Employment (RAIS

Identificada), available for formal workers in Brazil for the period 2006 to 2013, are used.

Due to the large number of observations in the original dataset, a sample of 1% of workers was considered. The selected individuals were followed over the eight years in analysis. A total of 545184 pooled and balanced observations was considered, corresponding to 68148 male workers, aged between 16 and 65 years old, with positive wage and employed on the 31st December in each year.

The (logarithm) of the wage per hour worked is the variable of interest. Overall, there is a positive growth trajectory for the average wage per hour, with evidence of significant variability by sector of activity (industry; building; services; public administration; agriculture); region of the country (North; Northeast; Southeast; South; Midwest); size of the company (<99; 100-499; >= 500 workers); occupation of the worker (manual; non-manual); education level (basic; high school; college); and tenure (number of months the worker has been at the current employer).

Classical fixed and random effects models are considered and estimated in STATA. Latent growth curve models, with different types of restrictions, estimated using Mplus, are then considered. The specifications required to make models (and estimated results) comparable are highlighted. The flexibility and advantages of the proposed structural equation modelling approach over the classical econometric fixed and random effects models are discussed.

References

Bollen, K.A. & Curran, P.J. (2006) *Latent Curve Models – A Structural Equation Perspective,* New Jersey, USA: John Wiley & Sons.

Bollen, K.A. & Brand, J.E. (2010) A General Panel Model with Random and Fixed Effects: A Structural Equations Approach. *Social Forces,* 89**,** 1-34.

Modelos com Variáveis Latentes – Sábado, 22 de Abril, Auditório A (9h50)

O efeito da escolha do método de estimação em modelos com variável latente contínua: Um estudo de simulação

Paula C.R. Vicente[1], Maria de Fátima Salgueiro[2]
[1] Universidade Lusófona de Humanidades e Tecnologias – Escola de Ciências Económicas e das Organizações, paula.vicente@ulusofona.pt;
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Lisboa, Portugal, fatima.salgueiro@iscte.pt

Sumário: Este trabalho apresenta os resultados de um estudo de simulação realizado recorrendo ao pacote estatístico Mplus 7, com o qual se pretende aferir o efeito da escolha do método de estimação em modelos de análise fatorial confirmatória e em modelos com trajetória latente, face a amostras de pequena dimensão. Os métodos de estimação considerados são a máxima verosimilhança (ML), os mínimos quadrados generalizados (GLS) e os mínimos quadrados ponderados (WLS).

Palavras-chave: Análise Fatorial Confirmatória, Máxima Verosimilhança, Mínimos Quadrados Generalizados, Mínimos Quadrados Ponderados, Modelos com Trajetória Latente.

Modelos de análise fatorial confirmatória (como o representado na figura 1) e modelos com trajetória latente (como o especificado na figura 2) são dois exemplos de modelos com variável latente que integram o *framework* dos modelos com equações estruturais (Bollen, 1989 e Bollen & Curran, 2006).

A utilização destes modelos nas áreas das ciências sociais tem sido crescente nos últimos anos. Todavia, a impossibilidade de dispor de amostras de dimensão tida por razoável é muitas vezes uma realidade com a qual os investigadores destas áreas se confrontam.

Assim, e com o objetivo de auxiliar os investigadores que utilizam este tipo de modelos, é apresentado um estudo de simulação em que se pretende testar a estabilidade das estimativas obtidas face à utilização de diferentes métodos de estimação. São usadas amostras de diferentes dimensões, com diferentes números de variáveis observadas ou momentos temporais (dependendo do modelo) e com diferentes valores para os parâmetros populacionais. O estudo de simulação é realizado recorrendo ao pacote estatístico Mplus 7, Muthén & Asparouhov (2002).

Neste estudo é considerado um modelo de análise fatorial confirmatória com dois fatores latentes, cada um deles medido por 2, 3 ou 4 indicadores, com os pesos fatoriais de 0.6 e 0.8, a que correspondem valores de fiabilidade de 0.36 e 0.64, respetivamente. Os dois fatores foram considerados correlacionados entre si a 0.1, 0.25 e 0.5.

É também considerado um modelo com trajetória latente com 3 e 4 momentos temporais, com os parâmetros populacionais: média do intercepto e do declive fixos a 0;

variância do intercepto igual a 1; variância do declive 0.2; covariância entre intercepto e declive 0; e variância dos termos residuais fixa a valores que permitem obter uma fiabilidade de 0.5 para todos os indicadores, em cada momento temporal.

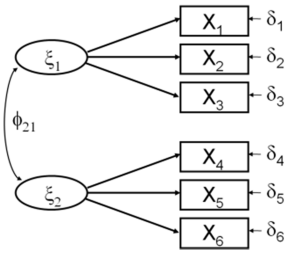


Figura 1: Modelo de Análise Fatorial Confirmatória com dois fatores, cada um deles medido por três indicadores

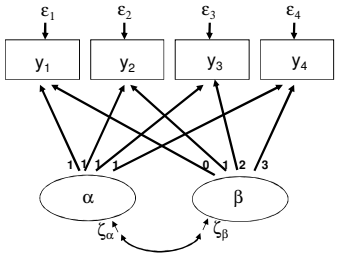


Figura 2: Modelo com trajetória latente com quatro momentos temporais

Para ambos os modelos em estudo são geradas 1000 amostras de dados com distribuição normal, com dimensão de 100, 250 e 200 observações. Os métodos de estimação considerados são ML, GLS e WLS, os recomendados na literatura da área e que se encontram disponíveis no pacote estatístico Mplus 7.

Os resultados obtidos mostram que, nos dois tipos de modelo considerados, em amostras de pequena dimensão, o erro quadrático médio é maior usando WLS e menor usando ML. Nos modelos de análise fatorial confirmatória com dois ou três indicadores e baixa correlação entre fatores, para amostras de pequena dimensão, existe um enviesamento no valor estimado dessa mesma correlação superior a 5%, qualquer que seja o método de estimação usado. Em modelos com trajetória latente o enviesamento obtido na estimação dos parâmetros é inferior a 5%.

Referências

Bollen, K. (1989) *Structural Equations with Latent Variables*, New York, USA: Wiley.

Bollen, K.A. & Curran, P.J. (2006) *Latent Curve Models – A Structural Equation Perspective,* New Jersey, USA: John Wiley & Sons.

Muthén, B.O. & Asparouhov, T. (2002) Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. *Mplus Web Notes,* 1**,** version 2.

Modelos com Variáveis Latentes – Sábado, 22 de Abril, Auditório A (10h10)

Relationship between sustainability and competitiveness of nations: a macro level analysis

Nikolai Witulski[1], José G. Dias[2], Catarina Roseta Palma[3]
[1] Instituto Universitário de Lisboa ISCTE IUL, BRU-IUL, Lisboa, Portugal, nwiii@iscte.pt;
[2] Instituto Universitário de Lisboa ISCTE IUL, BRU-IUL, Lisboa, Portugal, jose.dias@iscte.pt
[3] Instituto Universitário de Lisboa ISCTE IUL, BRU-IUL, Lisboa, Portugal, catarina.roseta@iscte.pt

Abstract: Sustainability is conceptualized either by three equally important dimensions – economic, social, and environmental – or by adding a fourth pillar, the institutional dimension. Results of the measurement model indicate that sustainability is better measured by four dimensions. A second-order confirmatory factor model reveals a positive correlation between sustainability and competitiveness.

Key-words: Competitive sustainability, confirmatory factor analysis, structural equation model, country-level analysis.

Climate change has become one of the greatest challenges of humankind, resulting for instance in higher sea levels and stronger storms. The magnitude of these impacts is expected to intensify, if no measures are taken soon. The 2015 United Nations Climate Change Conference in Paris was a milestone in this debate and defined the goal of a maximum increase of 1.5° C in global average temperature. This increasing global awareness encourages changes in current lifestyle towards more sustainable standards and development paths. The Brundtland report established the foundation for all research related to sustainable development, and it defines sustainability based on three equally important pillars (social, economic, and environmental). Yet, there are researchers who argue in favor of a fourth dimension: the institutional dimension. Thus, our first research question is: Is sustainable development better measured by three or four pillars, given empirical data?

Moreover, is it important that the development of nations is not only sustainable, but also competitive. This latter concept has evolved since the 18th century and tends to be equated with the productivity of nations. The interconnection between competitiveness and sustainability has now emerged as a relevant new focus for researchers. Many scholars and politicians realize that it is important not only to look at both topics separately, but also to combine them, i.e., explore further the relationship between them, which can be termed *competitive sustainability*. Results from the World Economic Forum (2014) and dos Santos & Brandi (2014) indicate a relationship between competitiveness and (environmental) sustainability. Thus, the second question is: Is there an empirical relationship between competitiveness and sustainability?

This paper explores the relationship between sustainability and competitiveness at a macro level (138 countries) using a second-order latent-variable model. Figure 1 depicts the conceptual model. Each dimension of the model is measured by a set of indicators, namely the economic dimension by four indicators (e.g., *GNI p.c.,* and *Household final consumption expenditure*), the institutional pillar by five indicators (e.g., *Voice and accountability,* and *Rule of law*), the environmental dimension by five indicators (e.g., *Greenhouse gases,* and *Water resources),* the social dimension by six indicators (e.g., *Sufficient food,* and *Safe sanitation*), and the competitiveness dimension by the twelve pillars of the Global Competitiveness Index. The fit of the measurement models of the four dimensions of sustainability (social, institutional, economic and environmental), the second-order measurement model (which measures the four-dimensional concept of sustainability), and the competitiveness dimension are all within the acceptable range regarding the goodness of fit. The SRMR is below 0.1, CFI and TLI are close to 0.9, and the chi square test is in the acceptable range. To answer the first research question, all indicators of the social and institutional dimension are combined into a new dimension 'social 2', thereby creating three instead of four pillars of sustainability (social 2, environmental and economic). The measurement model (social 2) has a poor model fit with all values out of the admissible range, which shows lack of reliability of the measures. The second order measurement model (three pillars) shows convergence problems, probably caused by the poor model fit of social2. Implying, that sustainability is better represented by four than three dimensions.

The goodness of fit of the model that includes competitiveness and sustainability (four pillars) is within the recommended threshold. Results reveal a positive correlation between sustainability and competitiveness, i.e., competitiveness and sustainability co-vary in the same direction.

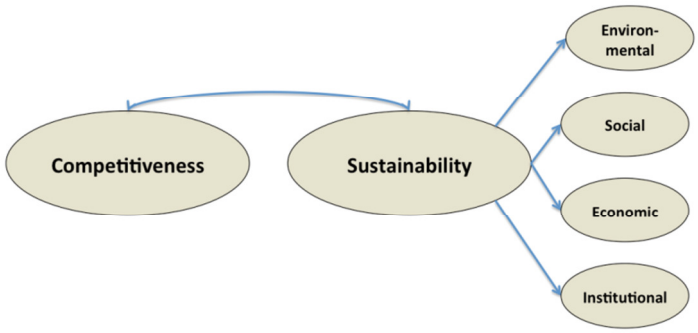


Figure 1: Conceptual Model

References

dos Santos, S. F., & Brandi, H. S. (2014). A canonical correlation analysis of the relationship between sustainability and competitiveness. Clean Technologies and Environmental Policy, 16(8), 1735-1746.

World Economic Forum. (2014). The global competitiveness report 2014-2015. Retrieved from http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2014-15.pdf

Aplicações em Ciências da Saúde II – Sábado, 22 de Abril, Auditório B (9h30)

Detectar anomalias a fim de facilitar o diagnóstico médico por meio de agrupamento de dados e abstrações temporais

Giovana Jaskulski Gelatti[1], André P.L.F. de Carvalho[2], Pedro Pereira Rodrigues[3]

[1] Laboratório de Análise de Massiva de Dados (ANALYTICS) - USP, giovanagelatti@usp.br;
[2] Laboratório de Análise de Massiva de Dados (ANALYTICS) - USP, andre@icmc.usp.br;
[3] Center for Health Technology and Services Research – CINTESIS, pprodrigues@med.up.pt

Sumário: A análise dos dados médicos está cada vez mais custosa devido ao montante de dados gerados. Como auxílio, os dados podem ser sumarizados por meio de técnicas de abstrações temporais. Esse procedimento, além de assistir a análise feita pelo especialista, facilita o processamento dos dados por outras técnicas como de detecção de anomalias. Anomalias podem representar uma nova doença, potencial problema de saúde ou pacientes que necessitam de uma análise criteriosa.

Palavras-chave: Abstração temporal, Detecção de anomalias, *Outlier.*

Detecção de anomalias (DA) é uma tarefa que visa identificar observações que se desviam das demais em um conjunto de dados, sendo também conhecida na literatura pelos termos *outlier,* exceção, anormalidade, entre outros. Um *outlier* em um conjunto de dados pode representar um novo evento a ser investigado (Tan, 2006). Por exemplo, uma nova doença ou um potencial problema de saúde. Um paciente que possui perfil de doente e é saudável também é encontrado pela detecção de anomalias. Numa análise atenta pode-se descobrir comportamentos ou características desse paciente que podem estar retardando doenças. É o caso da aplicação de detecção de anomalias em dados de pacientes com doença de Alzheimer (DA) e Comprometimento Cognitivo Leve (CCL). O CCL é preditor de DA. Indivíduos com CCL, se não diagnosticados, podem evoluir para DA. Quando identificados padrões que indicam a mudança de CCL para DA em estágio inicial pode-se retardar o avanço da doença ou até mesmo reverter o diagnóstico.

Com a superlotação em hospitais, atual problema mundial, e crescimento do volume de dados, torna-se impossível a análise criteriosa do especialista nos dados e identificação dos casos relatados. Por isso, são utilizados métodos de abstrações aos dados. Uma revisão sistemática das técnicas que exerçam essa função está sendo realizada.

As abstrações facilitam a compreensão e sumarizam dados que são relevantes para a análise. Tais técnicas resumem o estado do paciente ao longo de um período, identificam problemas (Shahar et al., 1996) e podem facilitar o processamento dos dados sumarizados para o uso desses por outras técnicas.

Técnicas que identifiquem observações anômalas por meio do encontro de padrões podem auxiliar o especialista na análise e diagnóstico, podendo inclusive gerar conhecimento novo. O objetivo deste estudo é encontrar tais dados que fogem ao padrão

e/ou representam uma mudança de estado ou caso a ser investigado. Os dados reais são, em sua maioria, não rotulados, cenário comum na DA (Eskin et al., 2002). Por esse motivo, foi realizada uma revisão sobre técnicas não supervisionadas de detecção de anomalias. No estudo e testes realizados, as técnicas que se mostraram promissoras foram as baseadas em agrupamento de dados, como pelos algoritmos *k-nearest neighbors* (k-NN) e *Density-based spatial clustering of applications with noise* (DBSCAN). Essas técnicas identificam estrutura de prováveis grupos (Mitchell, 1997), disponibilizando esse conhecimento adicional ao especialista, juntamente com as anomalias encontradas, para suporte ao diagnóstico. O software utilizado para os testes foi o R. Os próximos passos da pesquisa serão avaliar os resultados e incluir as abstrações temporais em fluxo de dados/dados temporais.

Para avaliar a abordagem proposta serão utilizados *benchmarks* disponibilizados publicamente para validar DA. Apesar de contrapor ao conceito do aprendizado não supervisionado, que não recorrem ao rótulo dos dados, bases de dados tipicamente utilizadas para avaliar técnicas de classificação são rotuladas a partir da aplicação de conceitos de anomalia para avaliar a detecção. Por exemplo, sabe-se que a quantidade de anomalias presentes em conjuntos de dados é pequena. Logo, rotula-se como anomalia a classe que possui o menor número de observações.

Agradecimentos: Project "NORTE-01- 0145-FEDER- 000016" (NanoSTIMA) is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

Referências

Tan, P. N. (2006) *Introduction to data mining*. Pearson Education India. 651-674.

Shahar, Y., & Musen, M. A. (1996) Knowledge-based temporal abstraction in clinical domains. *Artificial intelligence in medicine*, *8*(3), 267-298.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security* (pp. 77-101). Springer US.

Mitchell TM. (1997) Machine learning. WCB.

Aplicações em Ciências da Saúde II – Sábado, 22 de Abril, Auditório B (9h50)

Lúpus: Comparação de duas medidas da atividade da doença

Ana Cristina Matos[1], Carla Henriques[2], Diogo Jesus[3]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu; Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), amatos@estv.ipv.pt;

[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), carlahenriq@estv.ipv.pt;

[3] Clínica de Lúpus, Serviço de Reumatologia, Centro Hospitalar e Universitário de Coimbra, jesus.p.diogo@gmail.com.

