
Reunião Anual da

ASSOCIAÇÃO PORTUGUESA DE CLASSIFICAÇÃO

E ANÁLISE DE DADOS (CLAD)

# Livro de Resumos

**AS JOCLAD 2014 TIVERAM O APOIO INSTITUCIONAL DE:**


## Ficha Técnica

**Presidente das Jornadas**

Alda Carvalho (Presidente do INE)

**Secretário das Jornadas**

Fernanda Sousa (Presidente da CLAD e FEUP-Universidade do Porto)

**Comissão Organizadora**

Catarina Marques (ISCTE-Instituto Universitário de Lisboa)

Isabel Silva (FEUP-Universidade do Porto)

José Gonçalves Dias (ISCTE-Instituto Universitário de Lisboa)

Nuno Lavado (ISEC-Instituto Politécnico de Coimbra)

**Título:** XXI Jornadas de Classificação e Análise de Dados (JOCLAD 2014). Livro de Resumos.

**Produzido:** Instituto Nacional de Estatística

**Editores:** Fernanda Sousa, Catarina Marques, Isabel Silva, José Gonçalves Dias, Nuno Lavado, Carlos Marcelo

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

# Prefácio

Desde a sua constituição a Associação Portuguesa de Classificação e Análise de Dados (CLAD) tem vindo a desenvolver a sua actividade de acordo com a natureza e objectivos definidos na sua génese. Entre as diversas actividades desenvolvidas encontra-se a promoção das Jornadas Científicas, que têm tido lugar anualmente, sem qualquer interrupção. Para tal, têm contado com o precioso apoio de diferentes grupos de investigação, com actividade científica relevante nas áreas de actuação da CLAD, sediados em instituições universitárias. Das edições anteriores, onze tiveram lugar na zona da Grande Lisboa, três no Porto, e ainda Aveiro, Açores, Algarve, Vila Real, Tomar e Guimarães que receberam as Jornadas por uma vez.

Este ano a CLAD celebra vinte anos de existência e estas são já as XXI Jornadas de Classificação e Análise de Dados, JOCLAD 2014. Para assinalar esta data entendeu-se por bem dar visibilidade à forte cooperação existente, desde sempre, entre o Instituto Nacional de Estatística, INE, e a CLAD. Estas Jornadas, que pela primeira vez não ocorrem em seio universitário, têm lugar em Lisboa e contam com o valioso apoio logístico do INE.

O Programa das JOCLAD 2014 reflecte o carácter multidisciplinar das Jornadas, enquadrando de forma equilibrada a apresentação de trabalhos teóricos e aplicados e focando diversas temáticas da Análise de Dados em domínios transversais à sociedade.

A Comissão Organizadora agradece a todos os autores e moderadores de sessões, aos membros da Comissão Científica, bem como aos participantes, aos participantes convidados e aos colegas que procederam à revisão dos trabalhos que constam deste livro. Uma nota particular ao Professor Christian Hennig que lecciona o mini-curso, bem como ao Banco de Portugal e ao INE que, através dos seus corpos técnicos, organizaram as Sessões Temáticas que constam do Programa.

Por último, desejamos agradecer a todas as entidades que, directa ou indirectamente, apoiaram ou patrocinaram estas Jornadas.

O nosso obrigado a todos.

Lisboa, Abril de 2014

P'la Comissão Organizadora JOCLAD 2014

Fernanda Sousa

# ORGANIZAÇÃO

**Presidente das Jornadas**

Alda Carvalho (Presidente do INE)

**Secretário das Jornadas**

Fernanda Sousa (Presidente da CLAD e FEUP-Universidade do Porto)

**Comissão Organizadora**

Catarina Marques (ISCTE-Instituto Universitário de Lisboa)
Isabel Silva (FEUP-Universidade do Porto)
José Gonçalves Dias (ISCTE-Instituto Universitário de Lisboa)
Nuno Lavado (ISEC-Instituto Politécnico de Coimbra)

**Comissão Científica**

Conceição Amado (Universidade de Lisboa)
Helena Bacelar-Nicolau (Universidade de Lisboa)
Paula Brito (Universidade do Porto)
Jorge Cadima (Universidade de Lisboa)
Pedro Campos (Universidade do Porto)
Margarida Cardoso (Instituto Universitário de Lisboa)
José Gonçalves Dias (Instituto Universitário de Lisboa)
Susana Faria (Universidade do Minho)
Ana Sousa Ferreira (Universidade de Lisboa)
Carlos Ferreira (Universidade de Aveiro)
Adelaide Figueiredo (Universidade do Porto)
A. Manuela Gonçalves (Universidade do Minho)
Luís Miguel Grilo (Instituto Politécnico de Tomar)
Paulo Infante (Universidade de Évora)
Victor Lobo (Universidade Nova de Lisboa)
Catarina Marques (Instituto Universitário de Lisboa)
Manuela Neves (Universidade de Lisboa)
Fernando Nicolau (Universidade Nova de Lisboa)
Irene Oliveira (Universidade de Trás-os-Montes e Alto Douro)
Fátima Salgueiro (Instituto Universitário de Lisboa)
Pedro Duarte Silva (Universidade Católica Portuguesa)
Carlos Soares (Universidade do Porto)
Fernanda Sousa (Universidade do Porto)
Paula Vicente (Instituto Universitário de Lisboa)

# APOIOS


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


ISCTE IUL
Instituto Universitário de Lisboa


BANCO DE
PORTUGAL
EUROSISTEMA


PSE
PRODUTOS E SERVIÇOS DE ESTATÍSTICA

# PROGRAMA

## QUINTA-FEIRA, 10 DE ABRIL

**9:00** Registo e entrega de documentação

**9:30** **Mini-curso** – Salão Nobre

**Christian Hennig** - ***Clustering with the Gaussian mixture model***, p. 3.

**11:00 Pausa para café**

**11:30** **Mini-curso** (*cont.*)

**13:00 Almoço**

**14:00** **Sessão de Abertura das Jornadas** – Salão Nobre
**(Momento Musical: Grupo Coral "AD DIVITIAS" do Grupo Desportivo do INE)**

**14:30** **Sessão Plenária I** – Salão Nobre

**Mário A. T. Figueiredo** (IST, Universidade de Lisboa, Portugal)
***Sparsity and structured sparsity for feature selection in Machine Learning and Statistics***, p. 7.
Moderador: Margarida Cardoso

**15:30 Pausa para café**

**15:50 Sessões Paralelas**

| | Salão Nobre<br>***Classificação e Análise de Dados***<br>Moderador: Pedro Duarte Silva | Sala 316<br>***Análise de Dados em Medicina***<br>Moderador: Ana Sousa Ferreira |
|---|---|---|
| 15:50 | Finamore, A.C., Oliveira, M.R, Pascoal, C., Pacheco, A.: *Classifying a fairy tale: A case study*, p. 79. | Gaio, A.R., Felgueiras, O., Santos, R., Azevedo, E.: *Progression of carotid atherosclerotic plaques: speed and dependency from vascular risk factors*, p. 97. |
| 16:10 | Oliveira, M.R, Valadas, R., Pietrzyk, M., Collange, D.: *Impact of input variables' stability on the classification of Internet applications*, p. 85. | Guerreiro, J., Torre, C., Gomes, M., Costa, S.: *Impacto das normas de orientação clínica na evolução do padrão de prescrição de antidiabéticos orais e antihipertensores em Portugal – Exemplo prático da análise de regressão segmentada a uma série temporal interrompida*, p. 101. |
| 16:30 | Carrasquinha, E., Amado, C., Pires, A.M.: *On circulant matrix approximation to correlation matrix: an application to sounds*, p. 89. | Gaio, A.R., Costa, J., Severo, M.: *Equiparação das classificações dos cursos de Medicina*, p. 103 |
| 16:50 | Figueiredo, A.M., Figueiredo, F.O.: *Metodologia STATIS em controlo estatístico da qualidade*, p. 93. | Lourenço, V.M., Pires, A.M.: *M-regression, false discovery rates and outlier detection in genetic association studies*, p. 107. |

**17:10 Sessão Temática I – Instituto Nacional de Estatística**

| | Salão Nobre<br>***Sessão INE - Desafios nas Estatísticas Oficiais III***<br>Moderador: Carlos Marcelo |
|---|---|
| 17:10 | Moreira, F., Neves, C.: *SIOU - Fonte de atualização da Geografia do Ficheiro Nacional de Alojamentos*, p. 17. |
| 17:30 | Góis, E., Gonçalves, C., Figueiredo, E., Pereira, P.: *Inquérito às Despesas das Famílias: Porquê? Como? Para quê?*, p. 23. |
| 17:50 | Pereira, S., Correia, L., Campos, P.: *Estimação do desemprego ao nível NUTS III*, p. 29. |
| 18:10 | Mendonça, V.H.Q, Silva, A.B.: *Série longa do Índice de Preços no Consumidor (1948 – 2013)*, p. 33. |

**18:40 Porto de Honra**

**19:00 Reunião da Assembleia Geral da CLAD** – Salão Nobre

## SEXTA-FEIRA, 11 DE ABRIL

**9:00 Sessões Paralelas**

| | Salão Nobre<br>***Modelos Longitudinais***<br>Moderador: Fátima Salgueiro | Sala 316<br>***Análise de Estruturas de Covariância***<br>Moderador: Manuela Neves |
|---|---|---|
| 9:00 | Silva, I., Torres, C., Silva, M.E.: *Estimating bivariate integer-valued moving average models with the generalized method of moments*, p. 111. | Pral, C., Gonçalves, B., Marques, C.: *Depressão e risco de reincidência criminal face à delinquência juvenil*, p. 127. |
| 9:20 | Pereira, L.N., Ferreira, L.N.: *Modelação e previsão da procura turística doméstica em Portugal numa conjuntura de crise económica e financeira*, p. 115. | João, P., Lobo, V.: *Visual fraud detection with self organizing maps*, p. 131. |
| 9:40 | Vicente, P.C.R, Salgueiro, M.F.: *Modelo com trajetória latente com dados gerados a partir de um planned missing design: estudo de simulação*, p. 119. | Ribeiro, E.R., Marques, C., Correia, E.: *Os ginásios da Cidade de Maputo: Os determinantes da satisfação e da lealdade dos clientes*, p. 137. |
| 10:00 | Salgueiro, M.F., Vicente, P.C.R.: *The effect of observed data deviations from normality on the parameter estimates of a latent growth curve model: a simulation study*, p. 123. | Grilo, L.M., Coelho, C.A.: *Near-exact distributions for the statistic used to test the reality of covariance matrix in a complex normal distribution*, p. 141. |

**10:20 Pausa para café**

**10:40 Sessão Temática II – Banco de Portugal**

| | Salão Nobre<br>***Sessão Banco de Portugal***<br>Moderador: Filipa Lima |
|---|---|
| 10:40 | Lima, F., Correia, I., Batista, R.: *Non-financial sector indebtedness*, p. 39. |
| 11:00 | Gonçalves, H., Lourenço, M., Silveira, V.: *High-growth enterprises in Portugal*, p. 45. |
| 11:20 | Magalhães, C., Cordeiro, P., Poiares, R.: *Quarterly time-series from Central Balance Sheet Database*, p. 51. |

**11:40 Sessão Plenária II** – Salão Nobre
**Salvatore Ingrassia** (Università di Catania, Itália)
***Recent results in model based clustering via the Cluster-Weighted approach***, p. 9.
Moderador: José G. Dias

**12:40 – Almoço**

**13:45 – Passeio Pardal Monteiro**

**14:45 – "Um outro olhar: o edifício e as publicações do INE"** – Salão Nobre

**15:15 Sessão Plenária III** – Salão Nobre

**Georges Lemaître** (Formerly-OCDE, Paris)
***The Value-added of International Comparisons***, p. 11.
Moderador: Paulo Gomes

**16:15 Sessão de Posters I + Pausa para café**

Catalão, D., Gonçalves, A.M., Faria, S., Oliveira, J.: *Metodologia estatística para a avaliação de um recurso natural (Minho e Galiza)*, p. 183.

Dias, J.G., Tiago de Oliveira, I.: *Explaining contraceptive use by the wealth index in India: A latent variable approach*, p. 187.

Marques, C., Dias, J.G.: *The impact of population heterogeneity on factor analysis estimation*, p. 191.

Oliveira, R., Gonçalves, A.M., Vasconcelos, R.M.: *Estudo empírico do índice de satisfação da procura dos candidatos aos cursos superiores de engenharia*, p. 195.

Penalva, H., Nunes, S., Neves, M.: *Estimação paramétrica e semi-paramétrica do índice de cauda utilizando o R*, p. 199.

Pereira, L.N., Pedro, I., Carrasqueira, H.: *Fatores determinantes na manutenção da relação de compromisso entre os alumni e a alma mater: aplicação de um modelo de equações estruturais*, p. 203.

Santos, J., Faria, S.: *Modelação de contagens com excesso de zeros*, p. 207.

Vicente, P.: *Utilização de telemóveis entre a população sénior*, p. 211.

**16:45 Sessão Temática III – CLAD**

| | Salão Nobre<br>***Sessão 20 anos da CLAD***<br>Moderador: Fernanda Sousa e Helena Bacelar-Nicolau |
|---|---|
| 16:45 | O Movimento da Classificação e Análise de Dados (CLAD) em Portugal. |
| 17:05 | Sousa, A., Bacelar-Nicolau, H., Nicolau, F.C., Silva, O.: *Classes de objectos simbólicos: dados da indústria automóvel*, p. 57. |
| 17:25 | Ichino, M., Brito, P.: *A hierarchical conceptual clustering based on the quantile method for mixed data*, p. 61. |
| 17:45 | Ferreira, A.S.: *Avaliações internacionais e desempenho dos alunos portugueses*, p. 67. |
| 18:05 | Gomes, P.: *Índice de Bem-estar em Portugal – Contributos para a interpretação dos resultados baseada em classificação de variáveis*, p. 71. |

**20:00 Jantar das Jornadas – Restaurante Petra Rio**

## SÁBADO, 12 DE ABRIL

**10:00 Sessões Paralelas**

| | Salão Nobre<br>***Análise de Dados em Economia e Gestão***<br>Moderador: Adelaide Figueiredo | Sala 316<br>***Data Mining***<br>Moderador: Carlos Soares |
|---|---|---|
| 10:00 | Santos, F., Silva, A. L., Duarte, I.: *Fatores chave de sucesso das equipas virtuais de tecnologias de informação em regime de outsourcing: do ponto de vista dos membros da equipa*, p. 145. | Matos, D., Marques, N.C., Cardoso, M.G.M.S: *Agrupamento sobre uma matriz de distâncias UMAT – uma aplicação sobre dados financeiros*, p. 163. |
| 10:20 | Jerónimo, W., Amaro, A.: *Abordagem exploratória: análise hibrida de indicadores de sustentabilidade empresarial*, p. 149. | Gomes, L., Saleiro, P., Soares, C.: *Análise de tendências políticas no Twitter para previsão de sondagens*, p. 169. |
| 10:40 | Vicente, P., Marques, C., Reis, E.: *Resultados de uma sondagem CATI móvel*, p. 155. | Trigo, L., Brazdil, P.: *Análise de afinidades entre investigadores com text mining*, p. 173. |
| 11:00 | Duarte Silva, A.P., Brito, P.: *Discriminant analysis of interval data: Parametric versus distance-based approaches*, p. 159. | Costa, V., Saleiro, P., Soares, C.: *Active learning para análise de sentimento no Tweeter*, p. 177. |

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

Cabral, J., Carvalho, C.B., Silva, O.: *Análise fatorial confirmatória - Escala de integração comunitária de adultos com problemas psiquiátricos*, p. 215.

Dias, J.G., Ramos, S.B.: *Clustering European industries using longitudinal data*, p. 219.

Fernandes, L., Henriques, R., Lobo, V.: *Seleção de instâncias para algoritmos de aprendizagem não supervisionada: aplicação a dados de motores de aeronaves*, p. 223.

Frei, F., Netto, F.K., Juliana Alves Pegoraro, J.A.: *Avaliação do emprego da análise de agrupamentos nas revistas de saúde brasileiras no período de 1993 a 2011*, p. 227.

Gaio, A.R., Felgueiras, O., Dias, C., Paiva, J.-A., Czosnyka, M.: *Kidney-brain link in traumatic brain injury patients: A preliminary report*, p. 233.

Pereira, S., Lavado, N., Nogueira, L., Lopez, M., Abreu, J., Silva, H.: *Root resorption risk modeling*, p. 237.

Sousa, A., Batista, M.G., Medeiros, D.: *Motivação e satisfação na função pública: um exemplo dos Açores*, p. 241.

**11:50 Sessão Plenária IV** – Salão Nobre

**Christian Hennig** (University College London, Reino Unido)
***Measurement of quality in cluster analysis***, p. 13.
Moderador: Paula Brito

**12:50 Sessão de Encerramento das Jornadas**

# RESUMOS

# ÍNDICE

## ST II – Sessão do Banco de Portugal



## ST III – Sessão 20 anos da CLAD



## Sessões paralelas

### Classificação e Análise de Dados

## Análise de Dados em Medicina



## Modelos Longitudinais



## Análise de Estruturas de Covariância

## Análise de Dados em Economia e Gestão



## Data Mining



# POSTERS

## Sessão I

**Sessão II**

# MINI-CURSO

**5ª Feira, 10 de Abril – Salão Nobre (9:30 – 13:00)**

# Clustering with the Gaussian mixture model

**Christian Hennig**
*University College London, Reino Unido, Email: c.hennig@ucl.ac.uk*

Fitting a Gaussian mixture model is a popular and very flexible clustering method, sometimes referred to as "model-based clustering".

