Reunião Anual da

ASSOCIAÇÃO PORTUGUESA DE CLASSIFICAÇÃO

E ANÁLISE DE DADOS (CLAD)

PROGRAMA E LIVRO DE RESUMOS

Évora, Portugal

AS JOCLAD 2016 TIVERAM O APOIO INSTITUCIONAL DE:

CLAD
Associação Portuguesa de
Classificação e Análise de Dados

UNIVERSIDADE
DE ÉVORA

FICHA TÉCNICA

Presidente das Jornadas

José Dias (Presidente da CLAD)

Secretário das Jornadas

Paulo Infante (Universidade de Évora)

Comissão Organizadora

Anabela Afonso (Universidade de Évora)

Gonçalo Jacinto (Universidade de Évora)

Isabel Silva (Faculdade de Engenharia da Universidade do Porto)

Maria Filomena Mendes (Universidade de Évora)

Paulo Infante (Universidade de Évora)

Título: XXIII Jornadas de Classificação e Análise de Dados (JOCLAD 2016)
Programa e Livro de Resumos.

Produzido: Instituto Nacional de Estatística

Editores: Paulo Infante, Anabela Afonso, Gonçalo Jacinto,
Maria Filomena Mendes, Isabel Silva

ISBN: **978-989-98955-2-2**

PREFÁCIO

As XXIII Jornadas de Classificação e Análise de Dados (JOCLAD 2016) decorrem este ano na Universidade de Évora, a segunda Universidade a ser fundada em Portugal.

Foi criada pelo Cardeal D. Henrique, a partir do já existente Colégio do Espírito Santo, com a anuência do Papa Paulo IV, expressa na bula *Cum a nobis* de Abril de 1559, tendo a inauguração solene decorrido no dia 1 de Novembro desse mesmo ano. Ainda hoje, neste dia se comemora o aniversário da Universidade, com a cerimónia da abertura solene do ano académico. Duzentos anos após a fundação a Universidade foi encerrada, em consequência do decreto de expulsão dos jesuítas. A partir da Segunda metade do século XIX, instalou-se no Colégio do Espirito Santo o Liceu de Évora, ao qual a rainha Dona Maria II concedeu a prerrogativa do uso de "capa e batina", em atenção à tradição universitária da cidade e do edifício. Em 1973, por decreto do então ministro da Educação, José Veiga Simão, foi criado o Instituto Universitário de Évora que viria a ser extinto em 1979, para dar lugar à nova Universidade de Évora.

É nesta nossa casa do Honesto estudo com longa experiência misturado que com muita honra recebemos pela primeira vez as Jornadas promovidas pela Associação Portuguesa de Classificação e Análise de Dados (CLAD), fundada em 1994.

O volume de dados disponível tem aumentado muito nos últimos anos. Torna-se ainda mais fundamental saber extrair deles o essencial e ser capaz de sintetizar e transmitir esse essencial numa linguagem facilmente entendível. A classificação e análise de dados permite um avanço no conhecimento que temos sobre os diferentes fenómenos e constituiu uma importante ferramenta de auxílio à tomada de decisão. A sua importância nas mais diversas áreas está bem patente nas comunicações apresentadas nestas Jornadas.

Na sequência de uma já longa tradição, o Programa das JOCLAD 2016 reflete o carácter multidisciplinar das mesmas, enquadrando de forma equilibrada a apresentação de trabalhos teóricos e aplicados. Para além de trabalhos de proposta livre (em formato oral e poster) e as sessões temáticas, temos as sessões plenárias a cargo dos professores José Fernando Vera (Universidade de Granada, Espanha), Alfred Stein (Universidade de Twente, Holanda) e Antónia Turkman (CEAUL), a quem muito agradecemos.

Agradecemos também a todos os autores 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 deste livro. Um agradecimento particular ao Professor José Fernando Vera (Universidade de Granada, Espanha) e aos Professores Russell Alpizar-Jara e Anabela Afonso (Universidade de Évora) que lecionam os mini-cursos, bem como às Professoras Manuela Neves (ISA-UL) e Sónia Pintassilgo (ISCTE), Doutora Filipa Lima (Banco de Portugal) e Dr. Carlos Marcelo (Instituto Nacional de Estatística) que se disponibilizaram para organizar as Sessões Temáticas que constam do Programa.

Por último, desejamos agradecer as todas as entidades que direta ou indiretamente apoiaram ou patrocinaram estas Jornadas.

O nosso obrigado a todos!

Pel'A Comissão Organizadora JOCLAD 2016 Pel'A Direcção da CLAD

Paulo Infante José Dias

Évora, março de 2016

ORGANIZAÇÃO

Presidente das Jornadas

José Dias (Presidente da CLAD)

Secretário das Jornadas

Paulo Infante (Universidade de Évora)

Comissão Organizadora

Anabela Afonso (Universidade de Évora)

Gonçalo Jacinto (Universidade de Évora)

Isabel Silva (Faculdade de Engenharia da Universidade do Porto)

Maria Filomena Mendes (Universidade de Évora)

Paulo Infante (Universidade de Évora)

COMISSÃO CIENTÍFICA

A. Manuela Gonçalves (Universidade do Minho)
Adelaide Figueiredo (Universidade do Porto)
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)
Fátima Salgueiro (Instituto Universitário de Lisboa)
Fernanda Sousa (Universidade do Porto)
Fernando Nicolau (Universidade Nova de Lisboa)
Gilda Soromenho (Universidade de Lisboa)
Gonçalo Jacinto (Universidade de Évora)
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


UNIVERSIDADE
DE ÉVORA


CLAD
Associação Portuguesa de
Classificação e Análise de Dados


BANCO DE PORTUGAL
EUROSISTEMA


PSE
PRODUTOS E SERVIÇOS DE ESTATÍSTICA


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


CÂMARA MUNICIPAL
DE ÉVORA


UNIVERSIDADE DE ÉVORA
CIMA
UNIVERSIDADE da MADEIRA


UNIVERSIDADE
DE ÉVORA
CIDEHUS
Centro Interdisciplinar
de História, Culturas e Sociedades
da Universidade de Évora
UID/HIS/00057/2013


30 ANOS
évora
PATRIMÓNIO MUNDIAL


OLARIA PATALIM
A arte do barro


ROTA DOS
VINHOS DO ALENTEJO


ADEGA
DE BORBA


UNIVERSIDADE DE ÉVORA
ZEA
Sociedade Agrícola Unipessoal, Lda


cimac
COMUNIDADE INTERMUNICIPAL
DO ALENTEJO CENTRAL


1852
LUSO


EDIÇÕES SÍLABO, Lda.
Publicamos conhecimento
www.silabo.pt


mazda


COMPETE
2020


PORTUGAL
2020


UNIÃO EUROPEIA
Fundo Europeu
de Desenvolvimento Regional


FCT
Fundação para a Ciência e a Tecnologia

PROGRAMA

QUINTA-FEIRA, 31 DE MARÇO

9:00 Registo e entrega de documentação – Sala 134

9:30 **Mini-curso** – Sala 124

Amostragem em populações de difícil acesso

Russell Alpizar-Jara e Anabela Afonso, p. 9.

Moderador: Isabel Silva

10:45 **Pausa para café**

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

12:30 **Almoço**

13:30 **Mini-curso** – Sala 124

Introduction to Multidimensional Scaling Models

José Fernando Vera, p. 11.

Moderador: Gonçalo Jacinto

15:00 **Pausa para café**

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

16:30 **Sessão de Abertura das Jornadas** – Anfiteatro 131

17:00 **Sessão Plenária I** – Anfiteatro 131

Appreciating Spatial Statistics through case studies with a focus on health under environmental and natural conditions

Alfred Stein (Universidade de Twente, Holanda), p. 15.

Moderador: Russell Alpizar-Jara

18:00 **Pequena visita à cidade e Alentejo de Honra nas instalações da Rota dos Vinhos**

SEXTA-FEIRA, 1 DE ABRIL

9:00 Registo e entrega de documentação

9:30 **Sessão Paralela I**

	Sala 124 ***Modelos Estocásticos e Modelos Espaciais*** Moderador: Carlos Braumann	Anfiteatro 131 ***Classificação e Regressão*** Moderador: Fernanda Sousa
9:30	**Wavelet-based detection of outliers in time series of counts** Isabel Silva, Maria Eduarda Silva, p. 53.	**Discriminant analysis and classification of distributional and interval data** Sónia Dias, Paula Brito, Paula Amaral, p. 59.
9:50	**Políticas de pesca sustentáveis ótimas com esforço constante versus políticas de pesca ótimas com esforço variável: aplicação em ambiente aleatório com um modelo logístico** Nuno M. Brites, Carlos A. Braumann, p. 55.	**Modelos de regressão multinível no estudo do desempenho escolar** Susana Faria, João Silva, p. 61.
10:10	**Spatial data analysis: A comparison of ICM algorithms** Luís F. Domingues, José G. Dias, p. 57.	**A utilização dos Mídia Social para efeitos de viagem: Uma abordagem a partir da análise de agrupamento** Carla Henriques, Suzanne Amaro, Paulo Duarte, p. 63.

10:40 **Pausa para café**

11:00 **Sessão Temática I – Banco de Portugal**

	Anfiteatro 131 ***Economia e Finanças*** Moderador: Filipa Lima
11:00	**The indebtedness of Portuguese SMEs and the impact of leverage on their performance** Ana Filipa Carvalho, Manuel Perestrello, Mário Lourenço, p. 23.
11:30	**The usefulness of granular data - The new statistics based on Banco de Portugal's Central Credit Register** Rodrigo Batista, Isabel Alpiarça, p. 25.

12:00 **Sessão Plenária II** – Anfiteatro 131

Cluster Multidimensional Scaling for large proximity datasets

José Fernando Vera (Universidade de Granada, Espanha), p. 17.

Moderador: José G. Dias

13:00 **Almoço**

14:00 **Sessão Temática II – Instituto Nacional de Estatística**

	Anfiteatro 131 ***Desafios nas Estatísticas Oficiais V*** Moderador: Carlos Marcelo
14:00	**A construção de uma tipologia socioeconómica para as Áreas Metropolitanas de Lisboa e Porto: 2011 e evolução 2001-2011** Cátia Nunes, Francisco Vala, p. 27.
14:20	**SIMSTAT – Um modelo para a simplificação das estatísticas do Comércio Internacional** Cristina Neves, p. 29.
14:40	**Transmissão Automática de Dados para o INE** Luísa Pereira, p. 31.
15:00	**Acesso à informação estatística oficial para fins de investigação científica** Pinto Martins, p. 33.

15:20 **Sessão de Apresentação de Posters** – Anfiteatro 131

O impacto da cultura organizacional no desempenho financeiro das empresas da região Norte de Portugal

Flávia Araújo, Conceição Castro, Fernanda A. Ferreira, p. 103.

GDP per capita dynamics in the European Union

José G. Dias, p. 105.

Alterações climáticas na incidência de casos de dengue na cidade de Goiânia no período de 2008 a 2015

Susana Faria, Antônio Neco, Raquel Menezes, p. 107.

Simulating deterministic and stochastic SVEIR models to determine the disease elimination time for different vaccination rates

Luiz S. Freitas, Hyun Mo Yang, Carlos A. Braumann, p. 109.

O modelo de regressão logística na identificação de factores associados ao relato inconclusivo do rastreio de retinopatia diabética

A. Manuela Gonçalves, Inês Barros, João Reis, p. 111.

A análise estatística multivariada na avaliação da qualidade de águas subterrâneas

A. Manuela Gonçalves, Driano Rezende, Letícia Nishi, Fernanda O. Tavares, M. Teresa Amorim, Rosángela Bergamasco, p. 113.

A Análise Classificatória na caracterização da produção e consumo de produtos de origem animal a nível mundial

Manuel Minhoto, Luís Fernandes, p. 115.

Avaliação quantitativa do património edificado: o caso de estudo centro do Porto

Cilísia Ornelas, Fernanda Sousa, João Miranda Guedes, Isabel Breda-Vázquez, p. 117.

ANOVA a dois fatores não paramétrica com células omissas

Dulce G. Pereira, Anabela Afonso, p. 119.

Características psicométricas da Escala de Empatia de Jefferson em estudantes de Tecnologias da Saúde

Ana Reis, Helena Martins, Ana Salgado, Andreia Magalhães, Zita Sousa, Artemisa R. Dores, p. 121.

Efeito da alteração de medicamentos sujeitos a receita médica para não sujeitos a receita médica

Teresa Risso, Cláudia Furtado, p. 123.

Avaliação das perceções dos estudantes do 1º ano em relação à praxe académica

O. Silva, S. N. Caldeira, M. Mendes, S. Botelho, M. J. Martins, Á. Sousa, p. 125.

Perfis de estudantes no contexto do empreendedorismo: Análise de correspondências múltiplas e análise de clusters

Áurea Sousa, Gualter Couto, Nélia Branco, Osvaldo Silva, Helena Bacelar-Nicolau, p. 127.

O Teste da Razão de Verosimilhanças em Modelos com Equações Estruturais: Uma Abordagem Multi-grupos ao Estudo da Privação Material em Portugal com Dados do ICOR

Paula C. R. Vicente, Maria de Fátima Salgueiro, p. 129.

Assessment of sustainable development over time in OECD and BRICKS

Nikolai Witulski, José G. Dias, p. 131.

16:40 Pausa para café

17:00 **Sessão Paralela II**

	Sala 124 ***Controlo de Qualidade e Análise de Sobrevivência*** Moderador: Luís Grilo	Anfiteatro 131 ***Modelos com Variáveis Latentes*** Moderador: Fátima Salgueiro
17:00	**Uma aplicação da carta EWMA a dados correlacionados** Dora Carinhas, Paulo Infante, p. 65.	**Avaliação das propriedades psicométricas e validação do NEO-FFI para estudantes portugueses de tecnologias da saúde** Helena Martins, Ana Reis, Ana Salgado, Andreia Magalhães, Zita Sousa, Artemisa R. Dores, p. 71.
17:20	**A importância da adequabilidade do modelo no desempenho de cartas de controlo com risco ajustado** Maria João Inácio, Paulo Infante, Fernanda Otília Figueiredo, p. 67.	**Valor percebido do consumidor e comércio de retalho: uma abordagem multidimensional hierárquica de 2ª ordem à versão reduzida da escala PERVAL** João Saramago, Ana Sampaio, Elizabeth Reis, p. 73.
17:40	**Modelo de regressão de Cox robusto, uma aplicação a dados oncológicos** Eunice Carrasquinha, André Veríssimo, Susana Vinga, p. 69.	**Dados omissos em modelos de análise fatorial confirmatória não balanceados: estudo de simulação para detetar dimensão mínima da amostra** Maria de Fátima Salgueiro, Paula C. R. Vicente, p. 75.

18:00 Reunião da Assembleia Geral da CLAD – Anfiteatro 131
(visita à universidade para os não sócios)

20:00 **Jantar das Jornadas**

SÁBADO, 2 DE ABRIL

9:00 Registo e entrega de documentação

9:30 **Sessão Paralela III**

	Sala 124 ***Modelos de Mistura*** Moderador: A. Manuela Gonçalves	Anfiteatro 131 ***Robustez e Análise de Dados Composicionais*** Moderador: Conceição Amado
9:30	**The number of clusters on trust** Cláudia Silvestre, Margarida G. M. S. Cardoso, Mário T. Figueiredo, p. 77.	**Análise estatística das migrações internas usando biplots composicionais** Adelaide Freitas, Maria Cristina Gomes, Maria Luís Pinto, p. 83.
9:50	**How perfect is an imperfect test? A biomedical challenge** Ana Subtil, M. Rosário Oliveira, António Pacheco, p. 79.	**Nonparametric limits of agreement for vitamin B12** Luís M. Grilo, Helena L. Grilo, p. 85.
10:10	**Cluster-based conjoint models: An application to professional services marketing** José G. Dias, p. 81.	**Robust confidence intervals using minimum-distance method** Teresa Risso, Conceição Amado, Ana M. Pires, p. 87.

10:30 **Pausa para café**

10:50 **Sessão Temática III – Demografia**

	Anfiteatro 131 ***Demografia*** Moderador: Sónia Pintassilgo
10:50	**Da topologia à tipologia: padrões e tipos de divórcio em casais nacionais e binacionais** Madalena Ramos, Ana Cristina Ferreira, Sofia Gaspar, p. 35.
11:10	**Family dynamics in the relation between fertility and housing: diversity in the Southern European residential system** Alda Botelho Azevedo, Julián López Colás, Juan A. Módenes, p. 37.
11:30	**Imigração e mercado de trabalho: leituras de mobilidade a várias escalas nas regiões de Lisboa, Odemira e Algarve** Alina Esteves, p. 39.
11:50	**Envelhecimento demográfico: o desafio social do município de Évora** Filipe Ribeiro, Lídia P. Tomé, Maria Filomena Mendes, p. 41.

12:15 **Sessão Plenária III** – Anfiteatro 131

Religião, Fogos e Estatística: ponto de encontro

Antónia Turkman (CEAUL, Portugal), p. 19.

Moderador: Paulo Infante

13:15 **Almoço**

14:15 **Sessão Temática IV – Aplicações da Estatística em Ciências Biológicas**

	Anfiteatro 131 ***Aplicações da Estatística em Ciências Biológicas*** Moderador: Manuela Neves
14:15	**Time series analysis of remote sensing data** Clara Cordeiro, Sónia Cristina, Priscila C. Goela, Sergei Danchenko, John Icely, Samantha Lavender, Alice Newton, p. 43.
14:35	**Modelos mistos aplicados à análise da variabilidade genética intravarietal e selecção de castas antigas de videira** Elsa Gonçalves, p. 45.
14:55	**Uma abordagem bioestatística na caraterização do Envelhecimento Vascular Precoce** P. G. Cunha, J. Cotter, P. Oliveira, I. Vila, N. Sousa, p. 47.
15:15	**Spatio-temporal structure of ecological data: a three-way study** Susana Mendes, M.ª José Fernández-Gómez, Sónia Cotrim Marques, Ulisses Miranda Azeiteiro, Paulo Maranhão, Sérgio Miguel Leandro, M.ª Purificación Galindo-Villardón, p. 49.

15:35 **Sessão de Posters (sem apresentação)**

Conflito trabalho-família: Validação de um instrumento de medida para a Marinha Portuguesa

Sandra Veigas Campaniço, Dora Carinhas, Miguel Pereira Lopes, p. 133.

Factores influentes no sucesso vs. insucesso nas escolas da Província do Cunene

Palmira Caseiro, Helena Bacelar-Nicolau, Jorge Santos, Fernando da Costa Nicolau, p. 135.

Famílias estruturadas de matrizes estocásticas simétricas

Cristina Dias, Carla Santos, João Tiago Mexia, p. 137.

O sono das crianças do 1º ciclo: caso de estudo numa escola do concelho de Évora

Paulo Infante, Anabela Afonso, Gonçalo Jacinto, Teresa Engana, Filipe Gloria Silva, Rosa Espanca, p. 139.

A Bayesian LASSO method for replicated data

Jacinto Martín, Carlos J. Pérez, Lizbeth Naranjo, Yolanda Campos-Roca, p. 141.

Caracterização das explorações agrícolas e dos produtores de caprinos de raça Serpentina

Manuel Minhoto, António Fonseca, Luís Fernandes, António Cachatra, p. 143.

Estimação e condensação em modelos mistos, normais e não normais, com estrutura ortogonal por blocos

Carla Santos, Célia Nunes, Cristina Dias, João Tiago Mexia, p. 145.

Aprendizagem automática para a classificação da severidade da Doença Pulmonar Obstrutiva Crónica

Matheus Coppetti Silveira, p. 147.

What do you like in a hostel? Exploring the determinants of satisfaction

Paula Vicente, Rita Lima, p. 149.

15:50 **Pausa para café**

16:10 **Sessão Paralela IV**

	Sala 124 ***Regressão Logística*** Moderador: Paula Vicente	Anfiteatro 131 ***Gestão da Informação e Redes Neuronais*** Moderador: Susana Faria
16:10	**Avaliação do efeito do desenho de amostragem em modelos de regressão logística** Ana Laura Carreiras, Paulo Infante, Anabela Afonso, Maria Filomena Mendes, p. 89.	**A Cauda Longa: A sua existência no mercado de retalho Online Português** Juliana Rocha Costa, p. 95.
16:30	**A decisão de permanecer sem filhos a partir dos 30 anos de idade** Andréia Maciel, Rita Brazão Freitas, Maria Filomena Mendes, Paulo Infante, p. 91.	**Geração sintética de microdados utilizando algoritmos de data mining** Daniel Silva, Pedro Campos, Pavel Brazdil, p. 97.
16:50	**A ecografia como instrumento de diagnóstico – um estudo de caso** Ana Matos, Carla Henriques, Jorge Pereira, A. C. Afonso, J. Constantino, p. 93.	**Classificação acústica automática de espécies de morcegos** Bruno Silva, Gonçalo Jacinto, Paulo Infante, Sílvia Barreiro, Pedro Alves, p. 99.

17:10 Sessão de Encerramento das Jornadas

RESUMOS

Índice

MINICURSOS

Minicurso I - 5ª feira, 31 de março, sala 124 (9:30)

Amostragem em populações de difícil acesso

Russell Alpizar-Jara[1], **Anabela Afonso**[2]
[1, 2] *Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, e Departamento de Matemática, Escola de Ciência e Tecnologias, Universidade de Évora;*
[1] *alpizar@uevora.pt;*
[2] *aafonso@uevora.pt*

Um dos problemas fundamentais na amostragem de populações de difícil acesso é a inexistência de uma base de amostragem para enumerar as unidades da população. Esta restrição muitas vezes impossibilita a utilização dos métodos usuais de amostragem em populações finitas. Por outro lado, o tamanho destas populações é desconhecido e geralmente um dos objetivos do estudo é estimar a sua dimensão.

Neste minicurso, abordaremos algumas das técnicas de amostragem e métodos de estimação comumente utilizadas para estimar populações de difícil acesso, como por exemplo, imigrantes ilegais, trabalhadoras do sexo, os sem-abrigo, indivíduos com alguma doença crónica não comunicável, etc…. Incidiremos particularmente na amostragem por captura-recaptura, onde a estimação leva em conta o uso de listagens ou fontes de informação, geralmente incompletas, e os elementos comuns às várias listagens, para estimar as probabilidades de inclusão dos indivíduos da população e assim estimar a sua dimensão. As aplicações iniciais desta metodologia remontam às populações humanas, mas tem tido uma grande diversidade de utilizações nas ciências biológicas, ciências médicas, ciências sociais, controlo de qualidade e inspeção de *softwares*, entre outras. Outras metodologias, denominadas por amostragem indireta, em rede (*network*), adaptativa, conduzida pelo respondente (*response-driven*), e variantes destas abordagens serão também referidas no minicurso.

Palavras-chave: amostragem adaptativa, amostragem conduzida pelo respondente, amostragem em rede, amostragem indireta, captura-recaptura.

Agradecimentos: Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto «UID/MAT/04674/2013 (CIMA)».

Referências

Lavallée, P. (2007). *Indirect sampling*. New York, USA: Springer.

Seber, G. A., & Salehi, M. M. (2013). *Adaptive sampling designs: inference for sparse and clustered populations*. New York, USA: Springer Science & Business Media.

Thompson, W. (Ed.) (2004). Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters. Island Press.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (Eds.) (2014) *Hard-to-survey populations*. Cambridge, UK: Cambridge University Press.

Williams, B. K., Nichols, J. D., & Conroy, M. J. (2002) Analysis and management of animal populations: modeling, estimation, and decision making. San Diego, USA: Academic Press.

Minicurso II - 5ª feira, 31 de março, sala 124 (13:30)

Introduction to Multidimensional Scaling Models

José Fernando Vera[1]
[1] *Universidad de Granada, España, jfvera@ugr.es*

MDS was formally introduced in the 50s by Torgerson's metric method, known as classical MDS. It is a statistical technique originating in psychology and psychometrics. The data used for multidimensional scaling (MDS) are proximities between pairs of objects, usually dissimilarities.

The main objective of MDS is to represent these dissimilarities as distances between points in a low dimensional space such that the distances approach as closely as possible to the dissimilarities. The use of MDS is not limited to psychology but has applications in a wide area of disciplines, such as sociology, economics, biology, chemistry, archaeology, etc. Often, it is used as a technique for exploring the data. In addition, it can be used as a technique for dimension reduction.

In today's software, MDS comprises different models in terms of the type of geometry into which one wants to map the data, the mapping function, the algorithms used to find an optimal data representation, the treatment of statistical error in the models, or the possibility to represent not just one but several similarity matrices at the same time.

This short course will give an overview of the basic concepts about multidimensional scaling modelling, and will present examples of its use in diverse sciences. The practice of MDS with standard software, such R, will be discussed (other software such as SPSS is open to discussion upon request of the delegates).

1. Introduction to MDS
2. Classical MDS
3. Least squares MDS
4. The general MDS model for one-way one-mode data
5. MDS for three-way data sets
6. Extensions of MDS in connection with k-means, finite mixtures, log-linear analysis, etc.
7. Discussion

Bibliography

Borg I. & Groenen, P.J.F. (2005) *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.). New York, NY: Springer.

Borg I., Groenen, P.J.F., & Mair (2016) *Applied Multidimensional Scaling*. New York, NY: Springer.

Cox, T. F., & Cox, M. A. A. (2000) *Multidimensional Scaling* (2nd ed.). New York, NY: Chapman & Hall/CRC.

De Leeuw, J. & Mair, P. (2009) Multidimensional scaling using majorization: The R package smacof. *Journal of Statistical Software*, 31(3), 1-30, http://www.jstatsoft.org/v31/i03/

SESSÕES PLENÁRIAS

Appreciating Spatial Statistics through case studies with a focus on health under environmental and natural conditions

Alfred Stein[1]
[1] *University of Twente, Faculty of Geo-information Science and Earth Observation (ITC) PO Box 217, 7500 AE Enschede, The Netherlands, a.stein@utwente.nl*

Abstract: Spatial statistics is an emerging field in statistics, with references to the environment, health and natural conditions. In this presentation, typical cases studies are presented that all fall in this domain.

Keywords: Environment, Health, Natural conditions**,** Spatial statistics.

Spatial statistics has been an asset in analysis of health studies since long. Recently further interest is growing in health geography with novel opportunities for spatial statistics. In this presentation, some recent developments are given. First, attention is given towards relating *buruli ulcer* (BU) with soil and groundwater related variables. CAR based modeling was applied on villages in Ghana. The study revealed an association between (a) the mean As content of soil and spatial distribution of BU and (b) the distance to sites of gold mining and spatial distribution of BU. We concluded that both arsenic in the natural environment and gold mining influence BU infection. Second, attention is given towards spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area. Hierarchical modeling and Bayesian structured additive regression modeling were carried out. Both spatial and spatial temporal data were analyzed. Statistical modeling using OLS model reveals a significant negative association between (a) cholera prevalence and proximity to all the potential cholera reservoirs ($R^2 = 0.18$, $p < 0.001$) and (b) cholera prevalence and proximity to upstream potential cholera reservoirs ($R^2 = 0.25$, $p < 0.001$). The inclusion of spatial autoregressive coefficients in the OLS model reveals the dependency of the spatial distribution of cholera prevalence on the spatial neighbors of the communities. A flexible scan statistic identifies a most likely cluster with a higher relative risk ($RR = 2.04$, $p < 0.01$) compared with the cluster detected by circular scan statistic ($RR = 1.60$, $p < 0.01$). We conclude that surface water pollution through runoff from waste dump sites play a significant role in cholera infection. Finally, an important issue concerns a geospatial analysis of HIV-related social stigma that focused on females across Indian mandals. Data were available at a very detailed level. The spatial analysis shows that women in India move towards a different mandal for getting tested on HIV. Given the scale of study and different types of movements involved, it is difficult to say where they move to and what the precise effect is on HIV registration. Better recording the addresses of tested women may help to relate HIV incidence to population present within a mandal. This in turn may lead to a better incidence count and

therefore add to more reliable policy making, e.g. for locating or expanding health facilities.

References

Duker, A., Stein, A. & Hale, M. (2006) A statistical model for spatial patterns of Buruli ulcer in the Amansie West district, Ghana. *International Journal of Applied Earth Observation and Geoinformation,* 8, 126–136.

Kandwal, R., Augustijn, E.W., Stein, A., Miscione, G., Garg, K.P., & Garg, R.D. (2010) Geospatial analysis of HIV-Related social stigma: A study of tested females across Indian mandals. *International Journal of Health Geographics,* 2010, 9–18.

Osei, F. B., Duker, A. A., Augustijn, E. W. & Stein, A. (2010) Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. *International Journal of Applied Earth Observation and geoinformation*,12, 331–339.

Sessão Plenária II – 6ª feira, 1 de abril, Anfiteatro 131 (12:00)

Cluster Multidimensional Scaling for large proximity datasets

José Fernando Vera[1]
[1] *Universidad de Granada, España, jfvera@ugr.es*

Classification and spatial methods can be used in conjunction to represent the information of similar proximity data by means of groups. This methodology is particularly advisable in many practical applications in market research, psychology, sociology, environmental research, genomics, and information retrieval for the Web and other document databases, in which data consist of large similarity or dissimilarity measures on each pair of objects.

The combination of classification and spatial methods has the advantage that the objects or individual information is summarized by means of groups, which also significantly reduces the number of parameters to be estimated in the model. Therefore, not the original points as such but rather the cluster centres are located in a low-dimensional space. Several of these combined procedures have been proposed in a deterministic framework for two-way data (Heiser & Groenen 1997), but also for preference data (Vera, Macías & Heiser 2013).

From the perspective of a latent class models, several models for preference data have been proposed for clustering individuals while simultaneously the given groups of individuals and the objects are represented in a low dimensional space. Latent class models in combination with Simulated Annealing have been proposed for one-mode dissimilarity data (Vera, Macías & Heiser 2009b). For two-mode preference data, dual classification procedures have also been proposed (Vera, Macías & Heiser 2009b) that cluster individuals and objects while simultaneously both cluster centers are represented in a low dimensional space using Unfoldind.

References

Heiser, W. J., & Groenen, P. J. F. (1997) Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika*, 62, 63–83.

Vera, J. F., Macías, R., & Heiser, W. J. (2009a) A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika*, 74(2), 297–315.

Vera, J. F., Macías, R. & Heiser, W. J. (2009b) A dual latent class unfolding model for two-way two-mode preference rating data. *Computational Statistics and Data Analysis*, 53(8), 3231–3244.

Vera, J. F., Macías, R., & Heiser, W. J. (2013) Cluster differences unfolding for two-way two-mode preference rating data. *Journal of Classification*, 30(3), 370–396.

Religião, Fogos e Estatística: ponto de encontro

Antónia A. Turkman[3], José M. C. Pereira[1], Duarte Oom[1], Paula Pereira[2], K. Feridun Turkman[3]

[1] Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal;

[2] Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal;

[3] Centro de Estatística e Aplicações, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal, maturkman@fc.ul.pt

Sumário: Recentemente apareceu uma notícia interessante no Washington Post (Mooney, 2015) com o seguinte título: "Scientists may have just found a really surprising way to see religion from space". Este texto baseava-se nos resultados de um artigo (Earl *et al*., 2015) onde se conclui que, à escala global, os fogos exibem ciclos semanais, com significativamente menos fogos ao Domingo do que em outros dias da semana, atribuindo-se isto a padrões do comportamento humano onde a religião tem um papel preponderante. De facto, conclusão semelhante já tinha sido por nós obtida (Pereira et al, 2015) e previamente publicada, através de um estudo feito com dados do Continente Africano. Nesta comunicação apresenta-se esse estudo, onde estatísticos se aliam a investigadores especialistas em fogos florestais para tentar perceber até que ponto o comportamento humano, ditado pela prática religiosa, tem algum papel na justificação da existência desse padrão cíclico.

Palavras-chave: Fogos florestais, Modelos bayesianos hierárquicos, Regressão binomial negativa, Religião cristã, Religião muçulmana.

As queimadas (ver Figura I) são uma prática comum em África na gestão das áreas agrícolas, onde o fogo é usado para a caça, criação de gado, controle de pragas, colecta de alimentos, fertilização de terras cultiváveis, e prevenção de incêndios. Dado o forte controle antropogénico do fogo, testamos a hipótese de que estes fogos exibem ciclos semanais, e que o dia da semana com menor número de fogos depende da filiação religiosa regional predominante. Analisamos também o efeito do uso da terra como um factor significativo no ciclo semanal. A densidade de fogo (contagens/km^2) observada em cada dia da semana e em cada região foi modelada utilizando um modelo de regressão binomial negativo, com número de fogos como variável resposta, área de região como *offset* e um efeito aleatório estruturado para explicar a dependência espacial. O uso da terra (habitado, agricultura, natural, pastagens), religião (cristã, muçulmana, misto) dia da semana, e correspondentes interacções foram utilizadas como variáveis independentes. Os modelos também foram construídos separadamente para cada uso de terra, relacionando a densidade de incêndio com o dia da semana e filiação religiosa. A análise revelou uma interacção significativa entre religião e dia da semana, ou seja, regiões com diferentes filiações religiosas (cristã, muçulmana) exibem ciclos semanais

distintos de queima. No entanto, a interacção entre religião e dia da semana é apenas significativa para terras agrícolas, sendo que a actividade do fogo em terras agrícolas africanas é significativamente menor ao domingo em regiões predominantemente cristãs e na sexta-feira em regiões predominantemente muçulmanas. A magnitude da actividade do fogo não difere significativamente entre os dias da semana em pastagens e em áreas naturais, onde o uso do fogo está sob menor controle rigoroso do que em terras agrícolas. Estas conclusões são úteis na medida em que podem contribuir para a melhoria da especificação dos padrões de ignição em modelos globais/regionais de dinâmica de vegetação, e pode levar a uma previsão meteorológica mais precisa.

Figura I: Queimada em África

Fonte: *Gentilmente cedida por José M. C. Pereira*

Agradecimentos: Fundação para a Ciência e a Tecnologia (http://www.fct.pt/) PEst-OE/AGR/UI0239/2014 (JMCP e DO), PEst-OE/MAT/UI0006/2014 (KFT, AT e PP), PTDC/MAT/118335/2010 (KFT, AT, PP e JMCP) e bolsa de doutoramento SFRH/BD/4752/2008 (DO).

Referências

Earl Nick, Simmonds Ian & Tapper Nigel, (2015) Weekly cycles of global fires—Associations with religion, wealth and culture, and insights into anthropogenic influences on global climate. *Geophysical Research Letters, RESEARCH LETTER 10.1002/2015GL066383*. First Published on line 9 NOV 2015.

Mooney, Chris. (2015) Scientists may have just found a really surprising way to see religion from space. *Washington Post, Energy and Environment,* 17 November 2015.

Pereira José M. C., Oom Duarte, Pereira Paula, Turkman Antónia A. & Turkman K. Feridun (2015) Religious Affiliation Modulates Weekly Cycles of Cropland Burning in Sub-Saharan Africa. *PLoS ONE*, 10(9): e0139189. doi:10.1371/journal.pone.0139189, Published: September 29, 2015.

SESSÕES TEMÁTICAS

Sessão Temática I – Banco de Portugal – 6ª feira, 1 de abril, Anfiteatro 131 (11:00)

The indebtedness of Portuguese SMEs and the impact of leverage on their performance

Ana Filipa Carvalho[1], Manuel Perestrello[2], Mário Lourenço[3],
[1] *Banco de Portugal, mpvasconcelos@bportugal.pt;*
[2] *Banco de Portugal, mfllourenco@bportugal.pt;*
[3] *Banco de Portugal, afcarvalho@bportugal.pt*

Abstract: This paper aims to provide empirical evidence on the effect of indebtedness on the performance of Portuguese small and medium-sized enterprises (SMEs). Using *Banco de Portugal*'s Central Balance-Sheet Database, the analysis points to the fact that financial debt is not being used to increase these companies' profitability. Instead, higher debt levels seem to be increasingly linked to companies that eventually cease their activity.

Keywords: Enterprises, Indebtedness, Leverage, Profitability, SMEs.

Small and medium-sized enterprises (SMEs), which exclude microenterprises and holding companies, are a relevant part of the non-financial corporations (NFC) sector in Portugal. In 2014, although accounting for only 10% of the total number of enterprises, they represented 42% for turnover and 45% for the number of employees. Debt has played a significant role in these companies' sustainability. In 2014, 68% of SMEs' assets were funded by debt (average capital ratio of 32%) [Banco de Portugal 2015]. But is debt being used as a tool to expand the activity and achieve better performance or as a way to cover day-to-day activities? Addressing this issue seems to be of particular relevance in the context of the recent economic and financial crisis and considering the deleveraging effort recently undergone by the Portuguese economy.

Micro data available at *Banco de Portugal*'s Central Balance-Sheet Database were used in order to determine each company's financial leverage (through their capital ratio, i.e., the ratio between the company's equity and its total assets) and profitability levels (using the return on assets ratio, i.e., net profit on total assets) for each year in the 2006-14 period. SMEs were scored from 1 to 4 according to their level of financial leverage and their positioning within the quartile distribution of individual capital ratios. This score translates the company's performance as either being above 75% of its peers (score 4), above 50% but below 25% of its peers (score 3), above 25% but below 50% of its peers (score 2) or in the bottom 25% of the registered performances (score 1). The same procedure was carried out regarding the profitability ratio. The company's leverage score was then linked to the profitability score for the subsequent three years.

A special flag was considered in cases where the company ceased its activity, a situation determined using *Banco de Portugal*'s business register (which combines information from several databases managed by *Banco de Portugal*, as well as other

administrative sources) [Gonçalves *et. al.* 2013]. Results show that a significant share of Portuguese SMEs with the lowest capital ratio levels (hence, highest financial leverage levels) have ceased their activity in the following years (6%, considering a one year timespan; 20% if a three year timespan is considered) (Chart 1). These results contrast with the share of enterprises that ceased their activity after having registered low leverage levels (1% after one year and 5% after three years).

Among companies that did not cease activity, the share of SMEs with low profitability seems to increase with indebtedness (Chart 2): 42% of the companies with the highest leverage registered low profitability after three years, while for the least leveraged firms the percentage drops to 23%. On the other hand, only 17% of the most highly leveraged SMEs reached high profitability levels in three years, for the least leveraged firms the percentage increases to 30%. In fact, except for score 3, the profitability score seems to be linked to the same indebtedness score; for example, in the lowest leverage quadrant (score 4 at time T), the largest share of firms exhibits high profitability (score 4 at time T+3). These results seem to provide some evidence that high financial leverage is not associated with short/medium term profitability for Portuguese SMEs. Therefore, it could be argued that a significant share of Portuguese SMEs seem to be indebted (carrying non-profitable debt) rather than leveraged (debt leading to higher profitability levels).

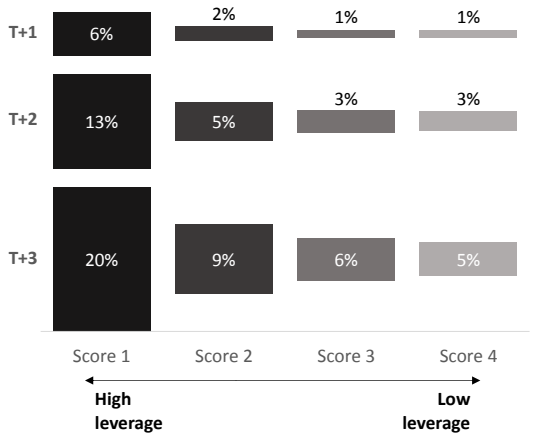


Chart 1: SMEs that ceased their activity according to financial leverage scores

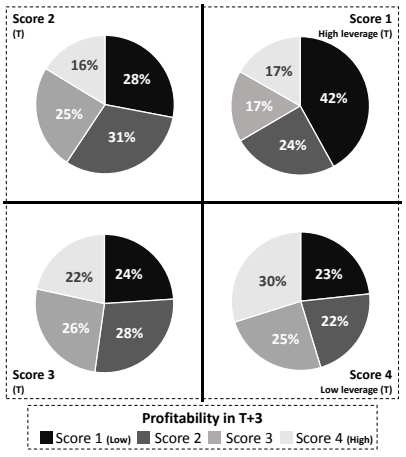


Chart 2: SMEs according to financial leverage scores and profitability scores

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) *Sectoral analysis of non-financial corporations in Portugal 2010-2015*. Central Balance Sheet Study, 23.

Gonçalves, H. & Lourenço, M. (2013) Building business registers to monitor entrepreneurial dynamics. *IFC Bulletin*, 37, 42-45.

Sessão Temática I – Banco de Portugal – 6ª feira, 1 de abril, Anfiteatro 131 (11:30)

The usefulness of granular data - The new statistics based on Banco de Portugal's Central Credit Register

Rodrigo Batista[1]**, Isabel Alpiarça**[2]
[1] *Banco de Portugal, rsbatista@bportugal.pt;*
[2] *Banco de Portugal, ifalpiarca@bportugal.pt*

Abstract: The constant need for information represents a challenge both in terms of how and when to acquire it. The Portuguese Central Credit Register contains monthly granular information on credit on a borrower-by-borrower basis, allowing detailed analysis unachievable to other databases as well as interlinkages with other micro-databases. This granular information has been a key factor in meeting most of the data demands Banco de Portugal has been faced with in this domain.

Keywords: Central Bank statistics, Central credit register, Granular databases.

The challenges posed to a national central bank represent a constant need for information. Predicting beforehand which information might be required and how to obtain it, so that users have it when needed, is the challenge laid to a statistics department. The Statistics Department of Banco de Portugal has the experience of managing several granular databases and of being able to surpass this challenge through their use.

One of this databases is the Central Credit Register (CCR), which contains granular data about all the loans above 50 euros granted by the financial sector. Besides statistical compilation, CCR's granular data is used for several purposes such as: reduction of the information gap between borrowers and lenders, supervision of financial entities, analysis of the stability of the financial system and conduction of monetary policy.

The CCR is able to fulfil all of its purposes due to the several variables that define each credit record but also due to an important advantage of granular data – the possibility of crossing its information with other internal granular databases. These interlinkages allow for the characterization of borrowers, in the case of non-financial corporations (NFC), by dimensions such as size, sector of economic activity or exporting company status, thus providing additional insights to the analysis of the loans' market in Portugal.

In February 2016, Banco de Portugal, using the potential of this granular data, published a new set of statistics based on the CCR. Charts 1 to 4, only achievable through the use of this potential, are examples of these statistics.

Chart 1 details the evolution of loans granted to NFC in default according to their sector of economic activity. The information regarding the sector of activity of the NFC is not directly available in the CCR, being acquired through a connection to the Central

Balance-Sheet Database of Banco de Portugal (CBSD). The analysis of Chart 1 depicts an increase both in the overdue loans ratio and in the share of borrowers with overdue loans across all sectors between 2010 and 2015.