Sumário: Este trabalho tem como objetivo comparar duas medidas de avaliação da atividade de uma doença crónica: a avaliação feita através de um clínico experiente e a avaliação produzida com base num índice obtido por soma de valores que pontuam a presença de manifestações clínicas e serológicas – o SLEDAI. O estudo envolve registos destas avaliações de pacientes que foram seguidos durante dois anos. É analisada a correlação e a concordância entre os resultados dos dois métodos. Adicionalmente é avaliada a capacidade do SLEDAI em detetar melhoras/pioras aferidas pela avaliação médica.

Palavras-chave: comparação de scores, estudo longitudinal**,** Lúpus Eritematoso Sistémico**.**

O Lúpus Eritematoso Sistémico (LES) é uma doença crónica imunomediada, muito heterogénea por apresentar uma grande variabilidade de manifestações clínicas e serológicas. A sua evolução é muito variável, podendo apresentar atividade persistente, ou evoluir com períodos de agudização intercalados com períodos de remissão. A avaliação da atividade de doença é deveras crucial para determinação da terapêutica a adotar, bem como monitorização da sua eficácia. Têm sido desenvolvidos vários índices de atividade de doença, valorizando de forma distinta diferentes manifestações clínicas. O SLEDAI – Systemic Lupus Erythematosus Disease Activity Index, é um dos índices mais utilizados na prática clínica, com uma pontuação que pode variar entre 0 e 105, resultando da soma de itens com diferentes ponderações, consoante a presença/ausência de determinadas manifestações clínicas e laboratoriais.

Como avaliação de referência (*gold standard*) temos a PGA- Physician Global Assessment, em que o médico através de uma escala visual analógica quantifica a atividade global de doença entre 0 e 30 pontos.

O nosso estudo envolve 279 pacientes que foram seguidos entre janeiro de 2014 a dezembro de 2016. Comparando os doentes em duas visitas, aferiu-se a evolução do doente de uma para a outra (melhoria, agravamento ou sem alteração). Estudou-se a associação e a concordância entre os scores em estudo quanto ao resultado desta

classificação, utilizando o teste do Qui-quadrado e o coeficiente Kappa de Cohen ponderado.

Sendo o intervalo entre consultas variável, a análise longitudinal da atividade da doença de cada paciente foi sumariada recorrendo à média ponderada do período em estudo. O cálculo desta estatística para os valores atribuídos ao PGA e para a pontuação do SLEDAI em cada paciente, permitiu calcular a associação entre os dois scores de avaliação.

Devido à natureza rígida do cálculo do SLEDAI, é por vezes posta em causa a capacidade deste índice em detetar melhoras/agravamentos identificadas pelo médico - PGA. Neste trabalho é feito um estudo de modo a avaliar a capacidade do SLEDAI em diferenciar os doentes que melhoram/pioram de acordo com o PGA. Adicionalmente, estuda-se também a habilidade do SLEDAI para detetar melhoras/agravamentos clinicamente relevantes, sendo estes caracterizadas por uma variação no valor do PGA de pelo menos 10% da escala. Este estudo envolveu a construção de curvas ROC e sua análise, de forma a aferir a capacidade do SLEDAI para diferenciar doentes identificados com melhoras/pioras no PGA, relevantes ou não, assim como obter uma indicação do valor de variação no SLEDAI que melhor identifica a evolução no PGA.

Conclui-se que apesar de haver uma correlação forte entre as duas escalas de avaliação (rho de Spearman=0.83, $p<0.0005$) a concordância não se pode considerar elevada. Tendo o PGA como referência, e considerando no SLEDAI uma diferença de pelo menos 1 ponto, o SLEDAI só deteta cerca de metade dos casos de melhoras e também cerca de metade dos casos de pioras. O SLEDAI apresenta um desempenho limitado na deteção de uma mudança clinicamente relevante, o que sugere a necessidade de otimizar este índice de atividade de doença.

Agradecimentos: Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/Multi/04016/2016. Agradecemos adicionalmente ao Instituto Politécnico de Viseu e ao CI&DETS pelo apoio prestado.

Referências

Twisk, Jos W. R. (2003) *Applied Longitudinal Data Analysis for Epidemiology,* Cambridge University Press.

Aplicações em Ciências da Saúde II – Sábado, 22 de Abril, Auditório B (10h10)

Avaliação do risco de morte em doentes operados com pancreatite aguda: É possível ir além dos scores clínicos?

Carla Henriques[1], Ana Cristina Matos[2], Jorge Pereira[3], Carlos Daniel[4], Tercio de Campos[5]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), carlahenriq@estv.ipv.pt;

[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu; Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), amatos@estv.ipv.pt;

[3] Centro Hospitalar Tondela-Viseu, docjota@netcabo.pt

[4]Centro Hospitalar Tondela-Viseu

[5]Hospital da Santa Casa de Misericórdia – São Paulo, Brasil

Sumário: Pancreatite aguda (PA) é uma urgência médica para a qual, em geral, não é fácil identificar doentes de alto risco. Existem vários scores que podem ser usados para avaliar a gravidade da PA, mas, no que respeita à previsão do pior desfecho, morte, nenhum se pode considerar ideal. Neste trabalho investiga-se a capacidade de alguns deste scores para prever morte numa amostra de doentes com PA tratados cirurgicamente. São também estudadas outras variáveis prognósticas e analisados modelos de regressão logística quanto à capacidade para distinguir os doentes com pior desfecho.

Palavras-chave: Curvas ROC, Máxima verosimilhança penalizada, Pancreatite aguda, Regressão logística, Validação de modelos.

A pancreatite aguda (PA) é uma doença relativamente comum e, em geral, sem gravidade, mas em alguns pacientes desenvolvem-se complicações que podem pôr em risco a vida do doente (Pereira, Constantino, Duarte, Pinho, & Pinheiro, 2015). Em doentes com PA é, portanto, crucial uma avaliação precoce da gravidade da doença, mas esta avaliação, apesar dos inúmeros scores de gravidade e do conhecimento atual sobre variáveis prognósticas, está longe de ser uma tarefa fácil (Gravante et al., 2009). Neste sentido, fez-se, neste trabalho, uma avaliação do poder preditivo de alguns scores clínicos no que respeita ao pior desfecho, morte, em doentes com PA tratados cirurgicamente. A amostra é constituída por 66 doentes operados entre 2000 e 2014, dos quais 26 faleceram. Tendo, pois, como objetivo, o de avaliar o risco de morte o mais cedo possível, estudaram-se várias variáveis e scores clínicos, quanto à aptidão para distinguirem os doentes falecidos dos sobreviventes. Para além disso, estabeleceu-se também como objetivo o de encontrar um modelo que, combinando informação de várias variáveis, pudesse melhorar a capacidade preditiva dos scores clínicos.

A comparação dos dois grupos de pacientes, falecidos e sobreviventes, permitiu identificar fatores/marcadores de risco para morte. Utilizaram-se curvas ROC, e a área sob estas curvas (AUC), para avaliar a capacidade discriminativa dos scores clínicos. A análise das curvas ROC obtidas para o APACHE II e Ranson (dois scores clínicos muito utilizados na pratica clínica) sugeria uma capacidade discriminativa aceitável para cada um destes scores, mas passível de se melhorada. De facto, para o Ranson registou-se uma AUC de 0.769 e para o APACHEII de 0,743, identificando-se, em ambos os casos, pontos de corte com sensibilidades de 0,65 e especificidades aproximadas de 0,7. Com recurso à regressão logística construíram-se e estudaram-se modelos que pudessem incrementar a capacidade preditiva de cada um destes scores clínicos. Verificou-se que juntando ao APACHE II, e ao Ranson, a informação de duas complicações (falência respiratória e renal), a AUC sofria um acréscimo significativo, passando para valores acima de 0,9, tanto no modelo construído com o APACHEII, como no modelo construído com o Ranson. Contudo, estes valores de AUC foram obtidos avaliando os dois modelos com a amostra que foi utilizada para os estimar e, como tal, é sabido, sofrem do otimismo caraterístico de todas as avaliações que são feitas desta forma. Neste trabalho utilizou-se o bootstrap para encontrar medidas de avaliação dos dois modelos corrigidas do otimismo. Aplicou-se também máxima verosimilhança penalizada como forma de recalibrar os modelos. Os modelos encontrados revelaram valores muito satisfatórios nos parâmetros de avaliação, nomeadamente, as AUC corrigidas apresentavam valores próximos de 0.9, sugerindo um excelente desempenho dos modelos de regressão logística para distinguir doentes em risco de morte. Estes resultados demonstram o valor preditivo dos modelos com a informação de falência respiratória e renal, acrescida ao score Ranson ou ao APACHE II, na predição de óbito.

Agradecimentos: Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/Multi/04016/2016. Agradecemos adicionalmente ao Instituto Politécnico de Viseu e ao CI&DETS pelo apoio prestado.

Referências

Gravante, G., Garcea, G., Ong, S. L., Metcalfe, M. S., Berry, D. P., Lloyd, D. M., & Dennison, A. R. (2009) Prediction of Mortality in Acute Pancreatitis : A Systematic Review of the Published Evidence. *Pancreatology*, 9, 601–614. http://doi.org/10.1159/000212097

Pereira, J., Constantino, J., Duarte, L., Pinho, H., & Pinheiro, L. (2015) Surgical treatment of severe acute pancreatitis : After 15 years of practice. *International Journal of Hepatobiliary and Pancreatic Diseases*, *5*, 74–81. http://doi.org/10.5348/ijhpd-2015-38-OA-13

POSTERS

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Avaliação de métodos de estimação da variância em amostras complexas

Eládio Muianga[1], Anabela Afonso[2]
[1] Mestrado em Modelação Estatística e Analise de Dados, eladio_muianga@yahoo.com.br;
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, aafonso@uevora.pt.

Sumário: Em amostras complexas nem sempre é possível obter uma expressão analítica para o estimador da variância dos estimadores, existindo na literatura alguns métodos para obter aproximações para esse estimadores. Neste trabalho, estudou-se o desempenho de algumas dessas aproximações (*Taylor*, *Jacknife* e *Bootstrap*). Com base em dados reais foram simuladas populações com características diferentes das quais foram sorteadas amostras com diferentes delineamentos. O estimador da variância da média amostral *Taylor* apresentou o melhor desempenho e o estimador *Bootstrap* foi o menos preciso.

Palavras-chave: Amostras complexas, inferência, estimadores da variância.

Uma amostra complexa consiste numa combinação de vários métodos probabilísticos de amostragem para a seleção de uma amostra representativa da população (Szwarcwald e Damacena). Estas amostras têm pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais, ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008).

Conhecer as propriedades dos estimadores da variância de medidas amostrais e as suas características constitui um passo primordial para intervenção em inquéritos amostrais, para garantir a precisão das estimativas consoante a estrutura do plano amostral. Mas estimar variâncias é um processo muito complexo, pois em amostras complexas com várias etapas de estratificação e conglomerados, a variância do estimador da média e do total deve ser calculada para cada nível e, em seguida, deve ser combinada segundo o delineamento do inquérito (Lohr, 2010). Uma forma de ultrapassar este problema é o uso de métodos que permitem obter aproximações para os estimadores da variância dos totais e médias estimados, dos quais destacamos o método de *linearização Taylor* e os métodos de reamostragem ou replicação (*Jackknife* e *Bootstrap*).

Este trabalho teve como objetivo principal de avaliar o desempenho, e respetivas propriedades, dos estimadores mais usuais da variância da média amostral em amostras complexas. Foram simulados três conjuntos de dados, a partir de dados reais sobre o índice de atividade económica de empresas do sector do comércio no território Moçambicano:

I. Com caraterísticas semelhantes às da população real;
II. Considerou-se uma igual distribuição de empresas por Região e CAE;

III. Introduziu-se uma maior dispersão na distribuição da variável de interesse.

De cada um desses conjuntos foram sorteadas 10 000 réplicas de amostras de acordo com quatro delineamentos de amostragem:

1. Aleatório (sem reposição) estratificado por uma variável;
2. Aleatório (sem reposição) estratificado por duas variáveis;
3. Por grupos a duas etapas;
4. Multietápico (estratificado e por grupos a duas etapas).

Os resultados empíricos do presente estudo mostraram que não existe um padrão consistente no comportamento dos estimadores da variância por população.

Nos delineamentos de amostragem estratificados (1 e 2), os resultados dos estimadores *Taylor*, *Jackknife* e o indicado na literatura foram iguais. Apresentaram melhores resultados quanto à eficiência e à precisão das estimativas do que o estimador *Bootstrap.* Verificou-se assim, à semelhança do referido por Pessoa e Silva, que em delineamentos estratificados o estimador *Jackknife* forneceu os mesmos resultados que o estimador *Taylor*.

Resultados diferentes podem ser visualizados nos planos multietápicos (delineamentos 3 e 4), que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais. O estimador *Taylor* foi o mais eficiente, seguido pelo estimador *Jackknife* sendo o estimador *Bootstrap* o que apresentou menor precisão.

Agradecimentos: Anabela Afonso é membro do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Referências

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks: SAGE Publications

Lorh, S. L. (2010). *Sampling: Design and Analysis*. Second Edition. Boston: Michelle Julet.

Pessoa, D. G. C. e Silva P. L. N. (1998). *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística.

Szwarcwald, C. L. e Damacena, G. N. (2008). Amostras Complexas em Inquéritos Populacionais: Planeamento e implicações na análise estatística dos dados*. Revista Brasileira de Epidemiologia,* 11(l1), 38–45.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

A comparison of team cooperation levels in different sports

Bruno Martins[1], Ana Cristina Costa[2]
[1] NOVA IMS, Universidade Nova de Lisboa, bruno.pt@gmail.com;
[2] NOVA IMS, Universidade Nova de Lisboa, ccosta@novaims.unl.pt

Abstract: This study compares cooperation levels of professional players from three sports: football, handball and rugby. Data from the samples of the three groups are normally distributed and homoscedastic. One-Way ANOVA was used to analyse the differences among group means. Using the Tukey's HSD test for multiple comparisons, a significant difference in the average cooperation levels of rugby and handball was found, at the 5% significance level.

Keywords: One-Way ANOVA, Sports Cooperation, Sports Psychology.

The sport cooperation paradigm arises from the need to explain the interactions within the team and the individual dilemma between cooperation and competition. It can be seen as a framework that contributes to the understanding of the dynamic of sports teams (Garcia-Mas, 2001; Garcia-Mas et al., 2006). Given the different psychological dynamics in different sports, the objective of this study is to investigate the differences between cooperation levels within a sample of professional players from three different team sports: football, handball and rugby. It is reasonable to propose that, based on both theory and research evidence, different sports may display different cooperation levels in their teams.

The study sample consisted of 111 male professional athletes of the same rank but from different teams, equally distributed in three different sports with 37 athletes on each subpopulation: football, rugby and handball. To measure sports cooperation, the "*Questionário de Cooperação Desportiva*" (QCD-p) was used. The QCD-p is a translated and adapted version of the "*Cuestionario de Cooperación Deportiva*" (Garcia-Mas et al, 2006) for the Portuguese population (Almeida et al, 2012). An exploratory data analysis was undertaken to summarise the main characteristics of the data, identify the most relevant patterns and uncover the underlying structure. The average cooperation levels of the three sports is greater than 50, which indicates that all three groups have at least medium levels of cooperation among them (Almeida et al, 2012). The average level of cooperation was smaller for rugby (59.68), followed by football (62.92), and then handball (64.22).

Since the objective is to compare three groups (handball, football and rugby) for one response variable (cooperation level), the relevant statistical test to use is the One-way ANOVA. First, it is necessary to check the ANOVA assumptions to guarantee the validity of its results: the three samples must be independent, the populations from which the samples were obtained must be normally distributed and have the same variance (homoscedasticity). The first assumption is covered due to the data collection method.

The three groups are composed of different individuals from different sports, thus the independence assumption is met. To test the normality of the samples, the Shapiro-Wilk test was chosen. Results show no evidence to reject the null hypothesis of normality in all three groups, at the 5% significance level (handball p-value=0.1386; football p-value=0.08465, rugby p-value=0.1723). The Levene's test was used to check the homoscedasticity assumption. It allowed concluding that there is no evidence against the hypothesis that the population variances from which the samples are drawn are equal, at the 5% significance level (p-value=0.058).