In this tutorial I will give an overview of the basics of fitting the Gaussian mixture model with R's mclust package, the EM-algorithm and some theoretical background. I will then discuss in more detail a number of issues connected to the use of Gaussian mixtures as a clustering method, particularly the implications of interpreting Gaussian mixture components as clusters, cluster validation and visualisation, estimation of the number of clusters, and merging of Gaussian mixture components in order to find non-Gaussian clusters.

# SESSÕES PLENÁRIAS

**5ª Feira, 10 de Abril – Sessão Plenária I, Salão Nobre (14:30)**

# Sparsity and structured sparsity for feature selection in Machine Learning and Statistics

**Mário A. T. Figueiredo**
*IST, Universidade de Lisboa, Portugal*

Sparsity is currently a major theme in statistics, machine learning, and signal processing, which can be seen in terms of the classical goals of feature selection and model selection. This talk will focus on methods which embed sparse model/feature selection into the learning algorithms. In such methods, learning is carried out by minimizing a regularized empirical risk functional composed of two terms: a "loss term," controlling the goodness of fit to the data (e.g., quadratic, logistic, or hinge loss), and a "regularizer term," which is designed to promote sparsity.

The simplest example is the now famous L1-norm regularization (often known as LASSO), which penalizes weight components individually, and has been explored in various machine learning and statistical applications. More sophisticated regularizers, such as those that use mixed norms and groups of weights, are able to promote "structured" sparsity: i.e., they promote sparsity patterns that are compatible with a priori knowledge about the structure of the problem. Some regularizers are even able to encourage structured sparsity, without prior knowledge about this structure. Sparsity-inducing regularizers require the use of specialized optimization routines for learning, some of which will be reviewed in this talk.

**6ª Feira, 11 de Abril – Sessão Plenária II, Salão Nobre (11:40)**

# Recent results in model based clustering via the Cluster-Weighted approach

**Salvatore Ingrassia**
*Department of Economics and Business, University of Catania, Italy*

Cluster-weighted models (CWMs) are a flexible family of mixture models for fitting the joint distribution of a random vector composed by a response variable and by a set of covariates. They act as a convex combination of the products between the marginal distribution of the covariates, and the conditional distribution of the response given the covariates, in each mixture component. In this talk, we introduce a wide family of CWMs where the component conditional distributions are assumed to belong to the exponential family and where the covariates are allowed to be of mixed-type (that is containing both categorical and continuous variables). Under the assumption of Gaussian covariates, sufficient conditions for model identifiability are provided. Moreover, maximum likelihood parameter estimates are derived using the EM algorithm. Parameter recovery and performance of some information criteria are both investigated by a wide simulation study. An application to real data is finally presented where the proposed model outperforms other well-established mixture-based approaches.

**6ª Feira, 11 de Abril – Sessão Plenária III, Salão Nobre (15:15)**

# The Value-added of International Comparisons

**Georges Lemaître**
*Formerly OECD, Paris*

Scarcely a week goes by these days without the publication of some statistic for which international comparisons are provided or referred to. At the most elementary level, such comparisons provide information on where countries stand on commonly reported indicators of economic and social performance. Sometimes these rankings are themselves surprising and revealing, showing, for example, that the extent of migration to France or Germany or even the United States is very low in relation to many other OECD countries, or that relatively more persons experience crime in a given year in Sweden than in Italy.

Until international microdatasets became available in recent years, international data could be characterized as being rather limited in scope compared to the richness that was available from national statistics. It has been impossible to assemble and disseminate this richness internationally in a sensible way, however, at once because of its volume and because the definitions in use varied from country to country. The requirements of international comparability also mean that many national specificities are lost in "forcing" national statistics into an international framework and common definitions. What is gained in return, however, is an additional source of variation, namely that by country, and variation, as we know, is often a source of information. What varies from country to country is, in addition to the outcomes, the history and traditions of each country but especially the institutions and policies in place. And it is the relationship between the latter and "success" or "failure" that is of particular interest.

What kinds of information can one extract from international comparisons?

The first, mentioned above, is a benchmarking of national outcomes to an international average which, if the national outcome differs significantly from the international average, leads to the question of why this is so. Ideally, one would like to be able to point to precise reasons; in practice, because social and economic outcomes tend to be multiply determined, it is rarely a simple matter to do so. The difficulty in explaining country-to-country differences in international student performance (the OECD PISA results), for example, are a case in point. But even if one cannot fully explain these differences, recent OECD assessment results for adults (PIAAC) show that in-country progress in reading skills across generations has been significant and that that there has been considerable catch-up to leading countries by many which were trailing over past decades.

A second use of international data is to shed light on the current policy and economic environment in a way that is not necessarily possible on the basis of national data alone. The trade-off between the degree of employment protection and the duration of unemployment is an example, as is the rapidity with which immigrants integrate into the labour market of a country.

A third use of international data is to make possible generalisations across countries. The lack of variation in a particular outcome across countries which differ considerably in their institutions and policy environment may suggest that one is dealing with a phenomenon which is fundamental and transcends policy intervention. Examples here are the relative employment outcomes of high- and low-educated immigrants compared to their native-born counterparts and the association between school outcomes of the children of immigrants and various measures of concentration of immigrant children or of disadvantage in schools.

A fourth use of international comparisons is to help confirm that one is asking the right research question, or to serve as a source of counterexamples. Here it will be seen that the concentration of immigrants in schools is not necessarily a problem per se for educational outcomes. This result for immigrants leads more generally into the broader question of the sources of inequality in outcomes for the educational system as a whole and what differentiates countries from one another in this regard. Among the comparisons of interest are the effects of parental education vs those of the concentration of advantage/disadvantage in schools on outcomes.

International comparisons can thus serve a variety of purposes other than that of simply showing where one's country stands in relation to others on particular statistical indicators. At their best, judicious use of international data can indeed illuminate and provide insights into some fundamental societal questions.

**Sábado, 12 de Abril – Sessão Plenária IV, Salão Nobre (11:50)**

# Measurement of quality in cluster analysis

**Christian Hennig**
*University College London, Reino Unido*

There is much work on benchmarking is supervised classification, where "quality" can generally be measured as a function of misclassification probabilities. In unsupervised classification (cluster analysis), the measurement of quality is much more problematic, because in reality there is no true class label which can be used for cross-validation and the like. Furthermore, there is no guarantee that in situations where there is a true classification (for example, where benchmark data sets from supervised classification are used to assess clustering methods, or where data is simulated from a mixture distribution), this classification is unique. There can be a number of different reasonable clusterings of the same data, depending on the research aim.

I will discuss the use of statistics for the assessment of clustering quality that can be computed from classified data without making reference to "the true clusters". Such statistics have traditionally been called "cluster validation indexes" (such as the average silhouette width), and sometimes been used for estimating the number of clusters. Most of the traditional statistics try to balance various aspects of a clustering against each other (such as within-cluster homogeneity and between-cluster separation), but in order to characterize what advantages and disadvantages a clustering has, it is useful to formalize different aspects of cluster quality separately. This can also be used to explain misclassification rates in cases where "true" clusterings exist as function of the features of these clusterings.

# SESSÕES TEMÁTICAS

**ST I – Sessão INE - Desafios nas Estatísticas Oficiais III**

**ST II – Sessão Banco de Portugal**

**ST III – Sessão 20 anos da CLAD**

**ST I – Sessão INE – 5ª Feira, 10 de Abril, Salão Nobre (17h10)**

# SIOU - fonte de atualização da Geografia do Ficheiro Nacional de Alojamentos

**Fátima Moreira[1], Cristina Neves[2]**
[1]*Instituto Nacional de Estatística, fatima.moreira@ine.pt;*
[2]*Instituto Nacional de Estatística, cristina.neves@ine.pt*

**Sumário**

O SIOU baseia-se no aproveitamento de dados administrativos das Câmaras Municipais relativos a operações urbanísticas. Em 2013 o SIOU foi alvo de uma reestruturação, a nível dos conteúdos, dos procedimentos e das funcionalidades, de entre os quais se destacam a recolha de informação sobre coordenadas de localização geográfica dos novos edifícios licenciados/concluídos e demolidos, assim como a identificação e caracterização dos fogos (novos ou intervencionados), para efeitos de atualização do Ficheiro Nacional de Alojamentos e da Base Geográfica de Edifícios do INE.

**Palavras-chave:** Georreferenciação, INE, Licenciamento, Obras, SIOU

## 1. Introdução

O Sistema de Indicadores de Operações Urbanísticas (SIOU) foi criado pelo Instituto Nacional de Estatística (INE), na sequência da publicação do Decreto-Lei nº 555/99 que aprovou o Regime Jurídico da Urbanização e da Edificação (RJUE), e que define a obrigatoriedade de envio mensal pelas Câmaras Municipais ao INE dos elementos estatísticos referentes a operações urbanísticas.

O SIOU assenta, fundamentalmente, no aproveitamento da informação administrativa associada ao novo regime jurídico, que distingue as diferentes formas de procedimentos: alvarás de licença, alvarás de autorização, comunicações prévias, pareceres prévios, projetos de obras municipais e cancelamentos.

A informação relativa a operações urbanísticas que as Câmaras Municipais de todo o país enviam mensalmente ao INE, integrada no SIOU, permite a compilação de indicadores estatísticos oficiais sobre a Construção e Habitação.

## 2. O SIOU como fonte de atualização da Geografia do Ficheiro Nacional de Alojamentos (FNA)

A informação obtida através do SIOU permite dar resposta ao Regulamento (CE) Nº 1165/98, do Conselho de 19 de maio, relativo às estatísticas conjunturais, ao Regulamento (CE) nº 99/2013 do PE e do Conselho, de 15 de janeiro, relativo ao Programa Estatístico Europeu e ao Programa Anual do Eurostat. Desse regulamento decorre a obrigatoriedade de

disponibilização mensal de informação à Comissão Europeia sobre o licenciamento de obras 40 dias após o mês de referência e de informação trimestral sobre obras concluídas, 75 dias após o trimestre de referência.

A informação mensal relativa ao licenciamento e obras concluídas, obtida através do SIOU, é de extrema importância na análise de conjuntura, constituindo um dos indicadores avançados de avaliação da evolução económica.

### 2.1. Reformulação do SIOU

Em janeiro de 2013 o SIOU foi alvo de uma reestruturação, a nível dos conteúdos, dos procedimentos e das funcionalidades, por força dos mais recentes normativos legais associados ao Regime Jurídico da Edificação e da Urbanização (nomeadamente decorrentes da Lei N.º 60/2007 de 4 de setembro e do Decreto-Lei N.º 26/2010 de 30 de março), das alterações introduzidas nos conceitos estatísticos relativos à Construção e Habitação e da inclusão de variáveis decorrentes de novas necessidades de informação.

No que respeita aos conceitos estatísticos, foram introduzidas alterações nos conceitos estatísticos relativos à Construção e Habitação, em consonância com o Decreto Regulamentar N.º 9/2009.

De entre as novas necessidades de informação destacam-se:

- A recolha de informação sobre a classificação energética dos novos edifícios construídos;

- A atualização do campo de morada (harmonizada de acordo com as premissas definidas no âmbito do projeto EURADIN (European Addresses Infrastructure) e a recolha das coordenadas de georreferenciação dos novos edifícios licenciados e dos edifícios demolidos, com o intuito de proceder à atualização da Base Geográfica de Edifícios (BGE), que foi construída pelo INE com a recolha da informação dos Censos 2011;

- A recolha de informação sobre a identificação e caracterização dos novos fogos ou fogos intervencionados, para efeitos de atualização do Ficheiro Nacional de Alojamentos (FNA) do INE.

### 2.2. O Ficheiro Nacional de Alojamentos (FNA)

Até 2012 as operações estatísticas (OE) às famílias, efetuadas pelo INE, tinham como base de amostragem uma amostra de elevada dimensão formada por unidades de alojamento designada por "Amostra-Mãe" (AM). A AM era selecionada após a realização de cada Recenseamento da população e habitação (Censos) e mantida ao longo de uma década, sendo atualizada com base nas OE correntes ou através de trabalho de campo específico.

A realização dos Censos 2011, a georreferenciação dos edifícios, o acesso a diferentes fontes administrativas (com diferentes atributos, campos-chave e desenhos de registo), o

projeto EURADIN constituiu, no seu conjunto, uma oportunidade para a mudança de estratégia na definição das bases de amostragem das OE dirigidas às famílias.

A nova estratégia consiste assim na constituição de um Ficheiro Nacional de Alojamentos, criado a partir dos microdados dos Censos 2011, sendo atualizado com base em diferentes fontes, a partir do qual o INE constitui um Universo de Referência donde são extraídas diferentes Bases de Amostragem. O FNA é assim constituído pela totalidade dos alojamentos familiares e respetivos edifícios e tem como objetivo principal servir de suporte à realização das OE dirigidas às famílias. A atualização do FNA constitui a fase subsequente à sua criação e sem a qual não é possível garantir o reforço da qualidade da informação produzida pelo INE. A concretização deste objetivo só é possível mediante a utilização de fontes de informação relevantes, atuais e com qualidade.

O FNA contém uma componente alfanumérica e geográfica que é em grande medida constituída pelos edifícios da Base Geográfica de Edifícios (BGE), um dos componentes primordiais da Infraestrutura de Dados Espaciais (IDE) do INE.

A primeira versão da BGE é referente ao momento censitário e integra todos os edifícios residenciais georreferenciados nos Censos 2011 (3 549 508 edifícios) em que os alojamentos do FNA se inserem. A BGE encerra um elevado potencial de valor no âmbito da racionalização de recursos e processos técnicos, designadamente no domínio da amostragem, daí que seja um instrumento privilegiado para a espacialização dos registos do FNA.

**Figura 1: BGE- representação geográfica e alfanumérica do edifício (natureza pontual)**

Assim, na componente geográfica do FNA há vários tipos de atualização, de entre as quais se destaca o SIOU, dado que contém informação sobre a dinâmica urbanística, nomeadamente através da inclusão dos pontos de localização dos novos edifícios licenciados pelas Câmaras Municipais e eliminação dos pontos dos edifícios demolidos.

## 2.3. O SIOU como fonte de atualização da Geografia do FNA

Com a criação do FNA, considerou-se que o SIOU poderia ser utilizado como uma fonte de atualização relativamente a novas construções e demolições. O aproveitamento desta informação esteve também na origem da reformulação do projeto, tendo-se desencadeado um trabalho significativo no sentido da adaptação do pedido de recolha da informação não só ao nível do edifício mas também do fogo (alojamento).

Deste modo, desde janeiro de 2013 que são recolhidas (também) no SIOU as seguintes variáveis para atualização do FNA:

- Morada de cada um dos edifícios licenciados de acordo com a estrutura do EURADIN;
- Coordenadas geográficas de cada edifício (construções novas e demolições);
- Atributos relativos a cada um dos fogos que constituem o edifício: andar e lado (para obtenção da morada completa do fogo), área e tipologia.

Considerando que o FNA é constituído, no momento da sua criação, pelos dados dos Censos 2011, tornou-se necessário, para efeitos de atualização, receber da parte de todas as Câmaras Municipais os dados do licenciamento (com os novos atributos necessários para efeitos de atualização do FNA) desde março de 2011 (momento censitário – data de referência dos Censos 2011).

Este processo de recuperação da informação sobre coordenadas geográficas e identificação dos fogos através do SIOU (Inquérito aos Projetos de Obras de Edificação e Demolição de Edifícios) está em curso (Tabela 1 e 2), correspondendo a obras de construção nova para habitação e demolições (no que respeita à recolha de coordenadas e identificação dos fogos) e a obras de alteração, ampliação e reconstrução para habitação (apenas no que respeita à identificação dos fogos).

**Tabela 1: Nº de processos recolhidos no SIOU com informação para atualização do FNA**

| | Licenciamento | | | Obras Concluídas | | |
|---|---|---|---|---|---|---|
| | Georreferenciação | Identificação de fogos | | Georreferenciação | Identificação de fogos | |
| | Nº Edifícios | Nº fogos construções novas, ampliações e reconstruções | Nº fogos alterações | Nº Edifícios | Nº fogos construções novas, ampliações e reconstruções | Nº fogos alterações |
| 2011 | 1 059 | 1 306 | 24 | 13 521 | 21 308 | 896 |
| 2012 | 1 053 | 1 000 | 44 | 13 349 | 20 411 | 988 |

Do total de processos enviados às Câmaras Municipais (28 966 licenças) para recuperação de coordenadas geográficas e identificação dos fogos, em obras concluídas, já foi possível recuperar 94,0% da informação, referentes a obras concluídas entre março de 2011 e dezembro de 2013 (27 216 licenças).

**Tabela 2: Nº de processos em recuperação de informação para atualização do FNA (março 2011 a dezembro 2012)**

| Licenciamento | Obras Concluídas | |
|---|---|---|
| Nº licenças para recuperação | Nº licenças enviadas | Nº licenças recuperadas |
| 17 410 | 28 966 | 27 216 |

## 3. Conclusão

A atualização do FNA constitui um enorme desafio pela sua dimensão e complexidade, assim como pelo facto de o INE recorrer essencialmente a dados administrativos com esta finalidade

As alterações recentes ao nível do SIOU vieram introduzir melhorias importantes através da adaptação do pedido de recolha da informação ao nível do edifício para o nível fogo (alojamento), para as novas construções e demolições.

A BGE, construída a partir da georreferenciação dos edifícios dos Censos 2011 constitui a componente geográfica mais importante do FNA, pelo que a utilização da informação do SIOU para a sua atualização representa um elevado potencial, dado que integra a componente relacionada com a dinâmica urbanística a nível nacional (com desagregação municipal).