Chart 2 presents information which aggregates households according to their indebtedness levels vis-à-vis the resident financial sector. Observing Chart 2 it is noticeable that the overdue loans ratio registers its highest value in the smallest brackets of credit amount (less than 25 000 euros) while the highest share of borrowers with overdue loans is verified in the case of borrowers indebted in more than 250 000 euros.

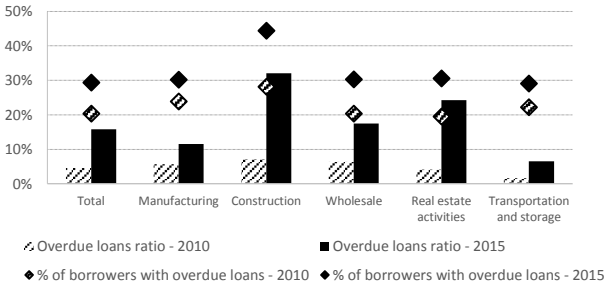


Chart 1: Evolution of the default indicators concerning loans granted to NFC by economic activity

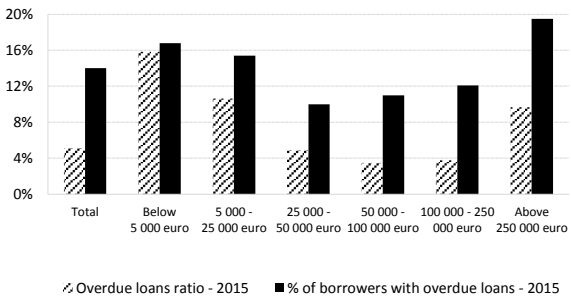


Chart 2: Default indicators concerning loans granted to households by bracket of credit amount

Charts 3 and 4 present relationship indicators between NFC and financial entities. These charts represent two examples of the advantages of granular data: firstly the NFC are classified according to their size, also through a connection to the CBSD, and secondly each borrower has to be evaluated regarding its indebtedness to each individual financial institution. Analysing Chart 3, it is observed that the average number of financial entities with which each borrower has a credit relation has decreased since 2010 while the analysis of Chart 4 shows that the concentration of loans in a single financial entity has increased.

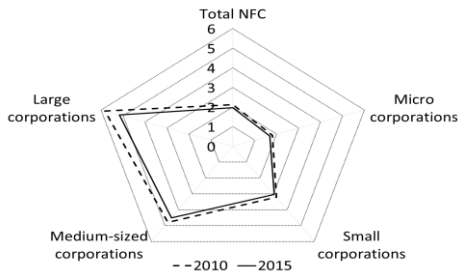


Chart 3: Average number of financial entities with which each borrower has a credit relation

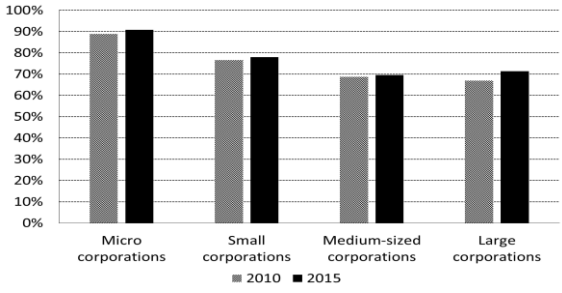


Chart 4: Average percentage of loans granted by the entity with the largest share

Disclaimer: The analysis, 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.

A construção de uma tipologia socioeconómica para as Áreas Metropolitanas de Lisboa e Porto: 2011 e evolução 2001-2011

Cátia Nunes[1], **Francisco Vala**[2]
[1] *Instituto Nacional de Estatística, catia.nunes@ine.pt;*
[2] *Instituto Nacional de Estatística, francisco.vala@ine.pt*

Sumário: Este estudo tem como principal objetivo a elaboração de um quadro de referência para a análise das características territoriais das áreas metropolitanas de Lisboa e Porto, através da produção de uma tipologia de padrões socioeconómicos com base em análises multivariadas aplicadas a indicadores censitários operacionalizados à escala da subsecção estatística. A definição de uma tipologia para 2011 é complementada com uma perspetiva de evolução temporal tendo em vista a caracterização dos processos de transformação ocorridos entre 2001 e 2011. Os resultados obtidos evidenciam as semelhanças e as diferenças entre os dois espaços metropolitanos e salientam as potencialidades da informação censitária no que respeita ao detalhe da escala territorial e, em particular, da Base de Georreferenciação de Edifícios 2011.

Palavras-chave: Análise multivariada, Áreas metropolitanas, Censos, Tipologia.

Com este estudo, foi dada continuidade ao trabalho iniciado pelo INE para as duas áreas metropolitanas, primeiro com base nos resultados dos Censos de 1991 e, seguidamente, com base na informação censitária de 2001. O presente estudo consiste numa análise similar mas atualizada em resultado da disponibilização dos dados dos Censos de 2011 e pelo facto de considerar um modelo de análise integrado dos dois territórios metropolitanos, tendo em consideração a sua configuração geográfica atual. Esta leitura comparada permitiu evidenciar padrões territoriais comuns e identificar características diferenciadoras entre a Área Metropolitana de Lisboa (AML) e a Área Metropolitana do Porto (AMP).

A tipologia produzida teve por base informação proveniente dos Censos de 2011, disponível à escala da subsecção estatística (Base Geográfica de Referenciação da Informação - BGRI). A recolha da informação de base procurou abarcar diferentes domínios socioeconómicos relevantes para a caracterização das duas áreas metropolitanas e, nesse sentido, foram considerados 23 indicadores abarcando as unidades estatísticas edifícios, alojamentos, famílias e indivíduos.

O modelo de análise consistiu, primeiramente, na aplicação de uma análise fatorial em componentes principais com o intuito de obter um número relativamente reduzido de variáveis latentes capazes de sintetizar grande parte da variabilidade presente na informação inicial. Esta análise permitiu identificar cinco dimensões socioeconómicas: *envelhecimento, qualificação, urbanização, imigração* e *mobilidade pendular.* Posteriormente, procedeu-se a uma análise de *clusters*, de acordo com o método não

hierárquico das K-médias e com base nas componentes extraídas da análise fatorial, que permitiu construir classes de subsecções estatísticas homogéneas do ponto de vista socioeconómico. A solução retida contemplou seis classes: *urbano consolidado, (sub)urbano novo qualificado, (sub)urbano não qualificado, espaços integrados de menor densidade, espaços autocentrados de menor densidade* e *espaços de imigração.*

Os resultados evidenciaram que as áreas metropolitanas são compostas por territórios heterogéneos e fragmentados. Na AML a fragmentação revelou-se especialmente notória em torno dos designados eixos de expansão suburbana que se formam a partir dos territórios limítrofes ao município de Lisboa. Na AMP a fragmentação traduziu-se na existência de espaços de (sub)urbanização que revelam uma oposição centro-periferia, evidenciando a centralidade do município do Porto e uma coroa de expansão suburbana circunferencial que abarca os municípios de Matosinhos, Maia, Valongo, Gondomar e Vila Nova de Gaia. A classe socioeconómica referente aos espaços de imigração evidenciou particularmente o território da AML, apresentando-se menos caracterizadora da AMP.

Esta análise foi complementada com uma perspetiva de evolução temporal com o intuito de retratar as transformações socioeconómicas ocorridas entre 2001 e 2011. Para este fim, e de modo a assegurar uma repartição territorial que permitisse a referenciação de informação dos Censos de 2001 e de 2011 numa base comum, foram consideradas as subsecções estatísticas de 2001 com informação estatística referenciada de 2001 e 2011. Este processo apenas foi possível a partir da georreferenciação dos edifícios em 2011. Foi desenvolvido um modelo de análise autónomo que consistiu em trabalhar conjuntamente os dados relativos aos dois anos e em aplicar as duas técnicas multivariadas de análise: uma análise fatorial em componentes principais, que gerou uma componente adicional (face às cinco iniciais), relacionada com a taxa de desemprego da população, e uma análise de *clusters* (k-médias) que resultou numa solução de seis classes socioeconómicas semelhantes às obtidas para o modelo de 2011. A leitura das classes socioeconómicas por período suportou o diagnóstico dos processos de reconfiguração dos dois territórios metropolitanos entre 2001 e 2011 e evidenciou a expansão da (sub)urbanização qualificada e o recuo da (sub)urbanização não qualificada.

Referências

INE (2004) *Tipologia sócio-económica da Área Metropolitana de Lisboa 2001*. Lisboa: INE.

Vickers, D. & Rees, P (2007) Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society*, 170(2), 379-403.

Sessão Temática II – Instituto Nacional de Estatística – 6ª feira, 1 de abril, Anfiteatro 131 (14:20)

SIMSTAT – Um modelo para a simplificação das estatísticas do Comércio Internacional

Cristina Neves[1]
[1] *Instituto Nacional de Estatística, cristina.neves @ine.pt*

Sumário: Com o objetivo de redução da carga estatística sobre as empresas, decorrente da recolha de informação sobre as transações Intra-UE de bens, está em análise a viabilidade de troca de microdados sobre exportações de bens entre os Estados-Membros da União Europeia, com vista à sua utilização na compilação das importações dos países parceiros, mantendo elevados padrões de qualidade na produção e divulgação das estatísticas do Comércio Internacional de Bens.

Palavras-chave: Exportações, Importações, Intrastat, Microdados, SIMSTAT.

A redução da carga estatística sobre as empresas e a racionalização dos recursos utilizados na produção de informação estatística são dois dos grandes objetivos estratégicos do Programa Estatístico Europeu, cuja implementação tem resultado na definição de projetos específicos com vista à simplificação dos procedimentos de recolha da informação.

O Sistema Intrastat corresponde a um sistema de recolha de dados junto das empresas relativos às transações de bens entre os Estados-Membros da U.E., em vigor desde 1993 (com o início do Mercado Único) e, por ser um dos projetos que mais carga estatística provoca sobre as empresas, foi um dos inicialmente escolhido para ser alvo de uma reestruturação, com vista à sua simplificação.

A redução da carga estatística sobre as empresas decorrente do Intrastat (Comércio Intra-UE de bens) é portanto um objetivo da Comissão Europeia, tendo sido lançado pelo Eurostat em 2012 o projeto SIMSTAT (*Single Market Statistics*). O SIMSTAT tem como principal objetivo analisar a possibilidade de implementação de um sistema de troca de microdados do Comércio Internacional entre os Estados-Membros, em que cada país partilhará os dados das suas exportações com os respetivos países parceiros, para que numa situação ideal deixe de ser necessária a recolha da informação relativamente às importações (evitando assim duplicação na recolha).

Uma das vertentes do SIMSTAT está relacionada com a implementação de um sistema de troca de microdados, entre as autoridades estatísticas dos Estados-Membros responsáveis pela compilação das estatísticas do Comércio Intra-UE, relativos às exportações, com vista à sua utilização pelo país parceiro para a compilação das suas importações.

O princípio subjacente ao SIMSTAT é o de que a informação estatística uma vez recolhida deve ser amplamente utilizada, evitando-se duplicação na recolha. Dado que,

em teoria, as exportações de um país serão iguais às correspondentes importações do país parceiro, o projeto assenta no princípio de que, uma vez recolhida a informação sobre as exportações de um país, esses dados poderão ser utilizados para compilar as correspondentes importações dos seus parceiros.

São objetivos estratégicos do projeto SIMSTAT a simplificação dos procedimentos de recolha de dados, a redução da carga estatística sobre os respondentes e a melhoria da qualidade da informação estatística a divulgar.

O projeto iniciou os seus trabalhos com a análise da viabilidade de implementação de um sistema de troca de microdados do Comércio Internacional entre os Estados-Membros, tendo-se posteriormente desenvolvido uma ferramenta para a troca de microdados. Após elaboração do estudo de viabilidade e acordados os procedimentos a ter em conta na troca de microdados (nomeadamente calendarização, tipo e formato de dados, conjunto mínimo de regras de validação a respeitar), entre abril e setembro de 2015, um total de 20 Estados-Membros (entre os quais Portugal) efetuaram um teste piloto que envolveu a troca mensal das suas exportações para os meses de janeiro a agosto de 2015, incluindo revisões, bem como os dados anuais de 2013 e 2014.

A análise comparativa das exportações declaradas pelos países parceiros como tendo por destino Portugal, com as importações Portuguesas declaradas pelas empresas nacionais no âmbito do Intrastat estão em curso, colocando-se o desafio futuro quanto à efetiva utilização desta informação como fonte complementar à atual recolha de dados sobre importações, mantendo elevados padrões de qualidade na produção e divulgação das estatísticas do Comércio Internacional de Bens.

Referências

Eurostat (2015) SIMSTAT – Exchange of microdata on intra-EU trade. *Statistics Explained*. http://ec.europa.eu/eurostat/statistics-explained/index.php/SIMSTAT_-_exchange_of_microdata_on_intra-EU_trade

Sessão Temática II – Instituto Nacional de Estatística – 6ª feira, 1 de abril, Anfiteatro 131 (14:40)

Transmissão Automática de Dados para o INE

Luísa Pereira[1]
[1] *Instituto Nacional de Estatística, luisa.pereira@ine.pt*

Sumário: O Webinq é um serviço *web* orientado para a recolha de dados por via eletrónica, promovendo a diminuição do esforço exigido às empresas para resposta ao INE, visando um melhor relacionamento com os respondentes, ao criar processos que reduzam e simplifiquem o seu trabalho.
A evolução tecnológica e a procura contínua de oportunidades para a modernização da produção de estatísticas oficiais, permitiu ao INE em 2013, introduzir um novo método de recolha, Transmissão Automática de Dados – TAD.

Palavras-chave: Recolha, TAD, Upload**,** Webservice, XML.

O INE dispõe de um sistema denominado Sistema Global de Gestão de Inquéritos – SIGINQ. Este sistema suporta atualmente o processo de recolha de dados de todas as operações estatísticas do INE às empresas.

O WebInq – Inquéritos do INE na Web, é um dos subsistemas do SIGINQ, sendo um serviço disponível na Internet, desde 2005, orientado para a recolha de dados por via eletrónica.

Este serviço é responsável pela receção de 95.5 % dos dados recolhidos (606353 respostas em 634929 possíveis para 2015), sendo os restantes 4.5 % recolhidos por outros métodos de recolha (papel, fax e e-mail).

Em 2013, o INE introduziu no WebInq um novo método de recolha denominado Transmissão Automática de Dados – TAD, o qual é disponibilizado de duas formas:

- Upload de ficheiros XML;
- Webservice para envio de ficheiros XML.

Com esta nova funcionalidade o INE alcança os seguintes resultados:

- Criar um ambiente mais amigável para as empresas;
- Reduzir os custos para os respondentes e para o INE;
- Motivar os respondentes;
- Melhorar a qualidade da produção estatística;
- Reduzir a carga estatística de reporte de dados sobre as empresas.

O serviço TAD atualmente está disponível para 14 inquéritos, sendo responsável pela receção de 23.15 % dos dados recolhidos (22282 respostas em 96234 possíveis para 2015).

A estrutura do ficheiro XML definida pelo INE foi concebida com o objetivo de incluir vários inquéritos, várias unidades estatísticas e várias ocorrências num único ficheiro. O ficheiro é composto por atributos de controlo e de dados.

O sucesso desde novo método de recolha foi potenciado pelo diálogo profícuo e dinâmico suscitado pelo INE com as empresas que demostraram interesse em participar.

Todo o desenvolvimento da TAD foi sempre efetuado sem perder de vista o sistema integrado de gestão dos processos de recolha que o INE possui.

Referências

Pereira, L. (2013) Data Collection, a shared responsibilities approach. *Grant Facilitation of data transfer from enterprises to NSIs*. Luxembourg: Eurostat.

Acesso à informação estatística oficial para fins de investigação científica

Pinto Martins[1]
[1] *Instituto Nacional de Estatística, pinto.martins@ine.pt*

Sumário: De acordo com a Lei do Sistema Estatístico Nacional (lei nº22/2008) é estabelecida a possibilidade da cedência de dados estatísticos individuais sobre pessoas singulares e colectivas para fins científicos, sob forma anonimizada.
Com a finalidade de cumprir este normativo, são apresentadas as condições e os procedimentos necessários para o acesso por investigadores a dados estatísticos individuais anonimizados constantes de bases residentes no INE, produzidas pelo INE e pelas entidades com delegação de competências, para fins científicos.

Palavras-chave: Acesso gratuito e privilegiado, Acreditação de investigadores, Bases de dados do INE, Dados estatísticos individuais anonimizados, Ficheiros de uso público.

O Instituto Nacional de Estatística (INE), consciente do facto da comunidade académica apresentar necessidades especiais no tocante à informação estatística, nomeadamente para o desenvolvimento de trabalhos de investigação e para a elaboração de teses de Mestrado e Doutoramento, estabeleceu um Protocolo com a Fundação para a Ciência e a Tecnologia (FCT) e a Direcção-Geral de Estatísticas da Educação e Ciência (DGEEC), com o objectivo de facilitar o acesso dos investigadores (acreditados) à informação estatística oficial de que necessitam para o exercício da sua actividade.

Na apresentação serão explicados os procedimentos necessários à acreditação científica (pela DGEEC) bem como o conjunto de bases de dados já preparadas especificamente pelo INE para utilização ao abrigo desta forma de acesso.

O acesso a esta informação é gratuito para os investigadores devidamente acreditados.

Complementarmente, e tendo como objetivo responder às necessidades dos demais utilizadores em aceder a informação mais detalhada, o INE preparou alguns ficheiros com informação ao nível da unidade de observação – os designados Ficheiros de Uso Público (FUPs).

Estes ficheiros (dados e metainformação) contêm registos anonimizados, tratados e preparados para que a unidade de observação não possa ser identificada direta ou indiretamente, exceto quando se trate de dados estatísticos individuais sobre a Administração Pública; são de acesso gratuito e estão conforme o princípio do segredo estatístico e de proteção de dados pessoais. Este acesso implica a aceitação prévia das condições de utilização e estão directamente disponíveis para *download* a partir do portal do INE (www.ine.pt).

Neste momento, encontram-se disponíveis três FUP's: Censos 2001 e 2011 (amostra de 5% relativa a indivíduos e alojamentos); Museus Públicos 2013 e Hospitais Públicos 2012.

Sessão Temática III – Demografia – sábado, 2 de abril, Anfiteatro 131 (10:50)

Da topologia à tipologia: padrões e tipos de divórcio em casais nacionais e binacionais

Madalena Ramos[1], Ana Cristina Ferreira[2], Sofia Gaspar[3]
[1] CIES-IUL/ISCTE-IUL , madalena.ramos@iscte.pt;
[2] DINAMIA/CET-IUL/ISCTE-IUL, cristina.ferreira@iscte.pt;
[3] CIES-IUL/ISCTE-IUL, sofia.gaspar@iscte.pt

Sumário: A Análise de Correspondências Múltiplas (ACM) permitiu identificar a existência de diferentes configurações ao nível dos divórcios ocorridos em Portugal, em 2013. Após a identificação destas configurações, formalizou-se a tipologia através de uma Análise de Clusters. Propomo-nos apresentar o processo de passagem da topologia à tipologia, bem como descrever quantitativamente os diferentes tipos de divórcio encontrados.

Palavras-chave: Análise de Clusters, Análise de Correspondências Múltiplas, Divórcio binacionais, Tipologia de Divórcios.

Os resultados que aqui se apresentam inserem-se num projeto desenvolvido entre 2014 e 2015, financiado pelo Alto Comissariado para as Migrações e intitulado, *Evolução e Perfis dos Divórcios em Casais Binacionais em Portugal (1995-2013).*

Este projeto tinha como objetivo desenvolver uma análise sobre a evolução e os padrões do divórcio em casais binacionais (ou exogâmicos) em Portugal entre 1995 e 2013. O facto das comunidades imigrantes em Portugal terem aumentado consideravelmente, sobretudo a partir dos anos 1990 do séc. XX, contribuiu para um crescimento notável dos casamentos entre indivíduos de distintas origens nacionais e, em particular de estrangeiros com portugueses. E se, por um lado, o casamento com um parceiro da nacionalidade do país de acolhimento pode ser um indicador de integração do imigrante, o divórcio não tem de ser necessariamente um sintoma de falhanço dessa integração. Pode até significar, pelo contrário, uma plena integração que proporcionou condições, quer materiais, quer emocionais, para o rompimento da união; ou seja, pode ser um sintoma de libertação, tanto mais facilitada quanto maior a integração. Este é, por isso mesmo, um tema central a ser analisado e compreendido.

A análise incidiu nos divórcios registados em Portugal, em 2013, entre casais nacionais (constituídos por dois portugueses) e entre casais binacionais (um cônjuge português e outro estrangeiro). A realização de uma Análise de Correspondências Múltiplas (ACM) permitiu perceber quais as características mais diferenciadoras destes casais e identificar quatro configurações distintas no que se refere aos padrões de divórcio.

Após a identificação destas configurações ou padrões de divórcio, formalizou-se a tipologia através da articulação com a Análise de *Clusters,* usando como *input* os scores dos objetos na ACM.

A Análise de *Clusters* permitiu validar os padrões encontrados através da ACM e agrupar os divórcios em quatro grupos/*clusters* com características distintas. Um grupo que se carateriza pelos divórcios com origem em casamentos mais longos, onde os ex-cônjuges são mais velhos e pouco habilitados e que, em termos de situação na profissão, se caracteriza por uma associação à situação de inatividade; um outro grupo constituído por divórcios que põem termo a casamentos mais curtos – até 9 anos - cujos ex-cônjuges são os mais jovens, com idades até aos 37 anos. É um grupo onde, tudo indica, o peso do ensino secundário e do ensino superior é importante; um terceiro grupo cujo perfil remete para divórcios de indivíduos cujo casamento tinha entre 10 e 14 anos, privilegiadamente com idades entre os 37 e os 42 anos e onde, tal como no perfil anterior, se regista uma associação às habilitações de nível secundário e superior; e por fim, um grupo que associa o fim dos casamentos longos (mas não os mais longos), entre os 15 e os 24 anos, a indivíduos com idades entre os 43 e os 50 anos.

Apesar de em todos os *clusters* predominarem os divórcios em casais nacionais, o que é compreensível já que a grande maioria dos divórcios ocorridos em Portugal têm precisamente como cônjuges dois cidadãos portugueses, a projeção em suplementar no plano da ACM do tipo de casal permitiu identificar uma associação entre o *cluster* com o perfil mais jovem e os divórcios em casais binacionais.

Nesta apresentação, iremos apresentar o processo de articulação entre a ACM e a Análise de *Clusters* que possibilitou a passagem da topologia à tipologia, bem como descrever os diferentes tipos de divórcio encontrados.

Referências

Carvalho, H. (2008) *Análise Multivariada de Dados Qualitativos. Utilização da Análise de Correspondências Múltiplas com o SPSS*. Lisboa: Edições Sílabo.

Gaspar, S., Ferreira, A. C. & Ramos, M. (2015) Marriage and Migration in Portugal: exploring trends and patterns of divorce in exogamous and endogamous couples. In Grassi, M. & Ferreira, T. (Eds.) *Places and Belonging. Mobility and Family Relations in Transnational Space*. Newcastle upon Tyne: Cambridge Scholars Publishing, Chapter 4 (no prelo).

Gaspar, S., Ramos, M. & Ferreira, A. C. (2013) Análise comparativa dos divórcios em casais nacionais e binacionais em Portugal (2001-2010). *Sociologia – Revista da Faculdade de Letras da Universidade do Porto*, 26, 81-111.

Ramos, M., Gaspar, S. & Ferreira, A. C. (2015) Padrões de exogamia em quatro comunidades imigrantes em Portugal (2001-2011). *Sociologia, Problemas e Práticas*, 77, 53-76.

Sessão Temática III – Demografia – sábado, 2 de abril, Anfiteatro 131 (11:10)

Family dynamics in the relation between fertility and housing: diversity in the Southern European residential system

Alda Botelho Azevedo[1], Julián López Colás[2], Juan A. Módenes[3]

[1] Instituto do Envelhecimento, Universidade de Lisboa; Centre d'Estudis Demogràfics and Departament de Geografia, Universitat Autònoma de Barcelona, aldazevedo@gmail.com;

[2] Centre d'Estudis Demogràfics, jlopez@ced.uab.es;

[3] Centre d'Estudis Demogràfics and Universitat Autònoma de Barcelona, jamodenes@ced.uab.es

Abstract: This study analyses the possible influence of homeowner-occupation on fertility behavior in southern Europe. Using the Household Finance and Consumption Network data, the likelihood of having a first-child birth was estimated using probit models, followed by a survival analysis testing the timing of the first-child birth. The study concludes with the finding that fertility trends would benefit from changes in the social and statutory treatment of homeownership and renting in southern Europe and from an assertive housing strategy in the fertility-oriented policies.

Keywords: Fertility, Homeownership, Parametric survival models, Probit regression models, Southern European housing system.

Several factors justify the importance given to homeownership when it comes to fertility in the southern European countries (SEC). First, the high prevalence of owner-occupied dwellings is a defining feature of SEC (Allen *et al*. 2004). Second, policies promoting access to housing by young adults might be more effective than explicit fertility-oriented policies (Bernardi 2005). Third, in SEC, homeownership is a long-term decision, strongly related to family and income stability and, consequently, to family formation and fertility (Clark, Deurloo & Dieleman 1994). Fourth, being a "difficult homeownership regime", the SEC are particularly unfriendly to family formation (Mulder & Billari 2010).

Thus, this study analyses if access to housing through homeownership increases the likelihood of first-child birth and if it accelerates the timing of the first-child birth in southern Europe. Using data from the Eurosystem Household Finance and Consumption Network, Wave 1, our sample consisted of the women aged 18-49 years, who changed to the dwelling where they currently live after turning 18 years old and who were childless at the time of the housing change. To test whether homeowners are more likely to have the first child birth than tenants we estimated probit regressions models. To verify if the housing tenure status impacts on the rate of transition to first child birth, we estimated parametric frailty models.

An exploratory analysis of the Kaplan-Meier estimates highlighted that after age 30 the housing tenure status increasingly influences the age when the first child is born,

particularly in Spain. Nevertheless, when taking into account by the educational attainment level of the women, the patterns of the survival functions clearly approach, even overlap in the case of Portugal. This emphasizes that there are substantial differences in the socioeconomic status of homeowners and tenants, and that the differences in fertility timing among tenure types are explained mainly by that divergent social composition.

Broadly, the results confirm our hypotheses. Thus, homeowner-occupation influences fertility behaviour in southern Europe through the increase in the likelihood of first child birth and as an accelerator of the timing of the first child birth. At Euro area and SEC levels, women living in an owner-occupied dwelling are roughly two times more likely to have the first child birth than tenants. Italy shows a similar pattern, but by far the biggest increase is observed in Spain, where being a homeowner increases the likelihood of having the first child, by almost three times over renters. A more moderated relationship is found in Portuguese, Greek, French and German models. When it comes to the effect of homeownership in the timing of having the first child, the time ratios confirm that being homeowner accelerates the event in five models: Euro area, SEC, Spain, Greece and France. Nevertheless, the differences between the two housing tenure statuses are very small, with the exception of the Spanish case where the time ratio indicates that being home owner is an acceleration factor of first-child birth.

Although the appreciable diversity among the SEC, they would all benefit from a quantitative change in the SEC housing system favouring the rental stock. At the same time, a qualitative change in statutory meanings of security associated with renting is needed in order to encourage fertility in the SEC. Renting is currently the housing tenure status with the highest growth rate among young adults; it is important to reduce the difference between housing tenure status and fertility, by encouraging parenthood among young couples, forming those future renter-occupied households. Finally, in the traditional sphere of fertility policies, the inclusion of a housing strategy could bring a change operating at the upstream of the fertility question.

Acknowledgments: The authors gratefully acknowledge the financial support from the Spanish Ministry of Economy and Competitiveness under the R+D+i project "Geographical mobility and housing: Spain in an international perspective" (CSO2013-45358-R), headed by Juan A. Módenes and Joaquín Recaño.

References

Allen, J., Barlow, J., Leal, J., Maloutas, T. & Padovani, L. (2004) *Housing and Welfare in Southern Europe*, Oxford: Wiley-Blackwell.

Bernardi, F. (2005) Public policies and low fertility: rationales for public intervention and a diagnosis for the Spanish case. *Journal of European Social Policy*, 15(2), 123–138.

Clark, W. A. V. & Dieleman, F. M. (1996) *Households and Housing: Choice and Outcomes in the Housing Market*, New Jersey: Center for Urban Policy Research.

Mulder, C. H. & Billari, F. C. (2010) Homeownership Regimes and Low Fertility. *Housing Studies*, 25(4), 527–541.

Sessão Temática III – Demografia – sábado, 2 de abril, Anfiteatro 131 (11:30)

Imigração e mercado de trabalho: leituras de mobilidade a várias escalas nas regiões de Lisboa, Odemira e Algarve

Alina Esteves[1]
[1] *Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa, alinaesteves@campus.ul.pt*

Sumário: Explorando os dados do projeto CRISIMI, a apresentação procura compreender algumas das trajetórias de mobilidade interna de um conjunto de nacionais de países terceiros a viver em Portugal, salientando as diferentes capacidades de atração e de retenção de 3 regiões (AML, concelho de Odemira e Algarve). O presente contexto de crise económica motiva igualmente a mobilidade internacional dos imigrantes residentes em Portugal, optando alguns por estratégias de re-emigração para países onde já têm familiares próximos.

Palavras-chave: Crise, Estratégias, Migrações, Mobilidade geográfica.

O mercado de trabalho e a família são duas áreas temáticas centrais nos estudos sobre migrações. Se a primeira permite a incorporação do imigrante enquanto elemento produtivo da sociedade de acolhimento, a presença da família é essencial para assegurar uma vida emocionalmente mais estável e equilibrada (Fonseca *et al.* 2005).

Recorrendo a informação qualitativa e quantitativa do projeto "CRISIMI - O impacto da crise económica sobre as condições de vida e dinâmicas de inserção laboral dos imigrantes em Portugal" (financiado pela Ação 3 do FEINPT, via Alto Comissariado das Migrações), esta comunicação procura explorar as estratégias de sobrevivência dos nacionais de países terceiros (NPT) num contexto de recessão económica, identificando opções de mobilidade interna e re-emigração, imaginada ou efetiva (Esteves *et al.* 2015). Os territórios de estudo são a Área Metropolitana de Lisboa (AML), o concelho de Odemira e o Algarve, onde foram aplicados 537, 65 e 80 inquéritos, respetivamente. Foram feitas 15 entrevistas semi-estruturadas a NPT na AML e 23 entrevistas em profundidade a dirigentes associativos, centrais sindicais, entidades empregadoras, autarquias locais, ONGs, IPSS nas 3 áreas de estudo. O trabalho de campo decorreu entre Março e Julho de 2015. Usou-se o método de bola de neve para aplicar os questionários em locais muito diferentes de modo a diversificar a amostra (locais de culto, embaixadas, cafés, associações culturais).

Procurando respeitar a importância relativa das várias nacionalidades nestes três territórios, 35,9% dos NPT inquiridos eram brasileiros, 18,7% ucranianos, 11,9% cabo-verdianos e 6,2% moldavos. As amostras recolhidas apresentam um equilíbrio entre os dois sexos na AML e no Algarve, mas em Odemira predomina o sexo masculino (72,3%), refletindo a realidade local. Neste município alentejano, as principais nacionalidades inquiridas foram indianos, bangladeshis e nepaleses.

Quanto ao contexto familiar dos imigrantes inquiridos, há diferenças entre os nacionais dos PALOP, com uma presença mais antiga em Portugal e os cidadãos asiáticos, chegados mais recentemente. Dois terços dos inquiridos vivem do salário do trabalho dependente, valor que chega aos 90% em Odemira. As profissões mais frequentemente exercidas enquadram-se nos grupos 5 (Serv. pessoais, proteção e segurança, e vendedores), 7 (Trab. qualif. ind., construção e artífices) e 9 (Trab. não qualif.) da CNP, tal como a nível nacional (Peixoto e Iorio, 2011).

Quanto à mobilidade geográfica interna, a AML é um importante ponto de entrada e fixação em Portugal, pois 89,9% dos imigrantes aqui inquiridos tiveram este território como 1º local de residência. Apesar do valor para o Algarve ser mais baixo, é de salientar que perto de 60% dos NPT aqui inquiridos tiveram como 1º local de residência outra região portuguesa e que migraram depois para esta região. A dinâmica do mercado de trabalho regional, virado para atividades menos dependentes do mercado interno, tem conseguido atrair trabalhadores estrangeiros que se adaptaram à elevada sazonalidade de algumas atividades. O concelho de Odemira é outro exemplo de um território com elevada capacidade de atração de trabalhadores estrangeiros, na medida em que 71,9% dos imigrantes aqui inquiridos tiveram como 1º local de residência outra região. Apesar da atual crise, o subsector da horto-fruticultura moderna para exportação, tem necessidades crescentes de mão-de-obra. Há diferenças de mobilidade residencial segundo a nacionalidade e o momento de chegada ao país.

A crise tem motivado a mobilidade internacional não apenas dos cidadãos portugueses, mas também dos estrangeiros que registam taxas de desemprego muito mais elevadas comparativamente aos nacionais (22,3% e 13,7%, respetivamente, em 2014). Desde que chegaram a Portugal, 5,6% dos inquiridos já trabalharam pelo menos uma vez noutro país e 40,9% já pensou em partir por motivos de trabalho. Entre estes últimos, 17,4% já fez alguma coisa no sentido de preparar essa re-emigração. É uma estratégia maioritariamente masculina (74%), em que a presença de familiares no estrangeiro e o estatuto legal em Portugal podem influenciar a tomada de decisão e o destino escolhido.

Agradecimentos: A toda a equipa do projeto CRISIMI que decorreu no Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa.

Referências

Esteves, A. (Coord.), Estevens, A., Amilcar, A., McGarrigle, J., Malheiros, J., Moreno, L., Fonseca, M. L. & Pereira, S. (2015) *Relatório final do projecto CRISIMI - O impacto da crise económica sobre as condições de vida e dinâmicas de inserção laboral dos imigrantes em Portugal*. Lisboa: IGOT, Universidade de Lisboa.

Fonseca, M. L., Ormond, M., Malheiros, J., Patrício, M. & Martins, F. (2005) *Reunificação Familiar e Imigração em Portugal*. Lisboa: Observatório da Imigração, ACIME.

Peixoto, J. & Iorio, J. (2011) *Crise, Imigração e Mercado de Trabalho em Portugal. Retorno, regulação ou resistência?* Cascais: Principia e FCG.

Sessão Temática III – Demografia – sábado, 2 de abril, Anfiteatro 131 (11:50)

Envelhecimento demográfico: o desafio social do município de Évora

Filipe Ribeiro[1], Lídia P. Tomé[2], Maria Filomena Mendes[3]
[1] CIDEHUS - Universidade de Évora, fribeiro@uevora.pt;
[2] CIDEHUS - Universidade de Évora, lidiatome@uevora.pt;
[3] CIDEHUS - Universidade de Évora, mmendes@uevora.pt

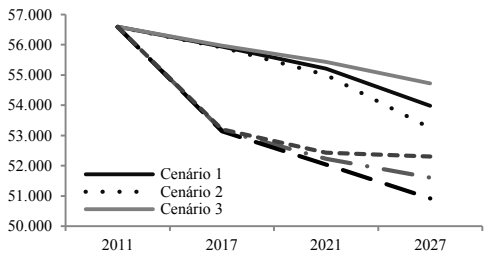
Sumário: Entre 1981 e 2011 o município de Évora perdeu aproximadamente 5 000 residentes, dos quais, cerca de 4 000 foram jovens até aos 18 anos. Évora apresenta-se em 2011 com uma estrutura demográfica irremediavelmente afetada pelo efeito de um duplo envelhecimento nas recentes décadas. Este trabalho pretende estimar, através de uma projeção por coorte e componentes combinada com uma componente probabilística, a evolução do crescimento populacional no horizonte de 2031. O trabalho pretende ainda sustentar cientificamente a intervenção dos decisores políticos da região.

Palavras-chave: Envelhecimento, Fecundidade, Migrações, Mortalidade, Projeções demográficas.

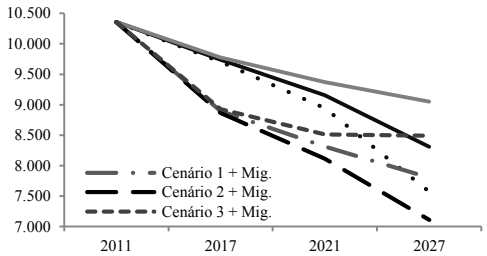
As sociedades contemporâneas um pouco por todo o mundo registam avanços significativos na sua esperança de vida, contudo, os baixos níveis de fecundidade registados nessas mesmas sociedades colocam em causa o seu equilíbrio demográfico. Esta realidade não é exclusiva de países avançados económica e socialmente, atingindo também regiões de menor dimensão e geograficamente isoladas. Em Portugal, essa realidade encontra-se especialmente marcada no interior e no sul do país, onde os aglomerados populacionais se encontram afastados entre si e das grandes metrópoles, e onde a capacidade económica para atrair os jovens adultos em idade fértil é muito reduzida.

No município de Évora, mas também em outras regiões do sul, a estrutura demográfica há muito que se encontra irremediavelmente afetada pelo envelhecimento da pirâmide populacional onde se denota um claro envelhecimento quer na base (estreita), quer no topo (alargado), ou seja, um duplo envelhecimento. Apesar de se poder admitir que um aumento generalizado dos níveis de fecundidade possa contribuir positivamente para alterar este contexto demográfico negativo, a verdade é que não será suficiente *per se*. À semelhança de Portugal (Mendes & Tomé 2014), a tendência (independentemente dos cenários de projeção elaborados) é de declínio da população residente. No entanto, são sobretudo os mais jovens, com idades até aos 18 anos, os mais afetados negativamente (Figura 1).

Figura 1: Possível evolução da população residente no município de Évora



a) População total



b) População até aos 18 anos

Fonte: Cálculo dos autores.

A realidade sociodemográfica atual e futura torna indispensável a intervenção urgente dos decisores políticos da região, procurando medidas que visem dar resposta às transformações demográficas caracterizadas pela inversão de uma tendência passada em que os jovens representavam uma elevada proporção quer quando comparados aos potencialmente ativos quer, principalmente, aos idosos.

O presente estudo, apresenta os resultados de projeções de população, idade a idade, diferenciadas por sexo, para um horizonte temporal de médio prazo (2031), através da utilização do método por coortes e componentes (e.g. Rowland 2003). Adicionalmente, foi introduzida uma componente probabilística, através da utilização do método de Lee-Carter (1992) que permite prever padrões futuros de mortalidade por idade. A introdução deste método na estratégia metodológica permite acrescentar aos resultados obtidos, em função dos diferentes cenários construídos, um intervalo de confiança e, consequentemente, delimitar aqueles resultados com um determinado grau de confiança, os resultados obtidos. Considerando as especificidades próprias de uma sede de concelho, pretendeu-se ainda identificar alterações em termos de dinâmica demográfica ao nível das freguesias, recorrendo ao "ratio method" (Rowland 2003), projetando igualmente a população ao nível da freguesia.

Em suma, com os resultados obtidos, pretendemos não só quantificar o declínio populacional inevitável no município de Évora, mas também identificar possíveis focos de intervenção urgentes ao nível sociodemográfico.

Referências

Lee, R.D. & Carter, L.R. (1992) Modeling and forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419), 659-671.

Mendes, M. F. & Tomé, L. P. (2014) Projecções: Resultados e Interpretação. In *Dinâmicas demográficas e envelhecimento da população portuguesa, 1950-2011 Evolução e Perspectivas*. Direcção Mário Leston Bandeira. Estudos da Fundação Francisco Manuel dos Santos, 451-490.

Rowland, D. (2003) *Demographic Methods and concepts*. Oxford: Oxford University Press.

Time series analysis of remote sensing data

Clara Cordeiro[1,2], Sónia Cristina[3,4],Priscila C. Goela[3,4], Sergei Danchenko[3,4], John Icely[3,5], Samantha Lavender[6], Alice Newton[3,7].

[1] *FCT, University of Algarve (UALg), PT, ccordei@ualg.pt;*

[2] *CEAUL, University of Lisbon, PT;*

[3] *CIMA-FCT, UALg, PT, cristina.scv@gmail.com, priscila.goela@gmail.com, danchenko-sergei@tut.by;*

[4] *FCMA, University of Cadiz, ES;*

[5] *Sagremarisco* Lda*., PT, john.icely@gmail.com;*

[6] Pixalytics Ltd, UK, *slavender@pixalytics.com*;

[7] *NILU-CEE, NO, anewton@ualg.pt*

Abstract: Satellite ocean colour remote sensing provides a valuable source of information on the status of marine ecosystems. The study, analysis and interpretation of such a data is based on statistical techniques for dependent data. Seeing beyond the data analysis and inferring valid and conclusive conclusions will help in the management and planning of marine economics, which will have a great impact in the Algarve's region.

Keywords: Climate change, Seasonal-Trend decomposition, stl.fit(), Structural breaks, Time series.

The analysis of time series is an important and valuable approach adopted in several studies, for its ability to improve the spatial and temporal resolution of the major seasonal and inter-annual patterns in biological and oceanographic data. These studies provide indicators about long-term changes in natural conditions, such as climate change.

One of the most challenging and difficult tasks in time series analysis is the selection of the statistical model that best describes the temporal behaviour of the time series. The models may be "useful" and "adequate" to describe a time series, but there is nothing definitive about it, because there are several models that could also be suitable. Decomposition is primarily useful for studying time series data, as it shows historical changes over time; indeed, classical decomposition methods have been widely used in marine sciences literature. However, these classical methodologies do not allow for a flexible specification of the seasonal component and the trend component is represented by a deterministic function of time that is easily affected by atypical observations. Therefore, the Seasonal-Trend decomposition of time series based on Loess (STL, local polynomial regression fitting) selected for the current study identifies a seasonal component that changes over time, is responsive to nonlinear trends, and robust in the presence of outliers. The authors have developed a procedure named stl.fit() that selects the best STL model for each combination of the seasonal and trend smoothing

parameters, based on an error measure. Satellite-derived data from the Algarve's region was used to evaluate the performance of the new procedure.

Most of the time series analyses consider some form of stationarity or continuity. However, some changes in the dynamic structure of the time series can be observed. These changes are known as structural breaks and are points in time at which statistical patterns change, generally due to non-climatic factors, such as changes in satellite instrumentation over time. Remote sensing data commonly used in several studies is the Sea Surface Temperature (SST). In the southwest coast of Portugal, the study of the SST is of major relevance. The consequent procedure considered a seasonally adjusted SST time series, where a structural break analysis was performed to detect potential climatic changes, such as an increase in warming trends off the Sagres region. Another point of view is the relationship between the upwelling and the SST in the study area. Important coastal processes such as upwelling, could be identified and associated with a decrease in SST within a short period after favourable wind stress conditions.