Since all ANOVA assumptions were validated, we tested the null hypothesis that the three sports have equal cooperation levels on average, against the alternative hypothesis that the mean cooperation levels of at least one of them is significantly different. The p-value of the F statistic was 0.004, thus the null hypothesis is rejected for significance levels greater than 0.4%. Consequently, there is strong evidence that the average cooperation levels of the three sports is not identical. Since the sizes of the samples are the same for all groups and the ANOVA assumptions are met, the Tukey's HSD (honest significant difference) test was applied for all pairwise comparisons. This test showed that the mean cooperation levels of rugby and handball are significantly different (p-value=0.004). There was no evidence to reject the hypothesis that the mean cooperation levels of football and handball (p-value=0.617), and rugby and football (p-value=0.0538), are identical at the 5% significance level. However, it is important to point out that there is evidence that the mean cooperation levels of rugby and football are not identical for significance levels greater than 5.4%.

According to the results, we can conclude that rugby has, in average, lower levels of cooperation than football and handball. Hence, sports psychologists should have a different approach when working on the cooperation of rugby teams for an increased performance level. Nonetheless, considering only a global cooperation level is a limitation of this study, since this construct can have different sub-constructs that may be important to differentiate cooperation issues for a more effective intervention.

References

Almeida, P. L., Lameiras, J., Olmedilla, A., Ortega, E., & Garcia-Mas, A. (2012) Avaliação da percepção de cooperação desportiva: propriedades psicométricas da adaptação portuguesa do CCD. *Laboratório de Psicologia*, 1(10), 25-37. doi: 10.14417/lp.622.

Garcia-Mas, A. (2001) Cooperation and competition in sports teams. *Análise Psicológica*, 1(XXI), 115-130.

Garcia-Mas, A., Olmedilla, A., Morilla, M., Rivas, C., García, E., & Ortega, E. (2006) Un nuevo modelo de cooperación deportiva y su evaluación mediante un cuestionario [A new model of cooperation in sport and its evaluation by questionnaire]. *Psicothema*, 18(3), 425-432.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Mapeamento das ONGD em Portugal. Um estudo sobre a comunicação estratégica

Cláudia Silvestre[1], César Neto[2], Mafalda Eiró Gomes[3]

[1] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt;

[2] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa e Plataforma Portuguesa das ONGD, cneto@escs.ipl.pt;

[3] Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, agomes@escs.ipl.pt

Sumário: Neste trabalho, pretende-se conhecer a realidade das Organizações Não Governamentais para o Desenvolvimento (ONGD) existentes em Portugal no que se refere à forma como comunicam com os *stakeholders*. Através de um inquérito por questionário foi feito um levantamento das práticas comunicacionais existentes nas organizações registadas no Ministério dos Negócios Estrangeiros. Os resultados deixam claro que é necessário reforçar o papel da comunicação nestas organizações.

Palavras-chave: Análise de Agrupamento, Comunicação Estratégica, ONGD.

As organizações da sociedade civil, no geral, e as organizações não governamentais para o desenvolvimento (ONGD) em particular, têm vindo a desempenhar um papel cada vez mais importante na nossa sociedade, ganhando desta forma mais visibilidade e reconhecimento.

Para as grandes organizações internacionais, como por exemplo, a Red Cross, a Caritas, a Oxfam, a Médecins du Monde ou a Medecins sans Frontiéres a comunicação estratégica tem sido considerada essencial para o conhecimento e reconhecimento destas organizações. No entanto, alguns estudos (por exemplo, Lourenço, 2009, Machado, 2009, Nunes, 2011) apontam para grandes lacunas existentes na forma como algumas organizações portuguesas e até mesmo congéneres portuguesas de organizações internacionais entendem, gerem e desenvolvem as suas políticas de comunicação.

Neste trabalho, com o objetivo de conhecer as ONGD em Portugal e fazer um levantamento das suas práticas comunicacionais, foram contactadas todas as organizações não governamentais para o desenvolvimento registadas no Ministério dos Negócios Estrangeiros em maio de 2016. Das 185 ONGD registadas, 77 colaboraram neste estudo. Foi enviado um questionário por correio eletrónico a cada uma destas 77 organizações, e agendada uma hora para que se pudesse contactar o responsável pela comunicação ou pela organização com o objetivo de responder ao questionário por telefone.

O questionário baseou-se em questões objectivas que permitem (i) caracterizar a organização (ex. data da criação da organização, número de trabalhadores e as principais atividades), (ii) avaliar o logótipo (ex. quando foi criado/alterado, as cores usadas e em

que suportes aparece), (iii) perceber como a organização comunica com os stakeholders (ex. se tem fachada do edifício, identificação à porta da sede, receção, veículos identificados, *site*, *facebook*, *email*, *roll-up*, *newsletter* e vídeo) e (iv) avaliar a importância da comunicação (ex. quais as ações de comunicação existentes, de quem é a responsabilidade e qual a sua formação).

Os resultados obtidos permitiram fazer um mapeamento das ONGD em Portugal continental, donde se destaca a concentração em Lisboa. Recorrendo a uma análise de agrupamento para dados binários baseada em modelos de mistura de binomiais (Silvestre et al., 2008) foi possível encontrar e caracterizar segmentos com práticas comunicacionais semelhantes. Para identificar as características mais relevantes em cada um dos segmentos também foram usados o teste de independência do qui-quadrado e a divergência de Jensen- Shannon.

Figura1: Distribuição das ONGD por NUTS II com indicação das cores mais usadas na comunicação da ONGD.

Agradecimentos: Este trabalho insere-se nos projetos de Investigação, Desenvolvimento, Inovação e Criação Artística do Instituto Politécnico de Lisboa -IDI&CA do IPL – 2016. Também agradecemos o apoio dado pela Plataforma Portuguesa das ONGD e a colaboração dos alunos 3º de Relações Públicas e Comunicação Empresarial: Inês Castelo, Pedro Rondão, Tânia Amaral e Vlada Domanska.

Referências

Bates, D. (1997) *Public Relations For Charities and Other Nonprofit Organizations*, Lesly, Philip - Lesly's Handbook of Public Relations and Communications. 5ª Edição, 569-590, Chicago: Contemporary Books.

Boyer, R. (1997) *Public Relations and Communications for Nonprofit Organizations*, Caywood, Clarke L. - The Handbook of Strategic Public Relations & Integrated Communications, 481-498, New York: McGraw-Hill Education.

Machado, T. (2009) ONGD: O Papel da Comunicação no seu Conhecimento e Reconhecimento, Tese de Mestrado, Escola Superior de Comunicação Social.

Silvestre, C., Figueiredo, M. A. T. & Cardoso, M. G.M. S. (2008) Clustering with Finite Mixture Models and Categorical Variables. IN BRITO, P., Physics-Verlang, 109-116. *Proceedings of the 18th International Conference on Computational Statistics.*

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Comparing data from CAPI and CAWI surveys

Paula Vicente[1], Elizabeth Reis[2]
[1] *Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, paula.vicente@iscte.pt*
[2] *Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, ear@iscte.pt*

Abstract: Mixing modes to collect survey data may reduce data comparability since different modes a) provide access to different types of people, b) attract different types of respondents, and c) elicit different responses. The decision to mix modes requires survey practitioners to evaluate and quantify the impact of mode on data quality. This study explores some of the issues surrounding the use of CAWI surveys (Computer-assisted web-based interview), in particular the extent to which data from a CAWI survey can be matched to data from a CAPI survey (Computer-assisted personal interview).

Key-words: Data quality, Mixed-mode survey, Mode effects

Survey researchers can choose from a wide range of methods of collecting information from sample members. The most commonly used methods or 'modes' of data collection include: (1) computer-assisted personal interviewing (CAPI) in which the questionnaire has been programmed into a laptop or handheld computer; (2) computer-assisted telephone interviewing (CATI), in which respondents participate in a survey either via a fixed-line telephone or mobile phone, and (3) self-administered questionnaires, in which respondents are invited to complete either a paper questionnaire, or a questionnaire hosted on the Internet, sometimes referred to as computer-assisted web interviewing (CAWI).

Choosing one mode over another involves an assessment of the strengths and weaknesses of each with respect to a range of different factors. Firstly, modes vary in the extent to which they provide access to different survey populations. For example, for surveys of nationally-representative samples, face-to-face interviews usually offer the most dependable method of accessing sample members (depending on the availability of comprehensive lists of addresses), because in theory, interviewers can visit any household in the country. Secondly, modes vary in the extent to which they are suited to the administration of questionnaires of different types (Czaja and Blair, 2004). Questionnaires differ in length and complexity, for instance in terms of the routing and skip patterns that guide respondents to applicable questions. Web-based questionnaires have the advantage of being programmed with sophisticated interactive designs to ensure only those questions that are applicable to the respondent (based on responses to previous questions) are presented. Thirdly, data collection methodologies can differ in what concerns the quality of the data collected. A range of different sources of bias, including coverage error, non-response error, and measurement error can affect data quality in surveys (Jäckle and Roberts 2007, Roberts 2007). Mode effects increase the

total amount of error in survey results and are especially problematic in surveys that use a combination of modes, because of their effect on the comparability of the data.

This study explores some of the issues surrounding the mixed use of CAWI and CAPI modes, namely i) demographic differences in the profile of the respondents, ii) differences in the distributions of the variables under study, iii) interviewer effect and social desirability bias in CAPI surveys, and iv) mode effects of CAWI and CAPI surveys, including how response scales are used. Data comes from parallel surveys on Consumption and Natural Environment conducted in Portugal using CAWI and CAPI methodologies; data were compared before weighting and following demographic weighting. CAWI sample was weighted to match the CAPI sample demographics.

Table 1: Effect of mode on response distribution of Q1

	CAPI	CAWI	
Q1: The solid waste generated in our households is part of man's history since its production is inevitable	Unweighted %	Unweighted %	Demographic Weighted %
1-Totally disagree	5.3	9.1	5.6
2	9.5	19.3	16.7
3	22.1	31.8	37.8
4	28.4	22.7	17.8
5-Totally agree	34.7	17.0	22.2
Base	95	88	90
Pearson $\chi 2$ test p-value	-	*0.020*	*0.034*

A significant association was found between demographics and survey mode, specifically in educational level, working status and household monthly income (outcomes not presented). Table 1 presents the outcomes of one of the analysis undertaken (Question 1) to assess the distribution of responses. The Pearson χ^2 test was performed to evaluate the independence between CAPI and CAWI unweighted and CAPI and CAWI weighted outcomes.

The perception about solid waste is significantly associated with the mode (p-value=0.02). Specifically, CAPI respondents are more likely to disagree with the statement about solid waste than CAWI respondents. This pattern did not change after weighting the CAWI sample by household monthly income (p-value=0.034), thus suggesting a non-ignorable mode effect.

References

Czaja, R., Blair, J. (2004). *Designing surveys: A guide to decisions and procedures*. Pine Forge: Sage Publications.

Jäckle, A., Roberts, C. (2007). Assessing the effect of data collection mode on measurement. *Proceeding of the 56th International Statistical Institute Session*. 22-29 August, Lisbon.

Roberts, C. (2007). *Mixing modes of data collection in surveys: A methodological review*. ESRC National Centre for Research Methods NCRM Methods Review Papers NCRM/008.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Mapas dos valores europeus

Margarida G. M. S. Cardoso[1], **Luís Chambel**[2]
[1] *Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, margarida.cardoso@iscte.pt;*
[2] *Sínese, luischambel@sinese.pt.*

Sumário: Atitudes, crenças e padrões de comportamento de diversas populações em mais de trinta nações são medidos de dois em dois anos pelo European Social Survey (ESS). Neste trabalho procede-se a uma análise exploratória dos dados ESS7 (dados de 2016) agregados por regiões, respeitantes aos valores com que se identificam os cidadãos europeus inquiridos. Procura-se, nomeadamente, ter uma perspetiva agregada dos valores partilhados pelas diversas regiões que participaram neste inquérito. O algoritmo de Kohonen é utilizado para proceder à segmentação das regiões.

Palavras-chave: European Social Survey, Kohonen, Segmentação, Valores

Os dados analisados referem-se ao European Social Survey, um inquérito transnacional dirigido maioritariamente a cidadãos Europeus e que se realiza de dois anos por toda a Europa, desde 2001.

São considerados para análise os dados respeitantes aos valores com que se identificam os 40.185 cidadãos europeus inquiridos. Valores como "Importante ter novas ideias e ser criativo" ou "Importante viver em ambientes protegidos e seguros" (por exemplo) são medidos em escala ordinal: 1= Muito parecido comigo, 2=Parecido comigo, 3= Um pouco parecido comigo, 4= Pouco parecido comigo, 5= Não parecido comigo, 6= Nada parecido comigo). Para a análise, estes dados são ponderados e agregados por região procurando-se determinar grupos de regiões com valores similares. Com o objetivo de obter uma identificação (/não identificação) mais clara das regiões com os valores, são consideradas base de agrupamento as percentagens de respondentes nos níveis 1+2 e nos níveis 5+6 da escala de resposta.

Para agrupar as regiões é usado o algoritmo de Kohonen – SOM - *Self Organizing Maps* (Kohonen, 2001) - sendo ensaiadas parametrizações alternativas da rede, procurando-se um compromisso entre simplicidade das configurações obtidas e a qualidade (coesão-separação) que se lhes associa. Deste modo a solução proposta inclui 9 grupos que se distribuem pelo mapa da Figura 1.

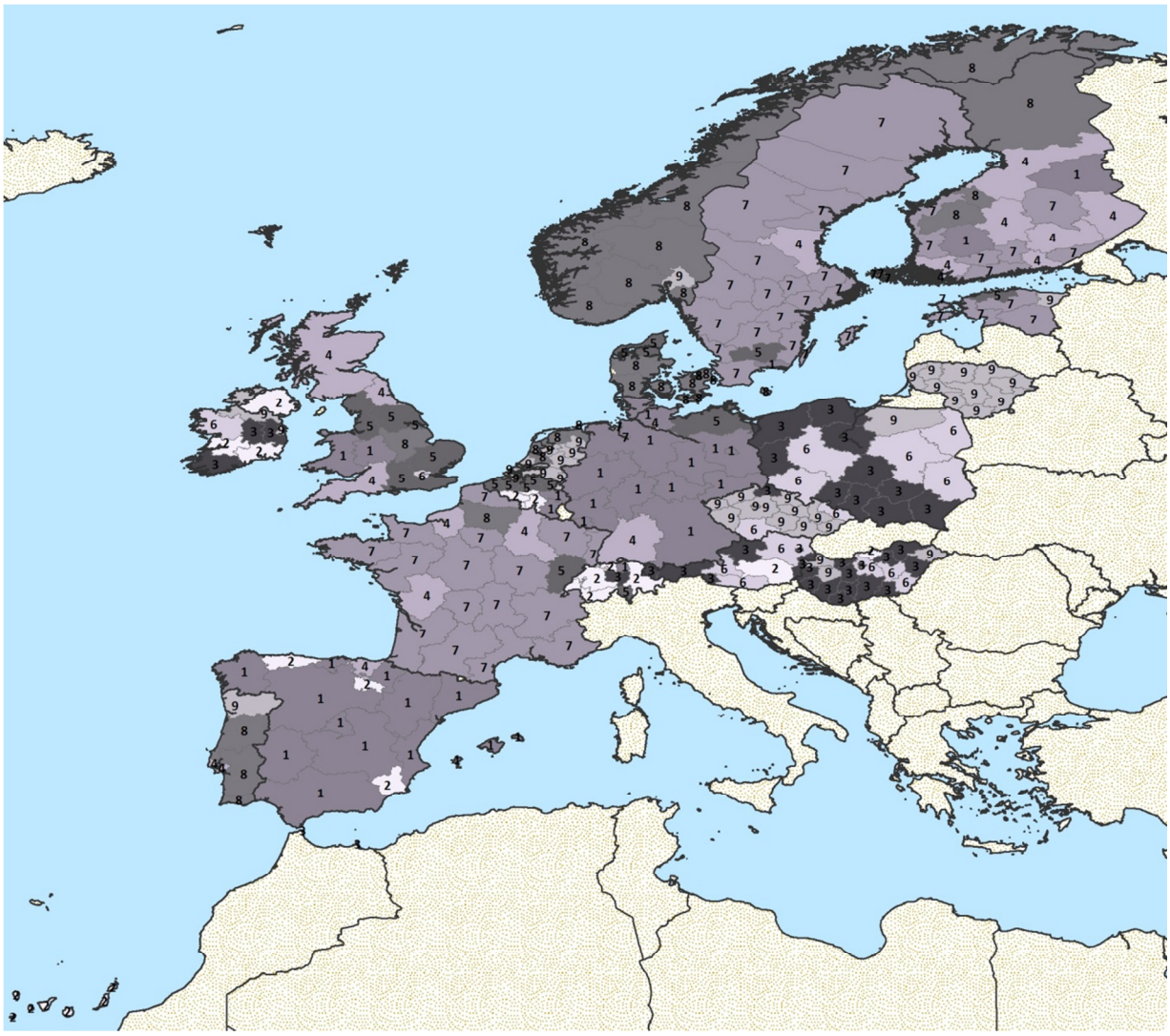


Figura 1: Distribuição dos grupos de regiões

A caracterização dos grupos constituídos atende às variáveis base de agrupamento que melhor discriminam entre eles, segundo a medida CART - *Classification and Regression Trees* de importância relativa (Breiman et a.l., 1984).