Ao longo do ano de 2013, ano de implementação do SIOU, não se verificaram constrangimentos no envio da nova informação pelas Câmaras Municipais, pelo que se garante uma cobertura completa ao nível das novas variáveis recolhidas. A implementação de métodos de recolha via formulários eletrónicos no SIOU, em 2013, contribuíram também para um maior controlo e melhor qualidade da informação recolhida.

A estratégia de atualização da BGE definida pelo INE, assente na gestão partilhada com as Câmaras Municipais representa um compromisso e um garante de qualidade da informação de base, permitindo ainda que a BGE possa assumir um carácter de base nacional oficial de referência, possibilitando a interoperabilidade e partilha de bases de dados nacionais entre diferentes instituições da Administração Pública e a apropriação de dados administrativos na atividade estatística.

## Referências

INE, DMSI (2013). *Modelo de Atualização do Ficheiro Nacional de Alojamentos* – DMSI, outubro 2013.

INE, DEE (2013). *Relatório sobre a efetiva apropriação de dados administrativos provenientes das Câmaras Municipais para o Sistema de Indicadores de Operações Urbanísticas* – INE, dezembro 2013.

**ST I – Sessão INE – 5ª Feira, 10 de Abril, Salão Nobre (17h30)**

# Inquérito às Despesas das Famílias: Porquê? Como? Para quê?

**Eduarda Góis[1], Cristina Gonçalves[2], Esperança Figueiredo[3], Patrícia Pereira[4]**

[1]*Instituto Nacional de Estatística, eduarda.gois@ine.pt;*
[2]*Instituto Nacional de Estatística, cristina.goncalves@ine.pt;*
[3]*Instituto Nacional de Estatística, esperanca.figueiredo@ine.pt;*
[4]*Instituto Nacional de Estatística, patricia.pereira@ine.pt*

**Sumário**

O Instituto Nacional de Estatística (INE) irá realizar em 2015 uma nova edição do Inquérito às Despesas das Famílias (IDEF), sobre a estrutura das despesas dos agregados familiares residentes em Portugal e distribuição dos rendimentos. Porquê a existência deste inquérito, como é realizado, qual a sua finalidade? São questões que pretendemos analisar neste artigo.

**Palavras-chave:** Agregados familiares, Análise de *clusters*, Despesas monetárias das famílias

## 1. Porquê?

O IDEF faz parte da série de dados estatísticos sobre orçamentos familiares iniciada em Portugal em 1967, sendo uma das operações estatísticas mais consolidadas do INE, com utilização crescente em vários sectores nacionais e entidades internacionais. Trata-se de uma operação estatística de grande dimensão, dirigida a uma amostra representativa da população residente no território nacional, estratificada regionalmente, e em que o questionário inclui a utilização de cadernetas para o preenchimento pelas famílias selecionadas de todas as despesas familiares e individuais durante duas semanas. Recolhe também dados demográficos, dados sobre rendimento e sobre as aquisições realizadas com frequência supra quinzenal, através de entrevista direta.

O seu principal objetivo é o apuramento quinquenal da estrutura de despesas familiares de acordo com a COICOP-IDEF (Classificação do Consumo Individual por Objetivo adaptada aos Inquéritos às Despesas das Famílias), concorrendo, deste modo, para a atualização dos ponderadores do Índice de Preços no Consumidor e para as estimativas de Consumo Privado das Contas Nacionais.

O inquérito permite ainda responder ao projeto europeu Household Budget Survey e a vários exercícios de aproximação à dieta alimentar dos residentes, através do estudo das quantidades de bens alimentares adquiridas.

**2. Como?**

Em 2010/2011, o questionário utilizado no IDEF organizava-se em diferentes módulos:

– Caracterização do alojamento, do agregado doméstico privado e dos seus membros, incluindo os rendimentos monetários e a disponibilidade de alguns bens de conforto;

– Um diário de consumo intensivo para o preenchimento das despesas quinzenais de todo o agregado, existindo ainda um diário de consumo intensivo individual para os membros do agregado que preferem fazê-lo separadamente;

– Recolha retrospetiva dos consumos geralmente realizados com frequência mensal, trimestral ou anual, apelando-se para a recordação das despesas efetuados durante os 30 dias anteriores à quinzena de entrevista, durante os três meses anteriores à quinzena de entrevista e no decurso dos doze meses anteriores à quinzena de entrevista; a recolha retrospetiva abrange ainda os recebimentos gratuitos e a título de salário.

A recolha de dados sobre a valorização do autoconsumo, do autoabastecimento, da autolocação e dos recebimentos em géneros e salários em espécie permite o desenvolvimento da caraterização e análise do rendimento total das famílias, complementando os indicadores habituais baseados na distribuição do rendimento monetário.

No inquérito realizado em 2010/2011 foi utilizado pela primeira vez o registo informático na recolha das despesas em bens e serviços de consumo corrente, através da integração da COICOP-IDEF na aplicação informática do inquérito, e no sentido de se obterem ganhos de qualidade, de proximidade local e temporal na relação entrevistador/família. A recolha dos dados sobre o alojamento, agregado, indivíduos, conforto e bens de equipamento, receitas monetárias líquidas e despesas de consumo supra quinzenais mantiveram o método de recolha utilizado nas edições anteriores: entrevista direta presencial com computador (CAPI).

Para a classificação das despesas com bens ou serviços é utilizada a Classificação do Consumo Individual por Objectivo adaptada aos Inquéritos às Despesas das Famílias (COICOP-IDEF), considerada na harmonização subjacente à CCIO (Classificação do Consumo Individual por Objetivo). A COICOP-IDEF associa a cada classe um período de referência específico (quinzenal, mensal ou trimestral de acordo com o que é mais frequentemente expectável), sendo necessário proceder à anualização dos valores das despesas indicados pelos respondentes através da aplicação de fatores multiplicativos que têm em conta o número de períodos no ano (26 no caso da periodicidade quinzenal, 12 no caso da periodicidade mensal, e 4 no caso de consumos a que está associada periodicidade trimestral).

## 3. Para quê?

Durante muitos anos os inquéritos aos orçamentos familiares constituíram em Portugal a grande, e talvez única, referência para o desenvolvimento de estudos e caraterização das famílias portuguesas em termos do rendimento, despesas e condições de conforto. Mais recentemente, a existência de dados anuais sobre a distribuição do rendimento monetário das famílias harmonizados no âmbito EU-SILC não comprometeu a procura de informação dos rendimentos do IDEF, pela sua dimensão e representatividade regionais, e pela possibilidade de análise integrada do rendimento e despesa e das vertentes monetária e não monetária. Possibilita em particular um exercício de consistência de outros instrumentos estatísticos realizados pelo INE.

Para além da publicação e apuramentos específicos disponíveis no Portal de Estatísticas Oficiais, os ficheiros de dados anonimizados disponibilizados pelo INE são amplamente utilizados para o desenvolvimento de análises detalhadas específicas sobre a distribuição do rendimento e despesas das famílias portuguesas. É neste contexto que se apresenta um exercício de análise de *clusters* para a classificação da população em grupos homogéneos segundo o seu padrão de despesas.

## 4. Caracterização de perfis de despesa: uma análise de *clusters*

Através de uma análise de agrupamentos, procedeu-se à agregação dos agregados familiares em *clusters*, com o objetivo de encontrar agregados que possuam um padrão de despesas monetárias homogéneo *intra-clusters* e heterogéneo *inter-cluster*.

As variáveis de agrupamento utilizadas correspondem às despesas monetárias para as onze primeiras Divisões da COICOP-IDEF, com aplicação do ponderador final do IDEF que integra um fator de correção das não respostas e o ajustamento por margens à população residente. Todas as variáveis utilizadas são numéricas, expressas em euros. Procedeu-se à transformação das variáveis originais em variáveis estandardizadas com vista a apresentarem a mesma amplitude.

Devido ao elevado número de casos (n=9 489 agregados; 11 variáveis analisadas) foram utilizados métodos de classificação não hierárquica, nomeadamente o método K-means, que consiste na transferência de um indivíduo para o *cluster* cujo centróide se encontra a menor distância (Reis, 2001). Considerou-se o método K-means com k=4, por ter resultado na melhor diferenciação, coerente e com interesse interpretativo no que se refere ao perfil de despesa monetária dos agregados familiares em estudo.

**Tabela 1: Caracterização dos *clusters***

| CLUSTER 1 | CLUSTER 2 | CLUSTER 3 | CLUSTER 4 |
|---|---|---|---|
| • **Despesa monetária anual média = 23 843€** | • **Despesa monetária anual média = 50 602€** | • **Despesa monetária anual média = 22 654€** | • **Despesa monetária anual média = 8 304€** |
| • **Estrutura da despesa monetária:**<br>• 28,8% em saúde<br>• 13,7% em alimentação<br>• 12,2% em habitação<br>• 12,2% em transportes | • **Estrutura da despesa monetária:**<br>• 21,8% em transportes<br>• 13,5% em hotéis e restauração<br>• 9,3% em alimentação<br>• 8,9% em lazer<br>• 8,7% em habitação | • **Estrutura da despesa monetária:**<br>• 20,9% em transportes<br>• 16,5% em alimentação<br>• 12,3% em habitação<br>• 11,4% em hotéis e restauração | • **Estrutura da despesa monetária:**<br>• 20,2% em alimentação<br>• 19,8% em habitação<br>• 14,3% em transportes |
| • Representavam 4% dos agregados<br>• 56,2% nos 4º e 5º quintis de rendimento<br>• 72,6% dos agregados sem crianças dependentes<br>• 44,2% com pelo menos 1 pessoa idosa<br>• Dimensão média do agregado: 2,6 pessoas | • Representavam 6% dos agregados<br>• 74% no 5º quintil de rendimento<br>• 73,2% dos agregados com crianças dependentes<br>• Dimensão média do agregado: 3,6 pessoas | • Representavam cerca de 30% dos agregados<br>• 58% nos 4º e 5º quintis de rendimento<br>• Dimensão média do agregado: 3,2 pessoas | • Representavam 60% dos agregados<br>• 55,6% nos 1º e 2º quintis de rendimento<br>• 75,1% dos agregados sem crianças dependentes<br>• 44,1% com pelo menos 1 pessoa idosa<br>• Dimensão média do agregado: 2,2 pessoas |

Da análise dos resultados deste exercício, complementada pelo confronto com os restantes dados anonimizados disponíveis, evidenciam-se alguns aspetos:

– *Cluster* 4

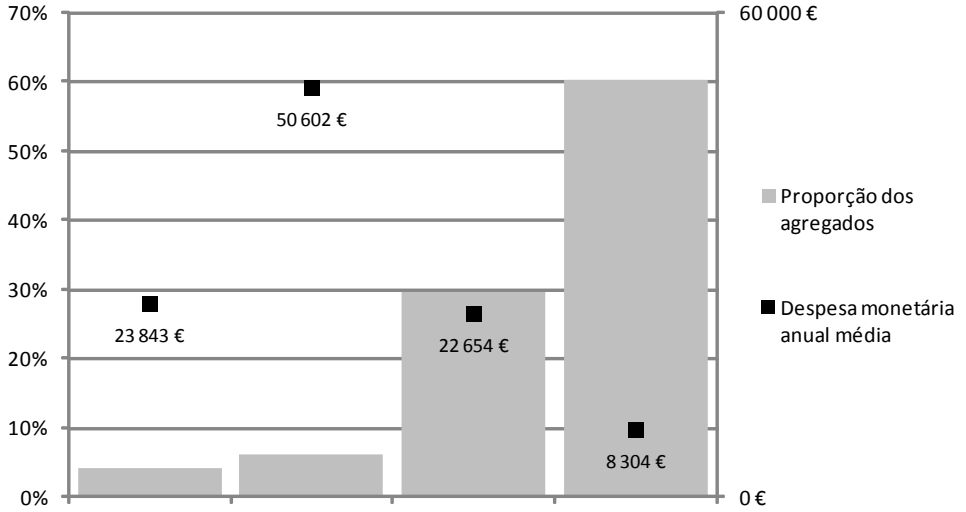
Um dos grupos (*cluster* 4) é muito numeroso − representa 60% das famílias residentes – e regista uma despesa monetária anual média reduzida: 8304€, que corresponde a cerca de metade da despesa monetária anual média no País, 15781€. São maioritariamente famílias sem crianças dependentes e em mais de 40% dos casos com a presença de pelo menos um idoso. Estes agregados familiares são também caraterizados por afetarem mais de metade da despesa monetária à aquisição de bens e serviços básicos (alimentação, habitação e transportes), com um valor médio destas despesas de 4516€, reduzido quando comparado com a média nacional de 7652€. Mais de metade destas famílias situa-se nos dois primeiros quintis de rendimento monetário por adulto equivalente.

– *Cluster* 2

No extremo oposto, encontra-se o grupo de famílias correspondente ao *cluster* 2, em que a despesa monetária anual média, 50602€, mais do que triplica a despesa monetária anual média no País. Ao contrário das famílias do *cluster* 4, o tipo de despesa predominante corresponde às despesas em transportes (21,8%), sendo que os três tipos de despesas em bens e serviços básicos não excedem 40% do total das despesas monetárias. Para estas famílias, o segundo tipo de despesa mais relevante corresponde às despesas com hotéis, restaurantes, cafés e similares. São famílias maioritariamente pertencentes ao quintil de rendimento por adulto equivalente mais elevado e com crianças dependentes, e representam apenas 6% das famílias residentes. O valor médio das despesas monetárias em alimentação, habitação e transportes é elevado (20140€) quando comparado com a média nacional.

– *Clusters* 1 e 3

Os *clusters* 1 e 3 registam despesas monetárias anuais médias de ordem de grandeza semelhante (respetivamente, 23843€ e 22654€), sendo todavia muito distintos na repartição desta despesas por Divisão e na sua representatividade populacional. O grupo de famílias do *cluster* 3, que representa 30% dos agregados familiares, afeta cerca de 50% da sua despesa à aquisição de bens e serviços básicos (alimentação, habitação e transportes). As famílias classificadas no *cluster* 1 representam apenas cerca de 4% dos agregados familiares e são caraterizadas pela predominância das despesas em saúde (28,8%) e por gastos com alimentação, habitação e transportes inferiores a 40%. Em mais de 40% dos casos incluem pelo menos uma pessoa idosa. Em média, as famílias do *cluster* 1 gastam menos 180€ por mês em alimentação, habitação e transportes (9077€) quando comparadas com as famílias do cluster 3 (11245€). Em ambos os *clusters* mais de metade das famílias situam-se nos dois últimos quintis de rendimento monetário por adulto equivalente.




| *Estrutura da despesa monetária* | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Alimentação | 13,7% | 9,3% | 16,5% | 19,8% |
| Habitação | 12,2% | 8,7% | 12,3% | 20,2% |
| Transportes | 12,2% | 21,8% | 20,9% | 14,3% |
| **Saúde** | **28,8%** | | | |
| **Lazer** | | **8,9%** | | |
| **Hotéis e restauração** | | **13,5%** | **11,4%** | |

**Figura 1: Proporção das famílias e Despesa monetária anual média por *cluster***

## Referências

INSTITUTO NACIONAL DE ESTATÍSTICA (2012). *Inquérito às Despesas das Famílias 2010/2011*, Lisboa, Instituto Nacional de Estatística, I.P.

MAROCO, J. (2007). *Análise Estatística com utilização do SPSS*, Lisboa, Edições Sílabo.

REIS, E. (2001). *Estatística Multivariada Aplicada*, 2ªEdição, Lisboa, Edições Sílabo.

**ST I – Sessão INE – 5ª Feira, 10 de Abril, Salão Nobre (17h50)**

# Estimação do desemprego ao nível NUTS III

**Soraia Pereira[1], Luís Correia[2], Pedro Campos[3]**
[1]*Instituto Nacional de Estatística, soraia.pereira@ine.pt;*
[2]*Instituto Nacional de Estatística, luis.correia@ine.pt;*
[3]*Instituto Nacional de Estatística, pedro.campos@ine.pt*

**Sumário**

A quantificação do desemprego assume uma enorme importância social e política nas sociedades contemporâneas. Em Portugal, o Instituto Nacional de Estatística (INE) publica estimativas trimestrais sobre o mercado de trabalho a nível nacional e regional (NUTS I e II). No entanto, para níveis mais desagregados não se consegue obter uma precisão aceitável utilizando o mesmo método de estimação. Neste estudo propõe-se uma metodologia de estimação da taxa de desemprego ao nível NUTS III com base num modelo de regressão logística utilizado pelo *Office for National Statistics* (ONS) do Reino Unido.

**Palavras-chave:** Estimação em pequenos domínios, IEFP, ONS, Regressão logística, Taxa de desemprego.

## 1. Introdução

As estimativas da taxa de desemprego são publicadas trimestralmente pelo INE ao nível regional NUTS II (Norte, Centro, Lisboa, Alentejo, Algarve, Região Autónoma dos Açores, Região Autónoma da Madeira). Estas características são calculadas usando um método direto com base nos dados do Inquérito ao Emprego.

Atualmente, as necessidades de conhecimento do mercado de trabalho impõem estimativas fiáveis para a taxa de desemprego a níveis mais desagregados, nomeadamente ao nível NUTS III (INE 2011).

As NUTS III são domínios de menor dimensão em que se subdividem as NUTS II (Figura 1) e, como tal, a informação sobre algumas das variáveis de interesse não é suficiente para se obter estimativas com precisão aceitável recorrendo ao método já referido. Este é um problema de estimação em pequenos domínios (RAO, J. N. K. 2003).

Este tipo de problemas surge em diversas áreas de aplicação, e originou vários projetos internacionais, tais como o EURAREA (CHAMBERS R. *et al* 2004), o SAMPLE (SAMPLE 2008), entre outros.