These are very interesting and challenging issues because one of the pillars of the Algarve's economy came from the exploration of the sea, and the fulfilment of this study will enrich the knowledge of marine resources.

Acknowledgements: Research of C. Cordeiro was partially funded by FCT - Portugal, through the project Pest-OE/MAT/UI0006/2014; P.C. Goela and S. Cristina were funded by PhD grants from FCT - Portugal (SFRH/BD/78356/2011 and SFRH/BD/78354/2011, respectively). J. Icely was funded by EU FP7 AQUA_USER (grant agreement no. 607325), www.aqua-users.eu, and Horizon 2020 AquaSpace (grant no. 633476). A. Newton was funded by EU FP7 project DEVOTES (grant agreement no. 308392), www.devotes-project.eu.

Sessão Temática IV – Aplicações da Estatística em Ciências Biológicas – sábado, 2 de abril, Anfiteatro 131 (14:35)

Modelos mistos aplicados à análise da variabilidade genética intravarietal e selecção de castas antigas de videira

Elsa Gonçalves[1]
[1] *Secção de Matemática/DCEB e LEAF, Instituto Superior de Agronomia, Universidade de Lisboa; Associação Portuguesa para a Diversidade da Videira-PORVID, elsagoncalves@isa.ulisboa.pt*

Sumário: Frequentemente, no melhoramento de plantas a análise de dados assenta na teoria dos modelos mistos, uma vez que os objetivos estão centrados na estimação de componentes da variância, na predição dos efeitos genéticos relativos a uma dada característica e do ganho genético de seleção. A metodologia de seleção da videira praticada em Portugal não foge à regra e os modelos mistos são a base da análise de dados conducente à quantificação da variabilidade genética intravarietal e à própria seleção. Neste trabalho discutem-se alguns resultados do ajustamento de modelos mistos a dados provenientes de ensaios de seleção da videira.

Palavras-chave: Modelos mistos, Seleção da videira, Variabilidade genética intravarietal.

A videira é uma cultura com grande importância económica e social em Portugal e os trabalhos conducentes à sua valorização económica, nomeadamente, os respeitantes à seleção são uma prática corrente, de modo a tornar disponível para os viticultores material de propagação que satisfaça os objetivos atuais da viticultura e do mercado. Para tal, é necessária uma metodologia de seleção da videira fortemente alicerçada na genética quantitativa e, consequentemente, na teoria dos modelos mistos.

Os ensaios de campo destinados à concretização de tais objetivos contêm, em geral, uma amostra aleatória de genótipos da população cultivada que se pretende estudar. A dimensão dessa amostra é normalmente da ordem das centenas, portanto, trata-se de ensaios agronómicos com elevado número de tratamentos (genótipos), considerados como de efeitos aleatórios. A grande dimensão destes ensaios exige que sejam estabelecidos segundo delineamentos experimentais orientados para o controlo da variação espacial (Gonçalves *et al*. 2010), idealmente, delineamentos pertencentes à classe dos blocos incompletos e linha - coluna, como alternativa aos clássicos delineamentos em blocos completos casualizados. Nestes ensaios, o primeiro objetivo passa por quantificar a variabilidade genética intravarietal, não só porque é um indicador da idade da casta, mas essencialmente porque é uma medida da matéria-prima disponível para a seleção (quanto maior a variabilidade genética intravarietal, maiores serão os ganhos de seleção). Por outro lado, quando se faz seleção é essencial que sejam avaliadas várias características, pois é importante compreender se ao selecionar uma característica com interesse económico não se está a prejudicar outra com igual importância. Também é importante avaliar se o comportamento comparado dos genótipos num ambiente é

semelhante ao observado noutro ambiente, etc.. Ou seja, todas as questões atrás referidas, essenciais num processo de melhoramento, encontram a sua resposta com o ajustamento de modelos mistos aos dados provenientes desses ensaios, sendo a escolha adequada da estrutura das matrizes de covariâncias um ponto crucial (Smith *et al*. 2005; Gonçalves *et al*. 2007; Gonçalves & Martins 2014). Desde logo, o interesse pelo conhecimento da estimativa da componente de variância genotípica é imediato: por um lado, surge como quantificador da variabilidade genética da casta para a característica em estudo, por outro, esta componente, bem como todos os outros parâmetros de covariância estimados, influenciam os resultados da seleção genética, já que os melhores preditores empíricos lineares não enviesados dos efeitos genotípicos, com base nos quais se faz a seleção, dependem diretamente da estrutura das matrizes de covariâncias.

Neste trabalho exemplifica-se a aplicação de modelos lineares mistos em várias fases do percurso de seleção, nomeadamente, (1) na quantificação da variabilidade genética intravarietal das características economicamente importantes, como o rendimento e características de qualidade do mosto, (2) no estudo da correlação genética entre características, (3) no estudo da interação genótipoXambiente.

Referências

Gonçalves, E., St. Aubyn, A. & Martins, A. (2007) Mixed spatial models for data analysis of yield on large grapevine selection field trials. *Theoretical and Applied Genetics*, 115, 653-663.

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.

Smith, A. B., Cullis, B. R. & Thompson, R. (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches*. Journal of Agricultural Science,* 143, 449-462.

Gonçalves, E. & Martins, A. (2014) Metodologias estatísticas para estudo da interacção genótipo×ambiente em clones de videira. In Estatística: A ciência da incerteza. *Actas XXI Congresso da Sociedade Portuguesa de Estatística*, 89-103.

Uma abordagem bioestatística na caraterização do Envelhecimento Vascular Precoce

P. G. Cunha[1], J. Cotter[2], P. Oliveira[3], I. Vila[4], N. Sousa[5]
[1] Centro Hospitalar do Alto Ave, IICVS, Universidade do Minho, pedrocunha@ecsaude.uminho.pt;
[2] Centro Hospitalar do Alto Ave, IICVS, Universidade do Minho, jorgecotter@gmail.com;
[3] Instituto de Saúde Pública, Universidade do Porto, pnoliveira@icbas.up.pt;
[4] Centro Hospitalar do Alto Ave;
[5] IICVS, Universidade do Minho, njcsousa@ecsaude.uminho.pt

Sumário: A população Portuguesa no Norte de Portugal regista prevalências elevadas de hipertensão arterial e de acidentes vasculares cerebrais. Não haviam sido medidos biomarcadores de lesão da parede arterial nesta população. Este trabalho apresenta os resultados de um estudo para identificar e caracterizar lesão arterial e Envelhecimento Vascular Precoce. Para esse efeito, a Velocidade de Onda de Pulso foi medida numa amostra aleatória dos residentes nas cidades de Guimarães e Vizela.

Palavras-chave: Regressão linear**,** Regressão logística, Velocidade de onda de pulso.

A população portuguesa no Norte de Portugal apresenta uma incidência elevada de acidentes vasculares cerebrais que se entende estar associada ao elevado consumo de sal, hipertensão e exposição a outros factores de risco cardiovascular. A rigidez arterial, avaliada pela Velocidade de Onda de Pulso (VOP), pode ser particularmente útil na compreensão dos efeitos acumulados dos diferentes riscos cardiovasculares no processo de Envelhecimento Vascular. Com o objectivo de identificar e caracterizar a distribuição dos valores da VOP e o Envelhecimento Vascular Precoce (EVP) foi desenhado um estudo de base populacional nas cidades de Guimarães e Vizela (Cunha *et al.* 2014). Com base nos registos dos utentes dos Centros de Saúde das duas cidades, uma amostra aleatória de 2542 indivíduos foi selecionada. Os sujeitos foram classificados como sofrendo de EVP se os respetivos valores de VOP estavam acima do percentil 97,5 ajustado para idade de acordo com os valores de referência Europeus. Considerou-se que apresentavam lesão arterial de grandes vasos quando a VOP era > 10m/s

Foram construídos modelos de regressão linear com a VOP como variável dependente, incluindo como variáveis independentes: idade, sexo, pressão arterial sistólica, frequência cardíaca, índice de massa corporal, anos de escolaridade, consumo de tabaco, tratamento antihipertensor, perfil lipídico, glucose em jejum, taxa de filtração glomerular, tratamento antidiabético, tratamento antilipídico e doença cardiovascular conhecida. De igual modo foram construídos modelos de regressão logística para estudar os efeitos das diversas variáveis no desenvolvimento de EVP e VOP >10 m/s.

A Velocidade de Onda de Pulso VOP média foi de 8.4m/s (homens: 8,6 m/s, mulheres 8,2 m/s). A prevalência global de EVP foi de 12,5% (26,1% dos sujeitos com menos de 30 anos; na população, 18,7% apresentava valores acima de 10 m/s). Modelos de regressão logística mostraram diferenças atribuíveis ao sexo nas razões de possibilidades para o desenvolvimento de lesão arterial, com as mulheres a atingirem a mesma razão de possibilidades para VOP acima dos 10 m/s dez anos mais tarde quando comparadas com os homens.

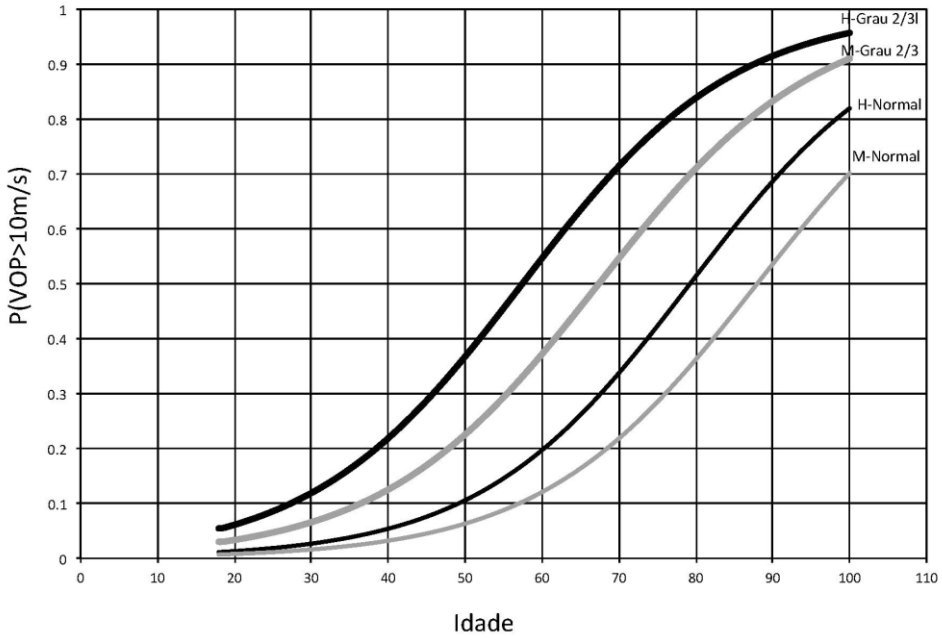


Figura 1: Probabilidade de VOP>10 em função da idade e hipertensão (H-Homens, M-Mulheres)

Neste trabalho é apresentado pela primeira vez em Portugal, num estudo de base populacional, a distribuição da VOP por idade e sexo. Verifica-se que os valores da VOP são marcadamente mais elevados do que os esperados de acordo com os valores de referência Europeus, em particular, para os indivíduos jovens do sexo masculino. Os modelos de regressão confirmaram, tal como em outros estudos, a associação da idade e da Pressão arterial Sistólica com a VOP. Os resultados têm relevância para a prevenção da doença cardiovascular e levantam várias questões sobre o perfil de risco cardiovascular da população mas também sobre a necessidade de medidas urgentes, quer clínicas quer de saúde pública.

Referências

Cunha, P. G., Cotter, J., Oliveira, P., Vila, I. & Sousa, N. (2014) The Rationale/Design of the Guimarães/Vizela Study: a multimodal population-based cohort study to determine global cardiovascular risk and disease. *J. Investg. Med.*, 62, 812-8290.

Spatio-temporal structure of ecological data: a three-way study

Susana Mendes[1], M.ª José Fernández-Gómez[2], Sónia Cotrim Marques[3], Ulisses Miranda Azeiteiro[4], Paulo Maranhão[5], Sérgio Miguel Leandro[6], M.ª Purificación Galindo-Villardón[7]

[1, 5, 6] MARE – Marine and Environmental Sciences Centre, ESTM, Instituto Politécnico de Leiria, 2520-641 Peniche, Portugal, susana.mendes@ipleiria.pt, pmaranhao@ipleiria.pt, sleandro@ipleiria.pt;

[2] University of Salamanca, Department of Statistics, mjfg@usal.es;

[3] University of Coimbra, CEF-Department of Life Sciences, scotrim@ci.uc.pt;

[4] Department of Sciences and Technology, University Aberta and University of Coimbra, CEF-Department of Life Sciences, ulisses@uab.pt;

[7] University of Salamanca, Department of Statistics, pgalindo@usal.es

Abstract: This work aims to characterize the different zooplankton species, with the target of analyzing the variations in species abundance and the different compositions at shelf and oceanic sites through its different dimensions (time vs space). All this is carried out using a Tucker3 model, through the use of core matrix for all modes. The preferred Tucker model showed the similarities between the successive data tables and proved to be useful for detecting spatial-temporal patterns in zooplankton distribution.

Keywords: Multiway data, Tucker3 model, Zooplankton.

Introduction

Multivariate analysis, and in particular, multivariate projection methods, have undergone an important development during the last few decades, and are widely applied in fields such as chemometrics or industry, traditionally associated to arrays that have a wide number of potentially related variables compared to the number of observations. These kind of problems are usual when studying environmental problems and N-way methods, such as Tucker3 (Tucker 1966; Kroonenberg 1983), had gained a lot of attention in the recent years in that field (Stanimirova 2006). The reasons for this are associated with rapid and incessant development of analytical techniques and the application of multicomponent analytical methods, which permit achieving large-scale sampling during monitoring. The aim of the present study was to analyze the variations in zooplankton species abundance and the different compositions at shelf and oceanic sites through its different dimensions (time vs space) by applying Tucker3 model. Furthermore, based on the results obtained from the N-way method, it is possible to determine the proportion of variability in the variables species that depends on space or on time. The data collected can therefore be arranged into three-way array as sites x species variables x time.

From February 2006 to February 2007, zooplankton samples were collected monthly from 6 stations (E1, E2, E3, E4, E5 and E6), located along a transect perpendicular to the coastline (between the Peniche coast and Berlenga islands). Zooplankton samples were collected through horizontal hauls and performed during day time from 1 m below the surface using a 500 μm mesh net.

Data Analysis

In order to investigate the temporal and spatial variability in the zooplankton community structure, the species densities were arranged in a tridimensional matrix (sites, species and dates) comprising a 6 x 50 x 12 data block of sites by species densities by dates. These data offered the possibility to study the spatial variability of the zooplankton community and its dynamics in time. In Tucker3 analysis, a (usually preprocessed) three-way data array $\mathbb{X}$, of order I x J x K, is described by the model $x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}$, i = 1,...,I, j = 1,...,J, k = 1,...,K; where a_{ip} and b_{jq} and c_{kr} denote the elements of the components matrices A (I x P), describing the sampling sites (objects), B (J x Q), describing the species variables and C (K x R), describing the sampling times (conditions), respectively, and g_{pqr} denotes the elements of core array $\mathbb{G}$ (P x Q x R), which weights the products between component p of the sampling sites (first mode), component q of the species variables (second mode), and component r of the sampling times (third mode) and explains the interaction among the p, q, r factors of each of the modes. The term e_{ijk} denotes an error term (or residual) associated with the description of x_{ijk} (in this case, the value of the measurement referring to the i*th* sampling site, j*th* species variable and r*th* sampling month). The values p, q and r are the number of components selected to describe the first, the second and the third mode, respectively, of the data array. The core efficiently describes the main relations in the data, and the component matrices A, B and C describe how the particular sites, species variables and sampling times relate to their associated components.

Results

The method optimizing the ''variance of squares'' of the core elements has allowed a meaningful and simple interpretation of the Tucker3 solution for the selected model. In particular, the procedure succeeded in decomposing nicely the overall temporal variations in four parts (considering the distribution of the dates and the arrangement of the species), thus highlighting August to November, June and July, April and May and February, March and December 2006 and January and February 2007. Besides the environmental considerations strictly connected to the present case study, in general terms, the suitability of the proposed procedure can be remarked to handle data coming from environmental studies that are typically collected on the basis of a multiway sampling plan. The ability to detect periodic patterns of multivariate time series within a pool of sampling points that can be hindered by both anomalous events and site specific features constitute a relevant advantage of the procedure.

References

Kroonenberg, P.M. (1983). *Three-mode Principal Component Analysis*. Leiden: DSWO Press.

Stanimirova, I., Zehl, K., Massart, D.L., Vander Heyden, Y., Einax, J.W. (2006). *Chenometric analysis of soil pollution data using the Tucker N-way method*. Analytical and Bioanalytical Chemistry, 385, 771-779.

Tucker, L.R. (1966). Some mathematical notes on 3-mode factor analysis. *Psychometrika* 31, 279–311.

SESSÕES PARALELAS

Wavelet-based detection of outliers in time series of counts

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

Abstract: In this work we consider the problem of analysing count time series contaminated with outliers. To address this problem we propose a wavelet-based approach that allows the identification of the time point of occurrence of an outlier in a time series of counts, by using the empirical distribution of the detail coefficient via resampling methods (parametric bootstrap). Results of a simulation study illustrating the effectiveness of the proposed method and a real dataset application are presented.

Keywords: Additive outlier, Discrete wavelet transform, Haar wavelet, INAR(1) models, Parametric resampling.

A problem of interest in time series modelling is to detect outliers which can be seen as discrepant observations. These observations originate biased model estimates and consequently invalid inferences. For conventional ARMA processes estimation and inference in the presence of outliers are well documented. However, the impact of outliers on the parameter estimation for time series of counts has not yet been completely addressed.

Time series of (small) counts are common in practice and appear in a wide variety of fields: social science, biology and environmental processes, economics and finance, telecommunications and insurance, among others. Different types of models that explicitly account for the discreteness of the data have been proposed in the literature. In this work we consider the INAR(1) (INteger-valued AutoRegressive) models, which are constructed by replacing the multiplication in the conventional AR models by an appropriate random operator called thinning operator (for details, see Scotto *et al.* 2015).

In the context of time series analysis we can consider additive outliers (AO), which are external errors or exogenous changes at a certain time point and affect only the observation at the time the disturbance occurs, and innovational outliers (IO), associated with internal changes or endogenous effects of the noise process that affect all the subsequent observations. Silva & Pereira (2015) suggested a Bayesian approach in order to detect additive outliers in Poisson INAR(1) models.

In this work we propose a method based on wavelets, which are a family of basis functions used to localize a given function in both space and scaling (Percival 2006), to address the problem of identifying the time point of the outlier in a time series of counts. The empirical distribution of the detailed coefficient derived from the discrete wavelet transform (DWT), using the Haar wavelet, is obtained by the parametric resampling

method of Tsay (1992). The outliers are identified as the observations outside of the acceptance envelope constructed with quantiles of this empirical distribution.

The proposed procedure is illustrated with simulated data and the results are compared with existing techniques. Furthermore, the method is also applied on an observed dataset.

Acknowledgments: This work is partially supported by Portuguese funds through the CIDMA and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013.

References

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

Percival, D., Walden, A. (2006) *Wavelet methods for time series analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. New York, USA: Cambridge University Press.

Silva, M. E. & Pereira, I. (2015) Detection of additive outliers in Poisson INAR(1) time series. In Bourguignon, J. P., Jeltsch, R., Pinto, A. A. & Viana, M. (Eds.) *CIM Series in Mathematical Sciences - Mathematics of Energy and Climate Change*. Springer, 377-388.

Tsay, R. S. (1992) Model checking via parametric bootstraps in time series analysis. *Journal of the Royal Statistical Society: Series C* (*Applied Statistics*), 41, 1-15.

Sessão Paralela I – Modelos Estocásticos e Modelos Espaciais – 6ª feira, 1 de abril, Sala 124 (9:50)

Políticas de pesca sustentáveis ótimas com esforço constante *versus* políticas de pesca ótimas com esforço variável: aplicação em ambiente aleatório com um modelo logístico

Nuno M. Brites[1], Carlos A. Braumann[2]

[1] Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, brites@uevora.pt;

[2] Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora & Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, braumann@uevora.pt

Sumário: A dinâmica de uma população sujeita a pesca pode ser modelada por equações diferenciais estocásticas. Aqui usamos um modelo logístico para o crescimento natural médio. As políticas de esforço variável de pesca que otimizam o lucro acumulado descontado são inaplicáveis. Aqui usamos uma política alternativa sustentável em que existe distribuição estacionária, otimizando o lucro estacionário por unidade de tempo para políticas de esforço constante e quantificando a resultante redução do lucro.

Palavras-chave: Equações diferenciais estocásticas, Esforço de pesca, Modelo logístico, Políticas ótimas, Políticas sustentáveis.

Para descrever o crescimento de uma população sujeita a pesca e a flutuações aleatórias do ambiente podemos usar modelos (ver, por exemplo, Braumann (1985) e Braumann (2002)) baseados em equações diferenciais estocásticas (EDE). Neste trabalho consideramos que o crescimento da população segue um modelo logístico ao qual subtraímos a taxa de capturas da forma $h(t) = qE(t)X(t)$, onde $q>0$ representa o coeficiente de capturabilidade, $E(t)\geq 0$ o esforço de pesca e $X(t)$ o tamanho da população no instante de tempo t. Obtemos a EDE (com $r>0$, $K>0$, $\sigma>0$, $X(0)>0$ e $W(t)$ um processo de Wiener)

$$dX(t) = rX(t)(1-X(t)/K)\,dt - qE(t)X(t)\,dt + \sigma X(t)\,dW(t)$$

Em Braumann (1985) otimizámos a taxa de captura para os modelos logístico e de Gompertz; Zou *et al.* (2013) faz um tratamento semelhante para o modelo de Gompertz. Aqui vamos, em vez disso, otimizar o lucro por unidade de tempo, que supomos ser da forma $L(t) = P(t) - C(t)$, onde $P(t) = p\,h(t)$ é o preço de venda (sendo $p>0$ o preço unitário) e $C(t) = c(E)E(t)$ são os custos de pesca. O custo por unidade de esforço $c(E)$ pode variar com o esforço aplicado; aqui vamos considerar o caso particular de $c(E) = c_0 + c_1 E(t)$.

Existem trabalhos anteriores relacionados com a escolha da política ótima de pesca com o objetivo de maximizar o lucro acumulado (descontado por uma taxa de desvalorização) ao longo de um horizonte de tempo finito (ver, por exemplo, Suri (2008)). As políticas ótimas conduzem a um esforço de pesca variável $E^*(t)$ que, sob certas condições, são do tipo 'arranca-pára', isto é, alternam constantemente entre períodos de

pesca e períodos de ausência de pesca, de acordo com os valores (que variam aleatoriamente) do tamanho da população. Este tipo de políticas pode ser aplicada, por exemplo, a activos financeiros, que dada a sua natureza podem ser quantificados quase continuamente, mas não são aplicáveis a populações sujeitas a pesca. Com efeito, a estimação do tamanho de uma população é difícil, dispendiosa e demorada, e a logística da atividade pesqueira não é compatível (quer do ponto vista prático quer do ponto de vista social) com alterações muito frequentes determinadas por variações aleatórias do esforço de pesca.

Uma abordagem alternativa foi proposta (ver Braumann (1985)) com base em políticas de pesca sustentáveis e aplicáveis, que também conduzem à sustentabilidade da população e à existência de distribuição estacionária para o tamanho da população. Neste trabalho determinamos a política de pesca de esforço constante ($E(t)\equiv E$) que otimiza o lucro sustentável (também constante) esperado por unidade de tempo. Seja E^{**} o esforço ótimo de pesca e L^{**} a taxa de lucro ótimo. Com recurso a simulações de Monte Carlo comparamos as duas metodologias, ou seja, comparamos os valores de E^{**} com $E^{*}(t)$ e L^{**} com o lucro ótimo variável esperado por unidade de tempo $L^{*}(t)$ correspondente a $E^{*}(t)$. Podemos verificar que, em situações típicas, há uma ligeira redução do lucro quando se escolhe esta nova política em vez da política inaplicável com esforço variável.

Terminamos com a comparação dos resultados obtidos pelas duas abordagens quando é utilizado o cálculo de Itô e o cálculo de Stratonovich.

Agradecimentos: Nuno M. Brites e Carlos A. Braumann são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT). O primeiro autor é financiado pela FCT através de uma bolsa de doutoramento com a referência SFRH/BD/85096/2012.

Referências

Braumann, C. A. (1985) Stochastic differential equation models of fisheries in an uncertain world: extinction probabilities, optimal fishing effort, and parameter estimation. *In Mathematics in Biology and Medicine* (V. Capasso, E. Grosso, and S. L. Paveri-Fontana, editors), Springer, Berlin, 201-206.

Braumann, C. A. (2002) Variable effort harvesting in random environments: generalization to density - dependent noise intensities. *Math. Biosciences*, 177 & 178, 229-245.

Suri, R. (2008) Optimal harvesting strategies for fisheries: A differential equations approach. PhD thesis, Massey University, Albany, New Zealand.

Zou, Z., Li, W. & Wang, K. (2013) Ergodic method on optimal harvesting for a stochastic Gompertz-type diffusion process. *Applied Mathematics Letters*, 26, 170-174.

Spatial data analysis: A comparison of ICM algorithms

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*

Abstract: This paper highlights the importance of spatial data modeling, namely in the context of restoration-classification models. We focus our attention on the Iterative Conditional Modes (ICM) algorithm proposed by Besag (1986) and the classification finite mixture model framework. The spatial groundwork is a restoration–maximization algorithm for hidden Markov models with parameter estimation via the Expectation – Maximization (EM) algorithm under the pseudo–likelihood assumption. Synthetic Gaussian data and empirical applications are used to compare the performance of both image-restoration algorithms.

Keywords: Classification EM, Hidden Markov random fields, Image restoration, Iterative conditional modes, Spatial data.

Finite mixture models have been used in many research areas and with distinct purposes. A popular area of application is market segmentation in which the goal is to cluster consumers into homogeneous groups. Whenever spatial information is available, the finite mixture model that assumes that observations are independent needs to be adapted, which may lead to a hidden Markov random field. Yet, the complexity of the underlying model structure makes the estimation of the field intractable and approximated expressions are needed, many of them leading to finite mixture models type-like specifications. A hidden random field for segmentation is a probabilistic model in which each observation is allocated into a subpopulation by taking into account the influence of its neighbors. The (spatial) Markovian property of the field is defined under the assumption that the interaction of each site or place with all other sites is restricted to be the interaction with its closest neighbors.

The goal of this work is to compare the ICM algorithm proposed by Besag (1986) with the one proposed by Alfò *et al.* (2008), which differ from the former by allowing that the β parameters of the Gibbs distribution is segment specific. We focus on decision–directed imputation methods and in particular in restoration–maximization algorithms (Qian & Titterington 1991). One of the most common approaches to the restoration step is to take the modal class (maximum probability) that either applies the ICM (iterated conditional modes) algorithm developed by Besag (1986) or computes the true mode and the maximum a posteriori (MAP) restoration. The latter implies computationally demanding estimation techniques such as the simulating annealing algorithm. For the maximization step we apply the EM algorithm. Using a pseudo-likelihood approximation, the local Gibbs probabilities for the hidden variables at segment s is given by

$P_G(Z_i = s \mid \check{z}_{\Delta_i}, \beta)$, where Δ_i represents the neighbors of site i, and captures the neighborhood influence of a given site.

One of the most intensive applications of probabilistic-based segmentation models under spatial neighborhood constraints is image segmentation. The advantage is that it allows grouping image pixels based on their properties such as gray level, color, and texture, and has been an active research area for a long time. Particular applications are in Brain Magnectic Resonance (MR) Images (in order to identify cancer) and satellite images (to identify e.g. crops and roads). Figure 1 depicts an image in which each color has a Gaussian distribution. Figure 2 shows the image recovery using 12 colors (segments).

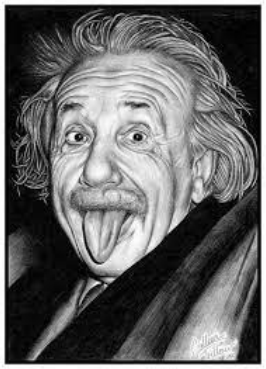


Figure 1: Original image

Figure 2: Restored image

Our results show that for weak spatial dependence both the ICM algorithm and the Alfò's variation have similar performance; but when spatial dependence is stronger, the former becomes better on clustering. For spatial data both algorithms provide better image recovery than the finite mixture model that assumes independent observations.

References

Alfò, M., Nieddu, L. & Vicari, D. (2008) A finite mixture model for image segmentation. *Statistics and Computing*, 18, 137-150.

Besag, J. (1986) On statistical analysis of dirty pictures. *Journal of the Royal Statistics Society B*, 48(3), 259-302.

Qian, W. & Titterington, D. M. (1991) Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Statistical Society of London A*, 337, 407-428.

Discriminant analysis and classification of distributional and interval data

Sónia Dias[1], Paula Brito[2], Paula Amaral[3]
[1] Escola Superior de Tecnologia e Gestão - Instituto Politécnico Viana do Castelo & LIAAD-INESC TEC, Universidade do Porto; sdias@estg.ipvc.pt;
[2] Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto; mpbrito@fep.up.pt;
[3] CMA & Faculdade de Ciência e Engenharia, Universidade Nova de Lisboa, paca@fct.unl.pt

Abstract: Histogram and interval-valued variables are particular types of variables studied in Symbolic Data Analysis. For these variables, each entity under analysis is described by a distribution or an interval, which may be represented by a quantile function under some distribution assumption. In this work a symbolic discriminant function is defined, allowing for the classification of a set of individuals, characterized by histogram or interval-valued variables, in two *a priori* classes.

Keywords: Distributional data, Fractional quadratic problems, Linear discriminant analysis, Mallows distance, Quantile functions.

Technological developments and applied research brought a lot of changes to the kind of data that is necessary to manage and analyze. Data tables where the cells contain a single quantitative or categorical value are no longer sufficient. Consequently, more complex data tables emerged, with cells that include more accurate and complete information. This is the case of data studied in Symbolic Data Analysis. The cells of symbolic data tables may contain finite sets of values/categories, intervals or distributions. Variables whose observations are distributions over a (finite) set of (sub-) intervals are named histogram-valued variables, whereas variables whose observations are intervals are named interval-valued variables; it should be noticed however that these latter constitute a special case of histogram-valued variables. To simplify the complexity of working with frequency distributions, and under some distribution assumption, quantile functions (the inverse of the cumulative distribution functions) are often used (Brito 2014).

In this work we develop a linear discriminant method for distributional data that may be particularized to interval-valued variables. The proposed model aims at obtaining a linear combination of features, now defined by distributions or intervals that characterize the individuals or classes of individuals, and that allows classifying them in different *a priori* groups.

Dias & Brito (2015) proposed a linear regression model - the Distribution and Symmetric Distributions (DSD) model - which is based on the definition of a linear combination of quantile functions, thereby obtaining distributions from other distributions. Using this linear combination, we define a discriminant function that allows

for the classification of a set of individuals in two classes. For each individual, a linear combination obtained as in the DSD Model is considered, which allows defining a score of the individual in the form of a quantile function. The observation is then assigned to the closest class.

The function to optimize to obtain the parameters of the linear discriminant function is based on the total inertia decomposition with respect to a barycentric histogram, defined with the Mallows distance. Irpino & Verde (2006) proved that total inertia so defined may be decomposed into within and between classes inertia, according to the Huygens theorem. The coefficients of the discriminant function are then obtained by maximizing the ratio of the between to the within classes inertia, subject to some constraints. This defines a constrained fractional quadratic problem. The BARON is used to solve this difficult optimization problem. A solution is obtained, but the optimality certificate is only possible using conic relaxation techniques (Amaral *et al.* 2014).

Some examples, both with synthetic and real data, illustrate the proposed method.

Acknowledgments: Este trabalho é financiado por Fundos FEDER através do Programa Operacional Competitividade e Internacionalização – COMPETE 2020 no âmbito do projeto «POCI-01-0145-FEDER-006961» e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia através do projeto UID/EEA/50014/2013

Referências

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

Dias, S. & Brito, P. (2015). Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 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 HH, Ferligoj A, Ziberna A (eds). *Data Science and Classification, Proceedings of the IFCS'2006,* Ljubljana, Slovenia, 185–192.

Amaral, P., Bomze, I. & Júdice, J. (2014). Copositivity and constrained fractional quadratic problems. *Math. Program.,* 146(1–2), 325–350.

Modelos de regressão multinível no estudo do desempenho escolar

Susana Faria[1], **João Silva**[2]
[1] *CMAT - Centro de Matemática, DMA - Departamento Matemática e Aplicações, Universidade do Minho, sfaria@math.uminho.pt;*
[2] *DMA - Departamento Matemática e Aplicações, Universidade do Minho, joaosilva17@sapo.pt*

Sumário: Neste trabalho, aplicando modelos de regressão multinível, pretende-se comparar os desempenhos escolares dos estudantes a Matemática em diversos países, e identificar quais os fatores que podem influenciar, positivamente ou negativamente, esse desempenho. Utilizam-se os resultados obtidos pelos estudantes relativamente à literacia em Matemática, nos testes do Programme for International Student Assessment (PISA) de 2012 da Organização para a Cooperação e Desenvolvimento Económico (OCDE).

Palavras-chave: Desempenho dos alunos, Modelo multinível de regressão, *Programme for International Student Assessment (PISA) 2012*.

Os modelos de regressão multinível são modelos de regressão que se diferenciam dos modelos clássicos de regressão linear, pois têm conta a estrutura hierárquica da população, incorporando as diferentes hierarquias observacionais dos dados, produzindo desta forma inferências mais fiáveis (Raudenbush & Bryk 2002). Por esse motivo, estes modelos são muito aplicados no contexto educacional: os alunos estão agrupados em turmas, as turmas em diferentes escolas, as escolas em agrupamentos,….

Neste trabalho, estes modelos foram aplicados a dados obtidos no âmbito do *Programme for International Student Assessment* (PISA) de 2012, na literacia de Matemática. Estimaram-se modelos de regressão multinível com três níveis (nível 1 – estudantes, nível 2 - escolas e nível 3 – países) para comparar os desempenhos escolares dos estudantes em vários países, com o intuito de identificar quais são os fatores que podem influenciar esse desempenho.

O PISA é um estudo a nível mundial desenvolvido pela Organização para a Cooperação e Desenvolvimento Económico (OCDE), que visa avaliar as capacidades dos estudantes de 15 anos, para utilizarem os conhecimentos adquiridos na resolução de problemas reais em três áreas principais: Leitura, Matemática e Ciências. Em 2012, para além destas áreas, alguns países ainda participaram na avaliação em outras duas áreas opcionais: Resolução criativa de problemas e Literacia financeira. Este programa tem uma grande utilidade na medida em que fornece dados comparáveis ao longo do tempo e entre países que permite a avaliação da qualidade dos sistemas educativos (Fuchs & Woessmann 2007).

Foi realizada uma análise exploratória dos dados e aplicadas metodologias estatísticas na área da Inferência Estatística (testes de hipóteses) e concluiu-se que os estudantes de Singapura e da Coreia do Sul, em média, foram os estudantes que apresentavam melhores desempenhos a Matemática (ANOVA, Tukey's test, $p < 0.01$). Por outro lado, os estudantes dos Estados Unidos e da Suécia foram os estudantes que apresentam piores desempenhos a Matemática. No caso particular de Portugal, verificou-se que os estudantes portugueses apresentavam um desempenho médio a Matemática de 486.465 pontos, um valor ligeiramente abaixo da média da OCDE (500 pontos).

A aplicação de um modelo de regressão multinível com três níveis permitiu concluir que os rapazes têm desempenhos médios a Matemática mais elevados que as raparigas e que os estudantes que vivem com ambos os pais têm desempenhos médios superiores aos estudantes que tenham outro tipo de estrutura familiar. O nível económico, social e cultural do estudante tem um impacto positivo no desempenho a Matemática dos estudantes, sendo este um dos fatores que mais influencia esse desempenho. Por outro lado, o facto de os estudantes serem imigrantes ou já terem repetido um ano escolar tem uma influência negativa no desempenho dos estudantes a Matemática.

Relativamente às variáveis da escola, a proporção de computadores ligados à *internet*, a proporção de raparigas na escola, o comportamento do estudante e o facto de os estudantes frequentarem escolas privadas independentes do governo, têm um impacto positivo no desempenho dos estudantes a Matemática. Por outro lado, o rácio entre o número de estudantes e o número de professores de Matemática tem uma influência negativa no desempenho dos estudantes a Matemática.

No que diz respeito às variáveis do país, a autonomia das escolas na alocação dos recursos escolares influencia positivamente o desempenho dos estudantes a Matemática mas, o rácio entre o número de computadores para fins educacionais e o número de estudantes tem um impacto negativo no desempenho dos estudantes a Matemática.

Referências

Fuchs, T. & Woessmann, L. (2007) What accounts for international differences in student performance? A reexamination using PISA data. *Empirical Economics*, 32, 433-464.

Goldstein, H. (2011) *Multilevel Statistical Models*. John Wiley and Sons.

OECD (2015) *Students, Computers and Learning: Making the Connection*. PISA, OECD Publishing.

Raudenbush, S. W. & Bryk, A. S. (2002) *Hierarchical Linear Models* (Second Edition). Thousand Oaks, Sage Publications.

Sessão Paralela I – Classificação e Regressão – 6ª feira, 1 de abril, Anfiteatro 131 (10:10)

A utilização dos *Mídia Social* para efeitos de viagem: Uma abordagem a partir da análise de agrupamento

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: As decisões tomadas no processo de compra de viagens são cada vez mais influenciadas pelos mídia social. Deste modo, é crucial que as empresas turísticas conheçam os seus potenciais clientes quanto ao seu perfil de uso de mídia social, por forma a adaptarem as suas estratégias. Neste trabalho, recorrendo a técnicas de análise de agrupamento, identificam-se cinco segmentos entre os viajantes, tendo em conta a sua utilização dos mídia social para efeitos de viagens.

Palavras-chave: Análise de agrupamento, Mídia social, Segmentação de mercado.

Está amplamente reconhecida a influência dos mídia social (MS) nas decisões de compra de viagens. Muitos estudos têm sido feitos à volta desta temática, indicando a importância crescente da utilização dos MS no planeamento e decisão de compra de viagens. Por exemplo, de acordo com o Tripadvisor (2014), na escolha do alojamento, 89% dos viajantes são influenciados por opiniões obtidas *on-line*. A mesma fonte indica-nos que 96% dos hotéis afirmam que as opiniões *on-line* são influentes na geração de reservas. Contudo, nem todos os viajantes têm o mesmo comportamento no uso de MS. Neste trabalho pretende-se identificar segmentos de viajantes com respeito ao seu perfil de uso de MS.

Os dados para este estudo foram recolhidos a partir de um questionário colocado *on-line* de 28 de julho a 25 de agosto de 2012, tendo sido obtidas 1732 respostas. Cinco questões foram incluídas para medir o consumo de mídia social (SMC) no contexto de viagens, e outras cinco para medir a criação de conteúdos (SMCR). Estas questões foram medidas numa escala de Likert de 5 pontos. O questionário incluía ainda outras questões relacionadas com viagens e com a forma como os viajantes percecionam o seu grau de envolvimento e de entretenimento no uso de MS. As variáveis de consumo e criação de conteúdos sobre viagens nos MS foram utilizadas numa análise de agrupamento para encontrar segmentos de mercado em termos de utilização de MS. A análise de agrupamento envolveu duas fases. Primeiro, aplicaram-se três métodos hierárquicos: o método do vizinho mais afastado o método da ligação média e o método Ward. Numa

segunda fase, aplicou-se o método das k-médias com soluções iniciais fornecidas pelos métodos hierárquicos. Tanto a estabilidade como a interpretação dos grupos finais levaram à escolha da solução de cinco grupos. De facto, comparando as soluções finais de cinco grupos, cada uma obtida por meio do método das k-médias com soluções iniciais dadas por um dos três algoritmos hierárquicos, menos do que 3% dos casos foram atribuídos a grupos diferentes nas três soluções de agrupamento. A análise da estabilidade da solução de 5 grupos foi também levada a cabo dividindo a amostra aleatoriamente em duas partes e aplicando a cada uma o método das k-médias com solução inicial dada pelo algoritmo Ward. Uma vez mais, observou-se uma percentagem muito baixa de alterações em membros dos grupos (5,4%), o que suporta a estabilidade da solução de cinco grupos. Refira-se que estas percentagens estão muito abaixo dos 10% sugeridos por Hair *et al.* (2010) para alterações nos membros de grupos em soluções muito estáveis. Os grupos encontrados estão representados na Figura 1.

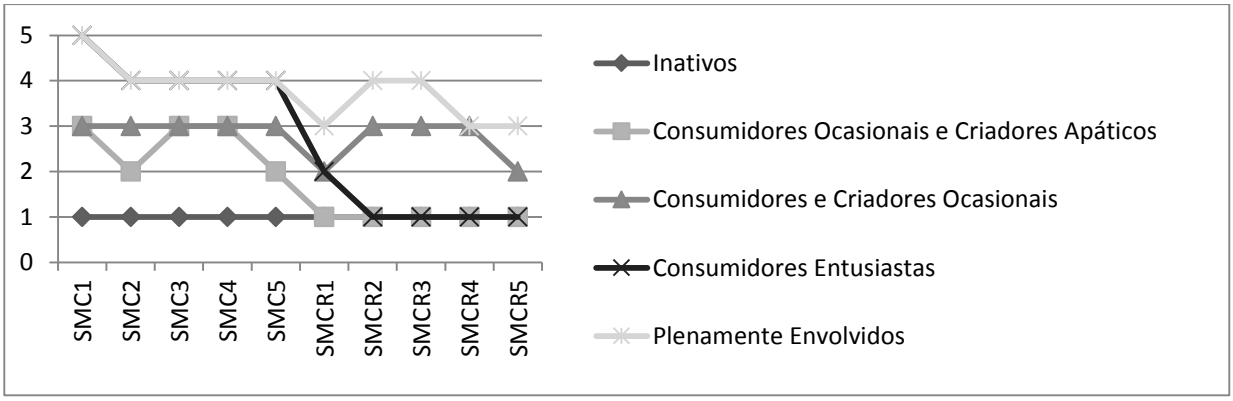


Figura 1 – Perfil dos 5 grupos

Variáveis externas (não utilizadas na segmentação) foram consideradas para caracterizar e validar a estrutura de agrupamento, identificando dimensões em que os grupos diferem significativamente. Para isso recorreu-se ao teste do Qui-quadrado para variáveis categóricas e ao teste de Kruskal-Wallis para variáveis ordinais. Os resultados obtidos fornecem informações úteis para agentes de viagens e provedores de sites de MS que precisam estar cientes dos diferentes segmentos para poderem personalizar sites em conformidade. O trabalho tem contudo algumas limitações, das quais se destaca o facto de os dados não serem recentes e a necessidade de se assumir que a diferença entre dois níveis consecutivos da escala teria sempre o mesmo significado, podendo ser quantificada da mesma maneira.