A solução proposta poderá ainda ser revista considerando novas metodologias de análise, em particular, de dados simbólicos. Estas deverão atender à natureza ordinal da escala de medida.

Referências

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. California: Wadsworth, Inc.

Kohonen, T. (2001) *Self-Organizing Maps*. Third Edition. Springer.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

A utilização da Internet para viagens: Comparação da Geração Y com os seus antecessores

Carla Henriques[1], Suzanne Amaro[2], Paulo Duarte[3]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), carlahenriq@estv.ipv.pt

[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu; Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), samaro@estv.ipv.pt;

[3] Universidade da Beira Interior, Research Center in Business Sciences (NECE), pduarte@ubi.pt

Sumário: A Geração Y distingue-se das gerações anteriores em muitos aspetos comportamentais e de atitudes, principalmente na afinidade com as novas tecnologias e o mundo digital. Neste trabalho, estudou-se esta geração, comparando-a com as gerações anteriores, no que diz respeito a atitudes e comportamentos on-line no contexto de viagens, nomeadamente na utilização dos social media e compra de viagens on-line. Para isso, recorreu-se à regressão logística e a uma abordagem que procura uma combinação linear que otimize a diferenciação da Geração Y em relação aos seus antecessores.

Palavras-chave: Regressão logística, Área sob a curva ROC, Social media, Millennials, Geração Y.

A geração Y, Millennials, ou "nativos digitais", engloba os nascidos entre finais da década de 70 e meados de 1990, não havendo, contudo, consenso na literatura quanto aos anos exatos de início e fim que marcam esta geração. Os Millennials cresceram com as tecnologias digitais e a internet, sendo caraterizados como utilizadores fluentes, e frequentes, destas ferramentas (Bolton et al., 2013). No que diz respeito aos social media, são claramente grandes utilizadores (Bolton et al., 2013), sendo mais propensos do que as gerações anteriores a valorizar opiniões e a compartilhar conteúdos em redes sociais (eMarketer, "Millennials Roundup", 2014). Este trabalho investiga e compara os Millennials com gerações anteriores em relação ao comportamento online no contexto de viagens, como por exemplo, a compra de viagens on-line e a procura de informação turística nos social media. Os resultados são úteis para os agentes turísticos on-line, uma vez que os Millennials representam um mercado importante.

Os dados utilizados neste estudo resultam de uma amostra constituída por 1732 utilizadores da Internet que responderam a um questionário on-line. Para investigar que características estão especificamente ligadas à geração Y, foram feitas comparações entre os Millennials e os seus antecessores através dos testes de Mann-Whitney, qui-quadrado e/ou teste exato de Fisher. Esta análise univariada permitiu a identificação de várias variáveis que distinguem os Millennials das gerações precedentes. Foi também aplicada a

regressão logística, tendo-se considerado como candidatas a incluir o modelo todas as variáveis com valor p <0,1 na análise univariada precendente. O modelo final foi depois obtido através de um método de seleção automática de variáveis. A capacidade discriminativa de um modelo de regressão logística é muitas vezes medida pela área sob a curva ROC (AUC), que mede a capacidade do modelo para distinguir os sujeitos de um grupo dos sujeitos do outro. A AUC para o modelo final foi de 0,734, indicativa de uma capacidade discriminativa aceitável. Além da regressão logística foi também aplicada uma outra metodologia proposta por Pepe e Thompson (2000) que visa encontrar uma combinação linear de variáveis que otimize a área sob a curva ROC. Aplicando esta metodologia, as primeiras variáveis a serem sucessivamente selecionadas para a combinação linear foram as mesmas selecionadas para o modelo de regressão logística.

Os resultados indicam que os Millennials, mais do que as gerações precedentes, se assumem proficientes e confiantes no uso da internet para compra de viagens online, e também indicam uma maior probabilidade de compra de viagens on-line, no caso de terem a intenção da compra. Contudo, viajam menos para o estrangeiro e, na verdade, compram menos viagens on-line. Estas conclusões não são contraditórias nem estão em desacordo com a literatura. De facto, encontram-se estudos que demonstraram que a geração X compra mais viagens on-line do que os Millennials (Tripadvisor, 2015) e que os Millennials têm rendimentos limitados, preferindo comprar outros produtos em vez de viagens (Barton et al., 2013). Na verdade, não foram encontradas diferenças significativas no número de viagens nacionais, nem no número de compras on-line, o que é indicativo de que eles viajam em seu próprio país e usam sites on-line para comprar outros produtos tanto quanto os seus antecessores. Apurou-se, também, que os Millennials são mais propensos à utilização dos social media, com maior número de sites de social media e com níveis mais elevados de criação de conteúdos on-line e de satisfação/interesse no uso dos social media.

Agradecimentos: Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/Multi/04016/2016. Agradecemos adicionalmente ao Instituto Politécnico de Viseu e ao CI&DETS pelo apoio prestado.

Referências

Barton, C., Haywood, J., Jhunjhunwala, P. and Bhatia, V. (2013), Traveling with Millennials, avaliable at: https://www.bcgperspectives.com/content/articles/transportation_travel_tourism_consumer_insight_traveling_with_millennials/#chapter1.

Bolton, R. N., Parasuraman, A., Hoefnagels, A., Migchels, N., Kabadayi, S., Gruber, T., Loureiro, Y. K. and Solnet, D. (2013) Understanding Generation Y and their use of social media: a review and research agenda, *Journal of Service Management*, 24, 3, 245-267.

Pepe, M. S. and Thompson, M. L. (2000) Combining diagnostic test results to increase accuracy, Biostatistics, 1, 2, 123-140.

Tripadvisor (2015), "6 key travel trends for 2016 ", available online at: https://www.tripadvisor.com/TripAdvisorInsights/n2670/6-key-travel-trends-2016

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Recommending a good classification workflow using metalearning

Maria João Ferreira[1], Pavel Brazdil[2], André Martinez[3]
[1] LIAAD – INESC.TEC, m.joao.fferreira@gmail.com;
[2] LIAAD – INESC.TEC, pbrazdil@inescporto.pt;
[3] FEP – Universidade do Porto, andremartinez92@gmail.com

Abstract: Our aim is to describe a system that helps the user to select a sequence of operations (workflow) for a document classification task. The aim is to identify a workflow that would obtain a good performance (e.g. in terms of accuracy) in a relatively short time. This work builds on previous work in the area of algorithm selection and metalearning, but extends it, as so far it has not been applied to quite complex text mining workflows.

Keywords: Algorithm selection, document classification, machine learning, metalearning, text mining.

Document classification is useful in many areas of human activity. We are normally interested in classifying given incoming documents into different categories so that we could correctly decide how to act on them. This is why the area automatic document classification drew a lot of attention over the last two decades. Normally, a supervised learning approach is adopted, which requires that unstructured text is transformed into a table (Feldman and Sanger, 2007) while applying various preprocessing operations. The user faces a problem regarding as to what to do. Many options exist how to compose the workflow of individual preprocessing and text mining operations. Table 1 shows a typical result obtained on text reviews of video games. Normally we would look for such combination of operations that lead to the best performance. We note that in this case neural networks (*nnet*) won across the different set-ups on accuracy comparisons.

But how can we give an advice regarding what to do on some new dataset without running many new experiments? This is where the methodology of metalearning can help. It uses results on past datasets to suggest a good solution on the new dataset.

			Representation					
			TF			TF-IDF		
			Information Gain			**Information Gain**		
			IG > 0	IG > 0.03	IG > 0.05	IG > 0	IG > 0.03	IG > 0.05
Sparsity correction	S = 90%	dtrees	79,35%	79,35%	79,35%	73,91%	73,91%	73,91%
		nnet	**94,02%**	**92,39%**	**90,22%**	**94,02%**	**94,02%**	**92,39%**
		svmRadial	84,24%	88,59%	86,41%	89,67%	91,30%	91,30%
	S = 95%	dtrees	79,35%	79,35%	79,35%	73,91%	73,91%	73,91%
		nnet	**94,57%**	**94,02%**	**90,76%**	**96,20%**	**96,20%**	**92,39%**
		svmRadial	87,50%	89,67%	85,87%	92,39%	91,85%	91,85%

Table 1: Accuracy scores of different algorithms with various preprocessing

Metalearning systems guide users to find the best algorithm for their specific problem without having to test all possible alternatives. Unfortunately, to our best of our

knowledge, no such system exists for text mining and document classification tasks. So the aim of this work is to fill in the gap.

Our objective is to show how a typical metalearning system (Abdulrahman *et al.*, 2017) can be adapted to text classification, which typically involves a chain preprocessing tasks, followed by application of a classification algorithm. That is, our problem is by no means trivial. This approach requires that various tests be carried out in a kind of preparatory phase. As it is not feasible to explore all possible combinations, we focus the study on a subset of all possible workflows. These include some classification algorithms, such as, *SVM*s, *Random Forests, neural nets*, among others. The preprocessing tasks will include various operations, such as, *stemming*, *sparsity correction* and *stop-words removal*. Different workflows will be tested on different datasets and ranked based on both accuracy and time. This process gives rise to *average ranking,* which will then be followed on the new dataset.

Evaluation is done with recourse to *loss-curves* (Fig. 1), which highlight the performance loss (or gain) resulting from following this ranking with respect to the *ideal* workflow. Our goal will be to compare different strategies by comparing loss-curves resulting from using different workflows.

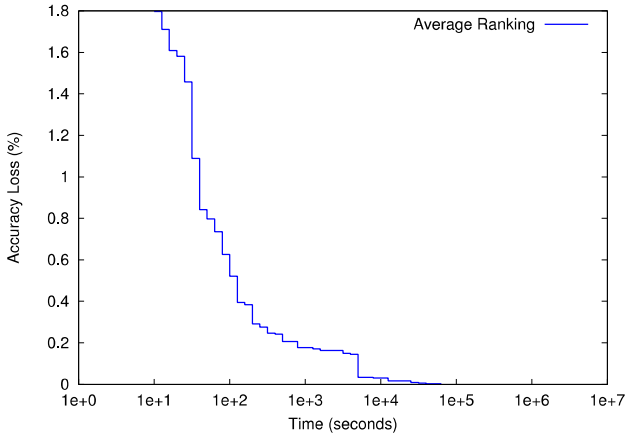


Fig. 1: Loss time-curve using the average ranking method (*adapted from* Abdulrahman *et al.* (2017))

Acknowledgments: The authors acknowledge the support of project *NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016*, which is financed by the *North Portugal Regional Operational Programme (NORTE 2020)*, under the *PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF)*.

References

Abdulrahman, S. M. and Brazdil, P. (2014). Measures for combining accuracy and time for meta-learning. In *Proc. of the int. workshop MetaSel-2014.*

Abdulrahman, S. M., Brazdil, P., van Rijn, J. N., and Vanschoren, J. (2017). Speeding up algorithm selection using average ranking and active testing by introducing runtime, to appear in *Machine Learning J*.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge University Press.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Firms' E-commerce adoption in Europe: A multilevel approach

Paula Gabriel[1], José G. Dias[2]
[1] *Instituto Universitário de Lisboa ISCTE IUL, Lisboa, Portugal, pclgl@iscte.pt*
[2] *Instituto Universitário de Lisboa ISCTE IUL, BRU-IUL, Lisboa, Portugal, jose.dias@iscte.pt*

Abstract: This study analyzes the *overall attitude towards e-commerce adoption* by European firms, linking specifics deterrents of e-commerce adoption and retailer's characteristics in a multilevel factor framework. Data comes from a Eurobarometer survey that was designed to enhance the general knowledge of national and European consumer and retailer e-commerce conditions in 28 countries of European Union. Results show that aspects such us firm size and the job position of the respondents do not influence the attitudes toward e-commerce adoption. Three covariates point to the aversion to e-commerce: telesales/call center retail, direct retail channels, and selling non-food products.

Keywords: E-commerce adoption, European firms, confirmatory factor analysis, multilevel analysis, country level analysis.

E-commerce adoption has been recognized as a new form of commerce and a fresh way to identify, target, and retain customers. Firms tend to face uncertainty about the decision to go online as investing in internal resources and in the service that may not lead to consumers' engagement in future online transactions. Studies in this area usually distinguish three dimensions: the environmental, the technological, and the organizational one (Rodríguez-Ardura & Meseguer-Artola, 2010; Zhu, Kraemer, & Xu, 2003).

Given the heterogeneity of contexts, factors that constrain the adoption of e-commerce are not the same across firms and countries (digital divide). Taking this into mind, our focus will be on firms that are not yet online as a result of a set of inhibitors of decision making and consider the differences within- and between-countries. The research question to be answered is: What are the inhibitors factors of e-commerce adoption in EU? Considering this, it will be analyzed the overall attitudes towards e-commerce adoption in the European firms, linking specifics deterrents of e-commerce adoption (e.g., higher delivery costs or the nature of business) and specific retailer's characteristics (covariates): size, type of product, retail channels, respondents' position in the firm and the engagement in distance selling. To explore this construct (*overall attitude towards e-commerce adoption)*, we propose a conceptual model that combines factorial and regression components.

The multilevel factor model considers two levels: the individual level, which models the attitudes towards online barriers within each country; and the country level that highlights the similarities (and differences) between EU countries. This model was estimated by the maximum likelihood method to handling missing values (which can be

tendentious in small samples) using the Mplus 6.12 (Muthén and Muthén, 2010). The ordinal specification of the model by thresholds relaxes the assumptions postulated by Gaussian models in the context of maximum likelihood. CFI is 0.989, TLI is 0.984, and RMSEA is 0.034; therefore, model fit is excellent, even when assuming the parameters of the measurement model are assumed invariant across countries.

Results show that the number of firms' employees and respondents' position do not contribute significantly to explain the attitude toward e-commerce adoption. Moreover, firms that sell non-food products, use a traditional way (direct retail), use call center, or Telesales channels tend to be more pessimistic about e-commerce adoption, which means that these covariates influence the latent variable negatively. Regarding the engagement in distance selling, firms currently engaged are more optimistic compared with the others, but the impact on latent variable is small. The EU28 countries globally present sensitivity to barriers in the e-commerce adoption (16 of 28 countries). Scores can be estimated at both levels. Using individual level scores, country averages can be compared with the grand mean. Denmark, Sweden, Estonia and Italy show an attitude above the European population average, considering not all important the majority of the deterrents; on the other hand, firms in countries such as Portugal, Poland and Luxembourg have on average an attitude below the European average.

References

Muthén, L.K. and Muthén, B.O. (2010). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Rodríguez-Ardura, I., & Meseguer-Artola, A. (2010). Toward a Longitudinal Model of e-Commerce: Environmental, Technological, and Organizational Drivers of B2C Adoption. *The Information Society*, 26, 209-227.