Os modelos mais utilizados para resolver o problema de estimação em pequenos domínios são baseados na regressão linear. Entre estes modelos, destaca-se o modelo usado pelo *Office for National Statistics* (ONS), Instituto de Estatística do Reino Unido, que aplicou um modelo de regressão logística para produzir estimativas da taxa de desemprego para domínios territoriais designados por *Parliamentary Constituencies* (ONS 2009).

Este estudo faz uma aplicação do modelo ONS, utilizando como informação auxiliar os dados do Instituto do Emprego e Formação Profissional (IEFP), que disponibiliza mensalmente o número de desempregados registados nos Centros de Emprego ao nível do município.

Ao longo do estudo foram testados outros modelos, embora não sejam descritos devido à limitação de espaço. Faz-se referência por exemplo a MARHUENDA, MOLINA, e MORALES (2013), que descrevem modelos espácio-temporais com base no modelo de *Fay-Herriot*.

**Figura 1: NUTS II e NUTS III de Portugal**

## 2. Metodologia baseada no modelo ONS

A informação do número de inquiridos e do número de desempregados na amostra do Inquérito ao Emprego é agregada em 224 grupos (cruzamento entre 8 grupos sexo x escalão etário e 28 NUTS III referentes a Portugal Continental).

Relativamente à informação mensal proveniente do IEFP, atendendo ao objetivo de calcular estimativas trimestrais é calculada uma média dos três meses do trimestre em causa. Esta informação é igualmente agregada em 224 grupos.

Considere-se então o modelo que relaciona a probabilidade de desemprego $p_{ij}$ para um grupo sexo x escalão etário i (i=1, ..., 8) na NUTS III j (j=1, ..., 28) com as variáveis auxiliares:

$$\log it(p_{ij}) = \beta_0 + \beta_i I_{idadesexo(i)} + \beta_{NUTSII(k)} I_{NUTSII(k)} + \beta_{13} X_{ij} + u_j$$

onde

$$\log it(p_{ij}) = \log_e \left( \frac{p_{ij}}{1 - p_{ij}} \right)$$

e

- $I_{idadesexo(i)}$ e $I_{NUTSII(k)}$ são indicadores *dummy* do grupo sexo x escalão etário i e da região NUTS II k (onde as NUTS III se inserem);

- $X_{ij}$ é o logit da proporção de desemprego registado em cada grupo sexo x escalão etário i na NUTS III j;
- $u_j$ é o efeito aleatório da área; e
- $\beta_0$, $\beta_i$, $\beta_{NUTSII(k)}$, $\beta_{13}$ são os coeficientes de regressão com i=1,...,7 e k=1,...,4 porque um grupo sexo x escalão etário (masculino com menos de 25 anos) e a primeira região NUTS II (Norte) correspondem a categorias *baseline*.

Quando agregadas ao nível NUTS II, as estimativas obtidas pelo modelo não coincidem com as estimativas diretas publicadas a este nível. Para contornar este problema, efetuou-se o seguinte ajustamento:

$$\hat{Y}_j^c = \frac{\hat{Y}_k^{INE}}{\sum_{j=1}^{D_k} \hat{Y}_j} \hat{Y}_j$$

onde $\hat{Y}_j$ é a estimativa do total de desempregados na NUTS III j obtida pelo modelo ONS, $\hat{Y}_k^{INE}$ é a estimativa direta do total de desempregados na NUTS II k (a que a NUTS III j pertence) e $D_k$ é o número de NUTS III da NUTS II k.

Pretende-se estimar a taxa de desemprego que é definida pelo quociente entre o número de desempregados e o número de indivíduos ativos. A metodologia adotada estima o número de indivíduos ativos através da soma da estimativa do total de desempregados com a estimativa direta do número de empregados.

## 3. Resultados e discussão

O modelo descrito foi aplicado usando a função *glmer* do software R.

Os resultados dos coeficientes de variação (Tabela 1), calculados segundo o método de reamostragem *Bootstrap*, permitem concluir que as estimativas obtidas pelo modelo ONS ao nível NUTS III são mais precisas do que as estimativas diretas calculadas a este nível.

Embora não tenham sido apresentados os resultados dos outros modelos testados, o modelo ONS foi o que mostrou maior adequabilidade aos dados.

Em conclusão, os resultados mostraram viabilidade na produção de estimativas do desemprego ao nível NUTS III usando os dados do Inquérito ao Emprego. De facto, este estudo revelou-se de grande importância dado que permitirá ao INE uma possível futura publicação de estimativas da taxa de desemprego ao nível das regiões NUTS III com boa precisão.

**Tabela 1: Taxas de desemprego do 1ºtrimestre de 2012 pelo modelo ONS e pelo estimador direto e respetivos coeficientes de variação, por NUTS III de Portugal Continental**

| | Taxa de desemprego (%) | | Coeficiente de variação (%) | |
|---|---|---|---|---|
| NUTS III | ONS | Método direto | ONS | Método direto |
| Minho-Lima | 13,0 | 13,2 | 6,5 | 18,5 |
| Cávado | 12,6 | 11,3 | 6,9 | 13,4 |
| Ave | 15,0 | 14,2 | 4,8 | 11,8 |
| Grande Porto | 17,4 | 18,2 | 3,5 | 6,4 |
| Tâmega | 16,1 | 17,4 | 4,8 | 10,4 |
| Entre Douro e Vouga | 12,6 | 14,6 | 7,0 | 15,6 |
| Douro | 13,8 | 8,9 | 6,9 | 22,2 |
| Alto Trás-os-Montes | 12,5 | 10,3 | 8,6 | 22,9 |
| Baixo Vouga | 11,8 | 11,4 | 5,2 | 15,4 |
| Baixo Mondego | 12,6 | 15,3 | 5,9 | 17,1 |
| Pinhal Litoral | 10,5 | 10,2 | 5,7 | 25,4 |
| Pinhal Interior Norte | 10,5 | 9,7 | 5,9 | 21,1 |
| Pinhal Interior Sul | 15,7 | 28,8 | 6,8 | 32,7 |
| Dão-Lafões | 11,4 | 13,7 | 6,0 | 20,6 |
| Serra da Estrela | 13,2 | 14,4 | 5,1 | 55,6 |
| Beira Interior Norte | 8,7 | 3,8 | 7,8 | 59 |
| Beira Interior Sul | 12,3 | 8,0 | 5,6 | 88,6 |
| Cova da Beira | 11,5 | 8,8 | 6,5 | 24,2 |
| Oeste | 13,8 | 12,8 | 5,5 | 16,2 |
| Grande Lisboa | 16,1 | 16,3 | 1,7 | 5,9 |
| Península de Setúbal | 17,4 | 16,9 | 4,0 | 8,8 |
| Médio Tejo | 11,9 | 10,7 | 5,7 | 18,2 |
| Lezíria do Tejo | 16,0 | 17,3 | 4,8 | 10,3 |
| Alentejo Litoral | 14,2 | 13,9 | 6,1 | 17,5 |
| Alto Alentejo | 16,6 | 16,2 | 5,3 | 19,9 |
| Alentejo Central | 14,9 | 14,8 | 5,0 | 14 |
| Baixo Alentejo | 14,8 | 12,7 | 6,5 | 19,8 |
| Algarve | 20,0 | 20,0 | 0,0 | 6,3 |

**Agradecimentos:** Agradece-se a contribuição financeira da FCT através da bolsa de investigação com referência SFRH/BGCT/51224/2010.

## Referências

CHAMBERS R., SAEI A., FALORSI S., *et al.* (2004). Linear models that borrow strength across time and space. *EURAREA Project Reference Volume*, Vol. 1.

MARHUENDA, Y., MOLINA, I., and MORALES, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

ONS (2009). Estimates of ILO Unemployment for Parliamentary Constituencies in Great Britain. *Guide for Users*.

RAO, J. N. K. (2003). *Small Area Estimation*, Wiley.

SAMPLE (2008). Small Area Methods for Poverty and Living Conditions Estimates (SAMPLE), http://www.sample-project.eu/ (acedido em 23 de Janeiro de 2014).

INE (2011). Censos 2011 Resultados Definitivos, http://mapas.ine.pt/map.phtml (acedido em 23 de Janeiro de 2014).

**ST I – Sessão INE – 5ª Feira, 10 de Abril, Salão Nobre (18h10)**

# Série longa do Índice de Preços no Consumidor (1948 – 2013)

**Vitor Hugo Quaresma Mendonça[1], Anabela Costa da Silva[2]**

[1] *Instituto Nacional de Estatística, vitor.mendonca@ine.pt;*

[2] *Instituto Nacional de Estatística, anabela.silva@ine.pt*

**Sumário**

O Índice de Preços no Consumidor tem por finalidade medir a evolução dos preços de um conjunto de bens e serviços representativos da estrutura de consumo da população residente, sendo reconhecido como indicador de síntese da evolução dos preços na economia. Este projeto permitiu compilar uma série mensal de índices com um grau de detalhe relativamente elevado para um período alargado, minimizando as dificuldades inerentes à heterogeneidade das diversas séries.

**Palavras-chave:** Índice de Preços no Consumidor, Inflação, Números índice, Série longa, Séries temporais.

## 1. Introdução: o que é o Índice de Preços no Consumidor?

O Índice de Preços no Consumidor (IPC) é um indicador que mede a evolução dos preços de um conjunto de bens e serviços, de qualidade constante, considerados representativos da estrutura de despesa do consumo privado da população residente num espaço geográfico delimitado, através da variação do nível de preços entre dois períodos. É uma estatística essencial para a formação de expectativas e para a tomada de decisão por parte dos agentes económicos. A taxa de inflação, medida pelo IPC, é um elemento fundamental na definição de atualizações salariais e na atualização de diversos tipos de contratos, nomeadamente de arrendamento, entre outros.

Para o cálculo do IPC são utilizadas ponderações que se obtêm a partir da estrutura das despesas monetárias do consumo final das famílias de referência do índice. Esta estrutura não é estável pois depende da evolução das preferências de consumo das famílias. É necessário, então, ter em conta as modificações da estrutura de consumo e assegurar a representatividade da amostra dos produtos. Por esta razão, a construção do índice assume bases diferentes, que evoluem no tempo. Essas bases resultam de inquéritos feitos às famílias com o propósito de atualizar as respetivas estruturas de despesa.

## 2. Cronologia do IPC

As diversas alterações metodológicas verificadas ao longo das várias bases do IPC impuseram dificuldades nas tentativas de compilação de séries longas homogéneas. Efetivamente, revisões periódicas de ponderadores, introdução de novas categorias de bens e serviços, ajustamentos de conceitos e de classificações, bem como de metodologias de

agregação e cálculo ou alterações nos processos de recolha de preços, tornam as séries de preços nas diversas bases muito heterogéneas, impossibilitando a sua simples união. Existem várias séries de preços que representam realidades e conceitos diferentes e reuni-las numa só série homogénea pressupõe um longo e complexo trabalho. As séries consideradas na compilação da série longa do IPC foram as seguintes:

- IPC das seis cidades (1948-1976) – O índice das seis cidades (Lisboa, Porto, Coimbra, Évora, Viseu e Faro) apenas considerava as despesas das "famílias cujos chefes de família fossem operários, empregados de escritórios, de comércio, funcionários públicos até à categoria de primeiro-oficial e professores primários", obtidas a partir dos Inquéritos às Condições de Vida. A amostra da maioria das cidades era composta por cerca de 250 produtos, sendo a cidade de Faro a única com menos de 200 produtos.

- IPC base 1976=100 (1977-1987) – Esta série tinha por base o Inquérito às Despesas Familiares 1973/1974, que teve por amostra as famílias com "uma a cinco unidades de consumo, rendimentos anuais entre os 30.000 e os 180.000 escudos e cujo elemento principal fosse trabalhador por conta de outrem". A amostra era composta por 286 produtos, sendo recolhidos mensalmente cerca de 18.000 preços. A cobertura geográfica desta série era o Continente e as Regiões Autónomas, não tendo sido calculados índices Nacionais.

- IPC base 1983=100 (1988-1990) – A principal inovação da série de base 1983=100 foi o alargamento do âmbito à "população total sem limitações dimensionais, de rendimento e de categorias socioprofissionais", com base no Inquérito às Receitas e Despesas Familiares 1980/1981. A cobertura geográfica manteve-se igual à da série anterior, passando a ser considerados 524 produtos, correspondendo a cerca de 25.000 preços recolhidos por mês.

- IPC base 1991=100 (1991-1997) – Com esta série começou a ser calculado o IPC Nacional para o total da população, sem restrições, com base no Inquérito aos Orçamentos Familiares 1989/1990. Eram recolhidos cerca de 65.000 preços, correspondentes a 577 produtos.

- IPC base 1997=100 (1998-2002) – A principal alteração face à série anterior foi a seleção de uma nova amostra com base no Inquérito aos Orçamentos Familiares de 1994/1995, resultando num aumento do número de produtos para 700, correspondentes a cerca de 75.000 preços recolhidos mensalmente.

- IPC base 2002=100 (2003-2008) – Os resultados do Inquérito aos Orçamentos Familiares 2000 foram integrados, permitindo um aumento do número de produtos considerados (mais de 800) e do número de preços recolhidos mensalmente (cerca de 95.000). A nível de cálculo, esta foi a primeira série do IPC em que os índices são do tipo encadeado, permitindo alterações anuais de amostra, bem como a atualização dos

ponderados a preços de dezembro de cada ano, seguindo a metodologia definida pelo Eurostat para o Índice Harmonizado de Preços no Consumidor (IHPC).

- IPC base 2008=100 (2009-2012) – Com base no Inquérito às Despesas das Famílias 2005/2006, a amostra foi reforçada, passando a ser considerados perto de 1.200 produtos, resultando em cerca de 110.000 preços recolhidos por mês em cerca de 11.500 estabelecimentos.

- IPC base 2012=100 – A série atual do IPC é a primeira cujos ponderadores de níveis superior são obtidos com base nas despesas monetárias de consumo final das famílias das Contas Nacionais Portuguesas, em linha com a regulamentação da Comissão Europeia e as recomendações do Eurostat sobre o IHPC. A um nível mais desagregado, a amostra foi selecionada com base nos resultados do Inquérito às Despesas das Famílias 2010/2011, sendo recolhidos preços para cerca de 1.200 produtos, com cerca de 120.000 recolhas mensais em mais de 13.000 estabelecimentos.

## 3. Construção da série longa

A construção da série longa do IPC assentou fundamentalmente na harmonização das nomenclaturas com base na estrutura atual do IPC ao nível mais desagregado possível, na agregação das diversas subséries com base em índices não-encadeados e ponderadores atualizados anualmente e na ligação dos índices obtidos de forma retrospetiva, partindo dos índices da base 2012=100.

As séries consideradas para a compilação da série longa do IPC são baseadas em estruturas que não correspondem à Classificação do Consumo Individual por Objetivo (COICOP) utilizada atualmente no IPC. Enquanto atualmente são consideradas, por exemplo, 12 classes de produtos, na série de base 1976=100 apenas existiam 5 classes de produtos. Assim, foi necessário harmonizar as nomenclaturas das diversas séries de modo a conseguir calcular índices para níveis de desagregação até agora indisponíveis.

Para as séries em que não tinha sido calculado o índice Nacional, foi necessário agregar as Regiões Autónomas com base nos dados disponíveis de despesas das famílias, Censos e outras fontes, de modo a replicar a metodologia atual. Além disso, foram integrados dados sobre a evolução dos preços das rendas de habitação, que não eram considerados nas séries mensais do IPC entre 1977 e 1997, apesar de existir um inquérito autónomo às rendas de habitação.

Finalmente, a utilização de índices não encadeados permitiu a construção da série longa de base única, ao contrário das diversas séries existentes cuja ligação apenas era efetuada com base em momentos de sobreposição temporal das diversas séries.

Para além dos aspetos metodológicos, o principal desafio na construção da série longa do IPC foi a localização, transcrição e validação dos dados desagregados considerados nas séries

originais (índices e estruturas de ponderação). Foi necessário um esforço de transcrição de dados de modo a ser possível utilizar ferramentas informáticas na compilação desta série longa. Este processo permitiu também melhorar a qualidade dos índices, pois foi possível aumentar a precisão dos cálculos e corrigir alguns erros de agregação detetados nos dados dos períodos mais antigos.

## 4. Apresentação da série e dos principais resultados

Como resultado deste trabalho, conseguiu compilar-se uma série do IPC para o período 1949-2013, consistente com a metodologia atualmente utilizada no cálculo mensal do IPC.

O comportamento dos preços ao longo deste período de 65 anos (Gráfico 1) refletiu vários ciclos económicos, nomeadamente, as transformações das características e do modo de funcionamento da economia portuguesa, as alterações significativas que se produziram no enquadramento internacional e as políticas económicas adotadas nesse período.

Entre 1948 e meados da década de 1960 a inflação manteve-se estável em níveis relativamente baixos. A partir de meados da década de 60, os preços evidenciaram aumentos mais acentuados.