Referências

Hair, J. F., Black, W. C., Babin, H. J. & Anderson, R. E. (2010) *Multivariate Data Analysis* (7th ed.). New Jersey: Prentice Hall.

Tripadvisor (2014) TripBarometer April 2014: Global Edition, acedido em fevereiro de 2016 a partir de http://www.tripadvisor.com/TripAdvisorInsights/n2200/tripbarometer-april-2014-global-edition

Uma aplicação da carta EWMA a dados correlacionados

Dora Carinhas[1], Paulo Infante[2]
[1] *Instituto Hidrográfico, dora.carinhas@hidrografico.pt;*
[2] *CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt*

Sumário: Este trabalho apresenta cartas de controlo EWMA e de observações individuais (X) utilizadas na monitorização de processos cujas observações podem ser descritas por determinados modelos que acomodem a autocorrelação dos dados. Tomando como base, dados reais de nutrientes em águas salinas, os resultados mostram que a correlação entre observações e o tipo de modelo utilizado afeta a interpretação e o desempenho das cartas de controlo.

Palavras-chave: Autocorrelação, Cartas de controlo, Controlo estatístico do processo.

As cartas de controlo são uma importante ferramenta da qualidade, que permite distinguir entre a variação inerente ao processo e a variação resultante de causas especiais, constituindo uma representação gráfica da estabilidade ou instabilidade de um processo ao longo do tempo. Estas ferramentas estatísticas são habitualmente planeadas e avaliadas assumindo que as observações do processo são independentes e identicamente distribuídas (i.i.d.), hipótese que é frequentemente violada na prática, pois em muitos processos as observações são correlacionadas.

Uma estratégia frequentemente utilizada, sugerida por Reynolds & Lu (1997), consiste em ajustar as observações da característica de qualidade a um modelo de previsão apropriado e monitorizar o processo com cartas de controlo para os resíduos i.i.d. resultantes.

Embora a carta X seja mais frequente quando trabalhamos com observações individuais, Montgomery (2009) recomenda a carta EWMA para medições individuais, em particular na fase de monitorização quando se pretendem detetar pequenas alterações do processo, independentemente da distribuição das observações. Neste trabalho comparamos o desempenho das duas cartas.

Para se perceber a importância de se poder assumir a independência entre as observações, considere-se como exemplo a participação do Instituto Hidrográfico em Ensaios de Comparação Interlaboratorial (ECI), organizados pelo laboratório Quasimeme, entre abril de 2004 e abril de 2013, nomeadamente para a determinação do nutriente nitrito em águas salinas. Neste caso é monitorizada a estatística z-score (diferença entre o resultado obtido pelo laboratório participante e o valor de referência dividida pelo desvio-padrão do ECI). Pelos gráficos de autocorrelação e autocorrelação parcial (ACF e PACF, respetivamente), verificou-se que a hipótese de independência das observações dos z-scores não era válida, pelo que se procedeu à obtenção do modelo que se mostrou mais adequado, neste caso particular, o modelo ARIMA (1, 1, 1). Em seguida aplicou-se a carta

EWMA aos resíduos com o objetivo de verificar se o processo estava, ou não, sob controlo estatístico (Figura 1).

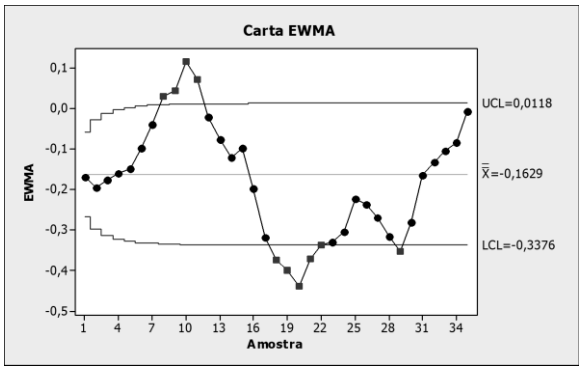


Figura 1: Carta EWMA aplicada aos z-scores

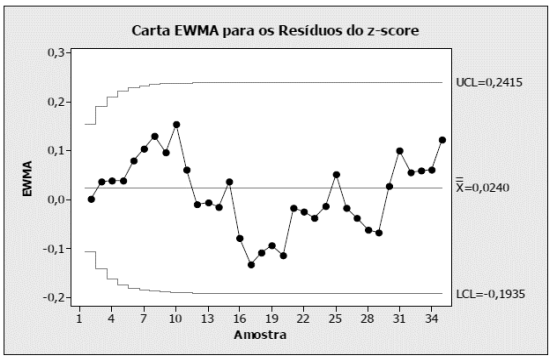


Figura 2: Carta EWMA aplicada aos resíduos do modelo X_t-0,6211X_{t-1} = -0,9554e_{t-1}

Verifica-se que, no primeiro caso, surgem algumas observações acima do limite superior de controlo indiciando um aumento da média dos z-scores e mais tarde algumas observações abaixo do limite superior de controlo indiciando uma diminuição da média dos z-scores. Portanto, a Figura 1 é reveladora de um processo fora de controlo estatístico. No segundo caso que pode ser observado na Figura 2, todos os pontos encontram-se dentro dos limites de controlo o que traduz um processo sob controlo estatístico com média muito perto do valor alvo (igual a zero) para a estatística z-score.

Em muitas aplicações de controlo estatístico de processos as soluções para as questões de autocorrelação não são fáceis de encontrar. Embora a carta de controlo aplicável aos resíduos seja uma ferramenta de inegável valor na melhoria dos processos, uma das suas desvantagens é a difícil interpretação por parte dos técnicos que atuam na operacionalização do processo. Outro inconveniente é ajustar e manter um modelo de séries temporais apropriado para cada variável do processo.

Agradecimentos: Ao Instituto Hidrográfico pela cedência dos dados, em particular à Eng.ª Pilar Pestana e ao Dr. Carlos Borges. Paulo Infante é membro do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Referências

Montgomery, D. C. (2009) *Introduction to Statistical Quality Control* (6ª Ed). New York, USA: Wiley.

Reynolds, M. R. & Lu, C.W. (1997) Control Charts for Monitoring Process with Autocorrelated Data. *Nonlinear Analysis*: *Theory, Methods & Applications*, 30, 4059-4067.

A importância da adequabilidade do modelo no desempenho de cartas de controlo com risco ajustado

Maria João Inácio[1], Paulo Infante[2], Fernanda Otília Figueiredo[3]
[1] Instituto Politécnico de Tomar - CIMA, mjantunes@ipt.pt;
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt;
[3] Faculdade de Economia da Universidade do Porto - CEAUL, otilia@fep.up.pt

Sumário: Várias cartas de controlo, tradicionalmente utilizadas na indústria, têm sido adaptadas para uma utilização mais eficiente na área da Saúde, incorporando na análise, para cada paciente, o risco de ocorrência do acontecimento que se pretende monitorizar. Neste estudo, fazemos uma breve abordagem às cartas de controlo por atributos com risco ajustado e avaliamos o impacto, no desempenho das cartas, do uso de um modelo não adequado.

Palavras-chave: ARL, Cartas de controlo por atributos, Controlo de qualidade.

As cartas de controlo foram desenvolvidas por Walter Shewhart em 1920 e posteriormente popularizadas por Deming. Ao longo dos anos têm sido feitas várias adaptações às cartas existentes e desenvolvidas novas metodologias, por forma a tornar esta ferramenta mais potente e/ou mais adequada em diversas situações.

Apesar de criadas inicialmente com o intuito de serem utilizadas em processos industriais, as técnicas de controlo estatístico do processo têm-se revelado muito úteis em todas as áreas, incluindo a área da Saúde. Mas, ao contrário dos processos industriais, onde os sujeitos são relativamente homogéneos, em aplicações médicas os sujeitos (pacientes) frequentemente variam muito em termos do risco de ocorrência do acontecimento que se pretende monitorizar.

Uma possível abordagem consiste em atribuir, a cada paciente, um risco de ocorrência do acontecimento que se pretende monitorizar. Para a construção de cartas de controlo com risco ajustado a probabilidade de ocorrência do acontecimento que se pretende monitorizar pode variar de paciente para paciente de acordo com um modelo probabilístico assumido à priori.

Ao longo dos últimos anos têm sido propostas várias cartas de controlo por atributos adaptadas para risco ajustado. Entre as mais populares salientamos as cartas ***RAP*** (*Risk-Adjusted P-charts*), ***VLAD*** (*Variable Life-Adjust Display),* ***CRAM*** *(Cumulative Risk-Adjusted Mortality*), ***RAEWMA*** *(Risk-Adjusted Exponentially Weight Moving Average)*, ***RASPRT*** *(risk-adjusted sequential probability ratio tests)* e ***RACUSUM*** *(Risk-Adjusted CUmulative SUM* control chart*).*

Neste trabalho damos uma atenção especial às cartas RACUSUM. As hipóteses subjacentes a estas cartas baseiam-se nas razões de chance R_0 e R_1 (especificadas nas

hipóteses nula e alternativa, respetivamente). Por exemplo, no caso do acontecimento que estamos a monitorizar ser a morte do paciente após ser submetido a determinado tratamento, para avaliarmos se a chance de morte aumentou, o que significa que os resultados são piores do que o que era esperado, consideramos $R_1 > R_0$.

Seja p_j o risco estimado de morte do paciente j no início do tratamento e seja y_j a variável dicotómica que assume o valor 1 se o paciente j faleceu. Assim, para o paciente j os scores são definidos da seguinte forma:

$$W_j = \begin{cases} ln\left(\frac{\left(1-p_j+R_0p_j\right)R_1}{\left(1-p_j+R_1p_j\right)R_0}\right) & \text{se } y_j = 1 \\ ln\left(\frac{1-p_j+R_0p_j}{1-p_j+R_1p_j}\right) & \text{se } y_j = 0 \end{cases}$$

A estatística a monitorizar é atualizada sempre que temos informação sobre um novo paciente e, no caso de querermos avaliar se houve uma deterioração do processo ($R_1 > R_0$), a estatística é dada por: $S_j^+ = max\left(0, S_{j-1}^+ + W_j\right),\ j = 1,2,3, \dots$, com $S_0^+ = 0$.

Para avaliarmos se houve uma melhoria ($R_1 < R_0$), a estatística é dada por:

$$S_j^- = min\left(0, S_{j-1}^- - W_j\right),\ j = 1,2,3, \dots, \text{com } S_0^- = 0.$$

Atendendo a que S_j^+ é sempre não-negativo e S_j^- é sempre não-positivo, podemos traçar os dois gráficos RACUSUM sobre o mesmo eixo horizontal.

Os limites de controlo são obtidos em função do ARL (A*verage Run Length*), sob a hipótese de que não houve alterações no processo, e cujos valores obtivemos por simulação de Monte Carlo.

O Risco estimado para cada paciente (p_j) é obtido assumindo um determinado modelo. A adequabilidade e a capacidade preditiva destes modelos afetam o desempenho das cartas. Neste trabalho, considerando uma base de dados reais de um estudo sobre cirurgia cardíaca apresentado em Steiner *et al.* (2000), estudamos a sensibilidade das cartas à não adequabilidade do modelo, aferindo o seu impacto no desempenho das mesmas, utilizando o ARL como medida estatística do desempenho.

Agradecimentos: Os autores agradecem ao professor Stefan Steiner da Universidade de Waterloo por ter facultado os dados que serviram de base para este trabalho. Paulo Infante é membro do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Referências

Cook, D., Duke, G., Hart, G., Pilcher, D. & Mullany, D. (2008) Review of the application of risk-adjusted charts to analyse mortality outcomes in critical care. *Critical Care and Resuscitation*, 10, 239-251.

Modelo de regressão de Cox robusto, uma aplicação a dados oncológicos

Eunice Carrasquinha[1], André Veríssimo[2], Susana Vinga[3]
[1, 2, 3] IDMEC, Instituto Superior Técnico, Universidade de Lisboa;
[1] eunice.trigueirao@tecnico.ulisboa.pt;
[2] andre.verissimo@tecnico.ulisboa.pt;
[3] susana.vinga@tecnico.ulisboa.pt

Sumário: A diferença principal do método robusto para o não robusto, no contexto do modelo de regressão de Cox, está na atribuição de pesos na função de verosimilhança do modelo semi-paramétrico de Cox. A robustez no modelo de Cox é mais sensível a indivíduos que tenham um tempo de vida demasiado curto ou demasiado longo, quando comparados com os restantes e no âmbito do modelo estimado. Com dados provenientes de doentes oncológicos fez-se um estudo comparativo do modelo de regressão de Cox com e sem robustez. Os resultados obtidos mostram diferenças nos valores dos parâmetros estimados.

Palavras-chave: Análise de sobrevivência, Modelo de Cox robusto**,** Modelo de regressão de Cox, Outliers.

A análise de sobrevivência é um método utilizado nas mais variadas áreas da ciência, em particular na medicina, onde o interesse reside no estudo do tempo até um determinado evento, o qual se denomina por tempo de falha ou tempo de sobrevivência. O evento de interesse pode ser a morte, o desenvolvimento de uma determinada doença, o aparecimento de um tumor, etc. Uma característica importante que distingue a análise de sobrevivência de outros métodos estatísticos é o facto de existir censura, ou seja, para alguns indivíduos em estudo não é observada a realização do evento de interesse durante o intervalo em que os indivíduos estão em observação (*follow-up*).

Existem na literatura vários modelos de regressão destinados à análise de dados de sobrevivência, no entanto o mais utilizado é o modelo de regressão proposto por Cox (1972), o qual se baseia na hipótese de riscos proporcionais. Com base nesta hipótese, a estimação dos parâmetros de interesse podem ser determinados com base numa função de verosimilhança semi-paramétrica na qual não é necessário especificar a função de riscos subjacente (*baseline*). Apesar da sua flexibilidade em modelar tempos de riscos instantâneos, mesmo na presença de observações censuradas, muitos têm sido os autores a propor métodos alternativos que tenham uma melhor robustez na presença de observações discrepantes (*ouliers*) e melhorem o *breakdown point* dos modelos de Cox de 1/n, o que significa que para um conjunto de dados com n observações, uma única observação discrepante é suficiente para que o valor obtido pelo estimador seja diferente do seu verdadeiro valor.

Huber (1964) introduz o conceito de estatística robusta na comunidade científica. Apesar do seu avanço nas mais variadas áreas científicas, a introdução de robustez em análise de sobrevivência não foi imediata, devido ao facto de os dados terem observações censuradas. O pioneiro em introduzir o termo robusto à análise de sobrevivência foi Bednarski (1993), sendo uma alternativa robusta ao modelo semi-paramétrico. Esta alternativa baseia-se na introdução de pesos na função de verosimilhança semi-paramétrica sem alterar a sua estrutura.

Sabe-se que os valores extremos nas variáveis explicativas podem alterar os resultados da variável resposta. Em análise de sobrevivência estes valores extremos são indivíduos que têm um tempo de vida demasiado longo ou demasiado curto, dadas as suas covariáveis. O modelo de Cox robusto consiste em atribuir duplos pesos no modelo semi-paramétrico de Cox, tornando-o mais sensível à presença de valores extremos.

De forma a comparar a performance do modelo de regressão de Cox com o modelo de Cox robusto, analisaram-se dados provenientes de doentes oncológicos tendo em conta dados clínicos, demográficos e bio-marcadores. Para o conjunto de dados analisaram-se ambos os modelos e identificaram-se algumas observações extremas. Concluiu-se que no modelo de Cox robusto surgiram novas covariáveis estatisticamente significativas, confirmando-se que apresenta uma maior sensibilidade na presença de observações discrepantes face ao modelo de Cox original

Agradecimentos: Projeto Europeu Horizon 2020, SOUND - Statistical multi-Omics UNDerstanding of Patient Samples, (European Union – No. 633974), e Fundação para a Ciência e Tecnologia com a bolsa de doutoramento com referência SFRH/BD/97415/2013.

Referências

Bednarski, T. (1993) Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*, 20, 213-225.

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

Huber, P. J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

Pinto, J. D., Carvalho, A. M. & Vinga, S. (2015) Outlier Detection in Survival Analysis based on the Concordance C-index. *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOSTEC 2015)*, 75-82.

Avaliação das propriedades psicométricas e validação do NEO-FFI para estudantes portugueses de tecnologias da saúde

Helena Martins[1], Ana Reis[2], Ana Salgado[2], Andreia Magalhães[2], Zita Sousa[2], Artemisa R. Dores[2]

[1] Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto, mhm@estsp.ipp.pt;

[2] Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto

Sumário: O modelo de 5 fatores (Big Five) organiza a personalidade humana em cinco fatores: neuroticismo, extroversão, abertura à experiência, amabilidade e conscienciosidade. O Neo-Five Factor Inventory (Neo-FFI) avalia a personalidade no modelo teórico referido. Neste trabalho apresentamos os esforços de validação da escala para estudantes portugueses de tecnologias da saúde, incluindo fiabilidade, análise fatorial exploratória e análise fatorial confirmatória.

Palavras-chave: Big Five, Instrumento, Personalidade, Psicometria, Validação.

O modelo de 5 fatores (Big Five) organiza a personalidade humana em cinco fatores: neuroticismo, extroversão, abertura à experiência, amabilidade e conscienciosidade.

Originalmente desenvolvido em 1932, o modelo de 5 fatores de personalidade tem vindo a ser modificado e melhorado desde então; apesar do debate a que o modelo tem sido submetido, a generalidade das alternativas sugeridas e este modelo na literatura incluem partes significativas deste modelo, que frequentemente se subdividem em fatores diferentes ou que sugerem fatores adicionais. Este modelo tem sido aplicado na investigação em diferentes áreas, incluindo a Educação e o Ensino Superior (e.g., Vedel 2016). Neste contexto, tem-se estudado sobretudo a influência da personalidade nas competências cognitivas dos estudantes, na escolha da área de estudo, bem como na melhor ou pior capacidade de adaptação/sucesso em determinados cursos.

O instrumento em estudo, o Neo-Five Factor Inventory (NEO-FFI) avalia a personalidade no modelo teórico de 5 fatores de Costa & McCrae (1992). Assim, o NEO-FFI é um instrumento com 60 itens de autorrelato num formato Likert de 5 pontos, de 0 (nada de acordo) a 4 (completamente de acordo).

O instrumento tem apresentado resultados de fiabilidade bastante estáveis em diferentes culturas, e a escala foi adaptada para a população portuguesa em 2014 por Magalhães *et al.* (2014), com resultados de fiabilidade entre 0,75 e 0,82, tendo o instrumento mantido a sua estrutura original.

O objetivo principal deste estudo foi analisar as características psicométricas do NEO-FFI (Costa & McCrae 1992) em estudantes portugueses de tecnologias da saúde. A amostra incluiu 703 estudantes de diversos cursos de licenciatura na área das tecnologias

da saúde da Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto (ESTSP – IPP).

A análise psicométrica do NEO-FFI compreendeu, num primeiro momento, o estudo da fiabilidade através do Alpha de Cronbach e a análise fatorial exploratória para testar as propriedades psicométricas do instrumento na amostra em estudo. Nesta primeira fase, foi possível confirmar a fiabilidade da escala, tendo sido obtidos alphas de Cronbach superiores a 0,70 em todas as dimensões.

Após levar a cabo procedimentos de análise fatorial exploratória com recurso ao SPSS v.21, as autoras encontram-se a concluir procedimentos confirmatórios que testam o ajustamento do modelo e o comparam com o modelo original (Costa & McCrae 1992) e Magalhães *et al.* (2014), usando o software AMOS v.20 e a mesma amostra.

O presente estudo insere-se num projeto de âmbito nacional que visa explorar a forma como os estudantes da área da Saúde podem desenvolver competências de comunicação no Ensino Superior. Este trabalho contribui para a literatura na área e para investigações futuras através da validação de um instrumento de medida importante para esta população.

Referências

Costa, P. T. & McCrae, R. R. (1992) Four ways five factors are basic. *Personality and Individual Differences*, 13(6), 653-665.

Magalhães, E., Salgueira, A., Gonzalez, A-J., Costa, J. J., Costa, M. J., Costa, P. & Lima, M. P. (2014). NEO-FFI: Psychometric Properties of a Short Personality Inventory in Portuguese Context. *Psychology/Psicologia Reflexão e Crítica*, 27(4), 642-657.

Vedel, A. (2016). Big Five personality group differences across academic majors: A systematic review, *Personality and Individual Differences,* 92, 1-10.

Valor percebido do consumidor e comércio de retalho: uma abordagem multidimensional hierárquica de 2ª ordem à versão reduzida da escala PERVAL

João Saramago[1], Ana Sampaio[2,3], Elizabeth Reis[3]
[1] Mestrado em Modelação, Estatística e Análise de Dados, Universidade de Évora, saramago_85@hotmail.com;
[2] Departamento de Matemática da Universidade de Évora, Ana Sampaio, sampaio@uevora.pt;
[3] Instituto Universitário de Lisboa, ISCTE-IUL, Business Research Unit, elizabeth.reis@iscte.pt

Sumário: Neste trabalho é adotada a escala de perceção de valor (PERVAL) de Sweeney e Soutar (2001) na versão reduzida de 12 indicadores (Walsh *et al.* 2014). No contexto multidimensional do valor percebido pelo consumidor da sua relação com o comerciante de retalho, são comparadas duas estruturas hierárquicas de 2ª ordem, formativa e refletiva, evidenciando as suas vantagens e desvantagens. Para avaliação e comparação da precisão e eficiência das estimativas de máxima verosimilhança, foram utilizados diferentes contextos de dimensões amostrais.

Palavras-chave: Escala PERVAL reduzida, modelos de equações estruturais, valor percebido pelo consumidor.

A investigação em torno do conceito do valor percebido pelo consumidor (VPC) e a utilização de modelos VPC em sectores da economia onde as vantagens competitivas representam estratégias cruciais para o desenvolvimento sustentável, tem vindo a aumentar significativamente nos últimos anos. Associada à área de pesquisa em marketing, a definição primordial de valor percebido pode ser estabelecida como uma avaliação global do consumidor da utilidade do produto (ou do serviço) baseado na perceção do que é recebido. Embora numa primeira fase do desenvolvimento do constructo tenha prevalecido o seu perfil unidimensional, apenas justificado por aspetos económicos e cognitivos, a identificação, em fases posteriores, de aspetos hedónicos e estéticos subjacentes ao processo de consumo e à avaliação do valor percebido pelo consumidor, motivou o interesse crescente pela vertente multidimensional do conceito e a crítica do modelo unidimensional. Preciosos contributos decorreram desta fase de desenvolvimento conceptual do VP, com várias propostas de modelos de 1ª ordem e refletivos. Os modelos conceptuais de indicadores e/ou dimensões de ordens superiores, trouxeram nova polémica acerca dos papéis de dependência protagonizados para os constructos. Neste trabalho é adotada a versão reduzida, com doze indicadores (Walsh *et al.* 2014) do modelo PERVAL (S&S), avaliado segundo uma escala de tipo Likert de 6 pontos (1:discordo totalmente, 6:concordo totalmente). São utilizados quatro fatores de 1ª ordem (emocional, social, preço e qualidade) subjacentes ao conceito de valor e duas estruturas hierárquicas fatoriais de 2ª ordem para a avaliação de VPC na sua globalidade. Numa primeira estrutura e na linha de S&S, foi considerado um modelo de medida, com

quatro variáveis latentes correlacionadas, de 1ª ordem e que refletem uma dimensão superior de 2ª ordem para VPC. Numa segunda estrutura fatorial, que difere apenas ao nível da conceptualização formativa da relação entre as variáveis latentes de 1ª ordem e VPC, foi considerado um modelo de medida com um constructo de 2ª ordem multidimensional formativo VPC. Participaram no estudo 595 consumidores com idades compreendidas entre 18 e 85 anos. A perceção dos consumidores sobre as diferentes dimensões da relação com o retalhista foi obtida a partir de um questionário estruturado, aplicado entre junho e novembro de 2013, na região da grande Lisboa.

Os resultados da análise fatorial confirmatória realizada indicam que o modelo de 1ª ordem apresenta índices de ajustamento adequados ($X^2 = 199{,}5; df = 48; p\text{-}value = 0{,}000; CFI = 0{,}96; TLI = 0{,}95; RMSEA = 0{,}07; SRMR = 0{,}06$), que os índices de validade convergente e discriminante para todos os fatores foram satisfatórios e que todos os indicadores apresentaram boa fiabilidade interna (pesos fatoriais superiores a 0,7). Os resultados obtidos para a avaliação dos ajustamentos realizados com os dois modelos hierárquicos de 2ª ordem, foram, igualmente, satisfatórios: ($\chi^2 = 272{,}5; df = 80; p\text{-}value = 0{,}000; CFI = 0{,}96; TLI = 0{,}95; RMSEA = 0{,}06; SRMR = 0{,}06; AIC = 15670{,}09; MFI = 0{,}867$) para o modelo formativo e ($\chi^2 = 219{,}6; df = 50; p\text{-}value = 0{,}000; CFI = 0{,}96; TLI = 0{,}95; RMSEA = 0{,}07; SRMR = 0{,}07$) para o modelo reflectivo. Todos os coeficientes estruturais foram estimados com significância estatística e sinais expectáveis em ambas as estruturas factoriais analisadas. Amostras adicionais de dimensões 50, 100, 150, 595 e 1000 foram obtidas por meio da técnica de reamostragem bootstrap sendo que se verificaram, nas amostras de menor dimensão (n=50 e n=100), diferenças significativas entre os modelos, ao nível da precisão e da eficiência das estimativas. Para o modelo refletivo obteve-se maior precisão com base nos valores do viés relativo, e menor precisão com base no desvio da raiz dos mínimos quadrados (RMSD). Relativamente à eficiência, a estrutura hierárquica formativa revelou ser mais eficiente que a estrutura hierárquica refletiva, também para amostras de menor dimensão (n=50, n=100 e n=150). Para amostras de n=595 e n=1000 as diferenças encontradas não foram significativas.

Os resultados sugerem que, apesar do modelo refletivo apresentar uma menor probabilidade de sobre ou subestimar os parâmetros (viés relativo), o modelo formativo apresenta menor distância entre os valores estimados e observados (RMSD) assim como maior eficiência uma vez que as amplitudes dos intervalos de confiança em torno dos valores estimados são menores (média do erro padrão). Estas conclusões são válidas para amostras de menor dimensão, sendo que para amostras de dimensão maior não se encontram diferenças significativas entre os modelos estruturais refletivos e formativos.

Referências

Walsh G., Shiu E., Hassan L.M. (2014) Replicating, validating, and reducing the length of the consumer perceived value scale. *Journal of Business Research,* 67(3), 260-267.

Sweeney, J.C. & Soutar, G.N. (2001) Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, 77(2), 203-220.

Dados omissos em modelos de análise fatorial confirmatória não balanceados: estudo de simulação para detetar dimensão mínima da amostra

Maria de Fátima Salgueiro[1], Paula C. R. Vicente[2]
[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, fatima.salgueiro@iscte.pt;
[2] ULHT – Escola de Ciências Económicas e das Organizações e Business Research Unit (BRU-IUL), Lisboa, Portugal, p951@ulusofona.pt

Sumário: Com recurso a um estudo de simulação é investigada a dimensão mínima da amostra que permite ao investigador detetar um modelo de análise fatorial confirmatória não balanceado em termos de número de indicadores por fator latente e fiabilidade dos mesmos, em particular em modelos com dados omissos resultantes do desenho do estudo.

Palavras-chave: Análise fatorial confirmatória, *Planned missing design*, *Split questionnaire design*.

Os modelos de Análise Fatorial Confirmatória (AFC), também conhecidos por componente de medida de um modelo com equações estruturais (Bollen 1989), são muito utilizados em várias áreas científicas, designadamente nas ciências sociais e comportamentais (Salgueiro 2012). O modelo de AFC estabelece as relações entre as variáveis latentes (fatores) e com as variáveis observadas, usadas como indicadores de medida. São parâmetros de especial interesse no modelo os pesos fatoriais (*factor loadings)* e a estrutura de correlações entre variáveis latentes.

São conhecidos na literatura resultados de estudos de simulação envolvendo modelos de AFC com dados completos (quando não existem não respostas); todavia pouco se sabe sobre modelos com dados omissos (por *item nonresponse* e/ou *unit nonresponse).* O mecanismo de omissão dos dados pode ser aleatório, completamente aleatório ou não ignorável, podendo as omissões resultar do desenho do estudo. Exemplos de desenhos de estudos que conduzem à existência de omissões incluem os *split questionnaire design*, utilizados no Marketing, e os painéis rotativos, utilizados em estudos longitudinais, com o intuito de minimizar o esforço de inquirição, aumentando assim a qualidade das respostas obtidas (Enders 2010).

Neste trabalho são apresentados os resultados de um estudo de simulação realizado em Mplus 7 (Muthén & Muthén 1998-2012) com o objetivo de detetar a dimensão mínima da amostra que permite identificar modelos de análise fatorial confirmatória não balanceados, isto é, com diferente número de indicadores por fator latente e diferentes graus de fiabilidade dos indicadores. Particular ênfase é dada a modelos de AFC com dados omissos, quer devido a um mecanismo de omissão completamente aleatório, quer

devido a um desenho planeado pelo investigador *(planned missing design)*, nomeadamente em estudos longitudinais com recurso a painéis rotativos.

É considerado um modelo de AFC com dois fatores correlacionados entre si a 0,1; 0,25 e 0,5. O número de indicadores por fator pode ser de 2; 3 ou 4, sendo considerados pesos fatoriais de 0,6 e 0,8, a que correspondem níveis de fiabilidade dos indicadores de 0,36 (valor abaixo do recomendável) e 0,64 (valor considerado aceitável), respetivamente. Recorde-se que a literatura recomenda valores mínimos de 0,7 para os pesos fatoriais, com o intuito de procurar garantir uma fiabilidade de 0,5. Modelos com igual número de indicadores por fator e fiabilidade igual para todos os indicadores, considerados como *baseline*, são comparados com modelos não balanceados com o intuito de aferir sobre os seguintes tipos de efeitos i) número de indicadores por fator; e/ou ii) magnitude dos pesos fatoriais. São geradas amostras de dimensão 100, 250 e 500, tanto para modelos com dados completos como considerando omissões em dois cenários: a) omissões completamente aleatórias de 25% de cada um dos indicadores (situação de *item nonresponse*); b) omissões planeadas pelo desenho do estudo – são consideradas diferentes combinações de 25% de omissão em todos os indicadores de um dos fatores, em particular no caso de modelos não balanceados (situação de *unit nonresponse* planeada por desenho).

No caso de modelos com dados completos, os resultados obtidos indicam que, para uma mesma proporção de réplicas (amostras geradas) para a qual é detetada a correlação especificada entre fatores, são precisas amostras de maior dimensão em modelos de AFC com menor número de indicadores por variável latente e com menores níveis de fiabilidade dos indicadores, particularmente em modelos não balanceados. Quando existem omissões, os resultados obtidos vêm agravados, sobretudo no caso dos *planned missing design.*

Referências

Bollen, K. A. (1989) *Structural Equations with Latent Variables*. New Jersey, USA: John Wiley & Sons.

Enders, C. K. (2010) *Applied Missing Data,* New York, USA: The Guilford Press.

Muthén, L. K. & Muthén, B. O. (1998-2012) *Mplus Users´Guide* (7th Edition*).* Los Angeles, CA: Muthén & Muthén.

Salgueiro, M.F. (2012) *Modelos com Equações Estruturais*. Edições Sociedade Portuguesa de Estatística.

The number of clusters on trust

Cláudia Silvestre[1], Margarida G. M. S. Cardoso[2], Mário T. Figueiredo[3]
[1] Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa, csilvestre@escs.ipl.pt;
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal; margarida.cardoso@iscte.pt;
[3] Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, mario.figueiredo@lx.it.pt

Abstract: In this work we analyze the performance of a new Expectation Maximization (EM) clustering approach. This method is based on the Minimum Message Length (MML) criterion and simultaneously yields clustering of categorical data and the number of clusters. We group European citizens based on their trust in institutions, using European Social Survey data. The results obtained illustrate the parsimony, the cohesion-separation and stability of the EM-MML solutions, when compared to traditional information criteria EM based approaches.

Keywords: Categorical data, Clustering, EM-MML, European Social Survey**,** Finite mixture models.

In this work, we address the performance of a new clustering approach, named EM-MML (Silvestre *et al.* 2008), which clusters categorical data and simultaneously determines the number of clusters. This approach assumes that the data comes from a finite mixture of multinomials and uses a minimum message length (MML) criterion to estimate the number of clusters (Figueiredo and Jain 2002).

We conduct experiments on real data and compare the performance of EM-MML with several alternative approaches, namely those using information criteria – e.g. Fonseca & Cardoso (2007). The comparisons address the number of clusters obtained (parsimony), their cohesion-separation, and also their temporal stability. Cohesion-separation is quantified using the *silhouette* index. The temporal stability of clusters (comparison between clusterings referring to two different periods of time) is quantified by means of the *adjusted Rand index* (Hubert & Arabie 1985).

An application to real data from the European Social Survey (ESS) illustrates the proposed approach. This ESS dataset originates from surveys that measure attitudes, beliefs, values, and behavior patterns of European populations. EM-MML is applied to sets of questions referring to the trust in some institutions: each country's parliament, legal system, police, politicians, political parties, European Parliament, and United Nations. For the purpose of our experiment, we aggregate the ESS data by region (taking into account ESS sampling weights) and recode the variables into binary indicators: "tend to trust" vs. "tend to not trust". Data arises from the most recent surveys: 2012 (round 6, covering 30

countries and 243 regions) and 2014 (round 7, covering 15 countries and 183 regions). We conduct clustering based on 153 regions that are considered by ESS both in the 2012 and 2014 surveys. The results obtained are summarized in Table 1. Intra-cluster results show incipient correlations between the trust variables (conditional independence holds).

Table 1: Experiments on trust data: results obtained

		BIC	AIC	CAIC	AIC3	ICL	EM-MML
2012	Number of clusters	9	17	9	17	9	5
	Silhouette index	0.213	0.199	0.213	0.199	0.213	**0.285**
2014	Number of clusters	9	18	9	16	9	5
	Silhouette index	0.227	0.193	0.227	0.221	0.234	**0.296**
2012 vs 2014	Adjusted Rand	0.313	0.324	0.313	0.329	0.313	**0.411**

The number of segments selected by EM-MML is much lower than for the remaining methods, AIC and AIC3, being the least conservative. This increased parsimony avoids estimation problems associated with very small segments and also improves the interpretability of the clustering solution. According to the silhouette index, the EM-MML solutions exhibit the highest cohesion-separation. The adjusted Rand index between the solutions for 2012 and 2014 also favors the EM-MML results stability. Finally, EM-MML attains the lowest computation times.

EM-MML yields 5 segments of European citizens, generally exhibiting the same ranking on trust – police (highest trust), legal system, UN, parliament, European parliament, politicians, and political parties – with segments being distinguished according to their levels of trust.

References

Figueiredo, M. A. T. & Jain, A.K. (2002) Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 381–396.

Fonseca, J. R. & Cardoso, M. G. (2007) Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(2), 155-173.

Hubert, L. & Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 2, 193-218.

Silvestre, C., Cardoso, M. & Figueiredo, M. (2008). Clustering with finite mixture models and categorical variables. IN BRITO. P. [Editor] Preceedings of the 18-th *International Conference on Computational Statistics-COMPSTAT2008*.

How perfect is an imperfect test? A biomedical challenge

Ana Subtil[1], M. Rosário Oliveira[2], António Pacheco[3]
[1, 2, 3] CEMAT e Departamento de Matemática, Instituto Superior Técnico, Univ. Lisboa;
[1] anasubtil@tecnico.ulisboa.pt;
[2] rosario.oliveira@tecnico.ulisboa.pt;
[3] apacheco@math.tecnico.ulisboa.pt

Sumário: Dichotomous diagnostic tests accuracy measures, such as sensitivity and specificity, should be estimated by comparison with a gold standard. Alternative methods relying on imperfect tests may be used if a perfect reference is missing. We compare some of these alternatives, under a theoretical approach, based on the deviations between the accuracy measures according to each method and the corresponding true values. An R interactive graphical application is made available for visualizing the outcomes. We discuss the methods validity and potential usefulness based on our findings.

Palavras-chave: Composite reference standard, Diagnostic test, Discrepant analysis, Imperfect gold standard, Latent class model.

In the biomedical context, dichotomous diagnostic tests are widely used for the detection of a target condition such as disease or infection. There is a strong need to evaluate the performance of diagnostic tests to determine their ability to discriminate between the presence or absence of a certain target condition and thus ensure that they are properly applied and interpreted.

Two commonly used diagnostic test performance measures are the sensitivity (Se), the probability that the test result is positive given that the subject has the target condition, and the specificity (Sp), the probability that the test result is negative given that the subject does not have the target condition. Ideally, these measures would be estimated by comparison with a perfect reference test or gold standard, an error-free procedure with Se = Sp = 1.

In practical situations, however, a gold standard may be difficult or impossible to obtain and alternative approaches must be used to estimate the test's accuracy. Some of the most widespread and undemanding among these methods are the comparison with an imperfect gold standard (IGS) or with a composite reference standard (CRS), discrepant analysis (DA) and latent class models (LCM), all of which rely on available imperfect diagnostic tests to estimate the accuracy of the test under study.

These methods are widely discussed, at times controversially, in the biomedical literature, and it has been shown that different approaches may lead to different performance estimates, which can be severely biased in some cases (Hadgu 1997; Schiller *et al.* 2015; Walter *et al.* 2012).

In the present work, we aim to systematically assess and compare the potential of using these methods and to produce recommendations on the most appropriate one for a particular situation. We adopt a theoretical perspective, based on the deviations between each method's analytical expressions for the sensitivity and specificity and the corresponding true values. We investigate the impact on each method's deviations of factors such as the accuracy of the tests and the prevalence. Furthermore, we study the effect of the violation of the Hypothesis of Conditional Independence, according to which the test results are independent conditional on the true status of the condition of interest.

Based on the derived analytical expressions, we built interactive plots in the R package Shiny (Fig. 1) to show how the sensitivity and specificity deviations of the methods under comparison change with the tests characteristics, prevalence and local dependence.

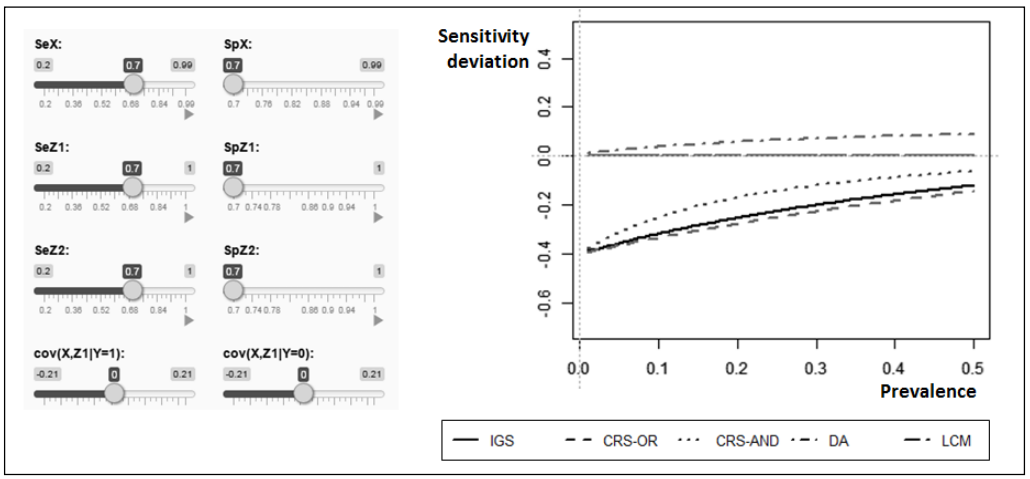


Figure 1: R package Shiny interactive plot showing the sensitivity varying with the prevalence, for the methods IGS, CRS with the "or" rule, CRS with the "and" rule, DA and LCM.

Agradecimentos: Ana Subtil tem a bolsa de doutoramento da FCT SFRH/BD/69793/2010.

Referências:

Hadgu, A. (1997) Bias in the evaluation of dna-amplification tests for detecting Chlamydia trachomatis. *Statistics in Medicine*, 16(12), 1391–1399.

RStudio, Inc. (2015) shiny: Web Application Framework for R. URL http://CRAN.R-project.org/package=shiny. R package version 0.10.1.

Schiller, I., Smeden, M., Hadgu, A., Libman, M., Reitsma, J. & Dendukuri, N. (2015) Bias due to composite reference standards in diagnostic accuracy studies. *Statistics in medicine*.

Walter, S., Macaskill, P., Lord, S. & Irwig, L. (2012) Effect of dependence errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Statistics in Medicine*, 31, 1129–1138.

Cluster-based conjoint models: An application to professional services marketing

José G. Dias[1]

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

Abstract: The paper proposes a general framework for conjoint studies that integrates multilevel modeling with experimental analysis and latent variable models. The application to the assessment of professional services of auditing identifies three benefit segments, which value different characteristics of the service being delivered. The segments can be profiled by the covariates included in the model.

Keywords: Conjoint analysis, Latent classes, Multilevel analysis, Professional services.

Conjoint analysis (CA) and discrete choice models have been frequently used to measure consumer preferences. In the last 40 years, conjoint analysis has been applied with great success, namely in new product development, competition analysis, pricing research, and benefit segmentation applications. It is the appropriate technique to assess consumer attitudes towards a product or service as it provides a measurement of the trade-offs implicit in the assessment (Green & Srinivasan 1978). Thus, CA allows product attributes to be considered jointly rather than in isolation, enabling trade-offs between attributes. Hence, it is based on the assumption that complex decisions are not based on a single factor or criterion, but on several factors assessed jointly. Lancaster's theory of utility provides theoretical support for CA as it assumes that a consumer's utility for a product can be decomposed into utilities for separate attributes or benefits provided by the product. In recent years a large strand of methodological research has increased the sophistication of conjoint analysis. In particular, the benefit segmentation purpose of conjoint analysis has further challenged new developments in latent class modeling. For instance, the multilevel structure of the data has been taken into account in latent class modeling, which allows covariance between different measurements from the same respondent (e.g., Qu *et al.* 1996). The proposed model is an extension to the conjoint analysis literature that assumes two levels of analysis: cards are clustered within respondents and, therefore, responses from the same respondent are not assumed to be independent.