Zhu, K., Kraemer, K. L., & Xu, S. (2006). The Process of Innovation Assimilation by Firms in Different Countries: A Technology Diffusion Perspective on E-Business. *Management Science*, 52 (10), 1557-1576.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

A GIS-based platform for data management in water resources systems

Bernardo Varela[1], Reydleon Paulo[2], Daniel Conde[3], Ricardo Canelas[4], Alexandre Gonçalves[5], Marta Castilho Gomes[6], Francisco Regateiro[7], Ana M. Ricardo[8]
[1,2,3,4,5,6,7,8]CERIS, Instituto Superior Técnico, Universidade Lisboa;
[1]bernardo.f.varela@tecnico.ulisboa.pt; [2]reydleon94@gmail.com;
[3]daniel.conde@tecnico.ulisboa.pt; [4]ricardo.canelas@tecnico.ulisboa.pt;
[5]alexandre.goncalves@tecnico.ulisboa.pt; [6]marta.gomes@tecnico.ulisboa.pt;
[7]francisco.regateiro@tecnico.ulisboa.pt; [8]ana.ricardo@tecnico.ulisboa.pt;

Abstract: The benefits of geospatial information systems (GIS) for geocomputations in water resources are countless. We target the development of a platform able to provide data management, visualization interfaces and tools to support researchers, engineers and policy and decision makers. This goal is achieved by a strong articulation of a relational database with a GIS-based graphical interface. This customized tool results from an integrated multidisciplinary effort.

Keywords: Decision support systems, Environmental data management, Geospatial information systems**,** Visualization**.**

Nowadays geographically based data visualization is nearly ubiquitous in water resources applications. Dixon & Uddameri (2016) defined geocomputation in water resources engineering and science as GIS-enabled analysis, synthesis, and design of water resources systems. The increasing volume and varying format of water resources data presents challenges in storing, managing, processing, visualizing and verifying the quality of data. For this reason, geospatial information systems (GIS) are often incorporated into workflows for analyzing such datasets (Tsihrintzis et al. 1996; Pons & Masó 2016).

The present work is aimed at developing a platform able to provide data management, visualization interfaces and tools to support researchers, engineers and policy and decision makers. It is a hydro-informatics platform in the sense that it provides a developing environment to deal with computerized information related to water resources systems. The main goal is the articulation of a flexible and efficiently structured database with a GIS-based graphical interface.

The database is structured in light of the relational model concept and the relational database management system used for its maintenance is PostgreSQL. The database includes time and space variables and additional metadata to maximize the efficiency of the database-interface communication. Temporal data is stored in different tables for better data management and to avoid data redundancy. In Portugal hydrological and meteorological data is mainly stored following its historical design, i.e., centered in the concept of time-series associated to monitoring stations. Following the new paradigm to organize water resources data (Dixon & Uddameri, 2016; Pons & Masó, 2016), we target a geospatial approach. For that PostGIS is used to manage geospatial information.

The graphical interface is rooted in Quantum GIS and tuned with customized tools and various web-services to couple data and analysis in a uniform environment. These tools, targeting the ease of use and workflow simplification, are developed for finding, displaying and processing data and are mainly designed as Python plugins. Common protocols for information transfer like Web Feature Service, Web Coverage Service, Web Map Service and Web Processing Service are also included. All the tools are coded in order to ensure the compatibility with the Open Geospatial Consortium (OGC) standards.

At the current stage of project development, the database stores precipitation data and a digital elevation model (DEM) of mainland Portugal. The application case study consists of the spatial and time characterization of the monthly precipitation in Mondego's catchment area. The Time Manager plugin is used for data visualization.

The platform and its applications are suitable for integration into decision support systems of water resources administrations and civil protection systems, contributing for increased resilience to floods and other nature disasters. This is ensured by close collaboration with the Portuguese Environment Agency, particularly with the Hydraulic Resources department, which is in charge of monitoring the hydrological network.

The present work is a multidisciplinary effort to create a novel, flexible, efficient and customized tool whose applications fall in the realm of water resources management and associated risk assessment. We started by integrating hydrologic and topographic data and in the near future other environmental datasets, like meteorological, sedimentological and geometric data, will also be integrated. The ultimate goal is the complete integration of the present platform with an existing mathematical model (Conde et al. 2015) to achieve a real-time, operational fluvial hydrodynamic modelling tool able to simulate solid material transport.

Acknowledgements: This work was funded by CERIS through the 2016 exploratory actions program (RD0475/01).

References

Conde, D. A. S.; Telhado, M. J.; Viana Baptista, M. A. & Ferreira, R. M. L. (2015). Severity and exposure associated with tsunami actions in urban waterfronts: the case of Lisbon, Portugal. *Natural Hazards*, 79, 2125-2144.

Dixon, B., Uddameri, V., & Ray, C. (2015). *GIS and Geocomputation for Water Resource Science and Engineering*. John Wiley & Sons.

Pons, X. & Masó, J. (2016). A comprehensive open package format for preservation and distribution of geospatial data and metadata. *Computers & Geosciences*, 97, 89-97.

Tsihrintzis, V. A., Hamid, R., & Fuentes, H. R. (1996). Use of geographic information systems (GIS) in water resources: a review. *Water resources management*, 10(4), 251-277.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

The van der Waerden control chart for monitoring the location

César Rocha[1], Fernanda Otília Figueiredo[2], Adelaide Figueiredo[3], Subhabrata Chakraborti[4]

[1] PDMA/FCUL, Universidade do Porto, and CEAUL, Universidade de Lisboa, cdrochaster@gmail.com;

[2] Faculdade de Economia, Universidade do Porto, and CEAUL, Universidade de Lisboa, otilia@fep.up.pt;

[3] Faculdade de Economia, Universidade do Porto, INESC TEC- Porto, and CEAUL, Universidade de Lisboa, adelaide@fep.up.pt;

[4] University of Alabama, USA, schakrab@cba.ua.edu;

Abstract: Control charts are among the classical tools useful in statistical process monitoring and control. The traditional control charts assume that the data distribution is normal. However, in many practical situations this assumption is not verifiable or simply not true. An alternative class of control charts is the nonparametric/distribution-free control charts that do not depend on the data distribution and can be applied under a minimal set of assumptions. In this paper we propose a nonparametric control chart for monitoring the location parameters of continuous distributions based on the van der Waerden test. Properties and performance of the proposed chart are examined and contrasted with existing charts.

Keywords: Average run-length, Control charts, Nonparametric tests**,** van der Waerden test.

In this paper we propose a nonparametric control chart based on the van der Waerden test (see, for instance, Gibbons and Chakraborti, 2003), which we will call VDW-chart. The in-control run-length distribution of the VDW-chart is the same for every continuous distribution, being therefore considered a distribution-free or nonparametric chart (Chakraborti *et al*, 2004).

Suppose we have available a reference sample, $(X_1,X_2,...,X_m)$, taken from an in-control stable process, and let $(Y_1,Y_2,...,Y_n)$ denotes an arbitrary test sample. Assume that the random samples $(X_1,X_2,...,X_m)$ and $(Y_1,Y_2,...,Y_n)$ are independent, and drawn from two populations with continuous cumulative distribution functions F_X and F_Y, respectively. Moreover, we will assume the data is collected such that the variables within these samples can be considered independent.

When the process is in-control, we have a set of m+n=N observations from a common but unknown population, $F_X=F_Y=F$. If we write the combined ordered sample as a vector of indicator random variables we have: $Z = (Z_1,...,Z_N)$, where $Z_i = 1$ if the ith random variable in the combined ordered sample is from population X, and $Z_i = 0$ if it is from population Y, for i=1,2,...,N, with N=m+n. The van der Waerden statistic is then defined by

$$V_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) Z_i,$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. The V_N statistic is symmetric about zero and has variance $\sigma^2_{V_N} = mn\frac{\sum_{i=1}^{N}\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2}{N(N-1)}$.

A two-sided VDW control chart uses V_N as the charting statistic and signals (out-of-control) whenever $\frac{V_N}{\sigma_{V_N}} < LCL \ \vee \ \frac{V_N}{\sigma_{V_N}} > UCL$, with LCL and UCL denoting the lower and upper control limit, respectively. These limits are obtained by simulation in order to attain an in-control ARL value of 500 and 1000.

In this study we computed the ARL values of the nonparametric VDW-chart as well as of the parametric chart based on the t-test for two samples, for comparison, when the underlying process data actually follows normal, uniform, Laplace, t(3) and Cauchy distributions, with mean zero and variance one after standardization. The out-of-control ARL was determined for shifts of size 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0 in the location parameter. These ARL values were computed using 10000 simulations, for sample sizes n= 10, 15, 20 and 25. Comparisons with other parametric and nonparametric control charts and other sample sizes will be done in a future work.

From this study we draw the following conclusions.

- Generally, the parametric chart is more efficient for the normal distribution, as it is expected, and the nonparametric chart performs better for uniform distribution.
- The nonparametric chart is more efficient for sample sizes n=15, 20 and 25 for all distributions except normal.
- For sample sizes n=10 and the Laplace distribution, the nonparametric chart has the smallest out-of-control ARL for a shift equal to 0.2, and for the Cauchy distribution this chart has a better out-of-control performance for shifts less or equal than 1.
- For sample sizes n=10 and the t(3) distribution, the parametric chart has always a smaller out-of-control ARL.
- For sample sizes n=15 and the Laplace distribution, the nonparametric chart has the smallest out-of-control ARL for shifts less or equal than 0.6, and for the Cauchy and t(3) distribution this chart outperforms the nonparametric chart for all shifts.
- For sample sizes n=20 and the Laplace distribution, the nonparametric chart has the smallest out-of-control ARL for shifts less or equal than 1, and for the Cauchy and t(3) distribution this chart has a smaller out-of-control ARL for all shifts.
- For sample sizes n=25 and the Laplace distribution, the nonparametric chart has the smallest out-of-control ARL for shifts less or equal than 1.4, and for the Cauchy and t(3) distribution this chart has a smaller out-of-control ARL for all shifts.

Acknowledgements: This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects POCI-01-0145-FEDER-00696 and UID/MAT/00006/2013 (CEA/UL). We thank the referee for the interesting comments and suggestions.

References

Chakraborti, S., van der Laan, P. & van de Wiel, M. A. (2004) A class of distribution-free control charts. *Journal of the Royal Statistical Society Series C- Applied Statistics*, 53, 443-462.

Gibbons, J. D. & Chakraborti, S. (2003) *Nonparametric Statistical Inference*, 5th edition, Marcel Dekker, New York.

Sessão de Posters I – 6ª Feira, 21 de Abril (11:20)

Clustering de relacionamentos entre entidades nomeadas em textos com base no contexto

Nelson Morais[1], Conceição Rocha[2], Alípio M. Jorge[3]
[1] INESC TEC, namorais@inesctec.pt;
[2] INESC TEC, conceicao.n.rocha@inesctec.pt;
[3] FCUP, amjorge@fc.up.pt

Sumário: Apresentamos uma abordagem para identificar/caracterizar, de forma automática, relações semânticas (RS) entre entidades nomeadas (EN) em textos com base no contexto (palavras envolventes) em que as mesmas coocorrem. A metodologia proposta é composta por três fases: identificação das EN nos textos; extração do contexto das diversas frases em que as EN coocorrem; e por último, classificação não supervisionada dos pares de EN, com RS semelhantes, com base nos diversos contextos associados a cada par de EN. Assumindo que pares de entidades com RS semelhantes ocorrem em contextos semelhantes conseguimos deste modo identificar e agrupar pares de EN com RS semelhantes.

Palavras-chave: Agrupamento, Entidade**,** Relacionamento**,** Similaridade, Texto.

Um problema importante na compreensão automática de texto é a identificação e caracterização de relações entre as entidades nomeadas (EN) nos mesmos (Collovini *et al.*, 2016). Através da interpretação humana do conteúdo do texto é possível saber, *p.e.*, que personalidades se relacionam com certas organizações (e de que forma). No entanto, obter uma ampla noção de milhares/milhões de diferentes relacionamentos, em poucos minutos, é um caminho só viável com recurso a técnicas computacionais.

Neste trabalho, propomos um processo não supervisionado para extrair e identificar, de forma automática, relações semânticas (RS) entre pares de EN que coocorrem na mesma frase. Tendo em consideração que dois pares de entidades com RS semelhantes tendem a ocorrer em contextos semelhantes, neste trabalho, as instancias que representam os pares de EN são caracterizadas pelas palavras que compõem o contexto em que estas coocorrem.

A metodologia proposta neste trabalho é composta essencialmente por três fases. Na primeira fase, são detetadas e localizadas as EN nos textos com recurso à ferramenta PAMPO (Rocha *et al.*, 2016). Numa segunda fase, formam-se pares de EN, com base na sua coocorrência na mesma frase, e extraem-se e agregam-se todos os contextos em que as mesmas coocorrem. Por último, agrupam-se os pares de EN, com base nos contextos em que as mesmas coocorreram, obtendo desse modo grupos de pares que têm RS semelhantes.

Na terceira fase, permite-se aplicar um de dois métodos de classificação: *clustering* hierárquico e *k*-Médias para diferentes valores de *k*. É de referir que neste trabalho, foram

ainda aplicadas variantes nos critérios de seleção dos pares de entidades bem como na caracterização do contexto dos mesmos. Na Figura 1 encontra-se representado um esquema desta abordagem e, como se pode observar, há um trabalho de limpeza bem como a aplicação de filtros na passagem da primeira para a segunda fase que regra geral é necessário aplicar.

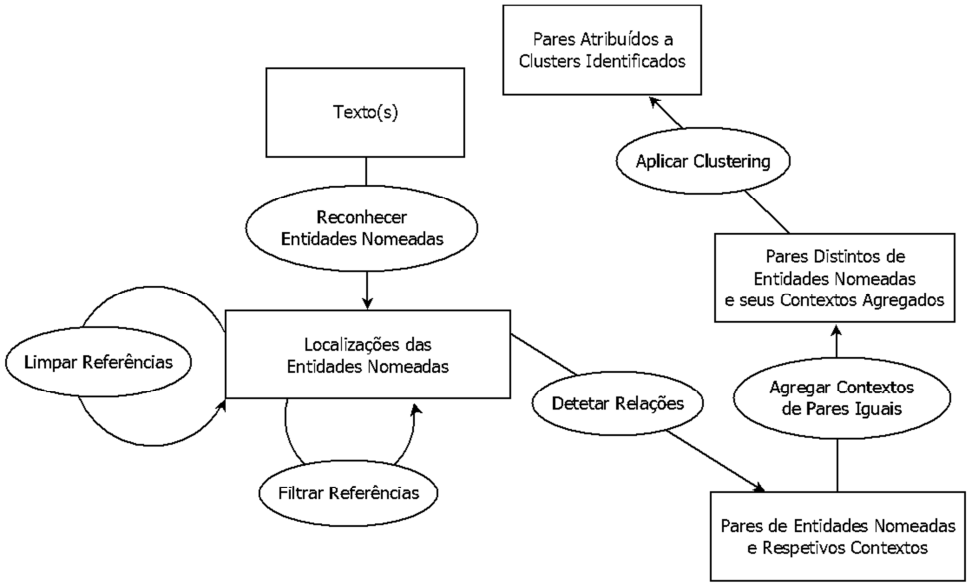


Figura 1: Visão geral sobre a abordagem apresentada.

A implementação da abordagem foi aplicada a um conjunto de 227 textos, em português, cedidos pela Sapo Labs, e encontra-se disponível para testes (Morais *et al.*, 2016). Foram testadas várias configurações de parâmetros: extração de contextos tanto intermédios como completos, pesagens de termos, medidas de distância e sete valores para *k*. Para avaliar a qualidade dos resultados usou-se a medida F1, que é média harmónica entre sensibilidade (percentagem de positivos corretamente classificados) e precisão (percentagem de casos corretamente classificados como positivos), para todos cenários em estudo. Verificou-se que com a diminuição do número de grupos o valor da medida F1 aumentava. Esta medida, num dos cenários, chegou mesmo a atingir um valor máximo de 60%.

Referências

Collovivi, S., Machado, G. & Vieira, R. (2016) "Extracting and Structuring Open Relations from Portuguese Text" in *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR*, p. 153-164.

Rocha, C. & Jorge, A. & Sionara, R. & Brito, P. & Pimenta, P. & Rezende, O. (2016) *"PAMPO: using pattern matching and pos-tagging for effective Named Entities recognition in Portuguese"*. CoRR abs/1612.09535

Morais, N., Jorge, A. & Rocha, C. (2016) *Framework for clustering of semantic relations* in https://github.com/LIAAD/EntityRelationsNelson.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Anova com dois fatores não paramétrica

Anabela Afonso[1], Dulce Pereira[2]
[1] CIMA/IIFA e DMAT/ECT, Universidade de Évora, aafonso@uevora.pt;
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, dgsp@uevora.pt

Sumário: Nos últimos anos têm sido propostas várias alternativas não paramétricas à Análise de Variância (Anova) com dois fatores. Neste trabalho, realizamos um estudo de simulação para analisar a probabilidade de erro de Tipo I de alguns desses testes alternativos, no caso de delineamentos equilibrados. Consideram-se amostras com diferentes dimensões amostrais e distribuições discretas. Concluímos que existem três testes que apresentam taxas de erro de Tipo I empíricas superiores ao nível de significância nominal, sendo por isso desaconselhada a sua utilização.