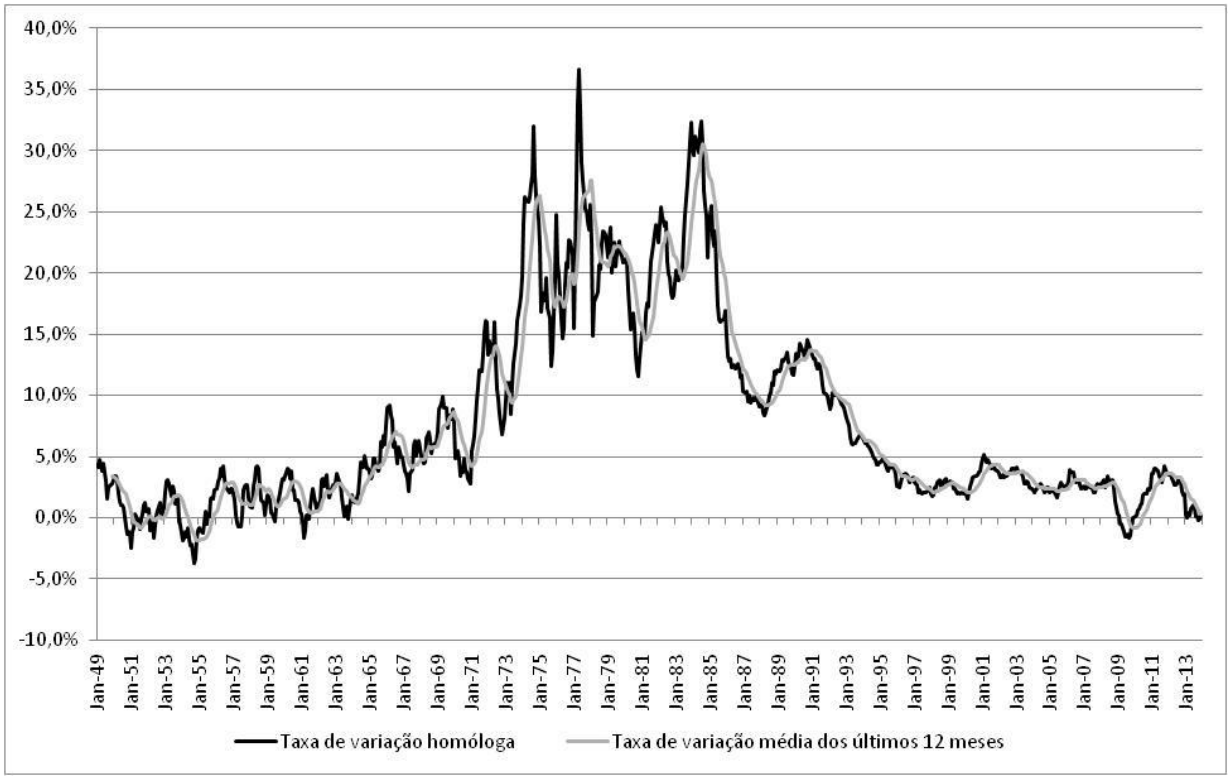



**Gráfico 1: Taxas de variação do Índice de Preços no Consumidor**

No início da década de 1970, o IPC registou uma aceleração para níveis mais elevados, com a taxa de variação média de 12 meses a ultrapassar os 10% em dezembro de 1971 e a

manter-se em valores superiores ou muito próximos nos anos seguintes. A desvalorização substancial da moeda nacional, a que se assistiu a partir da segunda metade da década de 70 e que se prolongou na primeira metade da década seguinte, foi outro fator decisivo para os elevados níveis de inflação que então se registaram. Entre abril de 1974 e dezembro de 1985 a taxa de variação média de 12 meses do IPC oscilou entre um máximo de 30,5% em agosto de 1984 e um mínimo de 14,5% em abril de 1981. Durante a grande maioria dos meses neste período a variação média anual situou-se próxima ou acima de 20%.

Na segunda metade da década de 80, o IPC desacelerou de forma significativa, num enquadramento externo marcado pela adesão à Comunidade Económica Europeia em 1986 e pela redução substancial nos preços do petróleo em 1984-1985.

Na década de 90, e no contexto da preparação da participação de Portugal na União Monetária, continuou a redução da taxa de inflação refletindo nomeadamente a alteração da política monetária que passou a orientar-se para a estabilidade cambial como objetivo intermédio para atingir a estabilidade de preços. Efetivamente, desde março de 1995, a taxa média de variação do IPC manteve-se sempre abaixo dos 5%.

Desde a introdução do euro que a taxa de variação média apresenta valores relativamente baixos, tendo mesmo ocorrido variações negativas durante o ano de 2009 na sequência da crise económica mundial iniciada no final do ano anterior.

**Agradecimentos:** Fundação para a Ciência e a Tecnologia: Bolsa SFRH/BGCT/51126/2010 (Ana Paula Diogo) e Bolsa SFRH/BGCT/51751/2011 (Anabela Silva).

**Referências:**

DIOGO, A. P. (2011). *Série longa de inflação em Portugal – Análise do período 1976-2010 com base no IPC*. Tese de Mestrado em Economia ISCTE/IUL.

INE (2012). *Destaque do Índice de Preços no Consumidor* – Abril 2012.

INE (2014). *Destaque do Índice de Preços no Consumidor* – Janeiro 2014.

SILVA, A. (2013). *Séries longas de inflação em Portugal 1977-1948*.

**ST II – Sessão Banco de Portugal – 6ª Feira, 11 de Abril, Salão Nobre (10h40)**

# Non-financial sector indebtedness

**Filipa Lima[1], Inês Correia[2], Rodrigo Batista[3]**
[1]*Banco de Portugal, slima@bportugal.pt;*
[2]*Banco de Portugal, ipcorreia@bportugal.pt;*
[3]*Banco de Portugal, rsbatista@bportugal.pt*

**Abstract**

The monthly publication of statistics concerning the non-financial sector indebtedness was one of the most outstanding achievements of the Statistics Department of Banco de Portugal in the recent years. Combining different dimensions of analysis, through the use and matching of the databases within the Department, it allows an innovative insight to the indebtedness of the sector. In this paper we briefly present the compilation methodology and some of the results that can be drawn from the data.

**Keywords:** Central bank statistics, Indebtedness, Micro-data, Non-financial sector.

## 1. Introduction

The thorough assessment of the current Portuguese economic and financial context by the three international organisations participating in the EU/IMF Financial Assistance Programme (FAP) proved to be quite demanding in terms of information requirements, with increasing requests for more detailed information. These requests focused in several areas of the economy with a special attention being drawn upon the indebtedness levels, not only for the general government, but for the entire non-financial sector. In February 2012, Banco de Portugal initiated the publication of the new chapter K and section A.20 of the Statistical Bulletin on the debt of the non-financial sector (Figure 1). This publication reflects the concerns of Banco de Portugal in making accessible to the public the information required by the international organisations.

This new chapter provides an innovative insight to the indebtedness of the non-financial sector since, for the first time, it provides data for several dimensions of analysis, namely: debtor and creditor sectors, type of financial instrument, original maturity, economic activity and size of the company. These dimensions are crossed between them offering information at an unprecedented level, even when comparing at an international level.

## 2. Methodological framework

This approach to data compilation requires some preconditions. Starting by the classification of the debtor according to its institutional sector, it is necessary to define the non-financial sector. The non-financial sector is composed of several entities that can be separated into two distinct groups, public or private entities, according to their ownership. Within the

private sector, the debtors can be allocated to the private corporations sector, in case they are a company, or otherwise to the private individuals sector. Within the public sector the allocation is made considering whether or not the entities are within the scope of the general government. It is important to mention that the public corporations can either be inside or outside the general government sector (Figure 1). The private corporations are also classified according to their sector of activity and their size.

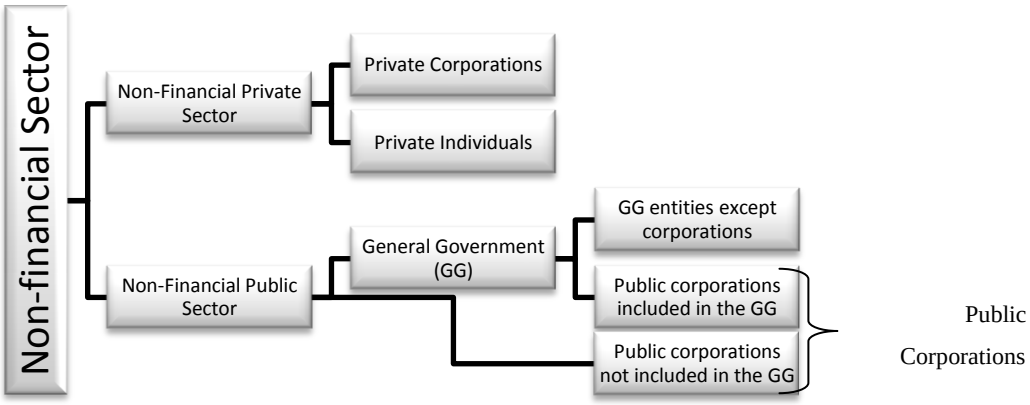



**Figure 1: Delimitation of the non-financial sector**

Information on the debtor side is crossed with that on the creditor's so that it is possible to measure how much funding is being channelled by which creditor to which debtor. Five categories of creditor are specified, four of them internal (general government, financial sector, corporations, private individuals) and the last concerning external creditors. The data is further broken down by financial instrument and original maturity.

As stated in the Statistical Press Release, "the concept of debt presented in this new chapter includes loans, debt securities and trade credits. In the case of general government, it includes also saving certificates, Treasury certificates and other Treasury liabilities. The values presented are based on end-of-period positions valued at nominal value, excluding accrued interest" [BANCO DE PORTUGAL (2012)]. Unless stated otherwise, the data refer to non-consolidated debt.

## 3. Databases and mapping

The combination of the several dimensions of analysis is only possible with the use and matching of data available from the several databases managed by the Statistics Department of Banco de Portugal. Besides the direct contribution given by the different data sources to the final output, there are information flows between the data sources themselves, where the aggregated data are complemented by data coming from micro-databases. For example we map information provided by the Central Balance Sheet Database (CBSD) to the Central Credit Register (CCR) so that we can breakdown bank loans by the private companies' size. This

upgrade to the information of the CCR micro-database will afterwards be used to allow an allocation of the information from Monetary and Financial Statistics.

Crucial in this process is the correct matching between the different databases, for which a unique key identifier is used in a common list of entities. This key identifier is the element that allows a coherent classification of the debtors (size, economic activity sector, institutional sector) in the different micro-databases (CBSD, CCR, securities database and external operations database).

## 4. Statistical analysis

Looking at the results, it is possible to see that the debt-to-GDP ratio of the Portuguese non-financial sector has increased from 346% in 2007 to 445% in 2013 (Figure 2).

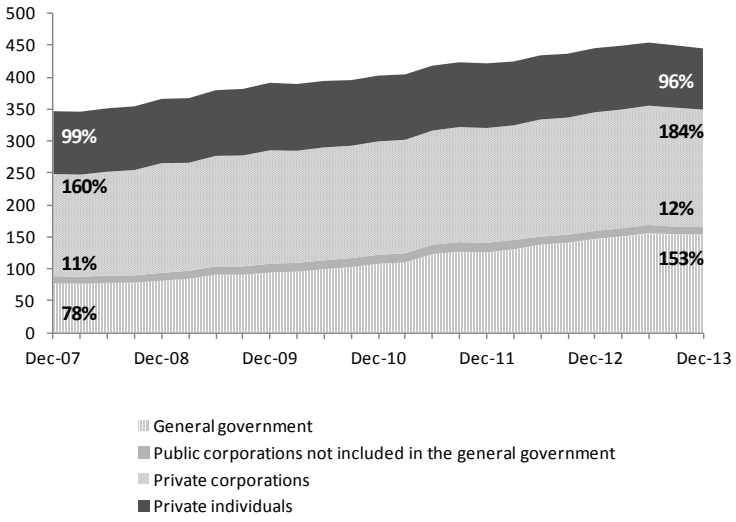



Figure 2 shows the composition of the non-financial sector debt. While the structure did not vary, both the general government and the private companies registered significant increases, +75 p.p. and 24 p.p., respectively. For private individuals, there was a small decrease from 2007 to 2013 from 99% to 96% of the GDP. Public companies not included in the general government remained stable at 12% of GDP.

**Figure 2: Debt-to-GDP Ratio**

Figure 3 shows, for each type of debtor, the structure of the creditors at end-2013. The detail by type of instrument is presented in Figure 4. Data for end-2007 are in parenthesis.

From the figure it is observable that the share of the external sector in the debt of the general government decreased from 67% in 2007 to 56% in 2013, whereas the weight of the financial sector increased by 12 percentage points (p.p.). Regarding private corporations, the share of the external sector almost doubled, from 12% to 21%, in the same period, whereas the financial sector decreased its weight by 10 p.p.. For public corporations, the funding distribution across creditor's sector changed considerably, with an increase from almost 0% to 25% in the share of the funding coming from the general government while the share coming from the external sector decreased by 26 p.p..

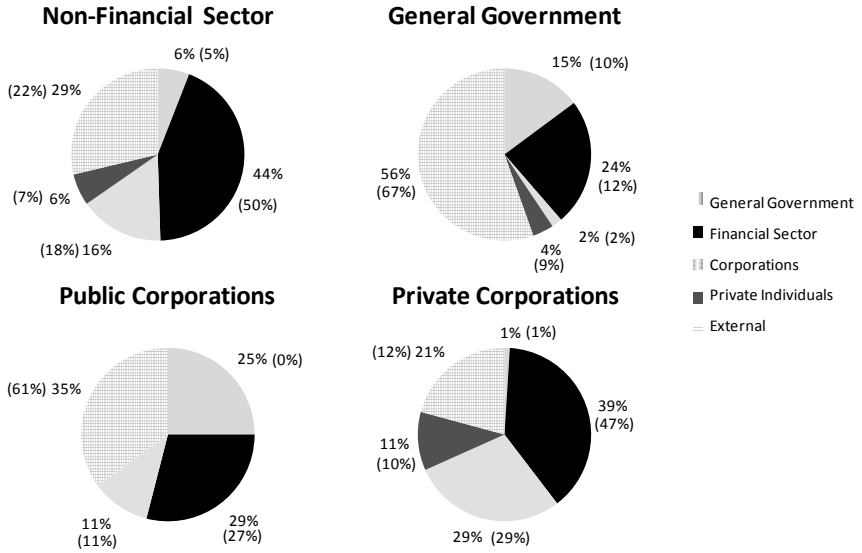



**Figure 3: Non-financial sector debt by creditor sector Dec-13 (Dec-07)**

The financing structure of the general government changed significantly between 2007 and 2013: while the external sector is still the main creditor, debt securities (59% at end-2007, 25% at end-2013) were replaced by loans granted under the FAP (31% at end-2013, 7% at end-2007); internal financing was mainly done through debt securities (+13 p.p.). For public corporations, domestic loans (mainly granted by the general government) represented at end-2013 the largest share (48%, 25% at end-2007) while at end-2007 external loans dominated (45%, 23% at end-2013). For private corporations, domestic loans decreased from 63% at end-2013 to 59% at end-2007, compensated by both external loans and debt securities held by non-residents (+4 p.p. each).

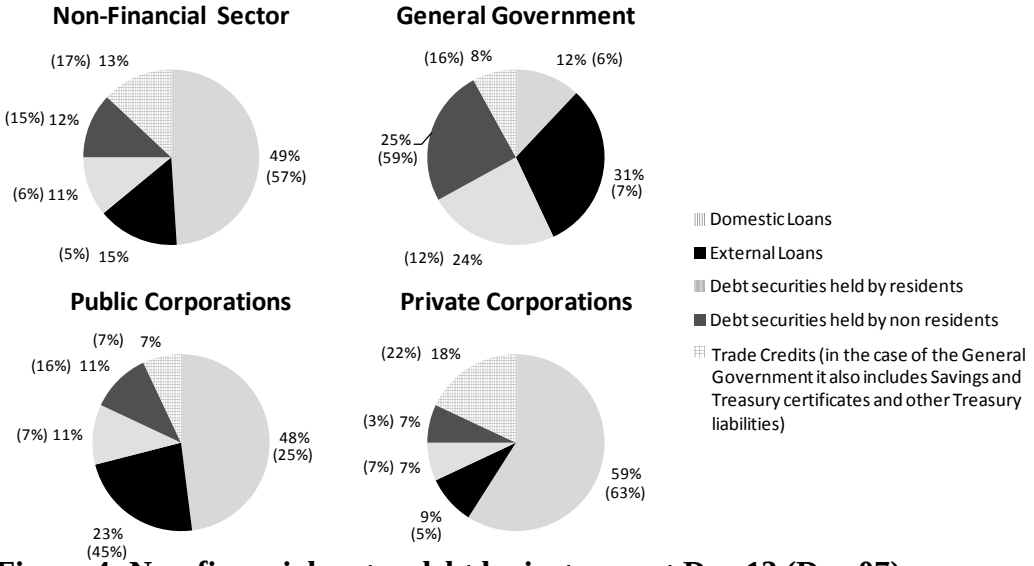



**Figure 4: Non-financial sector debt by instrument Dec-13 (Dec-07)**

**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 (2012), *New chapter on non-financial sector indebtedness (Statistical Press Release)*

**ST II – Sessão Banco de Portugal – 6ª Feira, 11 de Abril, Salão Nobre (11h00)**

# High-growth enterprises in Portugal

**Homero Gonçalves[1], Mário Lourenço[2], Vítor Silveira[3]**
[1] *Banco de Portugal, hgoncalves@bportugal.pt;*
[2] *Banco de Portugal, mfllourenco@bportugal.pt;*
[3] *Banco de Portugal, vfsilveira@bportugal.pt*

**Abstract**

In an economic crisis, such as the most recent one, examples of success and growth, counter-cyclical to the general recessive environment, are often used as beacons for other companies towards a path of recovery. Information available at Banco de Portugal allows the identification of a set of companies with high growth rates and of its distinctive features, as well as of the economic activities within which its presence is most noteworthy.

**Keywords:** Enterprises, Growth**,** Dynamics, Micro-data.

## 1. Motivation and conceptual framework

The Portuguese economy faced a significant contraction over the last few years during which unemployment grew to record levels. It is then crucial to restore growth and for that a key issue relates to the promotion of entrepreneurial dynamics. High-growth enterprises (HGE) play an important role in this matter as they are usually linked with innovation and job creation. A better knowledge about these firms would allow policy makers to develop appropriate approaches in order to maximize the chances of developing HGE [OECD (2010)].