The use of professional services has been increasingly investigated in marketing as a result of new challenges from different stakeholders. While research on professional services has increased in recent years, little research has been conducted on the relative important of attributes in professional services taking unobserved heterogeneity into account. For instance, Schroeder *et al.* (1986) conducted an exploratory analysis of the perceptions of the factors that influence the quality of external audits. This research takes an experimental perspective as trade-offs between attributes are included in the analysis.

Eight orthogonal stimuli plus four hold-out stimuli were assessed using the full-profile method by a sample of 161 respondents (both clients and not clients of the company) who were invited to a personal interview at the headquarters of the company. The selection of the number of latent classes was based on the BIC – Bayesian Information Criterion.

We estimated latent class models varying the number of latent classes from one to four classes. As the BIC decreases always and the solution with four latent classes has a latent class with less than 1% of the 161 respondents, the solution with three latent classes was chosen. Table 1 shows a summary of the main results: class' sizes and the relative importance of the attributes. The relative importance is a measure of an attribute's power to induce a variation of preferences. We conclude that at the aggregate level the most important attributes are Credibility (25.6%) and Accuracy (23.7%), whereas Price is the least important one (16.2%).

Table 1. Relative importance of the attributes

	Three-latent class model			Aggregate
	1	**2**	**3**	
Class sizes	0.48	0.13	0.39	1.00
Attributes				
Accuracy	0.246	0.216	0.306	0.237
Responsiveness	0.183	0.168	0.148	0.197
Price	0.168	0.189	0.111	0.162
Credibility	0.270	0.263	0.296	0.256
Brand recognition	0.133	0.163	0.139	0.148

At the segment level, results suggest heterogeneity about the relative importance of the attributes (and part-worth utilities): Segment 1 (almost half of the sample) gives an averaged importance to Accuracy and Credibility, and higher importance to Responsiveness. Segment two, the smallest (13%), is particularly sensitive to Price (18.9%) and Brand Recognition (16.3%). Finally, Segment 3 (39%) contains companies very sensitive to Accuracy (30.6%) and Credibility (29.6%), whereas Price and Responsiveness are not very important in the decision making.

References

Green, P. E., Srinivasan, V. (1978), Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research,* 5, 103-123.

Qu, Y., Tan, M., Kutner, M.H. (1996), *Random effects* models in *latent class* analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52, 797-810.

Schroeder, M., Solomon, I., Vickrey, D. (1986), Audit quality: the perception of audit-committee chairpersons and partners. *Auditing: A Journal of Practice and Theory*, 5(2), 86-94.

Análise estatística das migrações internas usando biplots composicionais

Adelaide Freitas[1], Maria Cristina Gomes[2], Maria Luís Pinto[3]
[1] Departamento de Matemática & CIDMA, Universidade de Aveiro, adelaide@ua.pt;
[2] Departamento de Ciências Sociais, Políticas e do Território, GOVCOPP, Universidade de Aveiro, mcgomes@ua.pt;
[3] Departamento de Ciências Sociais, Políticas e do Território, GOVCOPP, Universidade de Aveiro, mluispinto@ua.pt

Sumário: Consideram-se as respostas obtidas no último recenseamento sobre a mudança de residência 5 anos antes com vista a estimar os saldos migratórios internos dos 308 concelhos de Portugal e caracterizam-se os residentes envolvidos na mobilidade, usando biplots, analisando as composições por concelho, das idades, níveis de habilitação académica e situações profissionais dos imigrantes. É aplicada a transformação log-rácio centrada sobre os dados composicionais e construídos biplots tomando as duas primeiras componentes principais.

Palavras-chave: Biplot, Dados composicionais, Saldo migratórios internos, Transformação log-rácio centrada (*clr*).

Portugal não dispõe de instrumentos que permitam uma análise aprofundada dos seus movimentos migratórios internos apesar da sua importância na dinâmica populacional. De facto, a evolução populacional portuguesa foi impulsionada e condicionada por este tipo de mobilidade nomeadamente no que respeita à concentração urbana ou à litoralização, mantendo, ainda hoje, repercursões nas características e dinâmicas regionais

Uma análise de movimentos migratórios internos pode ser obtida por uma análise indireta da mudança de residência entre dois momentos distintos predefinidos. Neste trabalho, uma análise dos movimentos migratórios internos é desenvolvida com base nos dados obtidos através da pergunta do Recenseamento (2011) relativa à mudança de residência 5 anos antes do momento censitário. Perceber quem se move ajuda a compreender porque se move e permite ir mais longe na caracterização de uma realidade fluída e sobre a qual pouco se tem trabalhado, como é a das migrações internas.

Para este estudo, tendo em conta o interesse em caracterizar os residentes envolvidos na mobilidade, foram utiizados os dados dos migrantes segundo três variáveis: grupo etário, nível de habilitação e situação perante a profissão ao nível do munícipio. A recolha destes dados permitiu descrever a distribuição, em termos de frequências relativas, de cada uma daquelas três variáveis por concelho. Uma vez que a soma das frequências é unitária, tais dados induzem a uma singularidade no espaço simplex dos valores possíveis (soma das suas componentes é sempre constante; neste caso, igual a 1),

podendo ser tratados como dados composicionais. Essa singularidade nos dados implica que a aplicação de métodos estatísticos baseados em matrizes de correlações (por exemplo, análise de componentes principais, ACP) não proporciona uma solução óptima e podem conduzir a interpretações erróneas nomeadamente em biplots construídos com base na ACP sobre os dados composicionais (frequências). Novas metodologias baseadas em transformações de dados composicionais foram propostas permitindo obter biplots interpretáveis para dados composicionais (Aitchison & Greenacre 2002) e a deteção de valores atípicos com a aplicação de técnicas de ACP robustas (Filzmoser *et al.* 2009). Neste trabalho, exploram-se e analisam-se os biplots construídos sobre os dados com a transformação log-rácio centrada (*clr*, *center log-ratio*), onde cada elemento x_{ij} da matriz de dados composicionais, $X = \left[x_{ij}\right]$ foi transformado em

$$y_{ij} = \ln \frac{x_{ij}}{\sqrt[D]{x_{i1}x_{i2}\cdots x_{iD}}} = \ln x_{ij} - \frac{1}{D}\sum_{j=1}^{D} \ln x_{ij}$$

com *D* o número de partes de cada composição.

Dos biplots construídos, observa-se, por exemplo, que: i) a nível da qualificação académica dos migrantes, é evidente a existência de dois grupos de partes composicionais (habilitação até 3º ciclo e habilitação entre secundário e mestrado) com rácios mutuamente bastante estáveis, não se observando tal estabilidade entre a qualificação a nível de doutoramento e qualquer outra; ii) concelhos que registam uma maior atracção tendem a ter proporções de migrantes com níveis de qualificação superiores; e ainda iii) concelhos com saldo migratório positivo tendem a registar proporções de nível de emprego superiores.

Agradecimentos: AF parcialmente subsidiada por fundos portugueses através do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações) da Universidade de Aveiro e FCT (Fundação para a Ciência e a Tecnologia), dentro do projecto UID/MAT/04106/2013.

Referências

Aitchison, J. & Greenacre, M. (2002) Biplots of compositional data. *Appl. Statist*., 51, 375-392.

Filzmoser, P., Hron, K. & Reimann, C. (2009) Principal component analysis for compositional data with outliers. *Environmetrics*, 20, 621–632. doi: 10.1002/env.966.

Nonparametric limits of agreement for vitamin B12

Luís M. Grilo[1], Helena L. Grilo[2]
[1] *Unidade Departamental de Matemática e Física do Instituto Politécnico de Tomar, e Centro de Matemática e Aplicações (CMA), FCT-UNL, lgrilo@ipt.pt;*
[2] *Centro de Sondagens e Estudos Estatísticos, Instituto Politécnico de Tomar, helenagrilo56@gmail.com*

Abstract: The serum levels of vitamin B12 in the patients' blood is a continuous variable measured with two different clinical methods available in a Portuguese Hospital. To assess the closeness of both methods we estimate Limits of Agreement, but the assumptions of this parametric approach are in doubt. Thus, we also apply a nonparametric approach with the bootstrap resampling method to obtain robust confidence intervals for the median and for the Limits of Agreement. The conclusions seem to point out in the direction of the interchangeability of both methods.

Key-words: Bootstrap, Confidence intervals, Graphical techniques, Robustness.

In a Portuguese Hospital Laboratory the established medicine measurement method Radio Immune Analysis (RIA), with a lot of human intervention, is used to measure the continuous variable that represents the serum levels of vitamin B12 (nanograms per millilitre - ng/ml) in the blood sample of a patient. A new measurement method, Immunolite (IMM), which uses mostly machines, is also used with the aim to substitute the previous method RIA, without causing problems in clinical interpretations.

There are some interesting statistical procedures in the literature to evaluate the agreement between data produced by both measurement methods, which should reach equal clinical conclusions. In spite of the correspondent discussion about the advantages and disadvantages of each procedure (Barnhart *et al*. 2007), we decide to use the attractive and intuitive approach given by the Limits of Agreement (LoA) because medical researchers just have to compute simple statistics, to determine the 95% LoA and confidence intervals, posteriorly represented in graphs (Bland & Altman 1999; Grilo & Grilo 2012, 2016), which are easily obtained and interpreted. The parametric approach of LoA depends on some assumptions about the data, namely the mean and standard deviation of the differences of both methods (IMM–IMM) should be constant throughout the range of measurement and these differences must come from an approximately normal distribution, which is not the case in this particular study (since the *p*-value in Table 1 leads to the rejection of the hypothesis of normality, for a significance level of 5%). In fact, the empirical distribution of the variable difference is skewed and leptokurtic as well as it has some outliers that we should keep because they are possible values. Thus, the 95% LoA are also obtained after a logarithm transformation (which is a particular case of Box-Cox transformation) with the aim of normalizing the variable (Table 1). We also follow the suggestion of Bland & Altman (1999), for those cases where the

assumptions are in doubt, so 95% LoA are estimated for a regression approach and for a nonparametric approach (where we decide to apply a bootstrap resampling method in order to obtain robust 95% confidence intervals for the median of differences and for the nonparametric LoA).

Table 1. Results of the normality test for the difference: IMM - RIA.

	Kolmogorov-Smirnov[a]		
	Statistic	**df**	***p* - value**
IMM - RIA	0.212	78	0.000
Ln IMM – Ln RIA	0.091	78	0.167

a. Lilliefors Significance Correction

We analyse the agreement between both methods (RIA and IMM) with more than one statistical approach, verifying all the assumptions of each one. The application of the nonparametric approach seem to provide us a more accurate view of what really happens across the entire population. Despite some lack of agreement, maybe clinicians could consider the interchangeability of both measurement methods or even replace the reference method RIA by the new method IMM, which depends on their clinical purposes.

Acknowledgments: This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

Barnhart, H. X., Haber, M. J. & Lin, L. I. (2007) An overview on assessing agreement with continuous measurements, *Journal of Biopharmaceutical Statistics*, 17(4), 529-69.

Bland, J. & Altman, D. (1999) Measuring agreement in method comparison Studies. *Statistical Methods in Medical Research*, 8(2), 135-160.

Grilo, L. M. & Grilo, H. L. (2012) Comparison of clinical data based on limits of agreement. *Biometrical Letters*, 49, 1, 45-56.

Grilo, L. M. & Grilo, H. L. (2016) Robust statistical approaches to assess the degree of agreement of clinical data. *ICNAAM 2015*, *AIP Conf. Proc.* (in print).

Robust confidence intervals using minimum-distance method

Teresa Risso[1], Conceição Amado[2], Ana M. Pires[2]
[1] CEMAT – Instituto Superior Técnico, teresa.risso@tecnico.ulisboa.pt;
[2] CEMAT – Instituto Superior Técnico, conceicao.amado@tecnico.ulisboa.pt;
[3] CEMAT – Instituto Superior Técnico, apires@math.ist.utl.pt

Abstract: The purpose of this work is to provide a robust method for interval estimation based on minimum distance estimators. The usefulness of this new procedure is illustrated by several applications to parameter estimation in contaminated continuous and discrete distributions.

Keywords: Confidence intervals, Minimum distance method, Robust statistics.

Minimum distance estimation, introduced by Wolfowitz (1953), is a method for estimating parameters of a statistical model based on a distance measure. The main idea is to search for a distribution function that is close enough to the empirical distribution function of the data in terms of the distance under consideration. Several distances may be considered once the method is not restricted to any particular measure. One of the most popular ones are the Kolmogorov-Smirnov distance, the Pearson's Chi-square and the Hellinger distance.

Wolfowitz (1957) proved, under general conditions, the almost certain consistency of the minimum distance estimators. The method is found to be efficient in models contaminated with outliers (e.g., Parr & Schucany 1980). Hall & Wang (1999) described a method to estimating the end-point of a probability distribution using minimum-distance methods and, in this particular situation, they suggested exact interval estimator of that end-point.

In this work, we propose a new procedure to construct robust confidence intervals based on the minimum-distance approach. In order to study the properties of the method, a case study was carry out for estimating the parameters of some common distributions, namely, exponential, normal, symmetric and skewed beta. The Kolmogorov-Smirnov distance is used as a measure of the distance between the empirical distribution function and the model and Nelder and Mead optimization method was used to find the minimum. A review of methods for constructing exact, bootstrap and large sample intervals, is provided as a framework for comparison. All computations are performed in R software.

For the referred distributions, several sample sizes and parameters values where tested. As it can be seen in Figure 1, for a sample (size 100) from Normal distribution with location parameter 1 and scale 0.5, with 10% of contamination, the proposed confidence intervals contain the real values of both parameters, unlike usual confidence intervals.

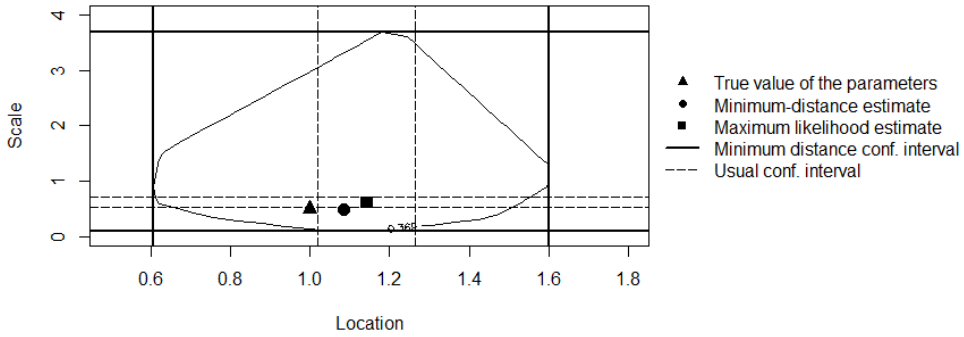


Figure 1: Contour level of the distance for N(1,0.5), n=100, with 10% of contamination, for a significance level of 10%.

In the case of an Exponential distribution with parameter equal to 0.5, and a sample size n=100 with 10% of contamination, it can be observed in Figure 2 that the proposed confidence interval also contains the true value of the parameter, unlike exact confidence intervals.

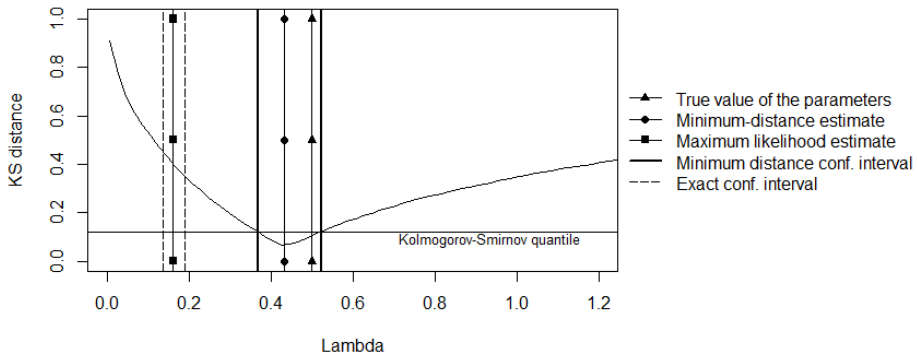


Figure 2: Distance for Exp(0.5), n=100, with 10% of contamination. KS quantile with a significance level of 10%.

A new procedure to construct confidence intervals is proposed, based on the minimum-distance method. For the considered distributions, the new confidence intervals have been shown a good behavior in presence of outliers and in many cases, they are also the only ones containing the real values of the parameters.

References

Hall, P. & Wang, J. Z. (1999) Estimating the end-point of a probability distribution using minimum-distance methods. *Bernoulli*, 5(1), 177-189.

Parr, W. C. & Schucany, W. R. (1980) Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75, 616-624.

Wolfowitz, J. (1953) Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics,* 5(1), 9-23.

Wolfowitz, J. (1957) The minimum distance method. *The Annals of Mathematical Statistics*, 28, 75-84.

Avaliação do efeito do desenho de amostragem em modelos de regressão logística

Ana Laura Carreiras[1], Paulo Infante[2], Anabela Afonso[3], Maria Filomena Mendes[4]
[1] Mestrado em Modelação Estatística e Análise de Dados, 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;
[4] CIDEHUS e DSOC/ECS, mmendes@uevora.pt

Sumário: Uma das estratégias mais usadas para corrigir as estimativas obtidas com base no pressuposto de uma amostra complexa ser proveniente de um esquema de amostragem aleatória simples é considerar os pesos normalizados corrigidos pelo efeito do desenho (*deff*). Neste trabalho, analisa-se o impacto de uma estimação incorreta do *deff* na significância das variáveis e no seu efeito em modelos de regressão logística.

Palavras-chave: Amostras complexas, Efeito do desenho, Pesos, Regressão logística.

As amostras complexas tendem a resultar da combinação de vários métodos de amostragem para a seleção de uma amostra representativa da população. Uma amostra complexa possui 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). Alguns autores ainda tratam este tipo de amostras sob a suposição de amostragem aleatória simples, ignorando o desenho de amostragem. Esta abordagem pode produzir incorreções, tanto para as estimativas, como para as respetivas variâncias, comprometendo os resultados e as conclusões da pesquisa (Osborne 2011).

Algumas vezes, por motivos de confidencialidade estas variáveis associadas ao desenho não são facultadas, mas em contrapartida poderão ser fornecidos os pesos de replicação (Sturgis 2004). Contudo, nem todas as ferramentas estatísticas que se pretendem utilizar estão implementadas em *software*, não sendo por isso possível utilizar os pesos de replicação no processo de inferência e modelação.

Existem várias estratégias para corrigir as estimativas obtidas com base no pressuposto da amostra ser proveniente de um esquema de amostragem aleatória simples (Osborne 2011). A mais usual é considerar os pesos normalizados corrigidos pelo efeito do desenho (*deff*), que produz erros padrão muito próximos dos que se obteriam considerando as variáveis de desenho. Este quantifica a perda de precisão na estimação devido ao desenho complexo e é definido pelo quociente entre a estimativa da variância determinada pelo plano amostral complexo e a estimativa da variância obtida por uma amostra aleatória simples do mesmo tamanho. Assim, o *deff* dependerá do número de observações e dos pesos de amostragem. Contudo, também nem sempre é fornecido o

efeito do desenho e este, usualmente, varia consoante a variável em estudo (Sturgis 2004).

Neste trabalho realizam-se análises de sensibilidade aos valores do efeito do desenho que são usados para corrigir os pesos, analisando o impacto de uma estimação incorreta deste efeito na significância das variáveis e no seu efeito.

Para tal, a partir dos dados do Inquérito à Fecundidade 2013, realizado no âmbito de um protocolo entre a Fundação Francisco Manuel dos Santos e o Instituto Nacional de Estatística, selecionou-se um conjunto de variáveis com características sociodemográficas dos indivíduos e, a partir destas, ajustaram-se modelos de regressão logística (Hosmer *et al.* 2013), usando como resposta a variável dicotómica tem filhos *vs.* não tem filhos, considerando diferentes estimativas do *deff*. Para avaliar o efeito da dimensão da amostra repetiu-se o mesmo procedimento, mas considerando a subpopulação dos indivíduos que não têm filhos, tomando como resposta a variável dicotómica quer ter filhos *vs.* não quer ter filhos.

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

Referências

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013) *Applied Logistic Regression*. 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.* Issue 43, Autumn 2004.

A decisão de permanecer sem filhos a partir dos 30 anos de idade

Andréia Maciel[1], Rita Brazão Freitas[2], Maria Filomena Mendes[3], Paulo Infante[4]
[1] Universidade de Évora, CIDEHUS.UÉ, deiabarroso@hotmail.com;
[2] Universidade de Évora, CIDEHUS.UÉ, ritabf8@gmail.com;
[3] Universidade de Évora, CIDEHUS.UÉ, mmendes@uevora.pt;
[4] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt

Sumário: Atualmente vive-se uma situação de muito baixa fecundidade na sociedade portuguesa, tonando-se fundamental identificar fatores que influenciam de forma significativa a decisão definitiva de não ter filhos. Neste trabalho pretende-se encontrar determinantes dos residentes em Portugal permanecerem sem filhos depois dos 30 anos. Procura também estender-se esta análise aos países do Sul da Europa (Espanha, Grécia e Itália).

Palavras-chave: Demografia, Fecundidade, Regressão logística.

Argumenta-se que a satisfação da parentalidade tem deixado de ser uma condição básica para se alcançar a autorrealização e que a opção por uma vida sem filhos (*childfree*) se tem tornado uma escolha cada vez mais comum e livre de estereótipos (Basten 2009; Tanturri & Mencarini 2008), fazendo com que a dimensão familiar desejada e, considerada ideal, seja um dos mais importantes determinantes da fecundidade futura. De igual forma, questiona-se a importância da educação e da condição perante o trabalho na opção por uma vida sem filhos.

Considerando-se que o atual declínio da fecundidade é, em grande parte, o resultado do adiamento da entrada na parentalidade (fenómeno que tem vindo a crescer ao longo dos últimos anos), olhamos para aqueles indivíduos que atingiram os 30 ou mais anos de idade, sem terem tido filhos, e procuramos traçar o perfil mais provável daqueles que deverão continuar sem filhos.

Dos residentes em Portugal à data do Inquérito à Fecundidade (IFEC2013), estima-se que 38,5 % dos indivíduos ainda não tinham filhos biológicos. Excluindo aqueles (1,1 %) que apesar de não terem filhos ainda não têm definida a sua decisão reprodutiva (indecisos, não sabem), 8,3 % dos residentes em Portugal esperam permanecer sem filhos no termo da sua vida reprodutiva (*childlessness* permanente*)*, enquanto 29,2 % esperam ainda vir a ter filhos (*childlessness* temporário). A grande maioria (75 %) dos indivíduos que ainda não tendo filhos pretende vir a tê-los tem menos de 30 anos de idade. De entre os indivíduos com idades entre os 30 e os 39 anos, quase um quarto (23 %) não pretende vir a ter filhos.

Com o objetivo de analisar e quantificar o efeito das características mais relevantes para os residentes em Portugal e nos países do Sul da Europa, com 30 ou mais anos,

permanecerem sem filhos, foram ajustados modelos de regressão logística, de acordo com a metodologia proposta em Hosmer *et al*. (2013), a partir dos dados do Eurobarómetro 2011 e do IFEC2013. Para estes modelos consideramos como variável resposta: 0 - *childlessness* temporário; 1 - *childlessness* permanente. Ambos os modelos mostraram boa capacidade discriminativa.

Tanto em Portugal como nos demais países da Europa do Sul, concluímos que o aumento da idade, a ausência de um companheiro e uma baixa dimensão familiar considerada ideal são os determinantes mais importantes na decisão de permanecer sem filhos. Adicionalmente, na análise para a Europa do Sul, concluímos que aqueles que residem em cidades grandes também são mais prováveis de não experimentar a parentalidade. Em Portugal verifica-se ainda que os indivíduos com baixos níveis de escolaridade tendem a fazer a sua transição mais precocemente, contudo, quando atingem os 30 anos de idade sem o ter feito, acabam por se tornar mais suscetíveis de continuar sem filhos.

Ao nível dos países, os gregos revelam menores possibilidades de permanecerem *childlessness* relativamente aos portugueses, espanhóis e italianos. Em Portugal, contribuem ainda para a decisão de continuar *childlessness* o facto dos indivíduos acharem que a parentalidade não é condição básica para se alcançar a autorrealização e, para os homens, o facto de não terem um trabalho a tempo inteiro.

Referências

Basten, S. (2009) *Voluntary childlessness and being Childfree. The Future of Human Reproduction*. Working Paper, No. 5.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013) *Applied Logistic Regression* (3rd edition). New Jersey: John Wiley & Sons.

INE (2013) *Inquérito à fecundidade.* Documento metodológico Versão 1.0, Lisboa, Instituto Nacional de Estatística.

Tanturri, M. L. & Mencarini. L. (2008) Childless or childfree? Paths to voluntary childlessness in Italy. *Population and development review*, 34(1), 51-77.

Sessão Paralela IV – Regressão Logística – sábado, 2 de abril, Sala 124 (16:50)

A ecografia como instrumento de diagnóstico – um estudo de caso

Ana Matos[1], Carla Henriques[2], Jorge Pereira[3], A. C. Afonso[3], J. Constantino[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] Serviço de Cirurgia 1, Centro Hospitalar Tondela-Viseu, docjota@netcabo.pt

Sumário: Equipas interdisciplinares estão na base do sucesso de muitos trabalhos de investigação. Um exemplo disso é a colaboração da estatística com as ciências da saúde. Neste trabalho, recorrendo a testes não paramétricos e a modelos de regressão logística, foi possível contribuir para o conhecimento da não acuidade, da ecografia no diagnóstico de colecistite aguda na presença de pancreatite aguda. Deste modo, este estudo contribui para um melhor diagnóstico clínico que será determinante na terapêutica a adotar.

Palavras-chave: Acuidade Ecográfica, Pancreatite Aguda, Regressão Logística.

A litíase vesicular sintomática (vulgarmente conhecida por Pedras na Vesícula) encontra-se entre as doenças mais vezes tratadas pelos cirurgiões, sendo a colecistite e a pancreatite as formas de apresentação aguda mais frequentes. A ocorrência simultânea de pancreatite e colecistite agudas é possível e, algumas vezes, a gravidade da colecistite sobrepõe-se à da pancreatite.

O diagnóstico de colecistite num contexto de pancreatite aguda assume crucial importância na medida em que influenciará determinantemente a terapêutica.

Clinicamente, a colecistite aguda partilha muitos dos sinais e sintomas da pancreatite aguda. A nível ecográfico a colecistite aguda é definida por alterações específicas que apresentam de forma isolada valores de sensibilidade e especificidade superiores a 90% (Yokoe *et al.* 2012). Seria pois de esperar que a ecografia fosse um instrumento de grande utilidade no diagnóstico de colecistite aguda, mesmo na presença de pancreatite aguda.

O estudo envolve 120 doentes com diagnóstico de pancreatite aguda que, entre 1998 e 2015, no Centro Hospitalar Tondela Viseu, foram sujeitos à retirada cirúrgica da vesícula biliar – colecistectomia. O exame histológico à vesícula removida durante a colecistectomia permitiu identificar os pacientes que verdadeiramente apresentavam colecistite aguda.

Neste trabalho comparam-se os resultados ecográficos de colecistite aguda com os resultados histológicos. A associação dos resultados foi analisada através do Teste do Qui-

quadrado. O estudo da associação entre o resultado ecográfico e as variáveis em estudo foi complementado com recurso à regressão logística.

Classificaram-se os 120 pacientes de pancreatite aguda em dois grupos: no primeiro grupo englobaram-se os pacientes que, de acordo com o diagnóstico ecográfico, não têm colecistite aguda (77 pacientes) e no segundo grupo aqueles que, de acordo com o diagnóstico ecográfico, apresentam colecistite aguda associada (43 pacientes).

Das variáveis em análise (género, comorbilidade, idade, tempo de internamento, tempo de espera entre a ecografia e a intervenção cirúrgica) apenas a idade se revelou significativamente diferente nos dois grupos, sendo maior no grupo dos pacientes com resultado ecográfico positivo para colecistite aguda.

Na amostra global, 36 doentes (30%) apresentaram exame histológico com resultado de colecistite aguda. Comparando os dois grupos em estudo, foram encontrados resultados semelhantes, com uma taxa de diagnóstico histológico de colecistite aguda de 31,2% no grupo 1 e 27,9% no grupo 2. Estes resultados revelam não haver relação estatisticamente significativa entre os dados da ecografia e os do exame histológico ($p=0,708$). A sensibilidade e a especificidade da ecografia para o diagnóstico de colecistite aguda são de 33,3% e 63,1% respetivamente.

Os fatores que se relacionam com o resultado ecográfico de colecistite aguda foram também investigados através de uma análise multivariada recorrendo à regressão logística. Verificou-se um efeito marginalmente significativo da idade ($p=0,067$) no resultado ecográfico de colecistite aguda, havendo uma tendência para mais resultados positivos em pacientes com mais idade. Mais, confirmou-se a ausência de relação estatisticamente significativa entre o tempo de espera, bem como do resultado histológico, com o resultado ecográfico. O modelo de regressão logística mostrou, pois, que o resultado ecográfico não ajuda a prever a existência ou a não existência de colecistite aguda, comprovada pelo exame histológico. Na presença de pancreatite aguda, a ecografia não tem acuidade suficiente diagnosticar colecistite aguda.

Referências

Yokoe, M, Takada, T, Strasberg, S, Solomkin, J, Mayumi, T, Gomi, H, *et al.* (2012) New diagnostic criteria and severity assessment of acute cholecystitis in revised Tokyo guidelines. *J Hepato-Biliary-Pancreat Sci.*, 19(5), 578–85.

Pereira, J., Afonso, A. C., Constantino, J., Matos, A., Henriques, C., Zago, M. & Pinheiro, L. (2015) Accuracy of ultrasound in the diagnosis of acute cholecystitis with coexistent acute pancreatitis. *European Journal of Trauma and Emergency Surgery*.

Sessão Paralela IV – Gestão da Informação e Redes Neuronais – sábado, 2 de abril, Anfiteatro 131 (16:10)

A Cauda Longa: A sua existência no mercado de retalho Online Português

Juliana Rocha Costa[1]
[1] *Faculdade de Economia do Porto, julianarochacosta@gmail.com*

Sumário: O crescimento dos mercados Online tem sido exponencial. A Internet permitiu o aparecimento do fenómeno da Cauda Longa, que vem pôr em causa a Teoria de Pareto, também conhecida como a Regra dos 80/20.
Com o objetivo de averiguar a existência e a tendência deste fenómeno, bem como, o custo de expansão da gama de artigos, no mercado de retalho Online Português, foi usada uma Base de Dados de uma grande empresa do retalho português. Recorreu-se a técnicas de modelização estatística de forma a verificar se o fenómeno segue a tendência Internacional.

Palavras-chave: Bens, Cauda longa, Curva de Pareto, Nicho, Online.

A teoria da Cauda Longa foi tornada célebre por Chris Anderson com a publicação do seu livro A Cauda Longa – Por que é que o futuro dos negócios é vender menos de mais produtos. Esta nova teoria que considera que a venda de poucas quantidades mas de muitos artigos, mais que compensa a venda dos artigos mais populares, vem pôr em causa a célebre Teoria de Pareto, que considera que 80% das vendas são obtidas com 20% dos produtos.

O comércio Online em Portugal tem registado um grande crescimento, de acordo com o estudo realizado pela SIBS e Datamonitor (2015). Em 2014, as compras na Internet em Portugal cresceram 1,5 mil milhões de euros face a 2009, em contrapartida, as vendas registadas em Lojas Físicas (lojas convencionais) apresentaram uma queda na ordem dos 2%. A empresa em análise tem mantido o número de clientes nas Lojas Físicas, enquanto que a Loja Online apresenta um crescimento médio mensal de 4%.

Os principais objetivos deste estudo são investigar a existência do fenómeno da Cauda Longa no mercado de retalho Online Português, e como se tem vindo a comportar. O estudo propõe-se a responder a duas questões "Será a Cauda Longa uma tendência na Loja Online Portuguesa em estudo?" e "Qual o benefício do alargamento da gama de artigos nas Lojas Online de venda de Bens Físicos?". Para além disso, este estudo propôs-se a investigar o benefício versus o custo de fazer um alargamento da gama de artigos disponíveis na loja Online. A metodologia de investigação utilizada foi uma análise quantitativa aplicada a um estudo de caso. Foi escolhida uma empresa líder deste sector, obtendo a informação sobre 3 Lojas Físicas e a Loja Online em dois períodos para 3 cabazes de artigos.

Para que a teoria da Cauda Longa se verifique é necessário que o gráfico da equação (1) em escala logarítmica contra a sua distribuição empírica apresente, aproximadamente, uma linha reta. Verificando-se, então a variável segue uma Power Law,

$$p(x) = Cx^{-\alpha}, \quad (1)$$

onde $p(x)$ é a função (densidade) de probabilidade da variável *X*. Neste caso, também a função cumulativa $P(X) = P(X \geq x)$ será uma Power Law. A distribuição cumulativa de uma variável é uma Power Law, se tem a expressão acima, o que significa que à medida que se aumenta o número de artigos disponíveis, as vendas acumuladas em escala logarítmica aumentam em linha reta. O α é uma constante, chamado o expoente da Power Law. O C é uma constante fixa que é determinada de forma que a distribuição de $p(x)$ tenha soma 1. Para averiguar sobre a existência da Power Law, no caso estudado, foi efetuada uma análise gráfica, fazendo o Diagrama de Pareto, que por definição é uma Power Law.

Para uma análise mais aprofundada sobre a existência de diferenças entre a concentração das vendas e, o respetivo crescimento das Caudas nas Lojas, foi estimada e testada a Curva de Pareto devidamente adaptada conforme equação (2):

$$\ln(vendas_i) = \beta_0 + \beta_1 \ln(rankvendas_i) + \varepsilon_i \quad (2)$$

em que $vendas_i$ representa as vendas do artigo *i*, e $rankvendas_i$ é o ranking de vendas de cada artigo vendido, numerado de 1 a *n*. β_1 mede a rapidez com que as vendas de *i* decrescem com o aumento do ranking. ε_i representa o erro aleatório para o artigo *i*. Estudos anteriores demonstram que a relação entre as vendas e o ranking de vendas são um bom ajustamento para a distribuição das vendas. Para que uma Loja tenha uma Cauda Longa superior a outra Loja necessita que o β_1 apresente o valor menos negativo (menor, em valor absoluto).

A Loja Online em análise apresenta uma Cauda com uma reduzida expressão quando comparada com a Amazon, que é a grande referência Internacional. Podemos concluir que considerando somente os cabazes de artigos constituídos por Bens Físicos se verifica que na Loja Online existe uma tendência de crescimento do comprimento da Cauda. Nas Lojas Físicas esta tendência não é linear, depende da Loja e depende do cabaz de artigos considerados. Estima-se que o benefício de alargar a gama de artigos disponíveis na Loja Online é compensador para a empresa, permitindo uma continuidade do aumento das vendas, verificando-se nos períodos em análise.

Referências

Anderson, C. (2006) *A Cauda Longa – Porque que é que o futuro dos negócios é vender menos de mais produtos.* (C. Pedro, Trad.) Lisboa: Actual Editora.

Brynjolfsson, E., Hu, Y. & Simester, D. (2011) Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the concentration of Product Sales. *Working paper, MIT Sloan School of Management,* Vol. 57, Nº 8, 1373-1386.

Sessão Paralela IV – Gestão da Informação e Redes Neuronais – sábado, 2 de abril, Anfiteatro 131 (16:30)

Geração sintética de microdados utilizando algoritmos de *data mining*

Daniel Silva[1], Pedro Campos[2], Pavel Brazdil[3]
[1] FEP-UP, 199804052@fep.up.pt;
[2] LIAAD/INESC TEC, FEP-UP, pcampos@fep.up.pt;
[3] LIAAD/INESC TEC, FEP-UP, pbrazdil@fep.up.pt

Sumário: Este trabalho utiliza dois algoritmos de *data mining* - Árvores de Decisão e *Random Forests* - para obter ficheiros totalmente sintéticos considerando uma base de dados de empresas com problemas de confidencialidade. Foram utilizadas duas operacionalizações distintas, ascendendo a quatro o número de métodos testados. Os resultados permitem concluir que os métodos que utilizam algoritmos de Árvores de Decisão reproduzem melhor as estatísticas univariadas e o que utiliza o algoritmo *Random Forests*, onde são consideradas todas as variáveis no modelo de imputação, retrata de forma mais consistente as estatísticas multivariadas.

Palavras-chave: Árvores de Decisão, Confidencialidade, *Data mining*, *Random Forests.*

Grande parte da informação obtida pelos detentores da informação é através da garantia da confidencialidade dos respondentes, que se encontram legalmente protegidos de modo a que a sua identidade não seja revelada. No entanto, a questão legal não é a única importante quando se fala em privacidade. Os respondentes que sentirem que a sua privacidade está em risco vão ser relutantes em fornecer informação sensível, podem fornecer informações incorretas ou até mesmo negarem-se a responder, com consequências devastadoras para a qualidade da informação.

Se por um lado o acesso à informação reveste características de bem público contribuindo para o desenvolvimento económico e social, por outro as questões relacionadas com a confidencialidade nunca foram tão pertinentes como atualmente.

Na literatura sobre o tratamento do segredo estatístico têm sido propostas diferentes técnicas de proteção e divulgação da informação. Estas técnicas assentam na ideia que a possibilidade de identificar os respondentes pode ser neutralizada através da redução da quantidade de informação facultada, mascarando os dados (por exemplo não facultar a informação na sua totalidade ou perturbando valores), ou pela divulgação de bases de dados sintéticas em que os valores da base de dados original são substituídos.

A Imputação Múltipla (Rubin 1993) é um dos métodos mais aplicados na geração sintética de informação devido à forte fundamentação teórica e aos bons resultados obtidos em diferentes trabalhos. Os métodos de imputação são divididos em abordagens paramétricas e não paramétricas. Apesar de alguns autores aplicarem com sucesso abordagens paramétricas, estes métodos implicam um conhecimento profundo da relação entre as variáveis e a sua implementação decorre de modelos sofisticados e tempos de

execução elevados. Já as abordagens não paramétricas conseguem lidar com diferentes tipos de informação de elevada dimensionalidade, obter relações não lineares e iterações que muito dificilmente são captadas pelas abordagens paramétricas (Drechsler & Reiter 2010; Raab *et al.* 2015).

Este estudo utiliza dois algoritmos de *data mining* (Árvores de Decisão e *Random Forests*) para gerar ficheiros totalmente sintéticos considerando uma base de dados de empresas composta por atributos qualitativos e quantitativos. Para a prossecução deste objetivo foram utilizadas duas operacionalizações distintas. Embora em ambas a geração das variáveis seja efetuada de forma sequencial, na primeira operacionalização são utilizadas todas as variáveis originais e previamente sintetizadas como previsores no modelo, na segunda apenas são considerados os atributos sintetizados em passos anteriores. Assim, no âmbito deste trabalho, foram testados quatro métodos diferentes (dois algoritmos e duas operacionalizações).

Os resultados permitem concluir que a utilização de algoritmos de *data mining* para a produção de ficheiros sintéticos é uma alternativa aos métodos paramétricos, uma vez que os resultados obtidos considerando o ficheiro original e os ficheiros sintéticos são semelhantes. Na avaliação da qualidade dos ficheiros gerados foi utilizada a sobreposição dos intervalos de confiança do ficheiro original com os ficheiros sintéticos (Karr *et al*. 2006). Em termos de algoritmos, constatou-se que o algoritmo de Árvores de Decisão consegue reproduzir melhor as estatísticas univariadas e o algoritmo *Random Forests*, considerando sempre todos os atributos no modelo de imputação, a relação entre as variáveis (estatísticas multivariadas).

A utilização de uma base de dados empresas, a geração de ficheiros totalmente sintéticos e a identificação e aplicação de quatro métodos são os elementos inovadores deste trabalho.

Referências

Drechsler, J. & Reiter, J. P. (2010) Sampling with synthesis: A new approach for releasing public use censos microdata. *Journal of the American Statistical Association*, 105(492), 1347-1357

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. & Sanil, A. P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3), 224-232.

Raab, Gillian, Beata Nowok & Chris Dibben (2015) *A simplified approach to generating synthetic data for disclosure control*. arXiv preprint arXiv:1409.0217.

Rubin, D. B. (1993) Statistical disclosure limitation. *Journal of official Statistics*, 9(2), 461-468.

Classificação acústica automática de espécies de morcegos

Bruno Silva[1], Gonçalo Jacinto[2], Paulo Infante[3], Sílvia Barreiro[4], Pedro Alves[5]
[1] CIBIO-UE e UBC da Universidade de Évora, bsilva@uevora.pt;
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, gjcj@uevora.pt;
[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt;
[4] PLECOTUS - Estudos Ambientais, sbarreiro@plecotus.com;
[5] PLECOTUS - Estudos Ambientais, pjalves@plecotus.com

Sumário: As estações de gravação automática de ultrassons de morcegos permitem recolher uma enorme quantidade de dados, não sendo viável a sua análise manual. Propomos neste trabalho um método totalmente automático de analisar e classificar este tipo de dados, através da combinação de um algoritmo para a deteção e medição do ultrassom e de um conjunto de redes neuronais artificiais para a sua classificação.

Palavras-chave: Bioacústica, Classificação automática, Morcegos, Redes neuronais artificiais, Ultrassons.

Os morcegos são os únicos mamíferos com capacidade de voo sustentado e muitas das espécies possuem um sistema de ecolocalização bastante avançado. Os seus hábitos predominantemente noturnos associados à ecolocalização através de ultrassom, tornam a sua identificação em voo bastante difícil e só através de métodos indiretos, principalmente estações de gravação automática de ultrassons, é possível estudar estes mamíferos. As gravações obtidas são posteriormente analisadas, maioritariamente de forma manual ou semi-manual, e por comparação com parâmetros conhecidos das várias espécies, são classificadas. Este processo é bastante moroso e existe muita incerteza e subjetividade já que as diferentes espécies exibem uma grande plasticidade e variabilidade nos parâmetros da ecolocalização normalmente usados para a classificação.

O desenvolvimento de um sistema completamente automático para processar milhares de gravações num curto espaço de tempo, não só tornará o processo mais objetivo como permitirá reduzir drasticamente o tempo de análise e os custos a ela associados, tornando viável monitorizações contínuas de populações de morcegos, com inegáveis vantagens ao nível da conservação destes.