Palavras-chave: Estatística de Wald, Testes de permutação, Transformação em ordens.

A Análise de Variância (Anova) foi introduzida por Fisher, com aplicações iniciais no domínio da Agronomia e Biologia, mas são várias as aplicações na área da Epidemiologia e das Ciências Sociais, por exemplo. Num delineamento fatorial pretende-se estudar a influencia de um ou mais fatores numa determinada variável resposta. Neste tipo de delineamento está implícito que as amostras são aleatórias e independentes. Na Anova para além de se assumir que a variável resposta é do tipo contínuo, também se pressupõe que as populações têm distribuição normal e as variâncias são homogéneas.

Quando se trabalha com dados reais encontramos vários problemas. A natureza dos dados nem sempre é contínua, por exemplo, é muito usual nas áreas da Biologia e da Ecologia obtermos dados de contagens e na área das Ciências Sociais predominam dados ordinais. A normalidade das distribuições é muitas vezes violada devido quer à forma da distribuição quer à presença de valores atípicos. Vulgarmente as amostras são pequenas, o que associado a dados não normais não possibilita a aplicação do Teorema do Limite Central, o que põe em causa não só o pressuposto da normalidade como também a validade da distribuição F. Nestas situações a Anova paramétrica não é a técnica mais adequada o que levou ao desenvolvimento de alternativas.

Na literatura encontram-se vários testes não paramétricos alternativos à Anova com dois fatores, sendo umas mais simples do que outras, não existindo uma alternativa que seja melhor que as restantes em todos os contextos. De forma muito resumida, há umas propostas que apenas substituem as observações pelas suas ordens ou pelos *scores* normais das ordens e aplicam uma Anova paramétrica usual, há outras que subtraem todos os efeitos que não sejam de primeiro interesse antes de se realizar a Anova, outras propõem as suas próprias estatísticas, algumas delas à custa dos modelos lineares, e há ainda alternativas baseadas nas permutações das observações.

Na literatura o desempenho destas alternativas foi analisado considerando distribuições contínuas, preocupando-se com a assimetria e a presença de valores atípicos, e/ou com variâncias heterogéneas, tanto para delineamentos equilibrados como desequilibrados. No entanto, não há estudos onde se utilizem distribuições discretas, que usualmente potenciam o surgimento de empates. O principal objetivo deste trabalho é colmatar esta lacuna, de forma a generalizar as vantagens e desvantagens de algumas dessas alternativas: transformação em ordens (RT), transformação normal inversa (INT), transformação em ordens alinhadas (ART), ART combinado com INT, combinação linear de estatísticas ordinais (estatística L do teste de Puri & Sen), teste de van der Waerden, estatística de tipo Wald (WTS), estatística de tipo Anova (ATS), teste de permutação de tipo Wald (WTPS) e testes de permutação sincronizada (CSP e USP) (Hahn *et al.*, 2014; Luepsen, 2016).

Foram considerados delineamentos equilibrados com 2 fatores, cada um com 3 níveis, e amostras com dimensão $n = 5, 10, 15$ e 20. Assumiram-se vários cenários distribucionais com base nas distribuições binomial, binomial negativa, poisson e uniforme, com diferentes graus de assimetria. Em todas as situações geradas foi considerada a ausência de interação e de efeitos principais. Para cada um dos cenários realizaram-se $M = 1000$ réplicas e contabilizou-se o número total de réplicas que ultrapassaram o nível de significância definido ($\alpha = 1\%, 5\%$ e 10%).

Os testes CPS e USP foram os que apresentaram maior dispersão nos valores p obtidos. Os testes WTS aplicados quer às observações originais quer às ordens tendem a apresentar valores p mais baixos do que os restantes testes. No teste à interação os testes USP, WTS e CSP apresentam taxas de erro de Tipo I empíricas superiores ao nível de significância definido. Para amostras de pequena dimensão ($n = 5$) o teste ATS aplicado às observações originais é o que apresenta pior desempenho, mas para amostras de maior dimensão reduz drasticamente a taxa de erro de Tipo I empírica. Nos testes aos efeitos principais, estes três testes também tendem a apresentar taxas de erro de Tipo I empíricas superior ao nível de significância nominal, embora com uma diferença inferior ao observado no teste à interação. Os testes de van der Waerden e de Puri & Sen são os que mostraram ter um melhor desempenho tanto no teste à interação como aos efeitos principais.

Agradecimentos: Anabela Afonso e Dulce G. Pereira são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Referências

Hahn, S., Konietschke, F. & Salmaso, L. (2014). A Comparison of Efficient Permutation Tests for Unbalanced ANOVA in Two by Two Designs and Their Behavior Under Heteroscedasticity. In MELAS, V.B., MIGNANI, S., MONARI, P., SALMASO, L. (Eds.) *Topics in Statistical Simulation: Research Papers from the 7th International Workshop on Statistical Simulation*, 257-269. New York, USA: Springer.

Luepsen, H. (2016). The Aligned Rank Transform and Discrete Variables - a Warning. *Communications in Statistics - Simulation and Computation*. doi: 10.1080/03610918.2016.1217014.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Investimento em Redes Rodoviárias Europeias: um Estudo Comparativo

Pedro Marcelino[1], **Marta Castilho Gomes**[2]

[1] *Laboratório Nacional de Engenharia Civil, pmarcelino@lnec.pt;*

[2] *CERIS, CESUR, Instituto Superior Técnico, marta.gomes@tecnico.ulisboa.pt;*

Sumário: O objetivo deste trabalho foi conhecer e comparar o investimento em infraestruturas rodoviárias na União Europeia, com base na recolha de um amplo conjunto de variáveis em bases de dados *online*. Para tal, foram aplicadas técnicas de estatística multivariada de análise fatorial, análise de *clusters* e regressão linear múltipla. Concluiu-se que os países em desenvolvimento são os que mais investem em infraestruturas rodoviárias. No caso de Portugal, verificou-se que os níveis de investimento excedem o previsto pelo modelo de regressão linear desenvolvido.

Palavras-chave: Análise de *clusters*, Análise fatorial, Infraestruturas rodoviárias, Redes rodoviárias, Regressão linear múltipla.

Este trabalho teve como objetivo o estudo do investimento em infraestruturas rodoviárias na União Europeia, tendo sido realizado no âmbito da unidade curricular de Modelação e Avaliação de Sistemas (do Mestrado Integrado em Engenharia Civil do Instituto Superior Técnico).

Para tal, foi recolhida informação sobre o tema em bases de dados disponíveis online (e.g. World Bank, Eurostat e OECD), tendo o país como unidade geográfica de base. A informação recolhida dividiu-se nos seguintes subtemas: Economia (produto interno bruto, investimento em infraestruturas rodoviárias, custos de manutenção com as mesmas, etc.); Social (população, acidentes rodoviários com feridos, índice de desenvolvimento humano, etc.); Ambiente (consumo energético e emissão de CO_2); Físico (dimensão da rede pavimentada, topografia, tempo de viagem, etc.).

Os dados (variáveis estatísticas) foram primeiro caracterizados através da estatística descritiva, tendo-se apurado medidas de síntese e traçado histogramas e gráficos de barras para cada variável. No estudo bivariado calculou-se a matriz de correlações e traçaram-se gráficos de dispersão.

Como resultado desta análise retiraram-se algumas conclusões, tais como a exclusão do 'Luxemburgo' da análise, após interpretação dos gráficos de dispersão; e a exclusão da variável 'tempo de viagem', devido a fenómenos de colinearidade. Assim, esta análise inicial acabou por constituir também um "pré-tratamento" dos dados que se revelou útil na aplicação de técnicas multivariadas, que foram: a análise fatorial de componentes principais, a análise de *clusters* e a regressão linear múltipla.

Com a aplicação da análise de componentes principais as variáveis foram sintetizadas em dois fatores (retendo 53% da variância): o primeiro interpretado como "Fator de desenvolvimento económico", englobando variáveis relacionadas com o desempenho económico do país, tais como o produto interno bruto, o investimento em rodovias e o fluxo de mercadorias; sendo o segundo fator interpretado como "Fator de dimensão física", diretamente relacionado com a dimensão do país a nível da população e do território, representado pelas variáveis população, dimensão da rede pavimentada e distância. A análise destes fatores permitiu compreender melhor as diferenças observadas entre países da UE a nível das redes rodoviárias.

Por sua vez, a análise de *clusters* agrupou países com características semelhantes, resultando três grupos: países com "redes em desenvolvimento", países com "redes consolidadas" e países com "redes de referência".

No que concerne à aplicação da regressão linear múltipla, foi desenvolvido um modelo que expressa a variável referente ao investimento em redes rodoviárias (roadInv) em função dos níveis de emissão de CO2 devido à circulação de veículos na rede (co2Emission) e dos custos de manutenção da rede (roadMnt). Este modelo apresentou um coeficiente de determinação aceitável e satisfez os vários testes de significância estatística realizados.

Finalmente estudou-se a realidade portuguesa à luz das conclusões retiradas ao longo do trabalho. Verificou-se que o País tem realizado investimentos na rede rodoviária acima da média europeia; também a previsão através do modelo de regressão linear múltipla desenvolvido ficou aquém dos valores do investimento em Portugal.

Com a realização deste estudo, foi possível constatar que, na União Europeia, os países em desenvolvimento são os que mais investem na rede rodoviária e que, no caso de Portugal, os valores de investimento excedem o previsto. Num plano mais lato, concluiu-se ainda que a disponibilização pública e livre de dados facilita a realização de estudos comparativos entre diferentes infraestruturas de transporte, algo que, habitualmente, é referido na literatura da especialidade como um dos desafios a ultrapassar.

Referências

Hair, J., Black, W., Babin, B. & Anderson, R. (2014) *Multivariate Data Analysis,* USA: Pearson.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Avaliação do desempenho dos alunos Portugueses em Matemática – PISA 2015

Susana Faria[1], Manuel João Castigo[2]

[1] CMAT – Centro de Matemática, Departamento de Matemática e Aplicações, Universidade do Minho, sfaria@math.uminho.pt;

[2] Departamento de Matemática e Aplicações, Universidade do Minho e Faculdade de Ciências Naturais e Matemática, Universidade Pedagógica de Moçambique, cmanueljoao@gmail.com

Sumário: Neste trabalho, aplicando modelos de regressão multinível, pretende-se comparar os desempenhos escolares dos alunos portugueses relativamente à literacia em Matemática, em diversas regiões portuguesas, com base nos dados disponibilizados pelo *Programme for International Student Assessment* (PISA) de 2015 da Organização para a Cooperação e Desenvolvimento Económico (OCDE). Pretende-se ainda identificar quais os fatores que podem influenciar, positivamente ou negativamente, esse desempenho.

Palavras-chave: Desempenho dos Alunos, Modelo Multinível de Regressão, *Programme for International Student Assessment* (PISA) 2015

O PISA – *Programme International for Students Assessement* é um estudo internacional da responsabilidade da Organização para a Cooperação e Desenvolvimento Económico (OCDE) que visa avaliar, em ciclos de 3 anos, a capacidade que os alunos de 15 anos de diferentes países e economias têm para mobilizar conhecimentos nos domínios da leitura, da matemática e das ciências e responder a situações comuns da vida quotidiana.

Este estudo internacional iniciou-se em 2000 e a edição de 2015 marca o sexto ciclo consecutivo do PISA. Pela primeira vez na história da participação de Portugal no PISA, os estudantes portugueses conseguiram resultados médios significativamente acima da média da OCDE em literatura e ciências e resultados médios que não diferem significativamente da média da OCDE em matemática.

Neste trabalho, usando os dados obtidos no âmbito do PISA 2015, estimam-se modelos de regressão multinível com três níveis (nível 1 – estudantes, nível 2 - escolas e nível 3 – regiões) para comparar os desempenhos escolares dos alunos portugueses por NUTS II, com o intuito de identificar quais são os fatores que podem influenciar esse desempenho.

Os fatores que, ao nível do aluno, serão considerados como podendo influenciar o seu desempenho são as caraterísticas sociais, económicas e culturais dos alunos (tais como, género, situação de imigrante, nível económico, social e cultural, idade que iniciou o

1º ciclo), a sua motivação e o seu empenho escolar, entre outros. Ao nível da escola serão consideradas variáveis relacionadas com infraestruturas da escola (tais como localização, tipo de escola, número de alunos, número médio de alunos por turma, número de computadores ligados à *internet*, rácio entre número de estudantes e o número de professores), variáveis relacionadas com os professores e alunos, entre outras. Ao nível da região, serão consideradas variáveis relacionadas com os recursos educativos, com indicadores de riqueza e com caraterísticas sócio-económicas, entre outras.

Em Portugal, o programa PISA 2015 foi aplicado a 7325 alunos de ambos os sexos, com uma média de idades de 15,8 anos, afectados a 246 escolas. Na distribuição de alunos por NUTS II, constatou-se que uma elevada proporção de alunos frequentava escolas no Norte (29,7%), na Região Autónoma dos Açores (19,1%), no Centro (18,7%) e na Área Metropolitana de Lisboa (17,9%). Cerca de 9,3% dos alunos frequentava escolas no Alentejo. Enquanto que nas escolas do Algarve e da Região Autónoma da Madeira, registaram-se percentagens menores de alunos participantes (2,8% e 2,4%).

Com base nestes dados serão estimados modelos de regressão multinível com os fatores relevantes nos três níveis de análise, que explicam o desempenho escolar dos alunos Portugueses em literacia Matemática.

Agradecimentos: Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - "Fundação para a Ciência e a Tecnologia", no âmbito do projeto PEstOE/MAT/UI0013/2017.

Referências

Agasisti, T., Cordero-Ferrera, J.M. (2013) Educational disparities across regions: A multilevel analysis for Italy and Spain. *Journal of Policy Modeling*, 35, 1079-1102.

Agasisti, T., Ieva, F. & Paganoni, A.M. (2017) Heterogeneity, school-effects and the North/South achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statistical Methods and Applications*, 26 (1), 157-180.

Goldstein, H. (2011) *Multilevel Statistical Models*. John Wiley and Sons.

Pereira, M. C., Reis, H. (2012) What accounts for Portuguese regional differences in students' performance? Evidence from OECD PISA. *Lisboa: Economic Bulletin*, Banco de Portugal.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

O Papel Mediador da Avaliação Cognitiva na Relação entre a Mútua Interferência Trabalho-Família e o B*urnout*

Jéssica Rodrigues[1], A. Manuela Gonçalves[2], Susana Faria[2], Clara Simães[3], A. Rui Gomes[4]

[1] DMAT – Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, pg32814@alunos.uminho.pt

[2] CMAT – Centro de Matemática, DMAT - Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, mneves@math.uminho.pt; sfaria@math.uminho.pt

[3] Escola Superior de Enfermagem, Universidade do Minho, Portugal, csimaes@ese.uminho.pt

[4] Departamento de Psicologia Aplicada, CIPsi, Centro de Investigação em Psicologia, Universidade do Minho, Portugal, rgomes@psi.uminho.pt

Sumário: Aplicando Modelos de Equações Estruturais pretende-se avaliar o papel mediador da avaliação cognitiva na relação entre mútua interferência trabalho-família e a experiência de *burnout* em professores portugueses que lecionam desde o jardim de infância até ao secundário. Os resultados evidenciam o efeito mediador da avaliação cognitiva nessa relação. A avaliação cognitiva é assim um importante mecanismo para explicar a adaptação no trabalho.