In order to identify such companies, it is common to consider the EUROSTAT-OECD (2007) approach according to which HGE are those achieving an average annual growth rate above 20% for a period of three consecutive years, having such growth evaluated either considering the number of employees or, as is the case of the present analysis, the turnover.

## 2. Methodological approach and data description

The need for in depth knowledge of the non-financial corporations (NFC) sector in Portugal led Banco de Portugal to the development of a business register combining data from the several databases it manages while also using other administrative sources. This business register, which is being constantly automatically updated with the most recent relevant information, gathers information on each enterprise's characteristics through a time-span of over 20 years [GONÇALVES et. al. (2013)].

Identifying HGE as a subset of this population is not straightforward. Potential HGE's turnover variations must be derived organically from their current activity, i.e., unrelated to other effects that could bias the analysis. Hence, in this exercise, enterprises involved in

mergers and acquisitions, or having registered recent operating activity shifts or other similar events, which may have led to an artificial growth, were excluded from the reference population of NFC.

Companies registered in Madeira's Free Trade Zone and firms with no employees were also excluded, as well as companies with a turnover below 50 thousand euros, for which the high turnover growth rates would only result from a relatively low initial level. Finally, since the criteria for a company to be defined as HGE derives from its turnover's average annual growth rate, enterprises active for less than four years were excluded, given that the turnover of its start-up years does not refer to a full year [BANCO DE PORTUGAL (2013)].

The set of companies which resulted from the implementation of these cut-offs is considered to be the population of potential HGE, i.e., the relevant population of NFC regarding the present analysis.

## 3. Some relevant findings

### 3.1. Population of potential HGE and HGE

The population of potential HGE in 2012 gathered 37% of the NFC operating in Portugal, a relative share considerably higher when taking in consideration the turnover and number of employees (about three quarters of the total, in both cases). Only 7% of such companies were considered to be HGE (3% of the reference population of NFC), a share 4 p.p. lower than the one registered in 2006, a decrease that could be explained by the recessive context of the last years (Figure 1). The impact of the economic crisis is also noticeable in the growing number of companies with negative annual turnover growth rates in this period (45% of the population of potential HGE in 2006 and 64% in 2012) (Figure 2).

### 3.2. Some characteristics of Portuguese HGE

The data reveal that companies are only HGE as a transitory phase in their life cycle. From over 50 thousand HGE identified between 2006 and 2012, more than 1/2 were considered to be HGE only once (56%), while companies considered HGE for more than four times represented only 1.5% of the referred total.

By company size, microenterprises are the class where HGE were most relevant (64% of total HGE in 2012), nevertheless below its weight in the population of potential HGE (74%), a situation opposing that of SMEs (35% of total HGE and a weight of 25% in the population of potential HGE) and large enterprises (0.9% e 0.6%, respectively). Analysed from another perspective, HGE stood for 6% of the microenterprises in the population of potential HGE in 2012, 10% of the SMEs and close to 11% of the large enterprises.

According to the geographic location, in 2012, around 42% of the HGE were headquartered in the northern region of Portugal, which compares to 28% in Lisbon region and 20% in Centre Portugal. When compared to 2006, the relevance of the northern region increased 7 p.p., while Lisbon and the Algarve went down by 7 p.p. and 1 p.p., respectively.

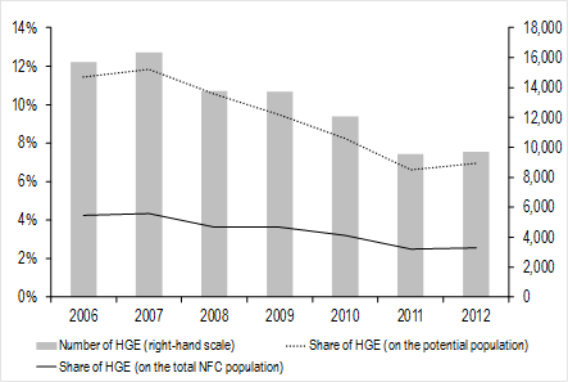



**Figure 1: HGE and the population of potential HE**

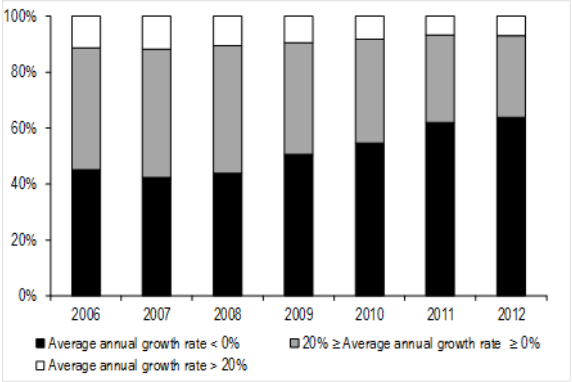



**Figure 2: Average growth rate in the population of potential HE**

HGE are usually young companies. In 2012, 56% of Portuguese's HGE had been operating for less than 10 years, a maturity class that weighted 34% in the population of potential HGE in the same year. Only 15% of the HGE had been active for more than 20 years (29% of the total population of potential HGE in 2012).

### 3.3. Economic activities where HGE are most relevant

Among the economic activities with a significant share of HGE in 2012, it is worth noticing the presence of some of the Portuguese traditional activities, such as the manufacture of leather (dominated by the manufacture of footwear), fishing and forestry. On the other hand, it is also interesting to identify in this group activities like air transport, scientific research and development and pharmaceutical products.

As distinctive characteristics these economic activities have:

- A growing number of active companies: an average increase of 15%, while the population of potential HGE registered an average decrease of 4% between 2009 and 2012;
- A relatively lower median age: around 11 years, which compares with 14 years in the population of potential HGE;
- A larger share of large enterprises: standing for 8%, on average, while in the population of potential HGE this class stood for only 0.6%, in 2012; and
- Diminishing turnover's concentration on large companies: on average, large companies' share of turnover within these activities decreased 2 p.p. between 2009 and 2012, while it increased 3 p.p. within the population of potential HGE.

### 3.4. HGE's role on employment and on loans from resident credit institutions

Although HGE's share of total employment in the NFC's sector is small (6% in 2012), these companies stand out by their ability to create jobs during the high-growth period (on average, by almost 60%). Along the high growth period, approximately 2/3 of these companies increased the number of employees, 21% maintained the employment level and only in 13% the number of employees diminished (Figure 3). The level of employment growth differs across firms, depending on the labour intensity of each company's production system, which is often related to its economic activity.

Information on loans granted by resident credit institutions suggests that the Portuguese financial system may be able to differentiate HGE from the remaining companies. Indeed, on average, throughout the high growth period of HGE, loans granted by resident credit institutions increased by about 7 times (Figure 4). This compares with an average null growth in the NFC sector as a whole in the 2006-2012 period.

Moreover, there is also some evidence that these companies pose less risk to the credit institutions, given that their non-performing loans ratio is very small (2% in 2012), a distinctive feature particularly relevant in recent years considering that the NFC's non-performing loans ratio has been steadily increasing, reaching 10% in 2012.

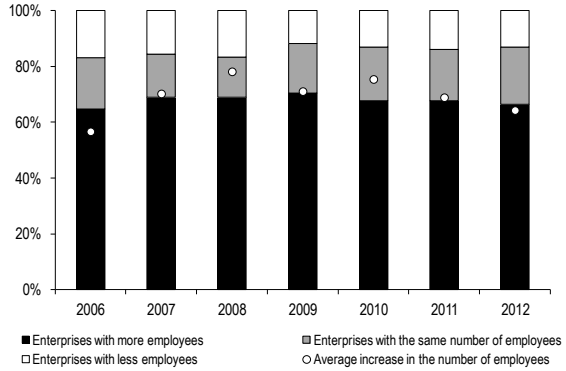



**Figure 3: HGE's employment**

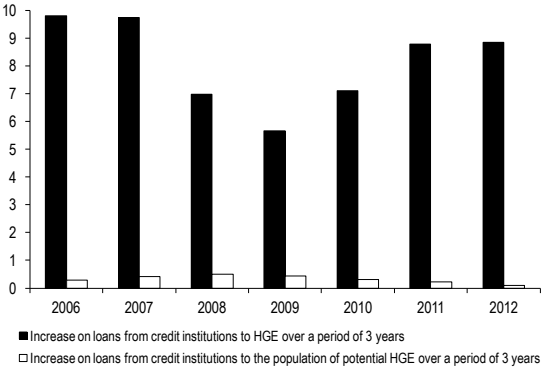



**Figure 4: HGE's loans from resident credit institutions**

### Disclaimer

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

### References

BANCO DE PORTUGAL (2013), *Estrutura e Dinâmica das Sociedades Não Financeiras em Portugal 2006-2012 (Portuguese version)*.

EUROSTAT – OECD (2007), *Manual on Business Demography Statistics*, Methodologies and Working Papers.

GONÇALVES, H. and LOURENÇO, M. (2013), Building business registers to monitor entrepreneurial dynamics, *IFC Bulletin*, 37, 42-45.

OECD (2010), High-growth enterprises – What Governments can do to make a difference, *OECD Studies on SMEs and Entrepreneurship*.

**ST II – Sessão Banco de Portugal – 6ª Feira, 11 de Abril, Salão Nobre (11h20)**

# Quarterly time-series from Central Balance Sheet Database

**Cloé Magalhães[1], Pedro Cordeiro[2], Rita Poiares[3]**
[1]*Banco de Portugal, clmagalhaes@bportugal.pt;*
[2]*Banco de Portugal, pccordeiro@bportugal.pt;*
[3]*Banco de Portugal, rmpoiares@bportugal.pt*

**Abstract**

The Central Balance Sheet Database of Banco de Portugal (CBSD) is a repository of annual and quarterly individual information of accounting and statistical nature, covering a wide range of non-financial corporations (NFC) for the period 1990-2013 (annual data until 2012). The methodology for the compilation of quarterly time-series has recently been improved, to obtain quarterly financial indicators representative of the population of NFC operating in Portugal [BANCO DE PORTUGAL (2013)].

**Keywords:** Ancillary information, Benchmarking, Imputation, Ratio estimator, Sampling

## 1. Sources of information

The process for obtaining quarterly time-series on NFC is based on a broad set of information. In addition to the databases that feed the CBSD – the reference population of NFC, the Simplified Corporate Information (IES) and Quarterly Survey on Non-Financial Corporations (ITENF) the process also uses information from the Central Credit Register (CCR) and the Securities Statistics Integrated System (SSIS) in the annual procedure.

### 1.1. Central Balance Sheet Databases

The CBSD incorporates three sources of information for the compilation of statistics on NFC. The annual data is collected through IES, which is, since 2007, a mandatory report from all entities operating in Portugal to four public entities – Tax Authority, Ministry of Justice, Statistics Portugal (INE) and Banco de Portugal. This database contains over 370.000 corporations a year, corresponding to a coverage rate above 95% of all NFC. Information collected through IES is chiefly of an accounting nature, based on the financial statements and the respective annexes set out in the accounting standards. It also comprises a range of data with additional detail on the activity and situation of the corporations, as necessary for statistical purposes [BANCO DE PORTUGAL (2008)].

As for the quarterly data, the ITENF is a statistical operation under the joint responsibility of Banco de Portugal and Statistics Portugal to collect quarterly information from a sample of around 4.000 NFC. Information collected is a subset of the variables collected through IES, and covers mainly a range of accounting variables relating to the activity and financial situation of corporations [INE (2012)].

Finally, the information on the population of NFC comes from the reference population of NFC, which is a business register that combines information from several sources from Banco de Portugal and other public entities, such as the Ministry of Justice and Statistics Portugal [GONÇALVES et al. (2013)]. This database comprises, for each non-financial corporation operating in Portugal, a set of structural information (legal person identification number, location of the head office and sector of economic activity) as well as economic variables (number of employees, turnover, total assets and equity, for each year).

### 1.2. Other databases

The CBSD also uses information from other databases managed by the Statistics Department. The SSIS is a database with data on securities issues and portfolios, on a "security-by-security" and "investor-by-investor" basis. As for the CCR, this is a database with information on actual and potential credit liabilities towards financial credit institutions granting credit in Portugal.

## 2. The annual procedure

The information of the reference population of NFC concerning the sector of economic activity and the annual turnover for each corporation is considered the foundation stone of this procedure. In order to fill information gaps existing in the annual database, a two-stage procedure has been defined: (i) the imputation of total assets, combining *cold-deck* and mean imputation methods in three sequential steps, in order to improve the information from the reference population of NFC to be used in subsequent procedures; (ii) with the outcome of the first stage, a treatment of non-response is applied over annual data in order to produce information relating to all NFC. In this stage, we have implemented a mean imputation process in three sequential steps.

As Figure 1 shows, this procedure allows the estimation of data for about 2% of the NFC with no response to IES in 2012.


**Figure1: Non-response rate, by sector of economic activity (2012) (%)**

## 3. The quarterly procedure

The quarterly procedure estimates the quarterly population totals for a set of variables, combining quarterly data from ITENF database with the ancillary information that results from the annual procedure. The latter set of information is incorporated in two different steps: first, by the post stratification of the sample, and then by the ratio estimator.

### 3.1. Post stratification of the sample

The extrapolation methodology includes the post stratification of the sample, by sector of economic activity and one out of two quantitative ancillary variables, each one directed to the estimation of a specific set of variables: (i) Turnover, for the activity variables, trade credits and inventories, or (ii) Total assets, for the remaining balance-sheet and interest expenses indicators. The post strata are defined with ancillary information from the reference population of NFC.

### 3.2. Ratio estimator

The ancillary information available in the annual database is useful to calibrate the quarterly estimates, through a ratio estimator, defined as follows:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \left(\frac{y_i}{\pi_i}\right) \times \frac{\sum_{j=1}^{N_h} AQV_j}{\sum_{i=1}^{n_h} \frac{AQV_i}{\pi_i}}$$

Where $\hat{Y}_h$ is the ratio estimate for the population total of variable Y in post stratum $h$, $n_h$ is the number of responses in post stratum $h$, $N_h$ is the number of corporations in post stratum $h$ in the population, $y_i$ is the variable Y for corporation $i$, $\pi_i$ is the probability of selection for corporation $i$, and $AQV_i$ is the ancillary quantitative variable for corporation $i$.

There are two possible ancillary quantitative variables: (i) Total income, for activity, trade credits and inventory indicators, and (ii) Total assets, for the remaining balance-sheet and interest expenses indicators.

Figure 2a uses, as an example, the Obtained funding for the *Manufacturing, mining and quarrying* sector to compare the Horvitz-Thompson estimator with the ratio estimator, as defined before, and we conclude that the latter has a smaller deviation from the annual.

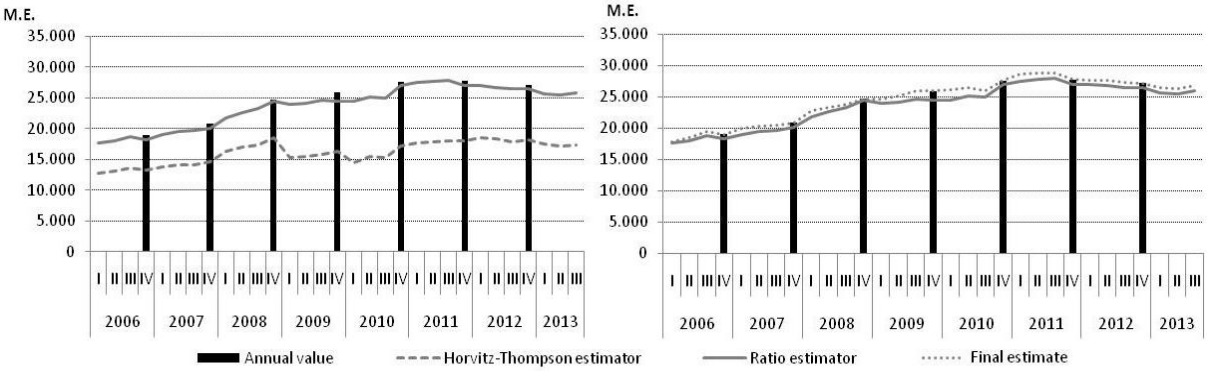



**Figure 2a: Quarterly procedure** **Figure 2b: Reconciliation procedure**

**Figure 2: Obtained funding – Manufacturing, mining and quarrying**

## 4. The reconciliation procedure

To reconcile the annual and quarterly series, it is applied an adjustment procedure referred as benchmarking that follows the *movement preservation principle* developed by DENTON (1971), which consists in obtaining adjusted series that keep the dynamics of the original quarterly series, through the minimisation of a quadratic loss function subject to a set of constraints that ensure (i) the temporal consistency between the adjusted quarterly time-series and the annual figures relating the population of NFC, that result from the annual procedure, and (ii) for each period, that the accounting equilibrium is met. Figure 2b continues the example mentioned in the previous section to illustrate the results from this procedure.

## 5. Final remarks

The outcome of the process described in this paper is a set of quarterly time-series, divided into Balance sheet and Profit and loss account indicators for all NFC, as well as a set or economic and financial ratios broken down by economic activity, size and capital holding sector. Given the amount of detail of these statistics, the use of ancillary information through the different techniques here described plays an important role in each of the procedures, as it improves the quality of the estimates, as well as it provides a link between the annual and quarterly databases.

**Acknowledgement and disclaimer:** We would like to thank Prof. Pedro Simões Coelho for his valuable contribution for development of the estimator for the quarterly data. 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 (2008), *Simplified reporting: Inclusion of the Simplified Corporate Information in the Statistics on Non-Financial Corporations from the Central Balance-Sheet Database*, Supplement 1/2008 to the Statistical Bulletin, Banco de Portugal.