Foram utilizadas 749 gravações de morcegos que foram capturados, identificados morfologicamente até à espécie e só depois gravados, permitindo desta forma a compilação de uma base de dados de referência de elevada fiabilidade contendo 2968 vocalizações das várias espécies presentes em Portugal continental. A análise foi implementada no R (R Core Team 2013). Cada gravação foi alvo de um filtro de corte abaixo dos 10kHz para eliminar insetos e ruído ambiental e o sinal de interesse (vocalização de morcegos) foi identificado na gravação através da intensidade. Depois de identificada a vocalização, o início e o final da mesma foram localizados. Este segmento da

gravação foi extraído e analisado com a transformada de Fourier. No total foram medidos 19 parâmetros nos domínios espectral e temporal. Para a classificação foi utilizado um conjunto 5 de redes neuronais artificiais da classe perceptrão multicamadas sem realimentação.

As redes foram desenvolvidas com uma camada escondida e com função de ativação logística na camada escondida e na camada de saída. Devido a limitações na amostra, as classificações foram efetuadas apenas para grupos de espécies. Foram definidos 5 grupos de espécies com semelhanças ecológicas. Do total das 2968 vocalizações, 2219 (70%) foram utilizadas para o treino das redes e 749 (30%) para avaliação da capacidade de generalização das mesmas. Os conjuntos de redes com melhor performance utilizaram 20 neurónios na camada escondida e apresentam um total de classificações corretas de 98.3%. A sensibilidade aos vários grupos variou entre 99.0% e 92.9%, com erros de classificação entre 0.2% e 10.3% (Tabela 1).

Tabela 1: Resultado da classificação do conjunto de redes neuronais para os 5 grupos. A taxa global de classificações corretas está assinalada a negrito. Sens - Sensibilidade; Erro – Taxa de erro.

		Observados						**Resultados**	
		Myo	Pip	Nyc	Bbar	Plec	Total est	Sens (%)	Erro (%)
Estimados	Myo	115	1	1	0	1	118	98.3	2.5
	Pip	1	473	0	0	0	474	99.0	0.2
	Nyc	1	0	66	0	1	68	97.1	2.9
	Bbar	0	4	0	56	0	60	96.6	6.7
	Plec	0	0	1	2	26	29	92.9	10.3
	Total obs	117	478	68	58	28	749	**98.3**	

Agradecimentos: Bat Conservation International, Inc e PLECOTUS – Estudos Ambientais Unip, Lda

Referências

Anastasiadis, A., Magoulas, G. & Vrahatis, M. (2005) New globally convergente training scheme based on the resilient propagation algorithm. *Neurocomputing*, 64, 253-270.

Walters, C. *et al.* (2012) A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, 49(5), 1064-1074.

Redgwell, R., Szewczak, J., Jones, G. &, Parsons, S. (2009) Classification of echolocation calls from 14 species of bat by support vector machines and ensembles of neural networks. *Algorithms*, 2(3), 907-924.

R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

SESSÕES DE POSTERS

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

O impacto da cultura organizacional no desempenho financeiro das empresas da região Norte de Portugal

Flávia Araújo[1], Conceição Castro[2], Fernanda A. Ferreira[3]
[1] ESEIG.IPP, flaviarc.araujo@gmail.com;
[2] ESEIG.IPP, Centro de Estudos da População, Economia e Sociedade (CEPESE), conceicaocastro@eseig.ipp.pt;
[3] ESEIG.IPP, Management Applied Research Unit (UNIAG), fernandaamelia@eseig.ipp.pt

Sumário: Este trabalho tem por objetivo estudar se a cultura organizacional influencia o desempenho financeiro das organizações, bem como o impacto de acordo com a cultura predominante. Em termos metodológicos, através de um inquérito por questionário identifica-se o tipo de cultura predominante nas empresas da região Norte de Portugal e recorre-se à análise de correlação e de regressão múltipla. Os resultados sugerem a existência de relações de pequena intensidade entre os tipos de cultura analisados e o desempenho financeiro das empresas daquela região.

Palavras-chave: Cultura organizacional; Desempenho financeiro; Regressão linear múltipla.

O conceito de cultura organizacional tem sido muito debatido na literatura económica. O objetivo deste artigo é o de avaliar o impacto da cultura organizacional no desempenho financeiro das empresas da região Norte, a mais industrializada de Portugal, onde predominam as microempresas. O presente estudo é baseado num modelo em que a cultura organizacional é a variável independente e o desempenho financeiro é a variável dependente. Em consonância com o modelo temos a seguinte hipótese: A cultura organizacional influencia o desempenho financeiro das empresas.

Para avaliar a hipótese definida, foram selecionadas aleatoriamente 380 empresas da região Norte de Portugal, às quais foi enviado um inquérito por questionário via *Lime Survey*. Obtiveram-se 127 respostas, das quais 66 foram consideradas válidas. A população alvo foram os gestores de topo. Aplicou-se, numa primeira fase, o inquérito por questionário desenvolvido por Cameron & Quinn (2006), para identificar o tipo de cultura (clã, adocrática, mercado e hierárquica) predominante nas empresas desta região. Posteriormente, efetuaram-se testes de correlação entre a cultura organizacional e o desempenho financeiro e, para averiguar a hipótese da cultura influenciar o desempenho financeiro, a especificação do modelo recai, tal como Yesil & Kaya (2013), no modelo de regressão linear múltipla. O tratamento de dados foi realizado com a ferramenta estatística SPSS.

Por uma questão de ajustamento dos dados foi usado o modelo *Log-linear*. O modelo de regressão múltipla pode ser representado pela eq.: $\log(DF) = \beta_0 + \sum_{i=1}^{4} \beta_i C_i + \sum_{i=1}^{7} \beta_i Z_i + \varepsilon$. A variável dependente desempenho financeiro (DF) é avaliada pelo Resultado Líquido (RL). As variáveis independentes C_{i} são variáveis qualitativas ordinais (avaliadas

por uma escala de resposta 5 pontos) e representam cada um dos tipos de cultura organizacional (clã, adocrática, mercado e hierárquica). z_i representam as variáveis de controlo (ramo de atividade económica, forma jurídica e número de colaboradores). Para verificação da hipótese do estudo aplicou-se o modelo de regressão linear múltipla. Foram testadas várias interações entre as variáveis e validados os pressupostos da RLM.

Tabela 1: Regressão linear múltipla - Variável dependente: log (Resultado Líquido)

VARIÁVEIS INDEPENDENTES	COEFICIENTES NÃO PADRONIZADOS (β)	t	Valor - *p*	FIV
Constante	13,014	3,625	0,001	
Cultura Adocrática	0,745	0,333	0,741	1,412
Cultura Mercado	1,144	0,799	0,427	1,103
Cultura Hierárquica	-1,902	-1,167	0,248	1,383
Número de colaboradores	0,001	4,002	0,000	1,090
R^2 ajustado	0,199			
F	0,001			

O resultado sugere que apenas 19,9% da variação total do resultado líquido é explicada pela relação entre as variáveis independentes (cultura adocrática, cultura mercado, cultura hierárquica e o número de colaboradores). Este valor é semelhante ao obtido por Yesil & Kaya (2013).

Os resultados sugerem uma organização com o tipo de cultura adocrática como predominante. Verificou-se, através de testes de correlação, que a cultura de mercado tem um impacto positivo mais forte sobre o resultado líquido do que a cultura adocrática. Ao contrário destas e, apesar da não significância do coeficiente estimado, uma organização em que predomine a cultura hierárquica, ou seja, que se foque na autoridade, nas regras e procedimentos na disciplina e que se feche à mudança potenciará uma influência negativa no seu desempenho.

Conclui-se que não se rejeita a hipótese de que a cultura organizacional não exerça influência no desempenho financeiro. É de ressalvar que apesar de os resultados não serem estatisticamente significativos, existem estudos como o de Yesil & Kaya (2013) que referem que a cultura organizacional pode ter um impacto indireto no desempenho através de outras variáveis como a conversão do conhecimento e gestão do conhecimento e a inovação.

Referências

Cameron, K. S. & Quinn, R. E. (2006) *Diagnosing and changing organizational culture: Based on the competing values framework* (Rev. Ed.). San Francisco, CA: Jossey-Bass.

Fekete, H. & Bocskei, E. (2011) Cultural Waves in Company Performance. *Research Journal of Economics Business and ICT*, University of Pannonia, 3(1), pp. 38-42.

Yesil, S. & Kaya, A. (2013) The Effect of Organizational Culture on Firm Financial Performance: Evidence from a Developing Country. *Procedia – Social and Behavioral Sciences*, 81(28), 428-437.

Zhang, Z. & Zhu, X. (2012) Empirical Analysis of the Relationship between Organizational Culture and Organizational Performance. *National Conference on Information Technology and C. S.*, 763-766.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

GDP per capita dynamics in the European Union

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

Abstract: This paper introduces a novel nonlinear model for panel data analysis that estimates jointly nonlinear trajectories. We apply the model to GDP per capita dynamics of the 28 European Union countries in the period from 1996 to 2013. We conclude that four regimes describe the dynamics of GDP per capita in each country. Moreover, we found that there is heterogeneity and time synchronization between most European countries.

Keywords: European economy, Hidden Markov model, Markov switching models, Nonlinear models, Panel data.

Linear modeling has been the rule in economics and other social sciences. Despite that, many phenomena in economics may not behave linearly. For instance, investors' attitudes towards risk and expected returns tend to be nonlinear. The way information is incorporated in the stock market prices and economic fluctuations (e.g., periods of expansion and recession) tend to be highly nonlinear. This nonlinear behavior may be explained by the bounded rationality and contagious phenomena as the decision makers tend to be partially emotional/irrational in their actions. This unobserved heterogeneity has hampered the specification of functional nonlinear links between the variables.

Recently there has been an increasing motivation in extending linear models by incorporating nonlinear components. This has been possible by the availability of computing power and simplified modeling tools. New generations of economists and other social scientists have also been more involved with mathematical and statistical modeling with increasing interest in capturing nonlinearities in economic and social phenomena by using dynamic systems, neural networks, and agent-based modeling. These new developments have extended the traditional approach built on the analytic and linear models to new frontiers. The Markov switching regime model, introduced by Hamilton (1989), is one of such extensions that defines a nonlinear generalization of a linear model by incorporating nonlinearity in the mean and variance of a dynamic process. It has been particularly useful in modeling and characterizing business cycles in most countries as nonlinear components are important to explain economic activity given the impact of shocks depends on which stage of the cycle the economy is, either expansion or recession period. The transition between states is modeled as a regime switching model, and this switching probability adds nonlinearity to the analysis. The most common specification assumes a two-regime model, generally named expansion and recession stages. However, some authors have shown that more regimes can further improve the retrieving of nonlinearities in empirical data. For instance, Artis *et al.* (2004) assumes

three states: recession, moderate growth, and high growth. They estimate Markov switching models individually for each country in the analysis: Germany, France, Italy, the Netherlands, Austria, Belgium, Spain, Portugal, and the United Kingdom.

We analysis the GDP per capita dynamics in the 28 European countries in the period 1996 to 2013 using the regime switching model for panel data (see, e.g., Dias *et al.* (2015)). We conclude using the BIC model selection criterion that the best solution has four regimes. Figure 1 summarizes the results from the model for the four regimes with its expected growth of GDP per capita.

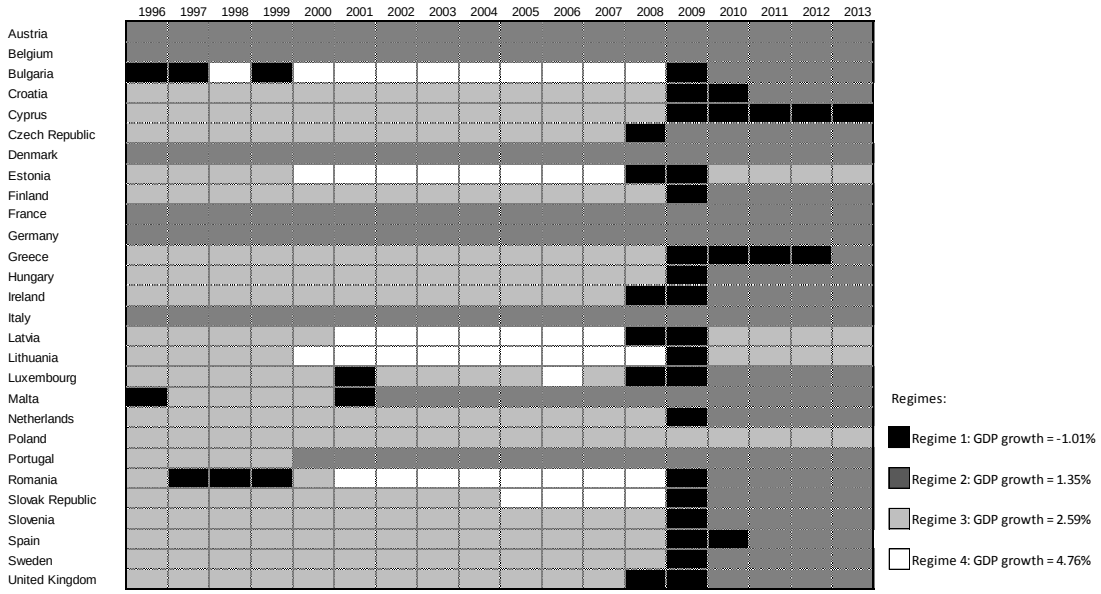


Figure 1: Posterior modal regimes estimates for each country

Our results show that we can trace and map the pattern of each country monitoring economic activity and identify the synchronization between the economies which are relevant results for policy decision making at European level. We observe distinct dynamics within specific countries with high progress in GDP per capita growth. After the Great Recession most of the countries showed a slowing down in the growth, and Cyprus and Greece entered a deep recession. These results can be updated as soon as new comparable data is made available for these countries (panel data).

References

Artis, M., Krolzig, H. M. & Toro, J. (2004) The European business cycle. *Oxford Economic Papers-New Series*, 56(1), 1-44.

Dias, J. G., Vermunt, J. K., & Ramos, S. B. (2015) Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research*, 243(3), 852-864.

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

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Alterações climáticas na incidência de casos de dengue na cidade de Goiânia no período de 2008 a 2015

Susana Faria[1], Antônio Neco[2], Raquel Menezes[3]
[1] CMAT - Centro de Matemática, DMA - Departamento Matemática e Aplicações, Universidade do Minho, sfaria@math.uminho.pt;
[2] Instituto Federal Goiano, antonio.neco@ifgoiano.edu.br;
[3] CMAT - Centro de Matemática, DMA - Departamento Matemática e Aplicações, Universidade do Minho, rmenezes@math.uminho.pt

Sumário: A dengue é uma doença viral transmitida pelo mosquito *Aedes aegypti,* que encontra nas variações climáticas ambientação para a sua proliferação, sendo as regiões tropicais e subtropicais as mais propícias. Neste trabalho pretende-se estudar a relação entre o número de casos de notificação de dengue e as variações climáticas no município de Goiânia, utilizando os modelos lineares generalizados, para prever períodos epidémicos e possibilitar ações preventivas.

Palavras-chave: Dengue, Modelos Lineares Generalizados, Precipitação, Temperatura.

A dengue é uma doença viral de rápida disseminação, devendo o agravo de notificação compulsória (Portaria GM/MS no 5 de 21/02/2006) e todos os casos suspeitos serem notificados à vigilância epidemiológica do município pelas unidades de saúde, via o Sistema de Informação de Agravos de Notificação (Sinan), para possibilitar o acompanhamento do padrão de transmissão da doença e permitir que ações preventivas possam ser tomadas. No Brasil, as primeiras epidemias de dengue ocorreram em 1981-1982 em Boa Vista e em 1986 no Rio de Janeiro e em algumas capitais do Nordeste. Desde então, ocorrem epidemias associadas com a introdução de novos sorotipos.

A expansão da dengue deve-se ao crescimento desordenado dos centros urbanos (onde o Brasil concentra mais de 80% da população), à falta de infra-estrutura de saneamento básico e às condições climáticas favoráveis que impedem a proposição de ações visando a erradicação do vetor transmissor, o mosquito *Aedes aegypti*, com um preocupante aumento de casos de dengue na faixa de jovens e crianças.

Em Goiânia (Goiás), a primeira epidemia foi registrada em 1994, com repetições recentes nos anos de 2008, 2010, 2013 e 2015. A cidade de Goiânia está localizada no planalto central do Brasil (Latitude: -15,91, Longitude: -50,13, Altitude: 512; 22m) com uma área de 740km^2, população estimada em 1,5 milhões de habitantes, temperatura média anual variando entre 18ºC e 26ºC e duas estações bem definidas: verão húmido (dezembro a março) e inverno seco (junho a agosto).

Neste trabalho pretende-se estudar a influência das alterações climáticas na incidência dos casos de dengue no município de Goiânia.

A base de dados foi construída a partir das informações de notificações de casos de dengue na cidade de Goiânia, as quais são registradas diariamente no Sistema de Informação de Agravos de Notificação (SINAM) e disponibilizadas pela Superintendência de Vigilância em Saúde do Estado de Goiás (SUVISA-GO). Essas informações permitiram quantificar o número de notificações semanais de casos de dengue. As informações meteorológicas foram obtidas a partir dos dados meteorológicos do Instituto Nacional de Meteorologia, o qual mantém os registros das observações que são realizadas diariamente na estação de monitoramento de Goiânia, para as variáveis precipitação (mm), temperatura mínima (ºC), temperatura máxima (ºC), humidade relativa do ar (%) e velocidade do vento (m/s). O período considerado neste estudo engloba os anos de 2008 a 2015.

Dada a natureza da variável resposta (dados de contagem), inicialmente foram aplicados modelos de regressão de Poisson. No entanto, como se identificaram problemas de sobredispersão, recorreram-se aos modelos de regressão com resposta binomial negativa. Foram criadas variáveis desfasadas em 1, 2,..., 10 semanas para as variáveis precipitação, temperatura máxima e mínima e humidade relativa. Posteriormente e uma vez que estes dados possuem uma dependência temporal, recorreu-se ao uso das Equações de Estimação Generalizadas (GEE) para modelar a correlação existente entre as observações.

Agradecimentos: Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Referências

Liang, K. & Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1), 13–22.

McCulloch, C. E. & Searle, S. R. (2001) *Linear and Generalized Linear Mixed Models.* Wiley, New York.

Ziegler, A. (2011) *Generalized Estimating Equations*. Springer.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Simulating deterministic and stochastic SVEIR models to determine the disease elimination time for different vaccination rates

Luiz S. Freitas[1], Hyun Mo Yang[2], Carlos A. Braumann[3]

[1] Laboratório de Epidemiologia e Fisiologia Matemáticas, Universidade Estadual de Campinas, luizfsf28@gmail.com;

[2] Laboratório de Epidemiologia e Fisiologia Matemáticas, Universidade Estadual de Campinas, hyunyang@ime.unicamp.br;

[3] Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora & Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, braumann@uevora.pt

Abstract: The stochastic SVEIR epidemic model with equal birth and death rates and constant vaccination rate ν is studied through Monte Carlo simulations and compared with the deterministic model. We assume an asymptomatic disease, so that we vaccinate among the individuals not yet vaccinated. This study, which includes the disease elimination time, is useful in public health. For example, one can determine the minimum ν for disease elimination before a time horizon with a prescribed high probability.

Keywords: Deterministic SVEIR model, Disease elimination time, Monte Carlo simulations, Stochastic SVEIR model, Vaccination rate.

We consider first a deterministic SVEIR epidemic model with equal birth and death rates and with compartments *X, V, H, Y, Z* (number of Susceptibles, Vaccinated, Exposed, Infectious, and Recovered, respectively). Let ν be the vaccination rate. From Yang & Hotta (2003), we know that there is a critical value of the vaccination rate ν_c such that, when $\nu > \nu_c$, the system converges to a disease-free equilibrium (in which $H = Y = 0$) and, when $\nu < \nu_c$, the system converges to an endemic equilibrium.

We consider the situation where we start with no vaccination (so the initial population will be at the endemic equilibrium corresponding to $\nu = 0$) and we introduce vaccination at a rate $\nu > 0$. We also assume that the disease is asymptomatic and we are not able to distinguish between susceptible and non-susceptible individuals, so that we will vaccinate among the individuals that have not yet been vaccinated, whether they are susceptible or not.

Then, we consider the more realistic stochastic model in which the transitions between compartments occur randomly according to a Markov chain with transition rates equal to the deterministic rates. We study this model through Monte Carlo simulations using the Gillespie algorithm (see Gillespie (1977)) and look at the evolution of the different compartments.

Contrary to the deterministic model, in the stochastic case a disease-free state will be reached whatever the value of ν is. We will look at the disease elimination time T (first time for which $H(T) + Y(T) = 0$), at the total number vaccinated until time T and at the number of "wasted" vaccines (vaccines administered to non-susceptible individuals). We will see how the distributions of these quantities vary with the vaccination rate ν. We will also find approximations to the interesting quantiles of these distributions by simple functions of ν.

These results have relevant applications in public health policy decisions. The consequences of adopting a given vaccination rate ν can be easily determined. The reverse problem of determining the minimum value of ν that will give a high probability (say 95% or 99% probability) of disease elimination (reaching a disease-free state) before some prescribed time horizon $T_{\max}$ is also tackled.

If, in the deterministic model, we consider that, for all practical purposes, the disease is eliminated if the number of exposed plus infectious is below 1 individual, then we may take the disease almost-elimination time T to be the first time for which $H(T) + Y(T) = 1$ and we can compare the deterministic and the stochastic models on the quantities referred to above.

Acknowledgements: Luiz S. Freitas acknowledges the Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, process 141084/2014-6. Carlos A. Braumann belongs to the Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), a research centre supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal).

References

Yang, H. M. & Hotta, L. K. (2003) Sobre a Erradicação de Doenças Infecciosas – Esforço de Vacinação. IN YANG, H. M. (Org.) *Matemática Aplicada a Fisiologia e Imunologia - Notas em Matemática Aplicada.* SBMAC & FAPESP, São Carlos, 7, 119-142.

Gillespie, D. T. (1977) Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 81(25), 2340-2361.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

O modelo de regressão logística na identificação de factores associados ao relato inconclusivo do rastreio de retinopatia diabética

A. Manuela Gonçalves[1], Inês Barros[2], João Reis[3]

[1] CMAT - Centro de Matemática, Universidade do Minho, Portugal, mneves@math.uminho.pt;

[2] DMA - Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, inesbarros18@gmail.com;

[3] ARS Norte – Administração Regional de Saúde do Norte, I.P., Portugal, joaoreis@arsnorte.min-saude.pt

Sumário: Este estudo teve como objectivo analisar a informação obtida no âmbito do Programa Regional de Rastreios da Retinopatia Diabética realizado em 19 ACES (Agrupamentos de Centros de Saúde) da região Norte de Portugal e, em particular, identificar os factores que influenciam o resultado do exame de Retinografia. A informação fornecida pela ARS Norte envolveu 149251 relatos, no período de 2009 a 2014. Foram aplicadas metodologias na área da Inferência Estatística e dos Modelos de Regressão Logística.

Palavras-chave: Modelo de regressão logística múltipla, Rastreio, Resultado do exame de retinografia, Retinopatia diabética, Testes de hipóteses.

A Retinopatia Diabética (RD) é uma complicação grave da Diabetes *Mellitus*, que é a principal causa de cegueira evitável na população entre os 20 e 64 anos de idade. Os programas de rastreio da RD baseados na realização de retinografias nos Centros de Saúde são a forma mais eficaz de detecção e tratamento da RD, permitindo aos doentes diabéticos a manutenção da visão útil. Na Região Norte de Portugal dos 24 ACES, apenas 19 ACES se encontram no Programa de Rastreio de Retinopatia Diabética (implementado em 2009). A informação deste estudo refere-se aos dados obtidos neste rastreio, de 2009 a 2014.

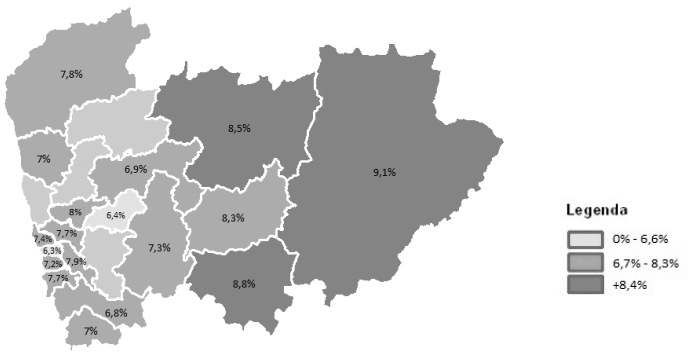


Figura 1: Distribuição espacial da prevalência de pacientes diabéticos inscritos nos Centros de Saúde, pelos ACES (19 no total de 24), na Região Norte.

A Figura 1 apresenta a distribuição geográfica, na Região Norte e a respectiva percentagem de prevalência dos diabéticos, no período observado. Da análise realizada concluiu-se que o ACES com maior número de pessoas inscritas e maior número de população diabética foi o ACES ULS Alto Minho, mas não foi o que apresentou maior percentagem de prevalência de diabéticos: o ACES Nordeste, com 9,1%. Este fenómeno pode estar relacionado com a estrutura etária da população, uma vez que é sabido que o aumento da idade induz um aumento da prevalência desta patologia (a população do interior da região é mais idosa). Conclui-se que nos 19 ACES da Região Norte há mais mulheres que homens a realizar o rastreio. O ACES Alto Ave foi o ACES com mais pessoas a realizar o rastreio da Retinopatia Diabética, apesar de só ter iniciado o programa de rastreio em 2011. Este ACES tem a segunda maior população diabética, com 18 891 doentes.

O Exame da patologia Retinopatia Diabética (com resultado Conclusivo (conclui que há ou não patologia) ou Inconclusivo (o exame efectuado não permite avaliar se existe ou não patologia)) é executado por um retinógrafo (há 3 tipos de aparelhos distribuídos pelas ACES) que produz uma retinografia a cores (depois lida por um médico oftalmologista) que é um método de diagnóstico simples, sensível e fiável, permitindo identificar os diabéticos que precisam de acompanhamento ou tratamento oftalmológico bem como um registo de evolução do estado da patologia da retina.

Neste trabalho foram desenvolvidos modelos de regressão logística simples e múltipla com o objectivo de identificar factores que influenciam o resultado do Exame da patologia Retinopatia Diabética. Os factores que foram estatisticamente significativos no resultado do Exame foram a Idade do Paciente, o Tipo de Retinógrafo, o Técnico e o Centro de Leitura (entre mais possíveis covariáveis que foram testadas). A capacidade discriminante do modelo ajustado foi avaliada pela área sob a curva ROC (o valor observado foi de 0,728), o que corresponde a uma discriminação aceitável.

Assim, identificaram-se e interpretaram-se os factores explicativos do resultado do relato, pois sabe-se o quanto é importante o resultado destes exames ser conclusivo tanto para a prevenção desta patologia, como para o tratamento atempado dos pacientes para que a sua patologia não evolua para estadios mais graves, como por exemplo a cegueira.

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 projecto PEstOE/MAT/UI0013/2014.

Referências

Christensen, R. (1997) *Log-Linear Models and Logistic Regression,* New-York, USA: Springer.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logistic Regression,* New-York, USA: John Wiley Sons.

http://portal.arsnorte.min-saude.pt/portal/page/portal/ARSNorte.

Sheather, S.J. (2009) A Modern Approach to Regression with R, USA: Springer.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

A análise estatística multivariada na avaliação da qualidade de águas subterrâneas

A. Manuela Gonçalves[1], Driano Rezende[2], Letícia Nishi[2], Fernanda O. Tavares[2], M. Teresa Amorim[3], Rosángela Bergamasco[2]

[1] CMAT - Centro de Matemática da Universidade do Minho, Portugal, mneves@math.uminho.pt;

[2] DEQ - Departamento de Engª. Química, Universidade Estadual de Maringá, Brasil, drirezend@gmail.com;leticianishi@hotmail.com; fernandaoliveiratavares@gmail.com; rosangela@deq.uem.br;

[3] DET - Departamento de Engª. Têxtil da Universidade do Minho, Portugal, mtamorim@det.uminho.pt

Sumário: Neste trabalho foram aplicadas metodologias estatísticas na área da Estatística Multivariada (Análise de *Clusters* e Análise de Componentes Principais) com o objectivo de avaliar e interpretar a qualidade de águas subterrâneas no município de Maringá, estado do Paraná, Brasil. As amostras de água foram recolhidas em 19 poços tubulares no período de um ano e submetidas a análises físico-químicas. Os resultados deste estudo permitiram identificar fontes de contaminação (naturais e antropogénicas).

Palavras-chave: Águas subterrâneas, Análise de Componentes Principais, Análise de *Clusters*, Contaminação, Variáveis físico-químicas.

Entre as dificuldades encontradas por muitos pesquisadores, em estudos relacionados com a qualidade de águas subterrâneas, destaca-se a leitura e interpretação dos dados obtidos no período de monitorização. Assim, torna-se necessário o emprego de métodos estatísticos como a Análise de *Clusters* (AC) e a Análise de Componentes Principais (ACP) com o objectivo de reduzir a dimensão do conjunto de informação, permitindo avaliar associações entre um vasto conjunto de variáveis em termos de um pequeno número de factores latentes sem perder muita informação (Johnson & Wichern 2007). Neste contexto, no presente trabalho foram aplicados estes métodos aos dados obtidos em 19 pontos de monitorização de águas subterrâneas no município de Maringá, estado do Paraná, Brasil.

Entre Agosto de 2014 a Agosto de 2015 foram realizadas colectas de água em 19 poços tubulares profundos, uma vez ao bimestre, totalizando 114 colectas no período considerado. As análises físico-químicas foram realizadas no laboratório de Gestão, Controle e Preservação Ambiental no Departamento de Engenharia Química da Universidade Estadual de Maringá, Paraná, Brasil. Foram consideradas 25 variáveis físico-químicas para avaliar a qualidade da água nomeadamente pH, nitratos, agro-tóxicos, metais pesados, etc.

A análise foi iniciada utilizando métodos da estatística descritiva, efectuando uma análise exploratória espaço-temporal das variáveis observadas e identificando padrões de comportamento ao longo do período observado. No período em estudo 10 variáveis físico-químicas apresentaram valores acima do valor recomendado pela Organização Mundial da Saúde (WHO 2011).

Seguiu-se uma Análise de *Clusters* e uma Análise de Componentes Principais (a análise estatística foi realizada utilizando o *software* R, versão 3.2.0 (Everitt & Hothorn 2011). Por meio da AC foi possível observar três tipos de águas subterrâneas na região de estudo (3 *clusters* obtidos por vários métodos hierárquicos aglomerativos (o método de Ward apresentou melhores definições) e pela utilização da distância euclidiana como medida de dissemelhança), as quais são características da geologia da região.

Por meio da ACP foram analisadas correlações entres os contaminantes, identificaram-se os factores responsáveis pelas variações da qualidade da água subterrânea nos diferentes pontos de amostragem, que estão maioritariamente associados com as possíveis fontes de contaminação da água subterrânea em estudo (de origem agrícola, urbana e industrial). Esta metodologia demonstrou ser uma ferramenta de grande importância para o conhecimento de diferentes composições hidroquímicas de águas subterrâneas e, também, correlacionar as variáveis físico-químicas com as principais fontes de poluição hidrológica (Voutsis *et al*. 2015).

Este trabalho insere-se num projeto em desenvolvimento pela equipa de investigadores, contribuindo para a interpretação da informação recolhida e relacionar o estado ecológico dos sistemas estudados, permitindo a identificação de fontes de contaminação das águas subterrâneas.

Agradecimentos: Este trabalho foi idealizado entre a Universidade Estadual de Maringá, PR, Brasil e a Universidade do Minho, PT, o qual foi financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES-Brasil) e também pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - "Fundação para a Ciência e a Tecnologia", no âmbito do projectos PEstOE/MAT/UI0013/2014 e UID/CTM/00264.

Referências

Everitt, B. & Hothorn, T. (2011) *An Introduction to Applied Multivariate Analysis with R*. USA: Springer.

Johnson, R.A. & Wichern, D.W. (2007) *Applied Multivariate Statistical Analysis*. New Jersey, USA: Prentice Hall.

Voutsis, N., Kelepertzis, E. Tziritis, E. & Kelepertsis, A. (2015) Assessing the hydrogeochemistry of groundwaters in ophiolite areas of Euboea Island, Greece, using multivariate statistical methods. *Journal of Geochemical Exploration*, 159, 79-92.

WHO (2011) *Guidelines for drinking-water quality* (Fourth edition). World Health Organization.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

A Análise Classificatória na caracterização da produção e consumo de produtos de origem animal a nível mundial

Manuel Minhoto[1], Luís Fernandes[2]
[1] CIMA/IIFA e DMAT/ECT, Universidade de Évora, minhoto@uevora.pt;
[2] Universidade de Évora – ICAAM, ladsf@uevora.pt

Sumário: Neste trabalho pretende-se comparar quer a produção, quer o consumo de produtos de origem animal destinados à alimentação humana nos últimos 50 anos. Selecionaram-se 70 países e 6 variáveis. Recorreu-se à Análise Classificatória Hierárquica Ascendente (ACHA) e Análise Classificatória Não Hierárquica para agrupar cada um dos países e à Análise Discriminante (AD) para visualizar esse agrupamento. Com base em dois índices multivariados efetuou-se uma ACHA para as 6 décadas envolvidas.

Palavras-chave: Análise Classificatória, Função linear de Fisher, Índice Multivariado GCD, Índice Multivariado RV, Produtos de Origem Animal.

O trabalho tem por objetivo analisar quer a produção, quer o consumo de produtos de origem animal destinados à alimentação humana nos últimos 50 anos. Recolheram-se os dados nos três primeiros anos de cada uma das décadas desde 1960. Assim, para a década de 60 consideraram-se os dados de 1961, 62 e 63. Determinou-se a média do triénio e procedeu-se de forma idêntica para as décadas de 70, 80, 90 e 2000. Na presente década recolheram-se os anos de 2009, 2010 e 2011 para o consumo porque os dados de alguns países ainda não estavam disponibilizados no *website* da *Food and Agriculture Organization of the United Nations* (FAO). Este estudo engloba um total de 70 países e 6 variáveis observadas com tratamento separado para produção e consumo. Tem-se então para cada década uma matriz $\mathbf{X}(70\times6)$ para a produção e uma matriz $\mathbf{Y}(70\times6)$ para o consumo.

Começou-se por efetuar uma ACHA para as seis décadas em separado, tendo em vista selecionar o número de grupos mais adequado que se pretendia que fosse o mesmo para todas as décadas. Fixado o número de grupos, passou-se a uma Análise Classificatória Não Hierárquica (algoritmo das K-Médias), para afetar cada país ao seu grupo. Os resultados obtidos com a referida análise não hierárquica fornecem-nos as médias de cada grupo para cada uma das variáveis observadas, permitindo, desta forma, a caracterização de cada um dos grupos, em função do seu afastamento da média global. Dispomos igualmente das distâncias entre os vários grupos permitindo a visualização destes grupos de países tomados como um todo, através de um dendograma. A visualização destes grupos com todos os seus elementos (países) pode também ser obtida em formato bidimensional com maximização das distâncias entre grupos. Com efeito, através do modelo de Análise Discriminante de Fisher podem ser obtidas as duas primeiras funções

discriminantes. Projetando todos os países no plano definido por estas duas primeiras funções discriminantes, obtém-se a representação bidimensional ótima para a matriz de dados com a divisão em grupos indicada. Desta forma, facilita-se a comparação dos grupos correspondentes em cada década, verificando-se os países que eventualmente tenham mudado de grupo, bem como verificar se os grupos mantêm as suas características relativamente às variáveis observadas e ainda verificar se os grupos mantêm os mesmos afastamentos entre si.

Finalmente, procedeu-se à comparação multivariada entre as matrizes de dados observados. Esta comparação é efetuada recorrendo a dois índices multivariados que variam entre 0 e 1, correspondendo o valor 1 a um ajuste perfeito. Um destes índices é o índice RV de Robert & Escoufier (1976), largamente utilizado não só na Metodologia Statis, mas de um modo geral para comparar matrizes de dados observados no espaço dos indivíduos. O outro índice utilizado é o Coeficiente de Determinação Generalizado (*Generalized Coeficient of Determination* (GCD)) de Yanay (Ramsey *et al.* 1984). Este índice é um indicador que mede o grau de semelhança entre dois espaços e como é referido por Cadima *et al.* (2004), este índice possui uma relação estreita com o conceito de distância entre subespaços abordada em Golub & Van Loan (1996).

Com base em cada um dos índices multivariados GCD e RV, obtiveram-se duas matrizes de semelhanças entre décadas (uma para produção, outra para consumo). Usando a relação $\sqrt{2(1-indice)}$, as semelhanças transformaram-se em distâncias. A partir da matriz de distâncias efetuou-se uma ACHA, com obtenção do respetivo dendograma, cuja análise permite agrupar as 6 décadas e verificar a continuidade e eventuais quebras no consumo/produção.

Referências

Cadima, J., Cerdeira, J. O. & Minhoto, M. (2004) Computational aspects of algorithms for variable selection in the contexto of principal componentes. *Computational Statistics & Data Analysis*, 47, 225-236.

Ramsay, J. O., ten Berge, J. & Styan, G. P. H. (1984) Matrix correlation. *Psychometrica*, 49(3), 403-423.

Robert P. & Escoufier. (1976) A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Applied Statistics*, 25, 257-266.

Golub, G. & Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Avaliação quantitativa do património edificado: o caso de estudo centro do Porto

Cilísia Ornelas[1], Fernanda Sousa[2], João Miranda Guedes[3], Isabel Breda-Vázquez[4]
[1] Faculdade de Engenharia, CONSTRUCT e CITTA, Universidade do Porto, cilisia@fe.up.pt;
[2] Faculdade de Engenharia e CITTA, Universidade do Porto, fcsousa@fe.up.pt;
[3] Faculdade de Engenharia e CONSTRUCT, Universidade do Porto, jguedes@fe.up.pt;
[4] Faculdade de Engenharia e CITTA, Universidade do Porto, ivazquez@fe.up.pt

Sumário: A reabilitação do património edificado exige um conhecimento multidisciplinar prévio a qualquer intervenção. Neste trabalho recorre-se a ferramentas de análise de dados, em particular a métodos de classificação, para avaliar quantitativamente o valor patrimonial e as condições de segurança e de habitabilidade dos edifícios, e a satisfação dos residentes do centro do Porto. Este estudo permite realizar diagnósticos detalhados e dirigidos, ferramentas essenciais de apoio à catalogação e criação de níveis de intervenção.

Palavras-chave: Análise de dados, Classificação, Edificado, Indicadores, Valor patrimonial.

A reabilitação do património edificado é um tema em debate entre várias entidades, na procura de soluções para a sua degradação. A falta de procedimentos de apoio à intervenção no edificado existente em Portugal é um dos motores deste trabalho, realçando a necessidade de avaliações holísticas e dirigidas. Em investigação de contexto mais alargado foi desenvolvida uma Metodologia de Avaliação do Património Edificado Habitado – MAPEH (Ornelas *et al*. 2014a), constituída por critérios heterogéneos, aplicáveis à análise de edificado antigo, combinando as dimensões patrimonial, técnica e social (Ornelas *et al*. 2014b). A aplicação desta metodologia a um contexto particular é realizada através de um instrumento de avaliação (inquérito), composto por critérios específicos.

O inquérito encontra-se dividido em três partes. A primeira parte contém 25 questões identificativas da presença/ausência de características físicas do edificado e de materiais/técnicas construtivas. Com a segunda parte do instrumento pretende-se avaliar as condições de segurança e de habitabilidade (65 variáveis). Os critérios de avaliação aqui usados são baseados em requisitos normativos, sendo necessário para cada característica decidir da sua aplicabilidade. A terceira parte visa caracterizar a satisfação residencial (7 variáveis). Este instrumento (num total de 97 variáveis binárias) foi aplicado a 42 edifícios habitacionais do centro do Porto, de matriz unifamiliar burguesa.

Tendo por objetivo medir quantitativamente as diferentes dimensões em análise, foram atribuídas ponderações de 1 a 3 aos itens intervenientes, de acordo com a importância conferida por especialistas nas várias áreas, o que permitiu calcular um

conjunto de indicadores quantitativos temáticos, denominados indicadores globais e definidos por médias ponderadas. A primeira parte do instrumento deu origem a um único indicador global, designado de "Valor Patrimonial". Com a informação da segunda parte construíram-se 3 indicadores globais relacionados com as condições de segurança e 8 relacionados com as condições de habitabilidade. A terceira parte deu origem a 5 indicadores, caraterizadores da satisfação residencial. A matriz de dados, definida pelos 17 indicadores globais, foi alvo de um tratamento exploratório multivariado, com recurso a várias técnicas multivariadas: de redução de dimensionalidade (análise em componentes principais) e de classificação (hierárquica e não hierárquica). O processo de classificação incluiu: i) classificação hierárquica com diferentes critérios de agregação, ii) estudo de comparação e validação para a escolha da melhor árvore e respetivo corte, iii) classificação não hierárquica (k-médias) para o número de classes escolhido. De entre os resultados obtidos salienta-se a classificação dos edifícios em quatro classes, A a D, cuja interpretação é uma mais-valia para a problemática abordada. A Figura 1 descreve estas classes para as dimensões patrimonial e técnica, evidenciando i) o caracter decrescente do valor patrimonial da Classe A para a D ii) não existir uma correspondência direta entre boas condições de segurança e habitabilidade e elevado valor patrimonial.

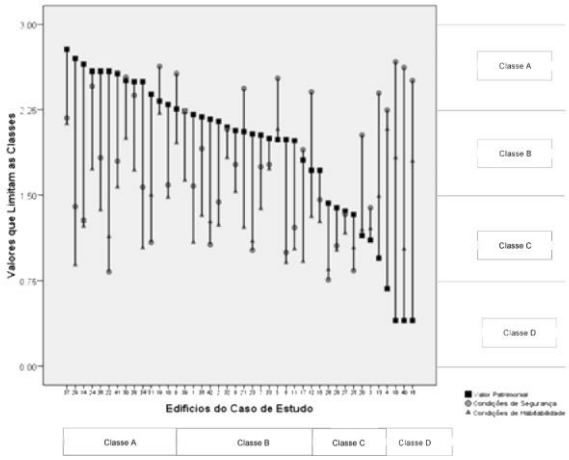


Figura 1: Classes de edifícios segundo o valor patrimonial.