Palavras-chave: Avaliação cognitiva, *Burnout*, Mediação, Modelo de equações estruturais, Mútua interferência trabalho-família

As profissões de ensino têm sido associadas pela investigação em saúde ocupacional a elevados níveis de *stress* crónico e de *burnout* (Gomes, Faria, & Gonçalves, 2013). Na atividade profissional dos Professores, porquanto maioritariamente feminina, numa sociedade onde os tradicionalismos de género ainda abundam e onde, o desempenho de múltiplos papéis familiares cabem à mulher, destaca-se a mútua interferência trabalho-família, apontada pela investigação neste domínio, como sendo uma das mais relevantes fontes de *stress* ocupacional. A dificuldade de conciliação da vida laboral com as responsabilidades familiares tem demonstrado significativos efeitos negativos sobre a saúde física, mental e social dos sujeitos (e.g., *distress*, depressão, esgotamento, fadiga crónica, insatisfação laboral e com a vida em geral), justificando a importância de analisar os mecanismos que podem explicar o aparecimento e desenvolvimento do *stress* ocupacional. Um destes mecanismos, refere-se aos processos de avaliação cognitiva que nos indicam até que ponto uma dada situação de *stress* pode ser avaliada pelo indivíduo como mais ameaçadora ou mais desafiadora. Assim, este estudo pretende averiguar se a avaliação cognitiva que os professores fazem da sua atividade laboral (Glaser & Tracy, 2013), poderá influenciar a relação entre a perceção de mútua interferência trabalho-família (e.g., conflito trabalho-família e conflito família-trabalho) e a experiência de *burnout* (e.g., fadiga física, fadiga cognitiva e exaustão emocional).

Neste estudo participaram 438 professores portugueses que lecionam desde o jardim de infância até ao secundário, sendo 304 do sexo feminino (69,4%) e 129 do sexo masculino (29,5%) com idades entre os 28 e os 67 anos (M = 46,85; DP = 7,884). A maioria dos professores é casada (70,8%), 13,5% é solteira, 11% divorciada e 2,7% pertence a outro estado civil. A nível académico, 6 professores têm bacharelato (1,4%), 351 licenciatura (80,1%), 68 (15,5%) mestrado, 6 (1,4%) doutoramento e 2 (0,5%) têm um outro grau académico. O número de anos de serviço dos professores varia de 2 a 42 (M = 22,63; DP = 8,399). A maioria dos professores não tirou baixa no último ano (94,5%), sendo o período máximo de baixa de 15 meses.

Todos os participantes responderam a um protocolo de avaliação constituído pelos seguintes instrumentos de autorrelato: a versão portuguesa do Work-Family Conflict & Family-Work Conflict Scales (WFC & FWC, Netemeyer, Boles, & McMurrian, 1996; adaptado por McIntyre, McIntyre, & Simães, 2006); o Questionário de Avaliação Cognitiva (EAC, Gomes, A.R., 2008); e a Medida de *burnout* de Shirom-Melamed (MBSM, Gomes, A.R., 2012).

O modelo de mediação da avaliação cognitiva foi estimado com o *software* AMOS 24 (Byrne, 2010). No ajustamento do modelo, utilizou-se uma estratégia em duas etapas: na primeira etapa ajustou-se um modelo de medida e na segunda etapa ajustou-se um modelo estrutural. Para avaliar a qualidade de ajustamento do modelo utilizaram-se a estatística χ^2, o índice TLI, NFI e CFI e o RMSEA. O método de reamostragem *Bootstrap* foi aplicado para se obterem os intervalos de confiança das estimativas dos parâmetros (Mackinnon, Lockwood & Wiliams, 2004).

Este estudo confirmou a importância da avaliação cognitiva na relação entre a perceção de mútua interferência trabalho-família e a experiência de *burnout* dos professores portugueses.

Agradecimentos: Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - "Fundação para a Ciência e a Tecnologia", no âmbito do projeto PEstOE/MAT/UI0013/2017.

Referências

Byrne, B. M. (2010) *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, New York: Routledge.

Glaser, W. & Tracy, D. H. (2013) Work-family conflicts, threat-appraisal, self-efficacy and emotional exhaustion. *Journal of Managerial Psychology*, 28(2), 164-182.

Gomes, A. R., Faria, S. & Gonçalves, A. M. (2013) Cognitive appraisal as a mediator in the relationship between stress and burnout. *Work & Stress*, 27(4), 351-367.

Mackinnon, D., Lockwood, C. & Williams, J. (2004) Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Estimação em Pequenos Domínios para obtenção de Estatísticas de Uso e Ocupação do Solo

Suelma Pina[1], Pedro Campos[2], Diana Almeida[3], A. Manuela Gonçalves[4]

[1]Escola de Ciências da Universidade do Minho, Portugal, suelmadepina@hotmail.com;
[2] INE-Instituto Nacional de Estatística, Portugal, pedro.campos@ine.pt;
[3] INE-Instituto Nacional de Estatística, Portugal, diana.almeida@ine.pt;
[4] CMAT-Centro de Matemática da Universidade do Minho, Portugal, mneves@math.uminho.pt.

Sumário: Com base em Estimação em Pequenos Domínios pretende-se avaliar o efeito das fontes de dados nacionais, para obter uma maior qualidade das estimativas a nível sub-regional no que respeita à proporção de área das classes de uso e ocupação do solo (*Land Use/Land Cover*). Este trabalho tem vindo a ser desenvolvido no âmbito do projeto harmonizado LUCAS do Sistema Estatístico Europeu.

Palavras-chave: Estimação em Pequenos Domínios, Informação auxiliar, LUCAS, Uso e Ocupação do Solo.

Introdução

O projeto LUCAS (*Land Use and Cover Area Frame Statistical Survey*), consiste num inquérito que fornece estatísticas harmonizadas e comparáveis sobre o uso e ocupação do solo em todo o território da EU. Com vista à validação e estimação a um nível mais desagregado dos dados provenientes do LUCAS, alguns dos países participantes (incluindo Portugal), desenvolveram no passado um estudo piloto (INE, 2013). Esse estudo visava avaliar a viabilidade da implementação de algum dos estimadores diretos e modificados para a obtenção de estimativas sobre o uso e ocupação do solo para as NUTS III, em Portugal, para estas operações estatísticas.

O objetivo deste estudo consiste em usar informação auxiliar das operações estatísticas tais como o Inquérito às Estruturas das Explorações Agrícolas, assim como das várias operações ligadas às Estatísticas da Produção Vegetal com vista a melhorar a qualidade das estimativas produzidas no âmbito do projeto LUCAS, tendo por base a COS2010 (Carta de Ocupação do Solo). Este estudo começa por apresentar os fundamentos teóricos associados à estimação em Pequenos Domínios. De seguida, apresenta-se a metodologia e descrevem-se alguns resultados preliminares.

Enquadramento e Metodologia

A divulgação de informação a níveis detalhados constitui um aspeto importante nos processos de tomada de decisão a nível regional e local. Publicar informação a níveis

geográficos ou de atividade económica desagregados permite uma significativa melhoria no conhecimento da realidade. No entanto, devido à insuficiente dimensão das amostras em algumas destas desagregações (aqui designadas por domínios), os habituais métodos de estimação não permitem a produção de indicadores com precisão adequada, conduzindo a resultados com grande variabilidade (Saei and Chambers, 2003). Os métodos de estimação em pequenos domínios permitem contornar os problemas associados à dificuldade em divulgar informação relativa a domínios de interesse para os quais não foi recolhida informação que permita obter estimativas fiáveis dos parâmetros que se pretende estimar. A introdução de informação espacial, através de efeitos aleatórios espacialmente correlacionados permite reduzir o enviesamento e estimar de forma mais eficiente os parâmetros do modelo. Assim, foi introduzido por Cressie (1992) um modelo com efeitos aleatórios espacialmente correlacionados e, posteriormente, foi considerada por Pratesi e Salvati (Pratesi and Salvati, 2008) uma extensão do modelo Fay-Herriot através de processos SAR (simultâneos autorregressivos). As abordagens espaciais no contexto da estimação em pequenos domínios são baseadas em modelos mistos. Neste trabalho foram comparados estimadores em pequenos domínios com o intuito de analisar a melhoria da precisão das estimativas ao nível das NUTS III das variáveis associadas ao Inquérito à Estrutura das Explorações Agrícolas (IEEA). Em seguida usou-se esta informação para melhorar a qualidade das unidades de superfície da COS2010 (Carta de Ocupação do Solo), de forma a assegurar um maior detalhe no nível 2 e 3 da nomenclatura LUCAS GT 2015.

Agradecimentos: Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT – Fundação para a Ciência e Tecnologia no âmbito do projeto PEstOE/MAT/UI0013/2014. Especial agradecimento ao Instituto Nacional de Estatística no qual está a ser realizado um estágio que conduziu ao presente trabalho.

Referências

INE (2013), Instituto Nacional de Estatística, Departamento de Estatísticas Demográficas e Sociais (2013), (Relatório Interno) Tipologia, integração e validação de dados de uso e ocupação do solo (sinergias LUCAS /Sistemas Nacionais de Informação), Lisboa.

Saei, A., Chambers, R. (2003). Small area estimation under linear and generalized linear model with time and area effects, Working Paper M03/15 Southampton Statistical Sciences Research Institute, University of Southampton

Cressie, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology* 18, 75–94

Pratesi, M., Salvati, N. (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat. Methods Appl.* 17, 113–141.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Clustering directional data from von Mises-Fisher distribution: a comparison of algorithms

Adelaide Figueiredo
Faculdade de Economia da Universidade do Porto, LIAAD – INESC TEC Porto e CEAUL, adelaide@fep.up.pt

Abstract: In the statistical analysis of directional data, the von Mises-Fisher distribution plays an important role to model unit vectors. In this study we consider a dynamic clusters type algorithm for clustering directions and we compare the solutions of this algorithm to those obtained with three variants of the *EM* algorithm. We also evaluate the solutions obtained with the algorithms through between-groups and within-groups variability measures. The comparison of the algorithms is made for simulated and real spherical data.

Keywords: Directional data, dynamic clusters algorithm, *EM* algorithm, von Mises--Fisher distribution.

Clustering data in the unit sphere is an important task in modern data analysis, for example, in clustering text documents when analyzing textual data. Several works have appeared in the literature for clustering directional data based on model-based clustering methods. In particular, Banerjee *et al.* (2005) applied a model-based clustering of directional data to text analysis. These authors considered the estimation of a mixture of von Mises-Fisher components using two variants of the Estimation-Maximization (*EM*) algorithm, denoted by soft-movMF and hard-movMF algorithms. The von Mises-Fisher distribution is one of the most used distributions in the statistical analysis of directional data and has probability density function defined by

$$f(\mathbf{x}|\mu,\kappa) = c_p(\kappa)\exp(\kappa\mu'\boldsymbol{x}), \quad \boldsymbol{x} \in S_{p-1},\ \mu \in S_{p-1}, \kappa > 0,$$

where the normalizing constant is given by $c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}\, I_{p/2-1}(\kappa)}$ and $I_\nu(.)$ denotes the modified Bessel function of the first kind and order ν and S_{p-1} denotes the unit sphere in $\mathbb{R}^p$. The parameter μ is the vector of the mean direction and κ is the concentration parameter around μ. This distribution is called Fisher distribution for spherical data.

In this study we develop a dynamic clusters type algorithm for estimating the parameters of a mixture of von Mises-Fisher components and for obtaining a partition of the sample into clusters. We considered this algorithm for the estimation of a mixture of Watson components defined in the unit sphere in $\mathbb{R}^p$ in Figueiredo & Gomes (2015).

Then we compare the solutions of the dynamic clusters type algorithm with three variants of the *EM* algorithm (soft-movMF and hard-movMF variants given in Banerjee *et al.*, 2005 and the stochastic *EM* proposed by Celeux & Govaert, 1992) in a simulation

study and in an example with data given in the literature. For evaluating the solutions obtained in the algorithms, we also determine between-groups and within-groups variability measures.

In the simulation study we generated samples from a mixture of equal proportions of two Fisher components, with a common concentration parameter. We supposed several sample sizes, various angles of separation between the two components and different values of the concentration parameter.

The simulations revealed that only for poorly or moderately separated components, the variants of the *EM* algorithm and the dynamic clusters type algorithm lead to different solutions in general. For very well separated components, the algorithms seem to originate the same result. Additionally, the larger concentration parameter associated with the components, greater is the tendency to obtain the same solution for the algorithms.

For each algorithm, as expected, the between-groups variability increases as the separation between components increases and for well separated components, the between-groups variability exceeds largely the within-groups variability. The between--groups variability also increases when the concentration of components increases, since these components are not badly separated.

Aknowledgments: This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project «POCI-01-0145-FEDER-006961».

References

Banerjee A., Dhillon I. S., Ghosh J. & Sra S. (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6, 1345-1382.

Celeux G. & Govaert G. (1992) A classification EM algorithm for clustering and two-stochastic versions. *Computational Statistics and Data Analysis*, 14, 3, 315-332.

Figueiredo, A. & Gomes, P. (2015) Clustering of variables based on Watson distribution on hypersphere: a comparison of algorithms. *Communications in Statistics - Simulation and Computation*. Special issue: Joint Meeting of y-BIS and jSPE, vol. 44, issue 10, 2622-2635.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Influence of the subprime crisis on public sector in European countries in the period 2002-2011

Catarina Soares[1], Adelaide Figueiredo[2], Fernanda Otília Figueiredo[3]

[1] NOS, catarina_lourenco@live.com

[2] Faculty of Economics and LIAAD-INESC Porto, University of Porto, and CEAUL, adelaide@fep.up.pt

[3] Faculty of Economics, University of Porto, and CEAUL, University of Lisbon, otilia@fep.up.pt

Abstract: The subprime crisis quickly became a global financial crisis, affecting a large number of countries, including the European economies. Europe also faces a crisis of public debt, particularly since the beginning of 2010, having the Greek debt served as a fuse. In this economic context and with the twenty-seven European countries being mentioned as having very different economies, it becomes interesting to identify, analyze and discuss the main weaknesses and similarities of the public sector of these economies. For this purpose, inspired by the early warning systems, eight variables are analyzed for the twenty-seven European countries during the period 2002-2011, through the STATIS methodology.

Keywords: Financial crisis, principal components analysis, public debt crisis, STATIS methodology.

The economic downturn, the bailouts and fiscal stimulus had a large budgetary impact in many countries (Mishkin, 2011). As Reinhart and Rogoff (2009) and Mishkin (2011) pointed out, after a financial crisis there was a considerable increase in public debt, increasing the risk of a sovereign default. Mishkin (2011) also states that *"having public accounts in order will be a top priority for governments all around the world."*

Through the analysis of some vulnerability indicators considered in early warning systems, which are intended to predict the occurrence of a crisis in a certain horizon of time, our aim is to analyze the evolutionary trends of the public sector in member states of the European Union during the period 2002-2011. For this purpose, we used the STATIS methodology (Lavit *et al.*, 1994), a three-way exploratory technique based on Principal Components Analysis (PCA).

In the present study twenty-seven member states of the European Union are considered and studied in the years 2002, 2004 and 2006 to 2011 and eight variables are analyzed. Indeed, recently, one of the problems that have haunted the majority of European countries relates to the stabilization and reduction of the public debt. This relative stabilization depends on the government budget, the difference between the interest rate required for the country and the growth rate of the product, as well as other adjustments on it. Additionally, countries have three possible ways of financing public deficits: via taxes, debt accumulation and monetary emission, whose effects can be felt

negatively on inflation. Thus, some of the variables identified above were considered, as public revenues and public expenses, in order to emphasize this issue. The eight variables considered were long-term sovereign interest rates, short-term interest rates, public expenses, public revenues, government budget, public debt, GDP growth rate and inflation rate.

After a preliminary analysis of data, we analyzed the results of STATIS methodology applied to the European countries described by the variables of public sector. The STATIS methodology allowed us to obtain quite interesting and relatively surprising findings about common features and dissimilarities in European countries. In particular, the methodology opposed the years before and after the financial crisis of 2007 and 2008, as expected, and led us to conclude that the years of the groups (2002, 2004), (2006, 2007) and (2009, 2010, 2011) are in general identified as being the most similar in what concerns to the similarities between these countries and the correlations between variables, but opposite when compared to others, being the year 2008 a "*turning point*".

Acknowledgements: This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia within projects POCI-01-0145-FEDER-00696 and UID/MAT/00006/2013 (CEA/UL).

References

Lavit, C., Escoufier, Y. , Sabatier, R. & Traissac, P. (1994) The ACT (Statis method). *Computational Statistics and Data Analysis*, 18, 87–119.

Mishkin, F. S. (2011) Over the Cliff: From the Subprime to the Global Financial Crisis. *Journal of Economic Perspectives*, 25, 49-70.