BANCO DE PORTUGAL (2013), *Estatísticas das Empresas Não Financeiras da Central de Balanços – notas metodológicas*, Supplement 2/2008 to the Statistical Bulletin, Banco de Portugal (Portuguese version).

DENTON, Frank T. (1971), Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. *Journal of the American Statistical Association*, 66 (Mar. 1971), 99-102.

GONÇALVES, H. and LOURENÇO, M. (2013), Building business registers to monitor entrepreneurial dynamics, *IFC Bulletin*, 37, 42-45.

INE (2012*), Documento Metodológico – Inquérito Trimestral às empresas não financeiras, versão 2.1, fevereiro de 2012* (Portuguese version).

**ST III – Sessão 20 anos da CLAD – 6ª Feira, 11 de Abril, Salão Nobre (17h05)**

# Classes de objectos simbólicos: dados da indústria automóvel

**Áurea Sousa [1], Helena Bacelar-Nicolau [2], Fernando C. Nicolau [3], Osvaldo Silva[4]**

[1]*Universidade dos Açores, Departamento de Matemática, CEEAplA, CMATI, aurea@uac.pt;*

[2]*Universidade de Lisboa, Faculdade de Psicologia (LEAD), ISAMB, hbacelar@fp.ul.pt;*

[3]*Univ. Nova de Lisboa, FCT, Dep. de Matemática e DataScience, geral@datascience.org*

[4]*Universidade dos Açores, Departamento de Matemática, CES, CMATI, osilva@uac.pt*

**Sumário**

Neste trabalho, é abordada a Análise Classificatória Hierárquica Ascendente (ACHA) de dados simbólicos ou complexos (generalizações de dados clássicos), com base no coeficiente de afinidade generalizado ponderado e em critérios de agregação clássicos e probabilísticos, estes últimos no âmbito da metodologia *VL*. São apresentados os principais resultados obtidos com a ACHA de 33 modelos de carros (dados simbólicos na área da indústria automóvel), com base no coeficiente de afinidade generalizado ponderado, centrado e reduzido pelo método de Wald e Wolfowitz, comparando-se os resultados obtidos com os de outros autores e com a partição definida *a priori* pelas categorias ("*Utilitário", "Berlina", "Desportivo", "Luxo")* a que os modelos de carros pertencem.

**Palavras-chave:** Análise classificatória hierárquica, Coeficiente de afinidade generalizado ponderado, Dados simbólicos ou complexos, Metodologia *VL*.

## 1. Introdução

Na sociedade actual, onde os avanços computacionais têm imperado, é cada vez mais frequente a utilização de bases de dados, sendo fundamental efectuar a síntese de conjuntos de dados de elevada dimensão em termos dos seus conceitos subjacentes, os quais têm de ser descritos por dados mais complexos, designados por dados simbólicos. Estes dados podem ser heterogéneos e são representados em tabelas, cujas células podem conter um ou mais valores, tais como subconjuntos de categorias, intervalos da recta real, ou distribuições de frequências (Bock and Diday, 2000; Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010; Sousa et al., 2013). As linhas da tabela de dados representam unidades de dados ou objectos simbólicos e as colunas representam variáveis simbólicas.

Muitas medidas de proximidade entre objectos simbólicos têm sido referidas na literatura (Bock e Diday, 2000). Uma vez obtida a matriz de proximidades entre os elementos do conjunto a classificar, podem ser aplicados critérios de agregação clássicos ou probabilísticos (Bacelar-Nicolau et al., 2009, 2010; Sousa et al., 2013). Neste trabalho, a ACHA foi efectuada com base no coeficiente de afinidade generalizado ponderado (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al., 2009, 2010) e em três critérios de agregação probabilísticos no âmbito da Metodologia *VL* (Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998), sobre um conjunto de dados retirado da literatura da análise de dados complexos.

## 2. Análise classificatória hierárquica de objectos simbólicos com base no coeficiente de afinidade generalizado ponderado

A Análise Classificatória (*Cluster Analysis*) tem como objectivo identificar grupos (classes) de entidades (indivíduos, objectos, etc.), relativamente homogéneos e, de preferência, bem separados, com base nas semelhanças ou dissemelhanças entre essas entidades. Os métodos hierárquicos aglomerativos começam por considerar um número de classes igual ao número de elementos a classificar e posteriormente, em cada etapa, efectuam a junção ou aglomeração de classes em classes maiores, obtendo-se na última etapa um único grupo contendo todos os elementos a classificar.

A partir do coeficiente de afinidade entre duas distribuições de probabilidade discretas, proposto por Matusita (1951), Bacelar-Nicolau (1980, 1988) introduziu o coeficiente de afinidade no domínio da Análise Classificatória, para avaliar a semelhança básica entre pares de colunas ou pares de linhas de uma matriz de dados, ou seja, entre variáveis ou entre indivíduos, conforme o conjunto que se pretende classificar. Este coeficiente foi estendido a diferentes tipos de dados, incluindo dados de tipo heterogéneo e de natureza complexa (ou simbólicos) (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010), frequentemente presentes em bases de dados de elevada dimensão. Os critérios de agregação probabilísticos usados, no âmbito da Metodologia *VL*, recorrem essencialmente a noções probabilísticas para a definição das funções de comparação (Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998). A extensão do coeficiente de afinidade para o caso de dados simbólicos é designada por coeficiente de afinidade generalizado ponderado (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010). O coeficiente assintoticamente centrado e reduzido, sob uma hipótese de referência permutacional baseada no teorema limite de Wald e Wolfowitz permite, por sua vez, definir um coeficiente probabilístico no contexto da metodologia *VL*, na linha iniciada por Lerman (1972, 1981) e desenvolvida por Bacelar-Nicolau (e.g.1980, 1987, 1988) e Nicolau (e.g. 1983, 1998). Aplicações da extensão do coeficiente de afinidade para o caso de dados simbólicos foram apresentadas, por exemplo, em Bacelar-Nicolau et al. (2009, 2010) e em Sousa et al. (2013).

## 3. Exemplo da indústria automóvel: "*Car data set*"

A matriz de dados simbólicos que aqui usamos, para exemplificar a metodologia, é referida na literatura da análise de dados simbólicos (e.g., De Carvalho et al., 2006*a*, 2006*b*; Souza et al., 2007) e contém trinta e três modelos de carros (objectos simbólicos), descritos por oito variáveis cujos valores são intervalos da recta real ("*Preço", "Cilindrada", "Velocidade Máxima", "Aceleração", "Passo", "Comprimento", "Largura"* e "*Altura")*, duas variáveis categóricas ("*Alimentação*" e "*Tracção*") com categorias não ordenadas que podem assumir múltiplos valores (subconjuntos de categorias) e uma variável nominal ("*Categoria do Carro").* Esta última variável, com as modalidades "*Utilitário", "Berlina", "Desportivo" e* "*Luxo"*, reflecte a classificação *a priori* dos modelos de carros, a qual pode ser encontrada, por exemplo, em De Carvalho et al. (2006*a*, 2006*b*). A Tabela 1 mostra uma parte da matriz de

dados simbólicos, sendo de referir que a matriz de dados completa está disponível no software SODAS (Symbolic Official Data Analysis System).

**Tabela 1-Parte da matriz de dados simbólicos -"Car data set"**

| ***Modelo*** | ***Preço*** | ***Cilindrada*** | ***Alimentação*** | ***Tracção*** | . . | ***Altura*** | ***Categoria*** |
|---|---|---|---|---|---|---|---|
| *Alfa 145* | [27806, 33596] | [1370, 1910] | *Gasoli, Diese* | *Anter* | . . | [143, 143] | *Utilit* |
| *Alfa 156* | [41593, 62291] | [1598, 2492] | *Gasoli* | *Anter* | . . | [142, 142] | *Berlina* |
| . . . | . . . | . . . | . . . | | . . | . . . | . . . |
| *Passat* | [39676, 63455] | [1595, 2496] | *Gasoli, Diese* | *Anter, Integ* | . . | [146, 146] | *Luxo* |

Nesta comunicação, são apresentados os principais resultados obtidos com a *A.C.H.A.* dos trinta e três modelos de carros, com base no coeficiente de afinidade generalizado ponderado, centrado e reduzido pelo método de Wald e Wolfowitz, e em três critérios de agregação probabilísticos, *AVL*, *AV1* e *AVB* (Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1972, 1981). Os principais resultados obtidos são comparados com os de outros autores (e.g., De Carvalho *et al*., 2006*a*, 2006*b*; Souza *et al*., 2007).

## Referências

BACELAR-NICOLAU, H. (1980) *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória,* Tese de Doutoramento, FCL, Universidade de Lisboa.

BACELAR-NICOLAU, H. (1987) On the Distribution Equivalence in Cluster Analysis. In Devijver, P.A. & Kittler, J. (Ed.) Pattern Recognition Theory and Applications, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, New York, Springer - Verlag, 73-79.

BACELAR-NICOLAU, H. (1988) Two Probabilistic Models for Classification of Variables in Frequency Tables. IN BOCK, H.-H. (Ed.) *Classification and Related Methods of Data Analysis.* North Holland, Elsevier Sciences Publishers B.V., pp. 181-186.

BACELAR-NICOLAU, H. (2000) The Affinity Coefficient. IN BOCK, H.-H. & DIDAY, E. (Ed.) *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, Springer-Verlag, 160-165.

BACELAR-NICOLAU, H. (2002) On the Generalised Affinity Coefficient for Complex Data. *Biocybernetics and Biomedical Engineering*, 22(1)**,** 31-42.

BACELAR-NICOLAU, H., NICOLAU, F.C, SOUSA, Á. & BACELAR-NICOLAU, L. (2009) Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets. Biocybernetics and Biomedical Engineering, 29 (2), 9-18.

BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á., BACELAR-NICOLAU, L (2010) Clustering Complex Heterogeneous Data Using a Probabilistic Approach. *Proceedings of* the *Stochastic Modeling Techniques and Data Analysis International Conference* (*SMTDA2010*), 85-93, (electronic publication).

BOCK, H.-H. & DIDAY, E. (2000) *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin: Springer-Verlag.

DE CARVALHO, F.A.T., BRITO, P. & BOCK, H.-H. (2006*a*) Dynamic Clustering for Interval Data Based on $L_2$ Distance. *Computational Statistics,* 21(2)**,** 1-19.

DE CARVALHO, F.A.T., SOUZA, R.M.C.R. de, CHAVENT, M. & LECHEVALLIER, Y. (2006b) Adaptive Hausdorff Distances and Dynamic Clustering of Symbolic Interval Data. *Pattern Recognition Letters,* 27 **(3),** 167-179.

LERMAN, I.C. (1972) *Étude Distributionelle de Statistiques de Proximité entre Structures Algébriques Finies du Même Type: Apllication à la Classification Automatique,* Cahiers du B.U.R.O., 19, Paris.

LERMAN, I.C. (1981) *Classification et Analyse Ordinale des Données,* Paris, Dunod.

NICOLAU, F.C. (1983) Cluster Analysis and Distribution Function. *Methods of Operations Research,* 45**,** 431-433.

NICOLAU, F.C. & BACELAR-NICOLAU, H. (1998) Some Trends in the Classification of Variables. IN Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Ed.) *Data Science, Classification, and Related Methods.* Springer-Verlag, 89-98.

MATUSITA, K. (1951) On the Theory of Statistical Decision Functions. *Ann. Instit. Stat. Math,* III**,** 1-30

SOUSA, Á., NICOLAU, F., BACELAR-NICOLAU, H. & SILVA, O. (2013) Clustering of Symbolic Data based on Affinity Coefficient: Application to a Real Data Set. *Biometrical Letters,* 50 (1), 27-38.

SOUZA, R.M.C.R., DE CARVALHO, F.A.T., & PIZZATO, D.F. (2007) A Partitioning Method for Mixed Feature-Type Symbolic Data Using a Squared Euclidean Distance. IN FREKSA, C., KOHLHASE, M., & SCHILL, K. (Ed.) *KI 2006: Advances in Artificial Intelligence.* 29th Annual German Conference on AI, KI 2006, Bremen, Germany, June 14-17, 2006, Proceedings. Series: Lecture Notes in Computer Science, Vol. 4314, Berlin Heidelberg, Springer-Verlag, 260-273.

**ST III – Sessão 20 anos da CLAD – 6ª Feira, 11 de Abril, Salão Nobre (17h25)**

# A hierarchical conceptual clustering based on the quantile method for mixed data

**Manabu Ichino[1], Paula Brito[2]**

[1]*School of Science and Engineering, Tokyo Denki University, Japan, ichino@mail.dendai.ac.jp;*
[2]*Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Portugal, mpbrito@fep.up.pt*

**Abstract**

We consider the case where each element of the set to be analyzed is described by variables of different types. The quantile method transforms each element to (m+1) numerical vectors, called the quantile vectors, for a given integer m, which controls the granularity of concepts generated. We define the concept size of a p-dimensional hyper-rectangle spanned by quantile vectors. The proposed hierarchical clustering method agglomerates clusters so as to minimize the concept size of the clusters formed, so as to achieve the compactness of the cluster descriptions. We introduce the weighted self-information (WSI), based on the concept size, to find informative clusters. Conjunctive logical expressions for clusters selected by the WSI are obtained.

**Keywords**: concept size, compactness, hierarchical conceptual clustering, quantile method, symbolic data analysis.

## 1. Quantile representation

Following the approach presented in Ichino (2008), the proposed method is based on a common representation model for data described by variables $Y_j$, j=1,….p, of possibly different types, which uses some predefined quantiles of the underlying distribution of the observed data values. We are interested in symbolic data, i.e., data that comprises intrinsic variability, and consider the different symbolic variable types, as defined in Noirhomme and Brito (2011).

Let S ={$s_1$, ….$s_n$} be the set of elements to be clustered. If the m-th quantiles $Q_1$…, $Q_j$, …, $Q_{m-1}$, are chosen, with ($Q_j$)= j/m, then each observation, for each element $s_i \in S$ and each variable $Y_j$, is represented by a (m+1) dimensional vector ($Q_0$=Min, $Q_1$,..., $Q_j$,…, $Q_{m-1}$, $Q_m$=Max). The value of m controls the level of granularity of data representation, and it is a user's choice. For the sake of simplicity, but also because of their broad utilization, let us suppose that we use the quartiles, i.e., m=4. The quartile representation is then defined by the 5-uple (Min, $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$, Max) where $Q_{0.25}$, $Q_{0.5}$, are $Q_{0.75}$ are the quartiles. In the case of multi-valued numerical variables, quantiles may be determined as in classical descriptive statistics, assuming equal weights for each observed value. Here we follow Mood *et al* (1994) and use the Empirical Distribution Function, without interpolation - defining the quantile $Q_j$ as the first observed value x, in the ranked list, where the condition $F(x) \geq j/m$ is met. For interval-valued variables, an underlying distribution must be assumed within each observed interval; this may be the Uniform distribution, as proposed by Bertrand and Goupil (2000); however, other distributions may be considered, e.g., the Triangular distribution, or even any parametric

distribution (Gaussian,...). If an Uniform distribution is assumed in an interval [$\ell$, u], we have, naturally, F(x)= (x-$\ell$)/(u-$\ell$), for x $\in$ [$\ell$, u], F(x)=0, x $\leq \ell$; F(x)=1, x $\geq$ u.

Example: Consider we are analyzing monthly temperatures in different meteorological stations, for which minima and maxima values have been recorded, and that we have Temp($station_1$) = [10,18]. Assuming uniformity within the interval, the quartile representation is then Temp($station_1$) = (10, 12, 14, 16, 18).

For a histogram-valued variable, quantiles of any histogram may be obtained by simply interpolation, assuming a Uniform distribution in each class (bid) of the histogram.

Consider now the case of a categorical variable Y, and let O={$c_1$,…, $c_k$} be the underlying set of categories. If O is ordered - i.e., Y is a categorical ordinal variable - then a rank may be associated with each $c_\ell$, $\ell$=1,…, k; then quantiles may then be determined on the rank values, using the same method as for multi-valued numerical variables. The case of categorical nominal variable is different; if the user wishes to also consider these variables along with those of the other types, and obtain a common quantile representation, a ranking must be defined on the category set; in Ichino (2008, 2011) it is proposed to define a raking based on the global observed frequency of each category. Once a ranking is defined, a multi-valued categorical nominal variable may be treated as an ordinal one. For categorical modal variables, weights (probabilities or relative frequencies) are associated with each observed category. Assuming that the categories are ranked, quantiles may be determined as for the multi-valued case, i.e., defining the quantile $Q_j$ as the first observed value x, in the ranked list, where the condition F(x)$\geq$ j/m is met.

Finally, each object $s_i \in S$, described by p variables, is represented by p $\times$ (m+1)-tuples, ($x_{ij(min)}$ , $Q_{ij1}$, $Q_{ij2}$,…,$Q_{ij(m-1)}$, $x_{ij(max)}$), j = 1, 2,…, p.

Example : Consider an airport A described by the number of passengers in arriving flights (an interval-valued variable), the delay-time in arrivals (an histogram-valued variable) and the size-type of aircrafts (a categorical ordinal modal variable), A= ([150,200] , {[0,10 [ (0.25) ; [10,30 [ (0.65) ; [30, 60] (0.10)} , {1 (0.40) ; 2 (0.40) ; 3 (0.2)} ). The quartile representation of A is then A = ((150, 162.5, 175, 187.5, 200) , (0, 10, 17.7, 25.4, 60) , (1, 1, 2, 2, 3) ).