Este estudo reflete o panorama real dos edifícios nas dimensões abordadas. Em particular, que a quantificação da informação constitui um utensílio essencial na criação de níveis de intervenção, alertando para a necessidade de se estabelecerem critérios de flexibilidade ajustados às características do edificado na aplicação da legislação.

Referências

Ornelas, C., Guedes, J. & Breda-Vázquez, I. (2014a) A Holistic Preservation and Maintenance of the Built Heritage: The Role of an integrated Methodology. In Tadeu, A., Ural, D., Ural, O. & Abrantes, V. (Eds.) *Proceedings of 40th IHAS World Congress on Housing - Sustainable Housing Construction,* Portugal.

Ornelas, C., Guedes, J. & Breda-Vázquez, I. (2014b) The Minimum Intervention in Built Heritage: comparing the potential role of codes for conservation In Lourenço, P., Haseltine, B. A. & Vasconcelos, G. (Eds.) *Proceedings of International Masonry Conference IMC*, Portugal.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

ANOVA a dois fatores não paramétrica com células omissas

Dulce G. Pereira[1], **Anabela Afonso**[2]
[1, 2] *CIMA/IIFA e DMAT/ECT, Universidade de Évora;*
[1] *dgsp@uevora.pt;*
[2] *aafonso@uevora.pt*

Sumário: Na análise de variância a dois fatores por vezes há combinações de níveis dos fatores que não são observadas, o que dificulta a análise dos dados pois depende do número de células omissas e da sua localização. Já foram propostas algumas abordagens paramétricas para lidar com esta situação. Muitas vezes a alternativa não paramétrica consiste em aplicar a abordagem paramétrica às ordens dos valores observados. Neste trabalho ilustraremos que na presença de células omissas este processo pode não ser o mais adequado como alternativa não paramétrica, sendo necessária mais investigação nesta área.

Palavras-chave: Hipóteses de Tipo IV, Modelo de efeitos fixos, Modelo de médias, Somas de quadrados.

Muitas experiências envolvem o estudo do efeito de dois ou mais fatores numa variável resposta. Admitindo que existem dois fatores R e C, com r e c níveis respetivamente, diz-se que o delineamento é fatorial quando cada réplica contém todas as rc combinações de níveis destes dois fatores (Montgomery 2013). No entanto, é usual na prática encontrarem-se situações em que o número de observações por célula não é igual (delineamento desequilibrado) e até podem não ter sido observadas algumas das combinações de níveis (células omissas). Este tipo de situações por vezes ocorre devido a várias razões, que podem ser por conveniência de delineamento devido por exemplo a custos associados, mas também porque o investigador não consegue controlar em absoluto a experiência.

Para a análise dos dados de um delineamento com células omissas foram propostas vários tipos de abordagens paramétricas. Montgomery (2013) propôs transformar o delineamento com dois fatores num delineamento a um fator com $rc - m$, onde m representa o número de células omissas e realizar uma análise de variância (ANOVA) a um fator. Para testar a interação entre os fatores, propôs a utilização de contrastes de interesse, linearmente independentes. Alguns *softwares* estatísticos, como sejam o SPSS e SAS, recomendam o uso das somas de quadrados de Tipo IV. As hipóteses de Tipo IV comparam a média de todos os níveis de cada fator com a média obtida a partir de um ou mais níveis comuns do outro fator. Estas somas coincidem com as somas de Tipo III quando não existem células omissas. Mas, vários autores têm salientado as hipóteses de Tipo IV podem não ser únicas, pelo que o investigador deve definir as suas comparações específicas. Além disso, as conclusões dependem do número de células omissas e da posição em que surgem. Milliken & Johnson (2009) propõe que em vez de se considerar a

análise através de um modelo de efeitos fixos se opte por considerar um modelo de médias (*means model*) o qual é muito simples e fácil de interpretar.

Quando as variáveis são de natureza ordinal, ou são violados os pressupostos de aplicação da ANOVA paramétrica a dois fatores, Hora & Conover (1984) propuseram um teste equivalente em que a estatística de teste *F* da ANOVA era obtida a partir das ordens das observações. Contudo, as conclusões referentes à possível existência de uma interação significativa não devem ser garantidas como corretas uma vez que os efeitos aditivos dos fatores nas observações originais não são extensíveis às suas ordens e vice-versa. Além disso, o estudo da interação origina probabilidades de erro de Tipo I elevadas (Shah & Madden 2004). Em alternativa foi sugerido o uso de outras estatísticas de teste com uma distribuição aproximada a um qui-quadrado, que conduzem às mesmas conclusões relativamente aos efeitos principais dos fatores (Shah e Madden, 2004). No entanto, em nenhuma destas alternativas foi tida em conta a possível presença de células omissas.

Neste trabalho vamos aplicar, quer aos valores observados quer às respetivas ordens, as três abordagens paramétricas referidas anteriormente para a ANOVA a dois fatores com células omissas. Recorrendo a alguns exemplos de aplicação, mostraremos que este processo pode não ser o mais adequado como alternativa não paramétrica. Deste modo é necessário mais investigação nesta área.

Agradecimentos: Este trabalho é financiado por Fundos acionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto «UID/MAT/04674/2013(CIMA)»

Referências

Hora, S. & Conover, W. J. (1984) The F-statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, 79, 668–673.

Milliken, G. A. & Johnson, D. E. (2009) *Analysis of Messy Data. Volume 1: Designed Experiments*, Second Edition. Chapman and Hall/CRC.

Montgomery, D. G. (2013) *Design and Analysis of Experiments* (8th Edition). John Wiley & Sons.

Shah, D. A. & Madden, L. V. (2004) Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, 94, 33-43.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Características psicométricas da Escala de Empatia de Jefferson em estudantes de Tecnologias da Saúde

Ana Reis[1], Helena Martins[2], Ana Salgado[2], Andreia Magalhães[2], Zita Sousa[2], Artemisa R. Dores[2]

[1] Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto, crr@estsp.ipp.pt;

[2] Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto

Sumário: O impacto positivo da Empatia no contexto das profissões de saúde é consensualmente aceite pela comunidade científica. Diversos estudos têm sido desenvolvidos na área da Medicina, mas os resultados têm-se revelado controversos. No entanto, são escassos os estudos que se debruçam sobre a Empatia entre os estudantes das Tecnologias da Saúde, pelo que é objetivo deste trabalho a validação da Escala de Empatia de Jefferson – versão para estudantes para o contexto das tecnologias da saúde.

Palavras-chave: Análise fatorial, Empatia, Ensino superior, Tecnologias da saúde.

O impacto positivo da Empatia no contexto das profissões de saúde é consensualmente aceite pela comunidade científica. A Empatia é considerada um atributo cognitivo e psicológico essencial dos profissionais de saúde, sendo reconhecida a sua relevância para a melhoria da prestação de cuidados de saúde. Diversos estudos têm sido desenvolvidos na área da Medicina; no entanto, são escassos os que se debruçam sobre a Empatia entre os estudantes de Tecnologias da Saúde (Williams *et al.* 2014).

O objetivo principal deste estudo foi analisar as características psicométricas da Escala de Empatia de Jefferson em estudantes de Tecnologias da Saúde.

A amostra incluiu 592 estudantes de diversos cursos de licenciatura na área das tecnologias da saúde, com média de idade de 20,5 anos (DP=2,7), da Escola Superior de Tecnologia da Saúde do Porto – Instituto Politécnico do Porto (ESTSP–IPP). Os participantes preencheram a Escala de Empatia de Jefferson – versão para estudantes (Magalhães *et al.* 2012) que inclui 20 itens, respondidos numa escala tipo Likert, que varia entre 1 (discordo totalmente) e 7 pontos (concordo totalmente). A escala é constituída por três dimensões: tomada de perspetiva (10 itens), compaixão (7 itens) e capacidade de se colocar no lugar do outro/paciente (2 itens). Da totalidade dos itens, 10 são formulados pela negativa. As pontuações obtidas (através do somatório de todos os itens da escala) podem variar entre 20 (mínimo) e 140 pontos (máximo), correspondendo uma maior pontuação a uma maior perceção do estudante acerca do seu comportamento empático na prestação de cuidados ao paciente/doente e da relevância da qualidade da relação médico-paciente (Hojat *et al.* 2007).

A análise psicométrica da Escala de Empatia de Jefferson compreendeu, num primeiro momento, o estudo da fiabilidade através do alpha de Cronbach e a análise

fatorial exploratória para testar as propriedades psicométricas do instrumento na amostra em estudo. Nesta primeira fase, foi possível confirmar a fiabilidade da escala, tendo sido obtidos alphas de Cronbach acima de 0,70.

A análise fatorial exploratória produziu resultados bastante semelhantes aos referidos em estudos similares, realizados com a versão original da Escala de Empatia de Jefferson (Hojat *et al.* 2001) e com a versão portuguesa de Magalhães *et al.* (2012). No presente estudo, os três fatores identificados foram perfeitamente idênticos aos fatores propostos pelos autores da versão original (Hojat *et al.* 2001). De sublinhar igualmente que, no presente estudo, não foram detetados indícios da existência de valores de saturação cruzada dos itens. A segunda fase de análise de dados compreende a análise fatorial confirmatória que, se encontra neste momento, em conclusão.

O presente estudo insere-se num projeto de âmbito nacional que tem como objetivo explorar a forma como os estudantes da área da Saúde podem desenvolver competências de comunicação no Ensino Superior e pretende contribuir para a literatura e investigações futuras, validando um instrumento de medida importante para esta população.

Referências

Hojat, M., Mangione, S., Nasca, T., Cohen, M., Gonnella, J., Erdmann, J., Veloski, J. & Magee, M. (2001) The Jefferson Scale of Physician Empathy: development and preliminary psychimetric data. *Educational and Psychological Measurement*, 61(2), 349-365.

Hojat, M. (2007) Empathy in patient care: Antecedents, development, measurement and outcomes. New York: Springer.

Magalhães, E., Costa, P. & Costa, M. (2012) Empathy of medical students and personality: evidence from the Five-Factor Model. *Medical Teacher*, 34, 807-812.

Williams, B., Brown, T., Boyle, M., McKenna, L., Palermo, C. & Etherington, J. (2014) Levels of empathy in undergraduate emergency health, nursing, and midwifery students: a longitudinal study. *Advances in Medical Education and Practice*, 5, 299-306.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Efeito da alteração de medicamentos sujeitos a receita médica para não sujeitos a receita médica

Teresa Risso[1], Cláudia Furtado[2]
[1] *Infarmed, teresa.risso@infarmed.pt;*
[2] *Infarmed, claudia.furtado@infarmed.pt*

Sumário: Passados 10 anos sobre a liberalização do mercado dos medicamentos não sujeitos a receita médica (MNSRM) é altura de fazer uma análise da política implementada. O presente trabalho tem como objetivo avaliar os resultados na acessibilidade do utente a estes medicamentos, focando a sua análise no efeito que a transição de medicamento sujeito a receita médica (MSRM) para MNSRM poderá ter tido nos consumos, utilizando o método de regressão segmentada.

Palavras-chave: Acessibilidade, Liberalização do mercado, Medicamentos não sujeitos a receita médica, Regressão segmentada**.**

Os MNSRM destinam-se ao alívio e tratamento de situações clínicas sem gravidade (Despacho nº 17690/2007 de 23 de julho). A sua utilização responsável pode constituir uma componente importante da autogestão do estado de saúde do indivíduo, evitando o recurso aos serviços de saúde em casos que não requerem acompanhamento. No ano de 2005 foram introduzidas alterações legislativas na área dos MNSRM que permitiram que estes fossem dispensados fora das farmácias, em locais de venda autorizados para o efeito (DL nº 134/2005, de 16 de agosto). Nesse mesmo ano ocorreu a liberalização do preço destes medicamentos.

Com o objetivo de avaliar o efeito desta política na acessibilidade, começou por caracterizar-se o mercado de MNSRM, a evolução dos preços e as variações geográficas neste segmento de mercado. Para tal, utilizaram-se os dados de vendas reportados mensalmente ao Infarmed pelos locais de venda livre e os dados de colocação nas farmácias disponibilizados pela IMS *Health*.

Com o objetivo de averiguar se o *switch* de MSRM para MNSRM teve influência nos consumos, utilizou-se o método de análise de regressão segmentada. Este método permite a avaliação dos efeitos provocados por intervenções na série dos dados. Irá estimar o impacto imediato do *switch* no consumo dos medicamentos estudados e a alteração da tendência dessa variável depois do *switch.* A especificação geral do modelo é a seguinte:

$$Y_t = \beta_0 + \beta_1 * tempo_t + \beta_2 * switch_t + \beta_3 * tempo\ após\ switch_t + \varepsilon_j,$$

em que Y_t é a variável de interesse, *tempo* é uma variável contínua que indica os meses do período de observação, *switch* é uma variável binária igual a 0 antes do *switch* e 1 após o *switch* e *tempo após switch* conta os meses a partir do momento em que se deu a alteração.

Foram selecionados para análise, de entre os MNSRM que fizeram o *switch*, alguns dos que apresentaram maior volume de vendas acumulado, pertencentes a vários grupos terapêuticos.

Conclui-se que, entre 2005 e 2014, o consumo de MNSRM aumentou cerca de 10,5%. Contudo, observou-se uma quebra entre 2009 e 2014, sendo as farmácias as maiores responsáveis, já que os locais de venda livre apresentam vendas cada vez maiores.

Observou-se ainda que o comportamento do mercado não é igual em todo o país, existindo diferenças regionais tando ao nível dos consumos, como no número de locais de venda livre e também índice de preços.

Dos 14 medicamentos selecionados para análise, na maioria não é evidente que o *switch* tenha provocado um efeito imediato no volume de vendas nem na tendência das mesmas. A título de exemplo, no caso do Analgésico e Antipirético, no início do período de observação as vendas situavam-se nas 435 mil embalagens. Os resultados não sugerem uma alteração imediata nos consumos deste medicamento no momento da passagem a MNSRM (valor p=0.493). Depois do *switch*, as vendas encontram-se praticamente estáveis (β_2=1.009, valor p=0.001).

Nos primeiros 10 anos da liberalização do mercado, vários medicamentos passaram de sujeitos a receita médica a não sujeitos, vendo a sua disponibilidade aumentar e o preço ser liberalizado. A análise efetuada aos medicamentos que sofreram essa transição e que apresentaram um maior volume de vendas não expôs evidências de que, na maioria dos casos, o *switch* tenha tido um efeito nas vendas.

A monitorização do mercado dos MNSRM torna-se essencial para avaliar as decisões tomadas ao longo dos anos e apoiar intervenções futuras.

Referências

Lagarde, M. (2012) How to do (or not to do)… Assessing the impact of a policy change with routine longitudinal data. *Health Policy and Planning*, 27, 76-83.

Wagner, A. K., Soumerai, S. B., Zhang, F. & Ross-Degan, D. (2002) Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27, 299-309.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Avaliação das perceções dos estudantes do 1º ano em relação à praxe académica

O. Silva[1], S. N. Caldeira[2], M. Mendes[3], S. Botelho[4], M. J. Martins[5], Á. Sousa[6]

[1] Universidade dos Açores, CICS.NOVA.UAc, Osvaldo.dl.silva@uac.pt;
[2] Universidade dos Açores, CICS.NOVA.UAc, suzana.n.caldeira@uac.pt;
[3] Universidade dos Açores, macmendes1@hotmail.com;
[4] Universidade dos Açores, susanapinhobotelho@hotmail.com;
[5] Instituto Politécnico de Portalegre, mariajmartins@esep.pt;
[6] Universidade dos Açores, CEEAplA, aurea.st.sousa@uac.pt

Sumário: O objetivo deste estudo é o de conhecer as perceções dos estudantes do 1º ano do ensino superior relativamente à forma como vivenciaram as praxes. O questionário utilizado contém, entre outras variáveis, a Escala de Avaliação das Situações de *Bullying* nas Praxes do Ensino Superior, com o intuito de avaliar o posicionamento (menos favorável ou mais favorável) dos estudantes em relação às praxes. São apresentados e discutidos os principais resultados obtidos com recurso à análise de dados.

Palavras-chave: Análise classificatória, Análise de dados**,** Análise em componentes principais categórica, Praxe académica.

As praxes académicas têm constituído um tema controverso, por, na perspetiva de uns, gerarem situações de afronta e humilhação aos estudantes recém-entrados na nova instituição de ensino por parte dos que já a frequentam (Silva 2013) ou, na perspetiva de outros, por constituírem um meio de ajuda aos novos estudantes, no sentido de estes conhecerem e se relacionarem com os colegas e a instituição através de atividades divertidas e num clima animado (Pimentel *et al.* 2012). Genericamente, os comportamentos de praxe podem ser enquadrados num "conjunto de costumes e tradições geradas entre estudantes do ensino superior, que se constitui como essência de uma vida muito própria, especial e diferente" (Loureiro *et al.* 2009, p. 89).

Os dados, referentes a 417 estudantes do 1º ano da Universidade dos Açores e do Instituto Politécnico de Portalegre, foram recolhidos através de um questionário (amostragem por quotas). O questionário contém variáveis de caracterização da amostra (e.g., sexo, idade, tipo de participação nas praxes, adjetivos que caracterizam as praxes, frequência da semana académica, frequência das praxes, frequência no cortejo) e 15 itens de auto-resposta, numa escala Likert (1-*Discordo totalmente*, 2-*Discordo*, 3-*Não concordo nem discordo*, 4-*Concordo*, 5-*Concordo totalmente*), referentes à Escala de Avaliação das Situações de *Bullying* nas Praxes do Ensino Superior (*EASBPES*), de Matos *et al*. (2010). Foi calculada a pontuação total (soma das pontuações obtidas nos quinze itens) da EASBPES, sendo de referir que quanto maior for a pontuação de um indivíduo mais favorável é a sua perceção relativamente às praxes.

Os dados foram analisados utilizando métodos estatísticos, de onde se destacam os gráficos *Zoom Star* a duas dimensões (*2D*), os diagramas de extremos e quartis, o coeficiente de correlação ordinal de Spearman, a Análise em Componentes Principais Categórica (CatPCA) e o método de análise classificatória não hierárquica das *k*-médias.

No que se refere ao tipo de participação nas praxes, verificou-se que 24,3% dos inquiridos não participaram e declararam-se "*anti-praxe*", 36% não participaram em "*quase nada*" mas não se declararam "*anti-praxe*", 22,1% participaram somente como caloiros, 5.8% participaram em apenas algumas atividades e 11,7% participaram ativamente em quase todas as atividades. A pontuação obtida na *EASBPES* pelos estudantes que participaram nas praxes de forma mais ativa foi mais elevada do que a obtida pelos que não participaram nas mesmas e se declararam "*anti-praxe*". Ainda com base na pontuação total obtida na *EASBPES*, foi aplicado o método das *k*-médias, considerando três classes, que engobam, respetivamente, os estudantes que obtiveram pontuações: mais baixas, intermédias e mais elevadas. Foi efetuada, ainda, a CatPCA, considerando como variáveis ativas alguns itens da *EASBPES*. A projeção das categorias no espaço bidimensional apontou também para a existência de três perfis de estudantes. Com base nas coordenadas dos indivíduos nas dimensões retidas foi efetuada, ainda, a articulação entre a CatPCA e a análise classificatória.

A pontuação obtida na *EASBPES* reflete o tipo de perceção dos estudantes em relação às praxes, sendo de referir que, como era expectável, os que têm pontuações mais elevadas são em geral os que participam mais ativamente nas suas atividades e os mais favoráveis às mesmas. Os aspetos que mais influenciam a opinião dos alunos em relação às praxes são a existência ou não de agressão por atos ou palavras no decurso das mesmas e a forma como estes conseguiram lidar com estas. Assim, as experiências percecionadas dependem, em muito, do modo como decorrem as praxes, pelo que devem ser envidados esforços para a prevenção de atos abusivos e para a existência de um acompanhamento institucional no decurso das mesmas.

Referências

Matos, F., Jesus, S. Simões, H. & Nace, F. (2010) Escala para avaliação das situações de bullying nas praxes do ensino superior. Psy@w@are, 3 (1), http://www1.ci.uc.pt/ipc/2002010/revista/c6944bceb08cb00930b00b6645171101.pdf

Loureiro, C., Frederico-Ferreira, M., Ventura, M., Cardoso, E. & Bettencourt, J. (2009) A praxe académica na escola superior de enfermagem de Coimbra. *Educação/temas e problemas*, 8, 89-97.

Pimentel, M., Mata, M. & Pereira, F. (2012) Práticas iniciáticas de integração no ensino superior. Um ritual institucionalizado ou um processo de (des) integração? In *Atas do V Encontro do CIED-Escola e Comunidade*. Lisboa: Escola Superior de Educação de Lisboa, 393-401.

Silva, A. (2013). *Bullying no ensino superior: Caso da universidade do Minho- O contributo do marketing social*. Dissertação de Mestrado em Marketing e Gestão Estratégica. Universidade do Minho. Braga.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Perfis de estudantes no contexto do empreendedorismo: Análise de correspondências múltiplas e análise de *clusters*

Áurea Sousa[1], Gualter Couto[2], Nélia Branco[3], Osvaldo Silva[4], Helena Bacelar-Nicolau[5]
[1] Universidade dos Açores, CEEAplA, aurea.st.sousa@uac.pt;
[2] Universidade dos Açores, CEEAplA, gualter.mm.couto@uac.pt;
[3] CMRG, nelia.cavaco.branco@gmail.com;
[4] Universidade dos Açores, CICS.NOVA.UAc, osvaldo.dl.silva@uac.pt;
[5] Universidade de Lisboa, Faculdade de Psicologia e ISAMB/FMUL, hbacelar@psicologia.ulisboa.pt

Sumário: Um dos objetivos deste trabalho é o de aprofundar o conhecimento referente à propensão empreendedora dos estudantes da Universidade dos Açores e aferir o modo como esta pode estimular o interesse dos alunos na criação de negócios. Pretende-se, ainda, relacionar algumas das dificuldades perspetivadas pelos estudantes em relação à inicialização de um novo negócio com a sua área científica. Apresentam-se as principais conclusões obtidas com base na análise dos dados recolhidos, com a aplicação da Análise de Correspondências Múltiplas (ACM) e da Análise de *Clusters*.

Palavras-chave: Empreendedorismo, Propensão empreendedora, Análise classificatória hierárquica, Método das *k*-médias.

O empreendedorismo desempenha um papel preponderante no desenvolvimento económico de uma região. Embora a obtenção de um grau de ensino superior não seja um pré-requisito para a criação de uma empresa, muitos empreendedores reconhecem a necessidade de receber formação em áreas tais como as de gestão geral, finanças, estratégia, marketing, liderança e comunicação (e.g., Cardoso *et al.* 2015). A educação para o empreendedorismo abrange a educação com vista ao desenvolvimento de atitudes e qualidades empreendedoras (e.g., criatividade, espírito de iniciativa, autonomia, transmissão de conhecimentos fundamentais/úteis para a iniciação de um novo negócio), as quais podem ser determinantes no que se refere à decisão de criar uma empresa (Sousa *et al.* 2015).

O questionário utilizado contém (entre outras variáveis) o "Género", o "Grupo etário", a "Área científica", a "Existência de familiares empreendedores" e quatro grupos de itens, que visam aferir: o conhecimento dos estudantes a nível do empreendedorismo e a sua familiaridade com este tópico; as principais dificuldades em relação à iniciação de negócios; e a sua opinião relativamente a algumas atividades e iniciativas que a Universidade pode desenvolver a nível da sensibilização e orientação para o empreendedorismo.

A amostra é constituída por 305 estudantes da Universidade dos Açores, inscritos em cursos de diferentes áreas científicas. Foram utilizados vários métodos de Análise de

Dados Multivariados: Análise Classificatória Hierárquica Ascendente (ACHA), com base no coeficiente de afinidade (Bacelar-Nicolau 1988) e em critérios de agregação probabilísticos (*AV1*, *AVB* e *AVL*), no âmbito da metodologia *VL* (e.g., Bacelar-Nicolau 1988; Nicolau & Bacelar-Nicolau 1998), Análise de Correspondências Múltiplas (ACM) e Método das *k*-médias (*k-means*).

A ACM, cujas três primeiras dimensões explicam cerca de 83.1% da variação dos dados, fez ressaltar três conjuntos de estudantes bem definidos e permitiu-nos estudar as associações entre as categorias de algumas variáveis relevantes (oito ativas e quatro suplementares). A aplicação do método não hierárquico das *k*-médias, considerando os *scores* dos indivíduos nas três dimensões resultantes da ACM, confirmou esses três perfis de estudantes, respetivamente, com baixa (56%), média (32%) e alta (12%) propensão empreendedora.

Os modelos de ACHA aplicados sobre a sub-matriz que contém a opinião dos estudantes (1 - *Discordo totalmente*, 2 - *Discordo em parte*, 3 - *Não concordo nem discordo*, 4 - *Concordo em parte*, 5 - *Concordo totalmente*) em relação a nove iniciativas académicas promotoras do empreendedorismo, suscetíveis de serem desenvolvidas pela Universidade, permitiram-nos obter uma tipologia dessas iniciativas.

No contexto económico e social atual, a promoção do empreendedorismo pelas instituições de ensino superior assume um papel cada vez mais importante. O incremento da percentagem de diplomados no perfil com elevada propensão empreendedora poderá contribuir também para o desenvolvimento económico regional.

Referências

Bacelar-Nicolau, H. (1988) Two probabilistic models for classification of variables in frequency tables. In Bock, H.-H. (Eds.) *Classification and related methods of data analysis*. North Holland: Elsevier Sciences Publishers B.V., 181-186.

Cardoso, I., Sousa, Á. & Lopes, F. (2015) Características empreendedoras dos técnicos de diagnóstico e terapêutica dos hospitais dos Açores. In Carvalho, L.C., Dominguinhos P., Baleiras, R.N. & Dentinho, T.P. (Eds.) *Empreendedorismo e Desenvolvimento Regional Casos Práticos*. Lisboa: Edições Sílabo, 51-74.

Nicolau, F. C. & Bacelar-Nicolau, H. (1998) Some trends in the classification of variables. In Hayashi, C., Ohsumi, N., Yajima, K., Tanaca, Y., Bock, H.-H & Baba, Y. (Eds) *Data Science, Classification, and Related Methods*, Springer-Verlag, 89-98.

Sousa, Á., Couto, G., Branco, N., Silva, O. & Bacelar-Nicolau, H. (2015) Entrepreneurship education: The role of the Higher Education Institutions in the Entrepreneurial Attitudes of the Students. *ICERI2015 Proceedings*, 707-714.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

O Teste da Razão de Verosimilhanças em Modelos com Equações Estruturais: Uma Abordagem Multi-grupos ao Estudo da Privação Material em Portugal com Dados do ICOR

Paula C. R. Vicente[1], Maria de Fátima Salgueiro[2]

[1] ULHT – Escola de Ciências Económicas e das Organizações e Business Research Unit (BRU-IUL), Lisboa, Portugal, p951@ulusofona.pt;

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

Sumário: É utilizada uma abordagem multi-grupos no âmbito dos modelos com equações estruturais, com o intuito de validar a estabilidade temporal do modelo de medida da privação material, bem como do impacte sobre a mesma da existência de crianças no agregado familiar. Para efeitos de comparação de modelos, é discutida a utilização do teste da razão de verosimilhanças quando não pode ser assumida a normalidade da distribuição dos dados e é utilizado o estimador *robusto Satorra-Bentler scaled-corrected chi-square*.

Palavras-chave: Análise Longitudinal**,** ICOR, LISREL, Modelo com Equações Estruturais, *Satorra-Bentler scaled-corrected chi-square*.

Embora o estudo da pobreza e da qualidade de vida das famílias seja grande parte das vezes feito com base no rendimento disponível, esta medida pode, por si só, não ser satisfatória. O nível e a qualidade de vida das famílias podem pois ser medidos através de outros indicadores, tais como a capacidade para aceder a um conjunto de necessidades básicas, a posse de bens duradouros, as condições de habitabilidade e mesmo ainda as condições ambientais do local onde residem, ou seja, através de um conceito de privação material (Guio 2005; Whelan & Maître 2007).

O Inquérito às Condições de Vida e Rendimento das Famílias (ICOR) é um painel anual implementado pelo Instituto Nacional de Estatística com o objetivo de garantir a participação da população portuguesa na base de dados estatística europeia EU-SILC (*European Statistics on Income and Living Conditions*).

Neste trabalho são utilizados os dados transversais do ICOR, correspondentes aos anos de 2007 e 2011, com o objetivo de analisar a privação material vivida pelas famílias, assumindo este conceito como medido em várias dimensões, designadamente i) constrangimentos económicos; ii) posse de bens duradouros e iii) condições da habitação. É proposta uma abordagem longitudinal à problemática da privação material, recorrendo a um modelo com equações estruturais multi-grupos (Bollen 1989). Assim, e para cada um dos anos 2007 e 2011, é especificado o modelo de medida da privação material e o impacte sobre a mesma da existência, ou não, de crianças no agregado familiar. Uma abordagem multi-grupos permite aferir da estabilidade temporal da estrutura fatorial

proposta (peso das diferentes dimensões da privação material) e da estabilidade do impacte da existência de crianças no agregado familiar em cada um dos dois momentos temporais considerados. Note-se que nesta abordagem as amostras utilizadas em cada um dos anos (2007 e 2011) são independentes (por desenho o ICOR tem uma rotatividade de 25%) não sendo considerado o desenho amostral complexo.

A modelação estatística foi realizada com recurso ao LISREL 8.80. Face à natureza ordinal dos dados em análise, e para efeitos de estimação dos parâmetros do modelo especificado, foi utilizado o estimador *robusto Satorra-Bentler scaled-corrected chi-square*. Tal estimador foi proposto para dados cuja distribuição se desvia da normal em termos de assimetria e/ou curtose, tanto para variáveis métricas como para variáveis ordinais para as quais se assume subjacente uma variável latente métrica. Não podendo ser assumida a normalidade da distribuição dos dados, e para efeitos de comparação de modelos, não pode ser diretamente aplicado o teste da razão de verosimilhanças, uma vez que, devido ao fator de correção, a distribuição da diferença de qui-quadrados deixa de ser uma qui-quadrado. Nestas circunstâncias, Bryant & Satorra (2012) recomendam a aplicação de um fator de correção sobre a estatística do teste da razão de verosimilhanças, para a qual disponibilizaram em 2013 uma macro em EXCEL para utilizadores de LISREL, EQS e Mplus.

Neste trabalho é ilustrada e discutida a utilização do fator de correção ao teste da razão de verosimilhanças quando não pode ser assumida a normalidade da distribuição dos dados, através da utilização de um modelo com equações estruturais multi-grupos construído para validar a estabilidade temporal do modelo de medida da privação material e o impacte sobre a mesma da existência de crianças no agregado familiar.

Referências

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

Bryant, F. B. & Satorra, A. (2012) Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling*, 19(3), 372-398.

Guio, A. (2005) Material Deprivation in the EU. *Statistics in Focus, Population and Social Conditions and Welfare*, 21.

Whelan, C.T. & Maître, B. (2007) Measuring material deprivation with EU-SILC: Lessons from the Irish survey. *European Societies, 9, 147-173*.

Sessão de Apresentação de Posters – 6ª feira, 1 de abril (15:20)

Assessment of sustainable development over time in OECD and BRICKS

Nikolai Witulski[1], **José G. Dias**[2]
[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*

Abstract: Increasing environmental and social pressure leads to the rethinking of the human behavior, and ultimately to the concept of sustainability. This paper discusses briefly the concept of sustainability and its measurement within developed and developing countries. In each year, from 2000 to 2011, a set of indicators is used to measure sustainability. Longitudinal trajectories of the sustainability factor are then modelled using a second order latent growth model. We conclude that developing countries have lower sustainability levels in 2000, whereas there are no significant differences in sustainability growth between developed and developing countries.

Key-words: Aggregated index, Latent growth model, Longitudinal analysis, Sustainable development, Sustainable development indicators.

The last UN climate conference in December 2015 delivered the pressing need to adapt our current life style towards a more sustainable standards and paths for development. The concept of sustainability has evolved in the last two centuries from the seminal contributions from Malthus and Mill. In 1987, the World Commission on Environment and Development published the report "Our Common Future", which is also known as the Brundland Report. In 2000, leading scientist from different areas pointed out the emerging of a new field 'Sustainability Science' (Kates *et al.* 2000). In recent years researchers have tried to conceptualize and measure the three pillars of sustainability (economic, social/human and environmental). Many international institutions (e.g., EU, World Bank, UN) use a different set of indicators to control sustainable development. For example, in the last year the Millennium Development Goals (MDG's) ended and the United Nations established the new Sustainable Development Goals (SDG's) with a time horizon of 15 years. The SDG's contain over 17 goal categories. To achieve this SDG's the OECD provides assistance. This includes tools, analysis and approaches, which is supported by the experience of the OECD. There are seven points in which the OECD will contribute to the achievement of the goals (e.g., improving policies and the way they work together and working with all stakeholders for better policies). In complement to OECD that represents mainly developed countries, we explore the BRICKS as they define a set of developing countries with important characteristics and impact on the world economy. This paper explores the heterogeneity of the trajectory of the sustainable development in OECD and BRICK countries in the period 2000-2011.

First, we have to decide how to conceptualize and measure sustainability. The used indicators are collected from the OECD database as well as from the World Bank for the period 2000 – 2011. The Sustainable Development Strategy (SDS) indicators to monitor Sustainable Development can be used for developed and developing countries. We used four important indicators to measure sustainable development: 1) gross domestic product per domestic material consumption accounts for sustainable consumption and production; 2) life expectancy at birth measures public health status; 3) total CO2 emissions (metric tons per capita) accounts for the greenhouse gas emission of the countries; and 4) official development assistance as share of gross national income (%) represents the global partnership development over time.

The sustainability for each year is measured by these four items. Results from the confirmatory factor analysis (each year) show that the model fit is good and these four indicators share the same underlying dimension. A second-order latent growth model is considered to describe growth trajectories of sustainability between 2000 and 2011 in OECD and BRICK countries. Then, the longitudinal sequence of measurements is linked by a latent growth model. The conditional model adds a dummy variable that indicates whether a given country is either developed or developing. The country classification comes from the official list of the International Monetary Fund. Results show that the intercept (sustainable levels in 2000) and slope (the difference between consecutive years in the sustainability level) are not associated. Sustainability growth trajectories are not significantly different for developed and developing countries between 2000 and 2011. However, sustainability levels in year 2000 are significantly lower for developing countries (-0.763, $p < 0.000$).

References

Böhringer, C. & Jochem, P. E. P. (2007) Measuring the immeasurable — A survey of sustainability indices. *Ecological Economics*, 63(1), 1-8.

Bolcárová, P. & Kološta, S. (2015) Assessment of sustainable development in the EU 27 using aggregated SD index. *Ecological Indicators*, 48, 699-705.

Mirshojaeian Hosseini, H., & Kaneko, S. (2011) Dynamic sustainability assessment of countries at the macro level: A principal component analysis. *Ecological Indicators*, 11(3), 811-823.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Gruebler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., Mooney, H., Moore III, B., O'Riordan, T. & Svedin, U. (2000) *Sustainability Science. Research and Assessment Systems for Sustainability Program Discussion Paper, 2000-33* Cambridge, MA: Environment and Natural Resources Program, Belfer Center for Science and International Affairs, Kennedy School of Government, Harvard University.

Conflito trabalho-família: Validação de um instrumento de medida para a Marinha Portuguesa

Sandra Veigas Campaniço[1], Dora Carinhas[2], Miguel Pereira Lopes[3]
[1] Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa; CINAV – Centro de Investigação Naval, sandra.patricia.campanico@marinha.pt;
[2] Instituto Hidrográfico, dora.carinhas@hidrografico.pt;
[3] Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa; CAPP – Centro de Administração e Políticas Públicas, mplopes@iscsp.ulisboa.pt

Sumário: O conflito trabalho-família resulta da dificuldade que o indivíduo tem em conseguir gerir as responsabilidades familiares e profissionais que surgem em simultâneo, podendo este conflito ser afectado por factores como o suporte do líder ou o ambiente de trabalho de suporte à família. Pretendeu-se avaliar, para a população de oficiais da Marinha Portuguesa, o ajustamento de um modelo de medida integrativo do conflito trabalho-família, suporte do líder e ambiente de trabalho de suporte à família a partir das escalas validadas para estes constructos existentes na literatura de referência.

Palavras-chave: Ambiente de trabalho de suporte à família, Conflito trabalho-família, Instrumento de medida, Marinha Portuguesa, Suporte do líder.

A Marinha Portuguesa, enquanto organização inserida na realidade dos dias de hoje e assente na condição militar dos elementos que integram as suas fileiras, não está alheia ao desenvolvimento do conflito trabalho-família. Fruto das especificidades das funções e missões desempenhadas, os militares deste ramo das Forças Armadas podem desempenhar as suas tarefas nos mais diferentes contextos, sendo a deslocalização geográfica em relação ao seu local habitual de residência uma realidade para estes militares. Face ao acima descrito, a avaliação da percepção quanto à existência do conflito trabalho-família reveste-se de grande relevância entre os militares da Marinha Portuguesa, sendo igualmente relevante avaliar em que medida factores como o suporte do líder ou o ambiente de trabalho de suporte à família podem ter impacto no desenvolvimento deste conflito. Para que tal seja possível, é necessária a existência de um instrumento de medida que permita, de forma integrada, avaliar a percepção individual quanto a estas três variáveis.

Com base nas escalas identificadas na literatura para o conflito trabalho-família (Carlson *et al.* 2000), suporte do líder (Thomas & Ganster 1995) e ambiente de trabalho de suporte à família (Allen 2001), foi construído um questionário para avaliação da percepção quanto a estas três variáveis entre a população de Oficiais dos Quadros Permanentes da Marinha Portuguesa. O questionário é constituído por três grupos de itens, de 1 a 18, de 19 a 32 e de 33 a 41 que correspondem, respectivamente, às escalas de Carlson (2000), de Allen (2011) e de Thomas & Ganster (1995) apresentados numa escala do tipo Likert de 6 níveis, que variam entre 1 (“discordo totalmente”) e 6 (“concordo

totalmente"). Os dados foram sujeitos a um processo de determinação do ajustamento do modelo de medida global com as aplicações informáticas IBM SPSS 23 e AMOS 23, através da avaliação do ajuste global do modelo e da fiabilidade dos resultados obtidos. A estrutura relacional subjacente às perceções foi avaliada a partir de uma Análise Fatorial Exploratória (AFE), sobre a matriz de correlações. Através da Análise Fatorial Confirmatória (AFC) foi verificada a hipótese de que determinados fatores latentes explicam o comportamento de variáveis manifestas. A avaliação da qualidade global do modelo envolveu a apreciação da capacidade que o modelo teórico proposto tem para reproduzir a estrutura correlacional das variáveis observadas na amostra (Maroco 2007). Assim a avaliação do ajuste global do modelo foi efectuada através da determinação dos índices de ajuste absoluto (raiz quadrada média do erro de aproximação, RMSEA), incremental (índice de ajuste comparativo, CFI) e de parcimónia (χ^2 normalizado). A fiabilidade interna, ou seja, a avaliação da qualidade do instrumento de medida, foi realizada através do coeficiente alfa de Cronbach. Os resultados referentes à avaliação do ajustamento do modelo de medida global construído com base nas três escalas acima referidas são apresentados na seguinte tabela.

Tabela 1: Determinação dos parâmetros de ajustamento e fiabilidade de resultados do questionário aplicado

Determinação	WFC	FSWE	LID	Modelo de medida global
RMSEA	0,050	0,09	0,099	0,048
CFI	0,977	0,815	0,969	0,933
χ^2 normalizado	1,583	2,862	3,251	1,517
Alfa de Cronbach	0,892	0,771	0,737	0,863

Legenda: WFC – Conflito trabalho-família, FSWE – ambiente de trabalho de suporte à família, LID – suporte do líder

Referências

Allen, T. D. (2001) Family-supportive work environments; The role of organization perceptions. *Journal of Vocational Behavior*, 58, 414-435.

Carlson, D. S., Kacmar, K. M. & Williams, L. J. (2000) Construction and initial validation of a multidimensional measure for work-family conflict. *Journal of Vocational Behavior*, 56, 249-276.

Maroco, J. (2007) Análise Estatística com utilização do SPSS (3ª Ed.). Edições Sílabo.

Thomas, L. T. & Ganster, D. C. (1995) Impact of family-supportive work variables on work-family conflict and strain: A control perspective. *Journal of Applied Psychology*, 80(1), 6-15.

Factores influentes no sucesso *vs.* insucesso nas escolas da Província do Cunene

Palmira Caseiro[1], Helena Bacelar-Nicolau[2], Jorge Santos[3], Fernando da Costa Nicolau[4]

[1] *caseiropalmira@hotmail.com;*
[2] *hbacelar@psicologia.ulisboa.pt;*
[3] *jmas@uevora.pt;*
[4] *geral@datascience.org*

Sumário: Esta apresentação faz parte de um projeto que estuda a situação do Sistema Educativo de Angola, através da Análise dos Dados fornecidos pelos mapas de aproveitamento escolar e questionários realizados nas Escolas. Foram aqui analisadas, por modelos de Análise Classificatória Hierárquica, Clássicos e Probabilísticos, as Escolas do Ensino obrigatório pertencente cada um dos 6 Municípios da Província do Cunene descritos por 23 variáveis definidas a partir dos Indicadores Básicos da Educação, Demográficos, Geográficos entre outros. Paralelamente fez-se a Análise da Eficiência das escolas.

Palavras-chave: Análise Classificatória, Análise da Eficiência, Analise Factorial das Correspondências, Indicadores Básicos da Educação, Indicadores Demográficos e Geográficos.

Considerando a vontade de realizar a Escolarização Obrigatória em Angola, o Ministério da Educação e Cultura e algumas Agências do Sistema das Nações Unidas estabeleceram o Plano – Quadro Nacional de Reconstrução do Sistema Educativo.

A estratégia integrada desse plano era melhorar a qualidade do Sistema Educativo para o período de 2001-2015. Na primeira etapa do plano, a Província do Cunene foi abrangida, possibilitando assim o levantamento de toda informação referente aos Indicadores Básicos da Educação, Indicadores Demográficos e Geográficos.

Com a finalidade de analisar a Estrutura Educativa de cada Município da Província do Cunene usamos a seguinte Metodologia: Analise Factorial das Correspondências (AFC) e Classificação Hierárquica Ascendente (CHA).

Foram analisadas 430 escolas do I Nível (Ensino primário: 1ª, 2ª, 3ª e 4ª classes), a fim de identificarmos os factores influentes no Sucesso vs Insucesso nas Escolas de cada Município, bem como procurar a relação existente entre Formação Académica dos Professores, Indicadores Básicos da Educação, Indicadores Sociais, Geográficos e Demográficos. Aplicou-se também a técnica de Data Envelopment Analysis para calcular a eficiência das escolas.