Reinhart, C. M. & Rogoff, K. S. (2009) The Aftermath of Financial Crises. *American Economic Review,* 99, 466-472.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Avaliação estatística da eficácia do projeto AdolesSer

Paulo Infante[1], Gonçalo Jacinto[2], Anabela Afonso[3], Ana Carla Coelho[4], Ana Gabriela Pontes[5], Isabel Fernandes[6], Edgar Palminhas[7]

[1] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt;
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, gjcj@uevora.pt;
[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, aafonso@uevora.pt;
[4] ARSA/ACES AC/ UCC Évora, AnaCarla.Coelho@alentejocentral.min-saude.pt;
[5] ARSA/ ACES AC/ UCC Évora, Ana.Pontes@alentejocentral.min-saude.pt;
[6] ARSA/ ACES AC/ UCC Évora, Isabel.Fernandes@alentejocentral.min-saude.pt;
[7] Porta Mágica, edpalminhas@hotmail.com.

Sumário: O projeto AdolesSer, da equipa de saúde escolar da Unidade de Cuidados na Comunidade de Évora, é dirigido a alunos do 6º e 9º ano de Escolas de Évora e pretende reforçar/promover a informação sobre as dimensões da sexualidade humana, crescimento e relações interpessoais, assim como a aquisição de competências. Neste estudo faz-se uma breve caracterização dos alunos do Concelho de Évora em termos de conhecimentos, comportamentos e atitudes sobre a sexualidade e avalia-se a eficácia das intervenções realizadas com base em questionários submetidos antes e após essas intervenções.

Palavras-chave: Análise classificatória, Análise de correspondências múltiplas, Modelos lineares generalizados, Teoria de resposta ao item.

A educação sexual foi integrada, por lei, na educação para a saúde, com vista à promoção da saúde física, psicológica e social dos indivíduos. O projeto AdoslesSer – Projeto de educação para a saúde sexual e reprodutiva - desenvolvido pelo Agrupamento de Centros de Saúde do Alentejo Central (ACES), em particular pela equipa de saúde escolar da Unidade de Cuidados na Comunidade de Évora, visa complementar a abordagem da educação sexual, com as turmas de 6º e 9º ano de escolaridade aderentes, articulando com o trabalho desenvolvido pelos docentes. Pretende apoiar de forma mais estruturada e continuada as escolas, no desenvolvimento do projeto de educação sexual de turma.

A avaliação da eficácia do projeto, bem como das melhorias a implementar, é concretizada pela aplicação de um questionário inicial de validação de conceitos e conhecimentos (pré-intervenção) antes da realização da primeira sessão de intervenção e a sua reaplicação imediatamente após a realização da última sessão (pós-intervenção). Aos alunos das turmas selecionadas foi solicitada a resposta anonimizada aos questionários na sala de aula.

Os questionários eram compostos, para além de diversas questões de caracterização do aluno, por outras questões que visavam avaliar os conhecimentos sobre as temáticas abordadas nas intervenções, que se podem agrupar em: corpo, sistema reprodutor, grupo e o próprio, para alunos do 6º ano; contracepção, infeções sexualmente transmissíveis,

violência, orientação sexual, comportamento na relação e pressão de pares, para alunos de 9º ano. Em algumas questões a resposta era do tipo Verdadeiro ou Falso, sendo avaliado se o aluno acertou ou não na resposta. Noutras questões o aluno indicava se Concordava ou Discordava da afirmação, tendo-se analisado se o aluno deu ou não a resposta esperada.

A percentagem média de respostas corretas e dadas de acordo com o esperado aumentou após a intervenção e diminuiu a percentagem de respostas "não sei", em ambos os anos de escolaridade. Existem pequenas diferenças entre sexos em cada uma das passagens. As raparigas inquiridas tiveram, em média, um desempenho superior ao dos rapazes embora existam questões em que os rapazes revelam um maior conhecimento. Não foram detetadas diferenças no desempenho dos alunos por tipo de freguesia de residência dos alunos nem por já terem ou não tido experiência sexual, nem pelo tipo e variedade de fontes de informação a que recorrem para obterem informação sobre a sexualidade.

Para averiguar associações entre as respostas dadas às questões (categorizadas apenas em duas categorias, resposta certa ou de acordo com o esperado *vs*. caso contrário), por ano de escolaridade, realizou-se uma análise de correspondências múltiplas e identificaram-se alguns perfis. Posteriormente foi usada uma análise classificatória (*K-means*) tendo-se identificado grupos de alunos, que correspondiam aos perfis, cujas principais características são: alunos que tenderam a acertar na maior parte das questões após a intervenção, alunos que de um modo geral não acertaram em quase todas as questões antes da intervenção, e os alunos que mostraram falha de conhecimentos em algumas questões específicas.

Recorrendo à teoria da resposta ao item (TRI) foi possível, em particular, identificar questões que poderiam ser substituídas por outras por pouco contribuírem ao nível da discriminação e informação que fornece sobre os conhecimentos dos alunos. Conforme os subgrupos de questões e o ano, os modelos que melhor se ajustaram foram o modelo de Rash (parâmetro dificuldade) e o modelo de 2 parâmetros (dificuldade e discriminação).

Recorrendo a testes de associação e a modelos lineares generalizados foi ainda possível identificar algumas características dos alunos que afetam o seu desempenho. Em particular, recorrendo à regressão logística, conclui-se que um bom desempenho dos alunos do 9º ano, antes da intervenção, é potenciado por viverem em freguesias urbanas, e obterem informação sobre sexualidade através da mãe, de pessoal de enfermagem e não de revistas, não sendo afetado por fatores como o sexo, a escolaridade dos pais, a situação perante o trabalho do encarregado de educação e a opinião que têm deles próprios.

Agradecimentos: Paulo Infante, Gonçalo Jacinto e Anabela Afonso são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Valores Humanos e Religião na Europa

Maria Paula Lousão[1], Cláudia Silvestre[2], José Luís Casanova[3]
[1] *Instituto Universitário de Lisboa (ISCTE_IUL), Centro de Investigação e Estudos de Sociologia (CIES-IUL),* mlousao@escs.ipl.pt, maria.paula.lousao@iscte.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.pt.

Sumário: O modelo de valores humanos desenvolvido por *Schwartz* (1992) considera a existência de dez domínios motivacionais reconhecidos em diferentes culturas de todo o mundo, em que os indivíduos apenas diferem entre si no grau de importância que atribuem a cada valor. Neste trabalho, que tem por base este modelo de valores contrastantes de *Schwartz,* pretende-se estudar os valores humanos em todas as denominações religiosas entre a população europeia, com o objetivo de avaliar se, e como, a relação entre religião e valores tem vindo a mudar nos diversos países da Europa.

Palavras-chave: Análise de Agrupamento, *European Social Survey*, Mudança Cultural, Religião, Valores Humanos

Neste trabalho pretende-se conhecer as tendências e as alterações de valores humanos nas diversas culturas nacionais e religiosas, o que muda e o que se mantém. Com esse objetivo, analisou-se os dados do *European Social Survey (*ESS*) e* tomou-se por base os trabalhos desenvolvidos por *Schwartz e Boehnke ((2004) e Brites (2011)*. Embora existam outras teorias sobre os valores humanos (por ex. *Ronald Inglehart*, 1997), como os valores humanos desenvolvidos por *Schwartz e Boehnke* estão na génese da construção do questionário do ESS, optou-se por considerar, neste estudo, os 10 valores humanos desenvolvidos por estes autores, a saber, benevolência, universalismo, realização, poder, autodeterminação, estimulação, hedonismo, segurança, conformismo e tradição.

Sendo que um dos objetivos deste estudo é a avaliar as diferenças existentes entre as diversas culturas religiosas, houve a preocupação de se escolher as rondas onde a Rússia estivesse presente, para se verificar se este país, maioritariamente ortodoxo, tem valores humanos diferentes ou não, dos restantes países europeus de maioria ortodoxa. Assim, neste trabalho, considerou-se as rondas 3 (2006), 4 (2008), 5 (2010) e 6 (2012) do ESS. De forma a ter uma análise consistente, optou-se por selecionar os 22 países comuns às 4 rondas. Segmentaram-se os 22 países de acordo com os resultados apresentados nestes 10 valores humanos. Para avaliar a (dis)semelhança entre os países foi usada a distância euclideana. Aplicaram-se 3 métodos hierárquicos: o método de Ward, o método da ligação dentro dos grupos e o método do vizinho mais afastado. Apenas houve um país, a Espanha, que não se manteve sempre no mesmo segmento. Nos métodos da ligação dentro dos grupos e do vizinho mais afastado a Espanha está no mesmo segmento, mas muda de segmento quando se usa o método de Ward. A segmentação das religiões, de acordo com os valores humanos, obteve-se também através dos 3 métodos anteriormente referidos. A segmentação, tanto dos países, segundo os valores, como das religiões

foi repetida para cada uma das 4 rondas (um exemplo de um dos resultados obtidos encontra-se nas figuras 1 e 2).

Para caraterizar os segmentos (de países e de religiões) obtidos, usaram-se variáveis externas à segmentação. Com o objetivo de encontrar diferenças significativas entre os segmentos recorreu-se ao teste do Qui-Quadrado e ao teste de Kruskal-Wallis.

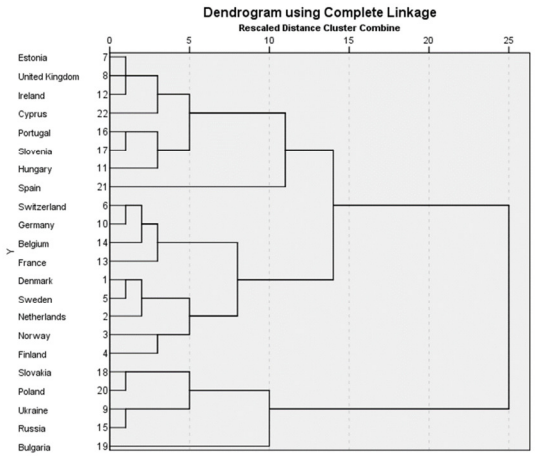


Figura 1

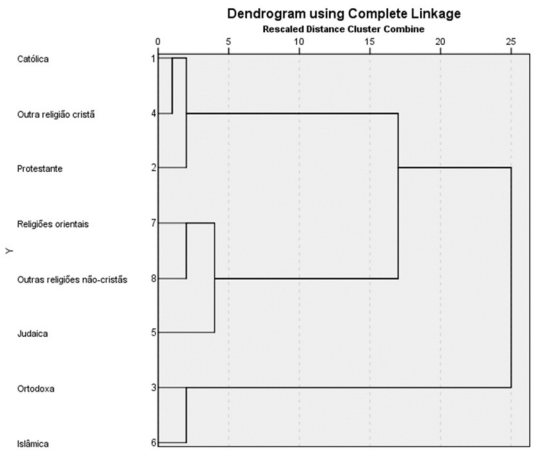


Figura 2

Pretende-se com este trabalho, o qual é o início de uma investigação mais aprofundada, contribuir para a melhor compreensão dos valores humanos, no que respeita à relação destes com as diversas religiões e, também com os diversos países.

Os resultados sugerem agrupamentos estáveis em todos os modelos hierárquicos utilizados, quer ao nível das religiões, quer ao nível dos países.

Referências

Brites, R. (2011). *Valores e Felicidade no Século XXI. Um Retrato Sociológico dos Portugueses em Comparação Europeia*. Lisboa. ISCTE-IUL (tese de doutoramento).

Inglehart, R. (1997). *Modernização e pós-modernização: A mudança cultural, económica e política em 43 sociedades*. Princeton: Princeton University Press.

Schwartz, S., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality*, 38, 230-255.

Sessão de Posters II – 6ª Feira, 21 de Abril (16:20)

Análise de textos sobre os hábitos dos agregados familiares portugueses em matéria de resíduos alimentares.

Conceição Rocha[1], Belmira Neto[2], Arian Pasquali[3], Alípio M. Jorge[4]
[1] INESC TEC, Universidade do Porto, conceicao.n.rocha@inesctec.pt;
[2] FEUP & CEMUC, Universidade do Porto, belmira.neto@fe.up.pt;
[3] INESC TEC, Universidade do Porto, arianpasquali@gmail.com;
[4] FCUP & INESC TEC, Universidade do Porto, amjorge@fc.up.pt

Sumário: Foram recolhidas informações relativas às práticas perante resíduos alimentares, através de um inquérito *on-line* realizado no âmbito de um estudo académico, onde os consumidores foram especificamente indagados sobre o desperdício de alimentos que ocorrem nos seus agregados familiares. No mesmo inquérito, foi colocada uma pergunta aberta cuja análise, por recurso a técnicas de *text mining*, resultou numa lista de medidas autopropostas pelos inquiridos para gerir os resíduos alimentares. Isto constitui uma fonte crucial de informação a ser utilizada pelas empresas do sector dos resíduos, permitindo compreender o papel dos consumidores na perda de alimentos.

Palavras-chave: Resíduos alimentares, Inquérito, Categoria morfossintática, *Text mining*, Tópicos.

O sector da gestão de resíduos alimentares procura compreender melhor as práticas atuais dos consumidores em matéria de resíduos alimentares produzidos a partir de refeições em casa. Isso servirá de base não só para as decisões estratégicas sectoriais como também para formular amplas campanhas de comunicação para a redução dos resíduos alimentares.

Com o objetivo de recolher informação sobre os hábitos dos agregados familiares portugueses em matéria de resíduos alimentares foi efetuado um estudo em parceria com a LIPOR (Divisão Intermunicipal de Gestão de Resíduos do Grande Porto – Portugal) que procedeu à divulgação de um inquérito (*on-line*) através da sua página web.

Foram capturadas as práticas de resíduos alimentares através de um inquérito *on-line*, onde os consumidores foram especificamente indagados sobre o desperdício de alimentos que ocorrem nos seus agregados familiares. O inquérito concentrou-se em 22 alimentos considerados os mais consumidos em Portugal. O trabalho realizado identificou os diferentes itens alimentares desperdiçados nas refeições consumidas em domicílios portugueses.

Atendendo a que o inquérito foi feito *on-line*, podemos considerar que a participação da população na pesquisa foi surpreendente uma vez que 457 indivíduos responderam ao inquérito. Os participantes são de ambos os géneros e distribuem-se por todas as faixas etárias como se pode observar pelos valores presentes na Tabela 1.

Tabela 1: Algumas características dos inquiridos.

Faixa etária	Género		Total
	Feminino	Masculino	
18 – 25	45	15	60
26 - 35	87	35	122
36 - 45	104	32	136
46 - 55	53	18	71
56 - 65	38	15	53
Mais de 65	9	6	15
Total	336	121	**457**

Neste trabalho, são analisadas as respostas à questão 'Sugira medidas para combater o desperdício alimentar ou indique o que já faz em sua casa'. Para explorar o conjunto de comentários produzidos espontaneamente pelos entrevistados são aplicadas técnicas de extracção de informação (Mooney & Bunescu, 2005). Em particular, é marcada de forma automática a categoria morfossintática de cada termo e, em seguida, são analisados os termos mais frequentes dentro das categorias mais relevantes, como verbos e substantivos. Como se pode observar, pela nuvem de expressões presentes na Figura 1, esta técnica conduziu a uma lista de medidas autopropostas pelos inquiridos para gerir os resíduos alimentares.

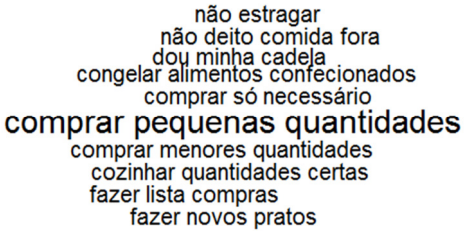


Figura 1: Nuvem de expressões, presentes nas respostas, com maior frequência.

Em uma segunda camada mais semântica, realizamos a descoberta automática de tópicos, utilizando LDA (Blei *et al.,* 2003), e analisamos os termos mais relevantes para cada tópico. Esta análise foi complementada com a identificação dos adjectivos frequentemente associados a termos como, *p.e.*, sobras e compras.

Os resultados obtidos constituem uma fonte crucial de informação a ser utilizada pelas empresas do sector dos resíduos, permitindo compreender o papel dos consumidores na perda de alimentos.

Referências

Mooney, R. J. & Bunescu, R.(2005) Mining knowledge from texto using information extraction. *ACM SIGKDD Explorations Newsletter*, 7(1), 3-10.

Blei, D., Ng, A. & Jordan, M.(2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research archive, 3, 993-1022*.

Índice de Autores