A partir das classes obtidas pelas classificações e também dos factores da AFC, podemos compreender que os resultados obtidos pelos métodos multivariados utilizados são concordantes ou complementam-se, traduzindo uma estrutura relativamente forte.

Um outro resultado importante refere-se ao índice elevado de alunos reprovados e desistentes em todos municípios da província do Cunene nomeadamente na 2ª, 3ª e 4ª classe. O motivo de insucesso escolar deve-se ao facto de muitos desses alunos em idade escolar ajudarem seus pais nas tarefas domésticos ou sustento da família. Um outro motivo de insucesso é devido à localização geográfica do município favorecendo a fuga dos alunos para outras localizações.

De entre os resultados e estudos acima citados podemos concluir que:

O processo de localização de escolas e a procura de áreas para construção de novas escolas revelou-se ser prioridade;

A análise comparativa entre os municípios permitiu compreender a relação entre as características Demográficas, Geográficas e a taxa de sucesso e insucesso, com base nos factores principais e nas classes das tipologias encontradas;

Os métodos da Análise de Dados Multivariados utilizados constituem um caminho metodológico importante para estudos futuros utilizando dados de natureza complexa.

A Análise da Eficiência complementa satisfatoriamente os resultados e conclusões dos métodos da Análise Multivariada.

Referências:

Bacelar-Nicolau, H., Nicolau, F., Sousa, A., Bacelar-Nicolau, L. (2014) Clustering of variables with a three-way approach for health sciences. *Testing, Psychometrics, Methodology in Applied Psychology (TPM)*, 21(4 Special Issue), 435-447.

Santos, J. & Caseiro, P. (2011) Introdução às Técnicas de Data Envelopment Analysis. *Actas do encontro: Modelos de Apoio à Decisão na Agricultura e Ambiente*, Universidade dos Açores.

Santos, J., Cavique, L. & Mendes, A. (2013) Super-efficiency and Multiplier Adjustment in Data Envelopment Analysis. In Mendes, A. B., Soares da Silva, E. L. D. G & Santos, J. M. A. (Eds.) *Efficiency Measures in the Agricultural Sector*. Springer. ISBN 978-94-007-5738-7.

Sousa, A., Bacelar-Nicolau, H. & Silva, O. (2014) Cluster Analysis of Business Data. Asian Online Journals: *Asian Journal of Business and Management*, 2(1), 18-26. ISSN: 2321-2802.

Sessão de Posters – sábado, 2 de abril (15:35)

Famílias estruturadas de matrizes estocásticas simétricas

Cristina Dias[1], Carla Santos[2], João Tiago Mexia[3]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Portalegre e CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa, cpsilvadias@gmail.com;

[2] Departmento de Matemática e Ciências Físicas do Instituto Politécnico de Beja e CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa, carla.santos@ipbeja.pt;

[3] Departamento de Matemática da Faculdade de Ciências e Tecnologia e CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa, jtm@fct.unl.pt

Sumário: Neste trabalho estudamos famílias estruturadas de modelos, cujas matrizes correspondem aos tratamentos de um delineamento base.
Consideramos ainda famílias de modelos divididas em subfamílias que correspondem a esses tratamentos. Estamos sobretudo interessados em modelos base com estrutura ortogonal. Apresentamos essa estrutura e mostramos como aplicar esses modelos no estudo de famílias estruturadas.

Palavras-chave: Estrutura ortogonal, Famílias estruturadas, Subfamílias.

As matrizes de uma família estruturada de matrizes estocásticas simétricas são todos da mesma ordem k e correspondem aos tratamentos de modelos base. Uma vez que as matrizes têm todas a mesma ordem, estamos perante o caso equilibrado em que temos o mesmo número de graus de liberdade para o erro para cada tratamento. A ANOVA e técnicas relacionadas são, no caso equilibrado, técnicas robustas para o caso de heterocedasticidade e ainda mais para o caso da não-normalidade, ver Scheffé (1959). Recorde-se que, para modelos individuais de matrizes simétricas, uma formulação interessante é apresentada em Areia (2009). Admitindo que a série de estudos tem uma estrutura comum de grau $r < k$ considerarmos o modelo

$$\mathbf{M} = \sum_{i=1}^{r} \boldsymbol{\beta}_i \boldsymbol{\alpha}_i^t + \overline{\mathbf{E}},$$

onde $\overline{\mathbf{E}} = \frac{1}{2}\left(\mathbf{E} + \mathbf{E}^t\right)$ com $vec(\mathbf{E}) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{k^2})$. Sejam $\theta_1 > \theta_2 > \ldots > \theta_k$ os valores próprios da matriz $\mathbf{M}$ que correspondem aos vetores próprios $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k$. Então podemos estimar o primeiro vetor de estrutura ver Oliveira e Mexia (1999) por $\boldsymbol{\beta}_1 = \theta_1 \boldsymbol{\gamma}_1$, e a soma dos quadrados dos erros é dada por

$$V = \|\mathbf{M}\|^2 - \sum_{j=1}^{r} \boldsymbol{\gamma}_j^2 ,$$

quando $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r$ são os primeiros vetores próprios.

Tem-se $g = \frac{k(k+1)}{2} - kr$, graus de liberdade para o erro, uma vez que se tem $\frac{k(k+1)}{2}$ componentes "livres" de $\overline{\mathbf{E}}$ e kr parâmetros, para estimar as componentes dos vetores $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_r$, vindo $V \sim \sigma^2 \chi^2_g$.

Para os modelos ajustados, temos o primeiro vetor de estrutura ajustado $\boldsymbol{\beta}_{1,1}, \boldsymbol{\beta}_{2,1},\ldots,\boldsymbol{\beta}_{m,1}$.

E a soma das somas dos quadrados dos erros

$$V(r) = \sum_{l=1}^{m} V_l(r),$$

com $V_1(r), V_2(r),\ldots,V_m(r)$ são a soma individual de quadrados dos erros.

Consideramos dois tipos principais de inferência:

Transversal: em que trabalhamos com as componentes homólogas dos vetores estrutura. Uma vez que existem *K* componentes para cada vetor realizamos esta análise *k* vezes.

Longitudinal: em que trabalhamos com vetores de contraste sendo nula a soma das suas componentes. Este tipo de análise é útil quando as componentes correspondentes dos vetores de estrutura são obtidas simultaneamente em momentos sucessivos.

Outro ponto de interesse será quando podemos fazer variar o grau *r* para as matrizes do modelo. Quando *r* é muito baixo as somas de quadrados para o erro terá um parâmetro de não centralidade bastante grande e a potência dos testes F irá diminuir (ver Mexia, 1989).

Acknowledgments: This work was partially supported by national founds of FCT-Foundation for Science and Technology under UID/MAT/00297/2013 and UID/MAT/00212/2013.

Referências

Areia, A. (2009) *Séries emparelhadas de estudo*. Tese de Doutoramento, Universidade de Évora.

Mexia, J. (1989) *Controlled Heterocedasticity, Quotient Vector Spaces and F Tests for Hypotheses on Mean Vectors.* Trabalhos de Investigação, nº2, Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

Oliveira M. M. & Mexia, J. (1999) F Tests for Hypothesis on the Structure Vectors of Series. *Discussiones Mathematicae*. 19(2), 345-353.

Scheffé, H. (1959) *The Analysis of Variance.* New York: John Wiley & Sons.

O sono das crianças do 1º ciclo: caso de estudo numa escola do concelho de Évora

Paulo Infante[1], **Anabela Afonso**[2], **Gonçalo Jacinto**[3], **Teresa Engana**[4], **Filipe Gloria Silva**[5], **Rosa Espanca**[6]

[1] *CIMA/IIFA e DMAT/ECT, Universidade de Évora, pinfante@uevora.pt;*
[2] *CIMA/IIFA e DMAT/ECT, Universidade de Évora, aafonso@uevora.pt;*
[3] *CIMA/IIFA e DMAT/ECT, Universidade de Évora, gjcj@uevora.pt;*
[4] *Associação de Pais e Encarregados de Educação da Escola Básica do Chafariz D`El Rei, eedochafariz2012@hotmail.com;*
[5] *Centro da Criança e do Adolescente, Hospital Cuf Descobertas, Lisboa, fs.sono@gmail.com;*
[6] *ACES Alentejo Central, Rosa.Espanca@alentejocentral.min-saude.pt*

Sumário: A privação do sono contribui para vários problemas de saúde, emocionais e comportamentais, bem como para o insucesso escolar. Neste trabalho avaliamos os hábitos e problemas do sono de um grupo de crianças do 1º ciclo e comparam-se alguns resultados com os obtidos com outros estudos realizados em Portugal, em idades similares.

Palavras-chave: 1º ciclo, Correlação, Regressão logística, Sono, Testes não paramétricos.

O sono está cada vez mais na ordem do dia. Para chamar a atenção para a sua importância foi definido que na sexta-feira da segunda semana completa de março se comemoraria o dia Mundial do Sono, que este ano será no dia 18 de março.

Vários estudos alertam para os efeitos negativos associados à privação do sono. Nas crianças para além das consequências na regulação emocional e do comportamento, nas funções cognitivas, no sucesso académico, na obesidade e no risco de quedas acidentais, também se repercutem na vida dos pais (Silva *et al.* 2012). Os padrões e ciclos de sono variam à medida que as crianças se desenvolvem, sendo influenciados por fatores de ordem cultural e ambiental. Para dormir bem é fundamental que a criança tenha bons hábitos de sono. Segundo a *World Association of Sleep Medicine*, uma criança entre os 6 e os 10 anos deve dormir entre a 10 a 11 horas por noite.

Para avaliar os hábitos e problemas do sono mais comuns de um grupo de crianças do concelho de Évora foi utilizada a "versão-curta" portuguesa do *Children's Sleep Habits Questionnaire* (Silva *et al.* 2012). Solicitou-se aos pais dos alunos do 1º ciclo de uma escola no concelho de Évora que preenchessem o questionário com base no padrão de sono do seu filho, durante a última semana. Foram ainda colocadas questões que permitissem caracterizar o agregado familiar e o encarregado de educação. Em Maio de 2015 foram distribuídos 98 inquéritos (1 por cada criança que frequentava a escola), tendo sido respondidos e considerados válidos para análise 88. A idade das crianças varia entre os 6 e os 11 anos, com uma média de 7,9 anos e um desvio padrão de 1,2 anos. A maior taxa de

não respostas por questão foi inferior a 8%.

Com base nos dados recolhidos, estas crianças dormem entre 8 a 12 horas por noite, sendo a média de 9 horas e 41 minutos e o desvio-padrão de 48 minutos. Apesar de 46,4% dormirem menos das 10 horas diárias recomendadas, apenas 6,3% dos pais "acha que o seu filho/filha tem algum problema com o sono". Verificou-se que existe relação entre o número de horas de sono das crianças (relatadas pelos pais) e a perceção destes sobre se a criança dorme pouco. A resistência da criança em ir para a cama e as parassónias (comportamentos peculiares que ocorrem durante o sono) influenciam negativamente o número de horas de sono das crianças. Contudo, não se verificou associação significativa do número de horas de sono com o sexo, com o horário de trabalho do encarregado de educação ser flexível ou não e com a entidade patronal possibilitar ou não horários adaptados aos da criança.

Com base nas respostas dadas às várias questões foi calculado o Índice de Perturbação do Sono. Este índice assume valores similares em ambos os sexos ($p=0,157$), conclusão diferente da obtida Silva *et al*. (2012), e não difere entre as idades ($p=0,739$). Também se obtiveram correlações positivas com a maior parte das escalas do questionário sendo, em geral, mais acentuadas que as obtidas em Silva *et al*. (2012).

Considerando o número de horas de sono da criança reportadas pelos pais, foi criada uma variável dicotómica que distingue entre as crianças "com défices nas horas de sono" (menos de 10 horas de sono diárias) das crianças "com número regular de horas de sono" (com pelo menos 10 horas de sono diárias). Recorrendo a um modelo de regressão logística concluiu-se que são fatores potenciadores de um sono deficiente, a criança ser do sexo masculino, ter um maior índice de sonolência diurna e ter um maior índice de resistência em ir para a cama.

Através da análise em componentes principais foi possível resumir a informação das 8 subescalas do Índice de Perturbação do Sono em 3 componentes principais e identificados 4 perfis de crianças: 1) 2 crianças que se destacam das restantes porque os seus problemas no sono são devidos à ansiedade e ao facto de terem muita resistência em ir para a cama; 2) 22 crianças que para além da ansiedade e resistência, sofrem de sonolência diurna; 3) 30 crianças que cujo principal distúrbio do sono é a sonolência diurna; e 4) 34 crianças "sem problemas do sono".

Concluímos que, apesar de pouco reconhecidos pelos pais, os problemas comportamentais do sono desta amostra são frequentes e têm consequências em termos de sonolência diurna, pelo que não podem ser negligenciados.

Agradecimentos: A. A., G. J. e P. I. são membros do CIMA (UID/MAT/04674/2013), financiado pela FCT.

Referências

Silva, F. G., Silva, C. R, Braga, L. B. & Neto, A. S (2013) Hábitos e problemas do sono dos dois aos dez anos: estudo populacional. *Acta Pediátrica Portuguesa*, 44(5), 196-202.

Sessão de Posters – sábado, 2 de abril (15:35)

A Bayesian LASSO method for replicated data

Jacinto Martín[1], Carlos J. Pérez[2], Lizbeth Naranjo[3], Yolanda Campos-Roca[4]

[1] Departamento de Matemáticas and Instituto de Computación Científica Avanzada (ICCAEx), Universidad de Extremadura (Spain), jrmartin@unex.es;

[2] Departamento de Matemáticas, Universidad de Extremadura (Spain), carper@unex.es;

[3] Facultad de Ciencias, Universidad Nacional Autónoma de México (Mexico), lizbeth@sigma.iimas.unam.mx;

[4] Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura (Spain), ycampos@unex.es

Summary: This work deals with the problem of selection and classification methods considering replicated data. The problem that motivates this work consists of discrimination of people suffering from Parkinson's disease from healthy subjects based on acoustic features automatically extracted from replicated voice recordings. The statistical model has been specified in such a way that it is possible to explore the posterior distribution of the parameters of interest via Gibbs sampling algorithm.

Keywords: Bayesian binary regression, Gibbs sampling algorithm, Parkinson's disease, Variable selection, Voice features.

People with Parkinson's disease (PD) exhibit a chronic neurological disorder caused by the progressive degeneration and death of dopaminergic neurons. These neurons play a key role in coordinating movements at level of muscular tone. An estimated 7 to 10 million people worldwide are living with this health condition.

Voice and speech, as dependent on laryngeal, respiratory and articulatory functions, are also affected in people with PD. Vocal impairment is hypothesized to be one of the earliest signs of the disease (Duffy 2005). Successfully addressing early diagnosis of people with PD is a key issue to improve patients' quality of life.

The voices of the subjects can be recorded to extract some specific features of the signals and can be used to classify individuals by using classification approaches. The Parkinson's Voice Initiative has played an important role in the spread of this topic.

In this context, it has been usual to make replicated voice recordings for each individual and treat the extracted features with independence-based classification methods (see, e.g., Little *et al.* (2009)). Since features are extracted from multiple voice recordings from the same subject, in principle, the features for each individual should be identical. Technological imperfections and the biological variability result in non-identical replicated features that are more similar to one another than features from different subjects. This within-subject variability must be statistically addressed which is not usually considered in the scientific literature.

Naranjo *et al.* (2016) proposed classification approaches for PD detection that took into account the underlying within-subject dependence. They conducted an experiment to discriminate PD subjects from healthy individuals by considering replicated voice recordings. A total of 40 people affected by PD and 40 healthy individuals were considered. The research protocol consisted of a brief questionnaire and three recording replications of the sustained /a/ phonation, leading to a total of 240 voice recordings. Each voice recording was processed to provide 44 acoustic features, i.e. a 44-dimensional vector per voice recording.

This paper extends the previous work considering variable selection. A regularization-based classification approach based on LASSO (Least Absolute Shrinkage and Selection Operator) is proposed. This approach integrates ideas from Genkin *et al.* (2007) and Naranjo *et al*. (2016). The LASSO method does not remove variables, but favours the best predictors and penalizes the worst ones through the regularization parameters. Computational difficulties are avoided by developing an easy-to-implement Gibbs sampling-based algorithm.

The proposed approach gains in interpretability with respect to the one in Naranjo *et al.* (2016), because many redundant variables have been regularized (penalized or favoured) through the regression parameters. This increases the estimation reliability. Furthermore, it exhibits a better chain mixing. It also allows a different number of replications per subject, which may be the case when not all voice recordings can be properly processed for some individuals.

Acknowledgments: Thanks to the anonymous participants and to Carmen Bravo and Rosa María Muñoz for carrying out the voice recordings and providing information from the people with PD. We are grateful to the *Asociación Regional de Parkinson de Extremadura* and *Confederación Española de Personas con Discapacidad Física y Orgánica* for providing support in the experimente development. This research has been supported by *Ministerio de Economía y Competitividad*, Spain (Project MTM2014-56949-C3-3-R), *Gobierno de Extremadura,* Spain (Projects GR15052 and GR15106), and *European Union* (European Regional Development Funds).

References

Duffy, J. R. (2005) *Motor Speech Disorders: Substrates, Differential Diagnosis and Management*. Elsevier.

Genkin, A., Lewis, D. D. & Madigan, D. (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291–304.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J. & Ramig, L. O. (2009) Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015–1022.

Naranjo, L., Pérez, C. J., Campos-Roca, Y. & Martín, J. (2016) Addressing voice recording replications for Parkinson's disease detection. *Expert Systems With Applications*, 46, 286–292.

Caracterização das explorações agrícolas e dos produtores de caprinos de raça Serpentina

Manuel Minhoto[1], António Fonseca[2], Luís Fernandes[3], António Cachatra[4]
[1] CIMA/IIFA e DMAT/ECT, Universidade de Évora, minhoto@uevora.pt;
[2] Universidade de Évora - ECT-DSD, dfonseca@uevora.pt;
[3] Universidade de Évora - ECT-DZoo, ladsf@uevora.pt;
[4] Associação Portuguesa de Caprinicultores de Raça Serpentina, associacao.serpentina@gmail.com

Sumário: O trabalho pretende analisar as principais características das empresas e dos produtores de caprinos da raça Serpentina, assim como as razões por estes indicadas para a continuidade desta atividade nas suas explorações. Os dados foram obtidos por inquérito aos produtores, recorrendo-se ao Escalonamento Ótimo (*Optimal Scaling*) englobado no SPSS e à seleção de variáveis observadas (package *subselect* do programa estatístico R).

Palavras-chave: Cabra Serpentina, *Optimal Scaling*, Produtores, Seleção de variáveis observadas.

O trabalho tem por objetivo mostrar algumas características das empresas agrícolas e dos produtores atualmente registados na Associação Portuguesa de Caprinicultores de Raça Serpentina (APCRS), assim como analisar as razões para a continuidade desta atividade de produção caprina nos planos de exploração das suas empresas agrícolas.

A informação foi recolhida através de inquéritos aos produtores realizados em 2014 por técnicos da APCRS, com base num modelo de questionário que havia sido preparado para um trabalho realizado na Associação de Criadores de Bovinos de Raça Marinhoa e que foi adaptado para a raça Serpentina. Foram registados 28 questionários válidos e um efetivo global de 3929 fêmeas reprodutoras da raça Serpentina, o que representava cerca de 72% das empresas e 82% das fêmeas reprodutoras que integravam a APCRS à data da realização dos inquéritos.

As respostas às questões relacionadas com a caracterização das empresas e dos produtores foram agrupadas em cinco níveis e nas razões para a continuidade da atividade foi aplicada a escala de *Likert*.

As variáveis consideradas ao nível das empresas foram a área da exploração agrícola, o número de fêmeas reprodutoras Serpentinas, o número de cabeças normais pecuárias e a densidade animal por hectare forrageiro; ao nível do produtor as variáveis foram a idade, a escolaridade, o tempo dedicado à empresa agrícola, o contributo dos apoios financeiros para os proveitos da empresa e o contributo da empresa agrícola para o rendimento total do agregado familiar do produtor.

Foram obtidas e interpretadas as matrizes de correlações para os seguintes grupos de variáveis: (i) informação relativa às explorações agrícolas e aos produtores, (ii) razões para a continuidade da atividade de produção de caprinos de raça Serpentina e (iii) características das explorações agrícolas e dos produtores e razões para a continuidade da atividade.

Para a Análise Multivariada foram constituídas duas bases de dados: a primeira englobando as características das explorações agrícolas e dos produtores e a segunda englobando as razões para a continuidade da produção de caprinos de raça Serpentina. Começou-se por recorrer ao programa estatístico SPSS. Escolheu-se a opção "*Optimal Scaling*" e como método "*CATPCA*" pois nem todas as variáveis são do mesmo tipo. Para critério de normalização selecionou-se o método das "*variáveis principais*". Procurou-se ainda interpretar os resultados nas duas bases através da seleção de variáveis observadas. Para tal optou-se pelo pacote *subselect* do programa estatístico R. Deste pacote selecionou-se o critério RM que toma valores entre 0 e 1, permitindo selecionar o subconjunto de variáveis observadas que melhor representa a totalidade destas. Dado um subconjunto de variáveis, o quadrado do valor deste critério (se expresso em percentagem) representa a percentagem de variabilidade explicada por esse subconjunto. Deste modo, num total de, por exemplo, dez variáveis observadas, pode-se comparar a percentagem de variabilidade explicada pelo melhor subconjunto de três variáveis observadas com a percentagem de variabilidade explicada pelas duas ou três primeiras componentes principais.

Agradecimentos: à Direção da APCRS e aos seus técnicos Engenheiros Zootécnicos Paulo Carreira e Victor Saraiva.

Referências

Cadima, J., Cerdeira, J. O. & Minhoto, M. (2004) Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47, 225-236.

Ferreira, E., Fernandes, L., Minhoto, M., Roquete, C. & Ferreira, P. (2014) Contributo para a Caracterização dos Criadores e Explorações Agrícolas Produtoras de Bovinos de Raça Marinhoa. In *Actas do 20º Congresso da Associação Portuguesa de Desenvolvimento Regional*, 1363-1370.

Lira, S. & Neto, A. (2006) Coeficientes de correlação para variáveis ordinais e dicotómicas derivados do coeficiente linear de Pearson, *RECIE*, Uberlãndia, vol. 15, n. 1/2, 45-53.

R Development Core Team (2012) *R: A language and environment for statistical computing*. R Fundation for Statistical Computing, Vienna, Austria, ISBN 3-9300051-07-0, URL http://www.R-project.org.

Sessão de Posters – sábado, 2 de abril (15:35)

Estimação e condensação em modelos mistos, normais e não normais, com estrutura ortogonal por blocos

Carla Santos[1], Célia Nunes[2], Cristina Dias[3], João Tiago Mexia[4]

[1] Departamento de Matemática e Ciências Físicas do Instituto Politécnico de Beja e CMA – Centro de Matemática e Aplicações, Universidade Nova de Lisboa, carla.santos@ipbeja.pt;

[2] Departamento de Matemática e Centro de Matemática e Aplicações, Universidade da Beira Interior, celian@ubi.pt;

[3] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Portalegre e CMA-Centro de Matemática e Aplicações, Universidade Nova de Lisboa, cpsilvadias@gmail.com;

[4] Departamento de Matemática da Faculdade de Ciências e Tecnologia e CMA – Centro de Matemática e Aplicações, Universidade Nova de Lisboa, jtm@fct.unl.pt

Sumário: Com base na estrutura algébrica dos modelos lineares mistos com estrutura ortogonal por blocos (OBS), assente em álgebras de Jordan comutativas, condensamos OBS obtendo novos modelos com menos observações que os OBS que lhe deram origem mas com os mesmos parâmetros. As propriedades dos estimadores para OBS, normais ou não-normais, mantêm-se para os modelos condensados uma vez que também estes são OBS.

Palavras-chave: Condensação, Estimação, Estrutura ortogonal por blocos, Modelos lineares mistos.

Um modelo linear misto, na sua forma usual, será representado por

$$\boldsymbol{Y} = \sum_{i=0}^{w} \boldsymbol{X}_i \boldsymbol{\beta}_i \qquad (1)$$

onde $\boldsymbol{Y}$ é o vector das observações da variável dependente, $\boldsymbol{X}_i$ é a matriz dos valores das variáveis explicativas, $\boldsymbol{\beta}_0$ é fixo e os $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_w$ são vectores aleatórios independentes com vector médio nulo e matrizes de covariância $\sigma_1^2 \boldsymbol{I}_{c_1} \dots \sigma_w^2 \boldsymbol{I}_{c_w}$, onde $c_i = car(\boldsymbol{X_i}), i = 1, \dots, w$.

$\boldsymbol{Y}$ tem vector médio $\boldsymbol{\mu} = \boldsymbol{X}_0 \boldsymbol{\beta}_0$ e matriz de covariância $\boldsymbol{V} = \sum_{i=1}^{w} \sigma_i^2 \boldsymbol{M}_i$, onde $\boldsymbol{M}_i = \boldsymbol{X}_i \boldsymbol{X}_i^T, i = 1, \dots, w$. Sendo Ω o espaço imagem da matriz $\boldsymbol{X}_0$, $R(\boldsymbol{X}_0)$, a matriz de projecção ortogonal (MPO) sobre Ω é $\boldsymbol{T} = \boldsymbol{X}_0 (\boldsymbol{X}_0^T \boldsymbol{X}_0)^+ \boldsymbol{X}_0^T$, onde $\boldsymbol{A}^+$ representa a inversa de Moore-Penrose de $\boldsymbol{A}$. Quando as matrizes da família $M = \{\boldsymbol{M}_1, \dots, \boldsymbol{M}_w\}$ comutam elas geram uma Álgebra de Jordan comutativa de matrizes simétricas, AJCS, $\mathcal{A}(M)$, isto é, um espaço linear constituído por matrizes simétricas que comutam e que contêm o quadrado das suas matrizes. Toda a AJCS tem uma única base, a sua base principal, constituída por matrizes de projecção ortogonal mutuamente ortogonais, MPOMO, (Seely 1971). Sendo $Q = \{\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_m\} = bp(\mathcal{A}(M))$, a base principal da AJCS $\mathcal{A}(M)$, temos $\boldsymbol{M}_i = \sum_{j=1}^{m} b_{i,j} \boldsymbol{Q}_j$, $i = 1, \dots, w$, com $\boldsymbol{B} = [b_{i,j}]$, a matriz de transição entre M e Q. Então $\boldsymbol{V} = \boldsymbol{V}(\gamma)$, com $\boldsymbol{V}(\gamma) = \sum_{j=1}^{m} \gamma_j \boldsymbol{Q}_j$ e $\boldsymbol{\gamma} = \boldsymbol{B}^T \boldsymbol{\sigma}^2$, com $\boldsymbol{\sigma}^2$ o vector das componentes de variância.

Um modelo misto é um OBS (modelo com estrutura ortogonal por blocos) quando $\sum_{j=1}^{m} \boldsymbol{Q}_j = I_m$, o que ocorre quando $\mathcal{A}(M)$ contém matrizes invertíveis e $\boldsymbol{\gamma}$ varre a família de vectores m dimensionais com componentes não negativas. Esta última condição

verifica-se quando a matriz $\boldsymbol{B}$ é invertível. Então temos $m = w$.

Consideremos $\boldsymbol{Y}_j = \boldsymbol{A}_j\boldsymbol{Y}$, $\boldsymbol{\mu}_j = \boldsymbol{A}_j\boldsymbol{\mu}$ e $\boldsymbol{X}_{0,j} = \boldsymbol{A}_j\boldsymbol{X}_0$, $j = 1, \dots, m$ e que os vectores linha de $\boldsymbol{A}_j$ constituem um espaço ortogonal para o espaço imagem de $\boldsymbol{Q}_j, j = 1, \dots, m$.

Sejam $\boldsymbol{P}_j$ $[\boldsymbol{P}_j{}^{\perp}]$ as MPO no espaço imagem de $\boldsymbol{X}_{0,j}$, Ω_j, [no complemento ortogonal de Ω_j], $j = 1, \dots, m$. Com $q_j = car(\boldsymbol{Q}_j)$ e $p_j = car(\boldsymbol{P}_j)$ teremos $p_j{}^{\perp} = q_j - p_j = car(\boldsymbol{P}_j{}^{\perp})$ assim como $\boldsymbol{Q}_j = \boldsymbol{Q}_j(1) + \boldsymbol{Q}_j(2)$, com $\boldsymbol{Q}_j(1) = \boldsymbol{A}_j{}^T\boldsymbol{P}_j\boldsymbol{A}_j$ e $\boldsymbol{Q}_j(2) = \boldsymbol{A}_j{}^T\boldsymbol{P}_j{}^{\perp}\boldsymbol{A}_j$, e $p_j = car\left(\boldsymbol{Q}_j(1)\right)$ e $p_j{}^{\perp} = car\left(\boldsymbol{Q}_j(2)\right), j = 1, \dots, m$.

Considerando $\boldsymbol{W}_j$ uma matriz cujos vectores linha constituem uma base para Ω_j, teremos $\boldsymbol{P}_j = \boldsymbol{W}_j{}^T\boldsymbol{W}_j$ e para $\boldsymbol{\mu}_j$ teremos o estimador de mínimos quadrados $\tilde{\boldsymbol{\mu}}_j = \boldsymbol{W}_j{}^T\tilde{\boldsymbol{\eta}}_j$, onde $\tilde{\boldsymbol{\eta}}_j = \boldsymbol{W}_j\boldsymbol{Y}_j$. Para o vector estimável $\boldsymbol{\Psi} = G\boldsymbol{\mu}$ temos o estimador centrado $\tilde{\boldsymbol{\Psi}} = G\tilde{\boldsymbol{\mu}}$, onde $\tilde{\boldsymbol{\mu}} = \sum_{j=1}^{m} G\boldsymbol{A}_j{}^T\tilde{\boldsymbol{\mu}}_j$ (Ferreira 2015), que é UBLUE, isto é, o melhor estimador linear centrado (BLUE) quaisquer que sejam as componentes de variância. Também temos o estimador $\tilde{\gamma}_j = \frac{\boldsymbol{Y}_j{}^T\boldsymbol{P}_j{}^{\perp}\boldsymbol{Y}_j}{p_j{}^{\perp}}$, onde $p_j{}^{\perp} > 0$.

Seja (1) um OBS normal.

Tomando $\boldsymbol{\eta}_j = \boldsymbol{W}_j\,\boldsymbol{\mu}_j$ temos $\begin{cases} \|\boldsymbol{Y}_j - \boldsymbol{\mu}_j\|^2 = s_j - 2\boldsymbol{\eta}_j{}^T\boldsymbol{z}_j + \|\boldsymbol{\mu}_j\|^2, \ j = 1, \dots, l \\ \|\boldsymbol{Y}_j - \boldsymbol{\mu}_j\|^2 = s_j\,, \ j = l+1, \dots, m \end{cases}$,

com $s_j = \|\boldsymbol{Y}_j\|^2$, $j = 1, \dots, m$. A densidade de $\boldsymbol{Y}$ tem as estatísticas suficientes $\boldsymbol{z}_1, \dots, \boldsymbol{z}_l$ e $s_1, \dots, s_m$. Estas estatísticas são completas, ver Seely (1971) devido ao espaço paramétrico ser aberto, pois $\boldsymbol{\eta} = [\boldsymbol{\eta}_1{}^T \ \dots \ \boldsymbol{\eta}_l{}^T] \in R^k$ e $\boldsymbol{\gamma} \in R^k_{\geq 0}$. O estimador $\tilde{\boldsymbol{\Psi}}$, que é UBLUE no caso geral, é UMVUE (estimador centrado de variância mínima) no caso de normalidade.

Consideremos agora que a MPO $\boldsymbol{Q} = \boldsymbol{A}^T\boldsymbol{A}$, onde $\boldsymbol{A}$ tem vectores linha que constituem uma base ortonormada para $R(\boldsymbol{Q})$, comuta com as $\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_m$, matrizes da $bp(\mathcal{A}(M))$.

Para o modelo condensado $\boldsymbol{Y}^0 = \boldsymbol{A}\boldsymbol{Y} = \sum_{i=0}^{m} \boldsymbol{X}_i^0\beta_i$, com $\boldsymbol{X}_i^0 = \boldsymbol{A}\boldsymbol{X}_i$, $i = 1, \dots, m$, temos as famílias de matrizes $M^0 = \{\boldsymbol{M}_1^0, \dots, \boldsymbol{M}_m^0\}$ e $Q^0 = \{\boldsymbol{Q}_1^0, \dots, \boldsymbol{Q}_m^0\}$ com $\boldsymbol{M}_i^0 = \boldsymbol{A}\boldsymbol{M}_i\,\boldsymbol{A}^T$ e $\boldsymbol{Q}_j^0 = \boldsymbol{A}\boldsymbol{Q}_j\,\boldsymbol{A}^T$, $i, j = 1, \dots, m$. Se as matrizes de Q^0 são MPOMO cuja soma é a matriz identidade e a matriz de transição entre M^0 e Q^0 é $\boldsymbol{B} = [b_{i,j}]$, então o modelo condensado é OBS com a mesma matriz de transição que o OBS inicial, ver Santos (2016). Esta possibilidade de obter um OBS condensado com menos observações e os mesmos vectores estimáveis do OBS inicial viabiliza o aproveitamento dos resultados obtidos para a estimação em OBS.

Agradecimentos: Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia através dos projectos UID/MAT/00297/2013 e UID/MAT/00212/2013

Referências

Ferreira, S., Nunes, C., Ferreira, D., Moreira, E. & Mexia, J.T. (2015) Estimation and Orthogonal Block Structure. *Hacettepe University Bulletin of Natural Sciences and Engineering, Series B: Mathematics and Statistics*, 45(58).

Santos, C., Nunes, C., Dias, C. & Mexia, J. T. (2016) Condensing normal mixed models with orthogonal block structure. *Discussiones Mathematicae - Probability and Statistics* (no prelo).

Seely, J. (1971) Quadratic subspaces and completeness. *The Annals of Mathematical Statistics*, 42, 710-721.

Aprendizagem automática para a classificação da severidade da Doença Pulmonar Obstrutiva Crónica

Matheus Coppetti Silveira[1]
[1] *Universidade de Évora, matheuscoppetti@gmail.com*

Sumário: É proposto aqui o desenvolvimento de uma ferramenta para a classificação automática da severidade da Doença Pulmonar Obstrutiva Crónica. Para tal, é necessário que anteriormente os dados sejam tratados em diversas etapas, sendo elas, a integração da informação armazenada em bases de dados heterogéneas, o processamento da língua natural e um entendimento semântico sobre a doença. Esta ferramenta possibilitará, através do uso de diretrizes da Global Initiative For Crhonic Obstructive Lung Disease, classificar a Doença Pulmonar Obstrutiva Crónica em seus níveis de severidade.

Palavras-chave: Aprendizagem automática, DPOC, Data analysis, PLN, Web semântica.

O uso de registos eletrónicos de pacientes para o armazenamento de informações clínicas permite que sejam utilizadas abordagens computadorizadas para analisar os dados, extraindo informações valiosas para a interpretação do quadro clínico dos pacientes. Uma vez que estas informações encontram-se, muitas vezes, sob a forma de narrativas livres, passivas de ambiguidades, é necessário que se utilize de técnicas de Processamento de Língua Natural (PLN) para identificar, nestes textos, os termos de interesse clínico e que transmitem informações para análise do quadro do paciente.

Contudo, mesmo com a identificação destes termos, ainda é necessário um entendimento sobre o domínio representado por aquela informação. O uso de ontologias permite expressar semanticamente esse conhecimento, e, através do uso destas ontologias, é possível a extração da informação de maneira mais completa, o que permite interpretações e tomadas de decisões quando se analisa um histórico de um paciente.

Propomos neste trabalho o uso de tecnologias de PLN e ontologias para classificar textos médicos e, com o uso de tecnologias semânticas, representar o conhecimento no domínio da Doença Pulmonar Obstrutiva Crónica.

Foram utilizadas duas bases de dados clínicas, a do OpenMRS e a do MIMIC2, para a obtenção do histórico clínico de pacientes. Estes dados são então formatados em um padrão de mensagens médicas, o HL7, para serem analisados pelo cTakes, uma ferramenta de PLN para a área médica, de modo que se obtenha a classificação dos termos de interesse clínico contido nestes registos médicos.

Usando uma ontologia própria para a representação do domínio da DPOC, é possível identificar as relações entre os termos classificados com características da DPOC, como os sintomas, fatores de risco, comorbidades, complicações e exacerbações. Deste modo, seguindo as diretrizes da Global Initiative For Crhonic Obstructive Lung Disease (GOLD)

para a classificação da severidade da doença e os termos classificados nos registos clínicos, utilizamos algoritmos de aprendizagem automática para classificar os pacientes nos quatro possíveis níveis de severidade da DPOC.

Com os termos já classificados e relacionados com a DPOC através da ontologia, é possível utilizar um modelo de dados que se assemelha com o proposto pelo COPD Assessment Test (CAT), teste que ajuda na identificação da severidade da DPOC. Ainda, para dar continuidade a classificação da severidade da doença, é necessário identificar o número de exacerbações apresentadas pelo paciente nos últimos dois anos, ou então, obter o valor dos testes de espirometria do paciente. Estas informações são organizadas em um dataset, são usadas como entrada em um algoritmo de classificação automática. Este algoritmo utiliza uma abordagem de Redes de Bayes, e é anteriormente treinado, de modo a ser capaz de classificar os novos dados e determinar a severidade da doença apresentada em cada paciente.

A adoção de uma abordagem bayseana se deu primeiramente devido ao número de dados para o treino, uma vez que este algoritmo apresenta bons resultados quando o corpo de treino é de menor dimensão. Busca-se ainda utilizar abordagens baseadas em redes neuronais e vetores de suporte à decisão, e, comparar o desempenho nas diferentes abordagens.

Referências

Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210). London: UCL press.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 507-513.

Vestbo, J., Hurd, S. S., Agustí, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A. & Stockley, R. A. (2013). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine*, 187(4), 347-365.

What do you like in a hostel? Exploring the determinants of satisfaction

Paula Vicente[1], **Rita Lima**[2]
[1] *Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, paula.vicente@iscte.pt;*
[2] *Instituto Universitário de Lisboa (ISCTE-IUL), rdlas@iscte.pt*

Abstract: This paper seeks to explore how service quality influences guests' satisfaction in hostels. The effect of seven dimensions of service quality is tested: location, ambiance & design, price, facilities & services, staff, security and cleanliness. The findings from the multiple regression model reveal that the two factors with the strongest effect on the overall satisfaction are the staff (their competence, friendliness and availability) and the cleanliness of both private and public areas of the hostel.

Key words: Lisbon, Multiple linear regression, Principal component analysis, Service quality.

The hostel segment in Portugal has grown significantly in recent years. The first (private) hostel opened in Portugal in 2005 and presently there are over 160 hostels, most of them are located in Lisbon. Moreover, several Portuguese hostels have been awarded prestigious international prizes (HostelWorld 2015) which is boosting the growth of the hostel market and fostering the refurbishment of existent hostels and the opening of new ones. Despite the growing relevance of this sector little research has been done about service quality in Portuguese hostels. The objective of this research is to identify the main dimensions of service quality that determine guests' satisfaction when staying in a hostel.

The research was conducted in 2014. A convenience sample of 223 guests was selected across 14 hostels in Lisbon. The questionnaire included: (a) a Likert-type scale with 27 items for measuring perceptions about the hostel's service quality (Mei et al. 1999); the participants rated their level of agreement with each of the items on a 7-point scale (1-strongly disagree to 7-strongly agree), (b) a question about overall satisfaction with the stay at the hostel (7-point rating scale from 1-totally dissatisfied to 7-fully satisfied), and (3) demographics. Principal Component Analysis was performed to reduce data dimensionality and Multiple Linear Regression using Ordinary Least Squares estimation was used to assess the determinants of satisfaction. Due to the convenience nature of the sample the p-values of the significance tests are not to be interpreted literally, but are merely standard values that say how large the difference between the realities under comparison needs to be so we can take note of it. SPSS 22 was used for the analysis.

Respondents were mostly female (56%), aged 15-25 years (44%) and from European countries (70%). The main purpose of the stay was Lisbon sightseeing, i.e.,

visiting the city's most emblematic places and major tourist attractions (69.1%), followed by exploring a different culture (30.9%), relaxing (27.8%) and having fun/entertainment (26.9%). In order to identity the dimensions of service quality, all 27 attributes were placed into an exploratory principal components analysis (KMO=0.922; p-value of the Sphericity Bartlett test<0.001). Seven components were retained justifying a combined total of 75% of the variance. All communalities were above 0.6. According to the highest loadings in each dimension (above 0.5) the new dimensions were labelled as: Staff, Cleanliness, Ambience & design, Location, Price, Facilities & services and Security.

The overall feeling of guests about their stay at the hostel was positive – 81.2% of the respondents rated overall satisfaction as 6 or higher. On average, the overall satisfaction rate was 6.1. Table 1 presents the estimates of the regression model explaining guests' overall satisfaction using dimensions of service quality.

Table 1: Estimates of the model explaining guests' overall satisfaction

	Standardised $\hat{\beta}$	t-statistic	p-value
Staff	0.460	12.749	0.000
Cleanliness	0.434	12.033	0.000
Ambiance & design	0.357	9.894	0.000
Location	0.192	5.316	0.000
Price	0.376	10.422	0.000
Facilities & services	0.183	5.060	0.000
Security	0.027	0.759	0.449
Constant	*6.051*	*168.133*	*0.000*

The estimated model is valid (all assumptions are verified) and well fitted to the data (adjusted R^2=0.722; p-value<0.001). It reveals a strong association between perception about service quality and guests' overall satisfaction. Staff ($\hat{\beta}$ =+0.460) and Cleanliness ($\hat{\beta}$ =+0.434) are the strongest determinants of guests' satisfaction. Only Security did not influence significantly overall satisfaction (p-value>0.05). The outcomes suggest that hostel guests, who used to accept an accommodation with limited services in exchange of a low-cost (Suvantola 2002) are giving way to a group whose satisfaction is less dependent on the price and more influenced by other attributes such as staff (their kindness, friendliness and availability to assist guests' needs) and cleanliness.

References

HostelWorld (2015) *Welcome to the 2014 Hostel Awards*. http://www.hostelworld.com/hoscars-2014.

Mei, A., Dean, A. & White, C. (1999) Analyzing service quality in the hospitality industry. *Managing Service Quality: An International Journal*, 9, 136–143.

Suvantola, K. (2002) *Tourist's Experience of Place*. Hampshire: Ashgate Publishing Ltd.

Índice de Autores