For each $s_i \in S$, we define p-dimensional numerical vectors, the quantile vectors, as follows:

$$\mathbf{x}_{i0} = (x_{i1(min)}, x_{i2(min)},\ldots, x_{ip(min)}) ; \mathbf{x}_{i1} = (Q_{i11}, Q_{i21},\ldots, Q_{ip1} ; \mathbf{x}_{i2} = (Q_{i12}, Q_{i22},\ldots, Q_{ip2}) ; \cdots ;$$

$$\mathbf{x}_{i(m-1)} = (Q_{i1(m-1)}, Q_{i2(m-1)},\ldots, Q_{ip(m-1)}) ; \mathbf{x}_{im} = (x_{i1(max)}, x_{i2(max)},\ldots, x_{ip(max)})$$

In other words, we divide each object $s_i \in S$ into m+1 sub-objects, the minimum sub-object $s_{i(min)}$, (m-1) quantile sub-objects, $s_{iQ1}$,…, $s_{iQ(m-1)}$, and the maximum sub-object $s_{i(max)}$. The quantile vectors are representations of these sub-objects in $\mathbf{R}^p$.

Let $\mathbf{J}(s_{it}, s_{hv})$ be a rectangle in $\mathbf{R}^p$ spanned by the vectors $\mathbf{x}_{it} = (x_{it1}, x_{it2},\ldots, x_{itp})$ and $\mathbf{x}_{hv} = (x_{hv1}, x_{hv2},\ldots, x_{hvp})$ associated with sub-objects $s_{it}$ of $s_i \in \mathbf{S}$ and $s_{hv}$ of $s_h \in$ **;** $\mathbf{J}(s_{it}, s_{hv})$ is defined by the following Cartesian product of p closed intervals:

$$\mathbf{J}(s_{it}, s_{hv}) = [\min(x_{it1}, x_{hv1}), \max(x_{it1}, x_{hv1})] \times \cdots \times [\min(x_{itp}, x_{hvp}), \max(x_{itp}, x_{hvp})]$$

We call $\mathbf{J}(s_{it}, s_{hv})$ as the Cartesian join (region) of sub-objects $s_{it}$ and $s_{hv}$ (Ichino and Yaguchi, (1994)).

## 2. Concept size

For each variable $Y_j$, j = 1,…, p, the domain $D_j$ of variable values is defined as the interval $D_j = [x_{jmin}, x_{jmax}]$, j = 1,…, p, where $x_{jmin} = \min\{x_{1j0}, \ldots, x_{nj0}\}$ and $x_{jmax} = ma\{(x_{1jm}, \ldots, x_{njm}\}$. Let $\mathbf{E} = E_1 \cdots \times E_p$ be a hyper-rectangle spanned by quantile vectors in the space $\mathbf{R}^p$.

We define the **concept size** of an interval $E_j$ in terms of variable $Y_j$ as $P(E_j) = | E_j |/| D_j |$, j=1,..., p, where $|D_j|$ is the length of interval $D_j$. We note that $0 \leq P(E_j) \leq 1$, j=1, …, p.

Then, we define the object size $P(\mathbf{E})$ of $\mathbf{E}$ by the arithmetic mean $P(\mathbf{E}) = (\sum_{j=1}^{p} P(Ej))/p$ . It is clear that $0 \leq P(\mathbf{E}) \leq 1$.

## 3. Hierarchical conceptual clustering based on compactness

In this section, we present a conceptual clustering method based on the compactness of cluster descriptions and the weighted self-information. The usefulness of proposed method has been shown in different examples.

### 3.1 Compactness

In the following, we represent a concept as a hyper-rectangle in $\mathbf{R}^p$. A rectangle is equivalent to a description by a conjunctive logical expression.

We define the **compactness** of the concept generated by two objects $o_h$ and $o_{h'}$, $C(o_h,o_{h'})$ as the concept size of the Cartesian join of $o_h$ and $o_{h'}$. Note that the objects $o_h$ and $o_{h'}$ may be elements of the (extended) set S, or already obtained by the Cartesian join of elements of S.

### 3.2 Weighted Self-Information

In a step of hierarchical clustering, let $P_{clust}$ be the cluster size defined as the proportion of elements of S belonging to the cluster: $P_{clust} = n_{clust} /n$, $n_{clust}$ being the cardinal of $P_{clust}$, $0 < P_{clust} \leq 1$. Interesting clusters are usually found in later steps of the clustering, and their cluster sizes rapidly increase towards the largest value. If these desirable clusters are mutually disjoint, their

cluster sizes cannot exceed 0.5 simultaneously, therefore, we have to pay attention to clusters of size below 0.5. We use the weighted self-information $WSI_{clust} = - P_{clust}Log_2 P_{clust}$ to detect informative clusters.

In the hierarchical method presented in the next section, we use the compactness, i.e., the concept sizes, as a measure of cluster quality. The $WSI_{clust}$ is merely a rough measure to evaluate informative clusters, since several clusters can have very different concept sizes in the representation space in spite of their cardinalities. For this reason, we use the WSI defined by the concept size P, i.e., $WSI = - P Log_2 P$ . The WSI function is not symmetric, and it has a round top for concept sizes 0.3 ~ 0.5. We use this measure to find mutually disjoint concepts with comparative sizes.

### 3.4 Algorithm of hierarchical agglomerative conceptual clustering

The proposed algorithm may now be described in the following steps:

1) For each pair of objects s and s' in (extended) S, calculate the compactness C(s, s') and find the pair $s_i$ and $s_h$ that minimizes C.

2) Generate the merged concept $s_{ih}$ of $s_i$ and $s_h$ in **S**, and delete $s_i$ and $s_h$ from **S** ; the new object $s_{ih}$ (a concept) is described by the Cartesian join $\mathbf{E}_{ih}$ in the representation space $\mathbf{R}^p$.

3) Repeat step 2 until **S** includes only one object (the whole concept).

4) Calculate the weighted self-information for each cluster obtained in 2.

5) Find a set of informative clusters that covers all objects in **S**.

6) Find a conjunctive logical expression for each cluster obtained in 5.

**Acknowledgements:** This research was supported by the Japan Society of the Promotion of the Science (Grant-in Aid for Scientific Research (C) 25330268) and also by the Project NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

### References

BERTRAND, P. & GOUPIL, F. (2000). Descriptive Statistics for Symbolic Data. IN: H.-H. Bock and E. Diday (Eds.), *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, 106-124. Springer, Heidelberg.

ICHINO, M. (2008). Symbolic PCA for Histogram-Valued Data. IN Proc *IASC'2008*., Yokohama, Japan.

ICHINO, M. (2011). The quantile method for symbolic principal component analysis. *Statistical Analysis and Data Mining* 4, 184-198.

ICHINO, M. & YAGUCHI, H. (1994).. Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Trans. Systems, Man, & Cybernetics*, 24 (4), 698-708.

MOOD, A. M., GRAYBILL, F. A., & BOES, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

NOIRHOMME-FRAITURE, M. & BRITO, P., (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining*, 4 (2), 157--170.

**ST III – Sessão 20 anos da CLAD – 6ª Feira, 11 de Abril, Salão Nobre (17h45)**

# Avaliações internacionais e desempenho dos alunos portugueses

**Ana Sousa Ferreira**[1]
[1]*Faculdade de Psicologia, ULisboa e UNIDE, asferreira@psicologia.ulisboa.pt*

**Sumário**

Portugal tem vindo a participar em diversos estudos internacionais de avaliação de alunos como o PISA, o TIMSS e o PIRLS. Todos estes projetos avaliam a proficiência de alunos de vários países em diversos domínios e níveis de ensino, contextualizada pela informação recolhida em diversos questionários ao aluno, à escola, ao professor e aos pais. A análise dos resultados destas avaliações sobre o desempenho dos alunos portugueses, aferidos a padrões internacionais, constitui o mote para esta comunicação.

**Palavras-chave:** *PIRLS, PISA, TIMSS,* Análise de Dados.

## 1. Introdução

A OCDE (*Organisation for Economic Co-operation Development*) e a IEA (*International Association for the Evaluation of Educational Achievement*) promovem programas internacionais de avaliação do desempenho de crianças e jovens do mundo inteiro em matemática, ciências e leitura. Portugal tem vindo a participar em diversos destes estudos internacionais como o PISA (*Programme for International Students Assessment), o* TIMSS (*Trends in International Mathematics and Science Study*) e o PIRLS (*Progress in International Reading Literacy Study*). *Todos estes projetos avaliam a proficiência de alunos de vários países em diversos domínios.*

O *PISA***,** desenvolvido pela OCDE, visa avaliar se os alunos de 15 anos, aqueles que na maior parte dos países participantes se aproximam do final da escolaridade obrigatória, estão bem preparados para enfrentarem os desafios da vida quotidiana. Os testes PISA foram concebidos para avaliar as competências de leitura, matemática, ciências e mais recentemente resolução de problemas, em situações relacionadas com a realidade, e ocorrem em ciclos de três anos. Em cada ciclo PISA, é selecionada, como domínio principal, uma das três áreas avaliadas (leitura, matemática, ciências). Em 2000, o domínio principal foi a leitura, domínio novamente avaliado em 2009. Em 2003 e em 2012 selecionou-se a matemática e, em 2006, as ciências. Em 2015, seguindo a rotatividade, o domínio principal será ciências. Portugal tem vindo a participar no projeto PISA desde o primeiro ciclo, em 2000, até ao último ciclo aplicado, em 2012.

O *TIMSS* é uma avaliação internacional do desempenho dos alunos do 4.º e do 8.º ano de escolaridade em matemática e ciências, desenvolvida pela IEA. O *TIMSS* A*dvanced* é uma versão do estudo que tem o objetivo de avaliar as tendências do desempenho dos alunos no final do ensino secundário em matemática e física. Desde 1995, os testes *TIMSS* são aplicados

de quatro em quatro anos com a finalidade de gerar informação de qualidade sobre os resultados do desempenho dos alunos e sobre os contextos em que estes aprendem. Portugal participou no primeiro ciclo do *TIMSS* em 1995 e voltou a participar na prova do 4º ano em 2011.

O *PIRLS* é uma avaliação internacional sobre a compreensão da leitura dos alunos do 4.º ano de escolaridade, desenvolvida pela IEA. Desde 2001, os testes *PIRLS* são aplicados de cinco em cinco anos com a finalidade de gerar informação de qualidade sobre os resultados do desempenho dos alunos em leitura e sobre os contextos em que estes aprendem. Em 2011, foi criada uma versão para aplicar em países cujos alunos estão aquém dos níveis de leitura estabelecidos para o *PIRLS* (*prePIRLS*). Portugal participou pela primeira vez no *PIRLS* em 2011.

Em 2011, a aplicação dos dois estudos da IEA, o *TIMSS* e o *PIRLS*, coincidiu, o que permitiu relacionar a literacia de leitura dos alunos do 4.º ano de escolaridade com o seu desempenho em matemática e ciências.

## 2. As Provas *PISA*, *TIMSS* e *PIRLS*

A prova do *PISA* é constituída por 13 cadernos de teste que combinam itens dos diferentes domínios avaliados, tendo cada aluno respondido a um único caderno de teste que lhe foi atribuído aleatoriamente. O teste *PISA* tem sido aplicado em papel e tem uma duração total de duas horas, ocupando os itens do domínio principal 2/3 do tempo total da prova. Em cada ciclo, a avaliação relativa ao domínio principal é mais detalhada. Por exemplo, no *PISA* 2012 participaram 65 países e economias e a avaliação em literacia matemática, considerou itens relativos a três processos cognitivos - *Formular*; *Aplicar* e *Interpretar* – e quatro conteúdos – *Quantidade; Incerteza*; *Mudança e relações*; *Espaço e forma*. O *PISA* não se limita a avaliar se um aluno reproduz eficazmente os conhecimentos adquiridos, procura antes aferir se os alunos conseguem aplicar, em contextos diferenciados, o que aprenderam. Uma abordagem desta natureza procura perceber se as sociedades contemporâneas reconhecem e valorizam os indivíduos não por aquilo que eles sabem mas por aquilo que eles conseguem fazer com o que sabem (OCDE, 2012).

Neste projeto, são também recolhidas informações que permitem contextualizar os resultados, através da aplicação de questionários aos alunos, aos pais e às escolas. O *PISA* permite, pois, identificar os fatores que influenciam os níveis de desempenho nos vários domínios de literacia, nomeadamente, elementos sociodemográficos dos alunos e da sua relação com a aprendizagem e, ainda, diversas características das escolas participantes, como a sua organização e recursos.

A prova do *TIMSS* é constituída por um conjunto de cadernos, numa composição de itens de matemática e de ciências que abrangem os três processos cognitivos (A*plicar*, *Conhecer* e *Raciocinar*). Cada aluno responde a um único caderno de prova. A avaliação em matemática

considera duas dimensões: uma relativa ao conteúdo – *Números, Formas Geométricas e Medida, Apresentação de Dados* – e a outra ao domínio cognitivo. A avaliação em ciências segue o mesmo desenho, sendo contempladas as seguintes áreas: *Ciências da Vida, Ciências Físicas e Ciências da Terra*. Neste estudo, a par dos testes, são também aplicados questionários visando recolher informação de contexto que permite descrever as situações e os fatores que influenciam a aprendizagem da matemática e das ciências.

A prova do *PIRLS* foi desenhada de modo a contemplar duas finalidades de leitura, a *literária* - ler como experiência literária - e a *informativa* - ler para adquirir e utilizar informação -, assim como os processos de compreensão da leitura - *Reter, Fazer inferências diretas, Interpretar e integrar* e *Avaliar* -. A prova é constituída por um conjunto de cadernos, numa composição de itens que envolvem as diferentes finalidades e processos de compreensão da leitura. Cada aluno responde apenas a um caderno de prova. Também neste estudo, a par dos testes, são aplicados questionários visando recolher informação de contexto que permite descrever as situações e os fatores que afetam a literacia de leitura.

Nestes três estudos, os itens não são públicos, possibilitando-se deste modo a comparação de resultados ao longo das várias edições da prova e a identificação de tendências. A OCDE e a IEA disponibilizam, em cada ciclo de cada um destes estudos, alguns itens que deixam de fazer parte das provas e que ilustram o tipo de situações apresentadas aos alunos.

## 3. O desempenho dos alunos portugueses

A análise do desempenho dos alunos portugueses pode ser efetuada por valores de referência internacionais (*benchmarks*), ou por conteúdos e domínios cognitivos, ou ainda usando a informação de contexto recolhida por qualquer dos estudos internacionais em que Portugal participa. Os dados recolhidos e os resultados obtidos por estes estudos são tornados públicos pelos consórcios internacionais após a divulgação do relatório internacional e, permitem aos investigadores o acesso a um alargado leque de informação, contribuindo para que estes instrumentos de avaliação internacional possam ser reguladores do sistema de ensino-aprendizagem.

No estudo *PISA 2012* participaram 65 países/economias sendo 34 países membros da OCDE e 31 países parceiros. Neste ciclo PISA, Portugal obteve 487 pontos na escala da matemática, representando uma progressão de 21 pontos relativamente ao resultado alcançado em 2003 – ano em que a matemática também foi domínio principal. Esta pontuação coloca Portugal, pela primeira vez, desde o início do Programa *PISA*, na média da OCDE. No domínio da leitura Portugal alcançou 488 pontos, sendo a média dos países da OCDE neste domínio de 496 pontos e no domínio das ciências, 489 pontos, sendo a média dos países da OCDE neste domínio de 501 pontos.

No *TIMSS 2011*, participaram 63 países e 14 participantes em *Benchmarking.* Neste estudo, o desempenho médio dos alunos portugueses em matemática foi de 532 pontos, numa

escala de 0-1000, com um ponto médio de referência de 500 e um desvio padrão de 100. Este resultado coloca Portugal entre os 15 países com melhor desempenho em matemática para o 4.º ano. Nesta prova e usando a mesma escala, o desempenho médio dos alunos portugueses do 4.º ano em ciências foi de 522 pontos o que o coloca entre os 19 países com melhor desempenho em ciências para o 4.º ano.

No *PIRLS 2011*, participaram 49 países e 9 participantes em *Benchmarking*. Neste estudo, Portugal obteve um desempenho médio de 541 pontos o que o coloca entre os 19 países com melhor desempenho em leitura no 4.º ano.

Nestes três estudos, em cada país participante, a amostragem é realizada em duas fases: num primeiro momento são selecionadas aleatoriamente as escolas e, num segundo momento são selecionados aleatoriamente os alunos ou as turmas de alunos de cada escola participante. Este tipo de metodologia, acarreta um cuidado especial na análise destes dados, uma vez que as técnicas estatísticas disponíveis nos *softwares* estatísticos mais comuns não são adequadas. Assim, estes estudos lançam um verdadeiro desafio aos investigadores não só pela colossal informação educacional gerada mas também pelos métodos estatísticos que devem ser usados na sua análise.

## Referências

MARTIN, M.O., MULLIS, I.V.S., FOY, P., & STANCO, G.M. (2012). *TIMSS 2011 International Results in Science*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College

MULLIS, I.V.S., MARTIN, M.O., FOY, P., & ARORA, A. (2012). *TIMSS 2011 International Results in Mathematics*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College

MULLIS, I.V.S., MARTIN, M.O., FOY, P., & DRUCKER, K.T. (2012). *PIRLS 2011 International Results in Reading*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I*), PISA, OECD Publishing.

ProjAVI (2012), *PIRLS 2011 - Principais resultados em Leitura,* www.projavi.mec.pt/np4/179/, (acedido em 21 de fevereiro de 2014).

ProjAVI (2012), *TIMSS 2011 Principais resultados em Ciências*, www.projavi.mec.pt/np4/179/, (acedido em 21 de fevereiro de 2014).

ProjAVI (2012), *TIMSS 2011 Principais resultados em Matemática,* www.projavi.mec.pt/np4/179/, (acedido em 21 de fevereiro de 2014).

ProjAVI (2013), *PISA 2012 Portugal – Primeiros resultados*, www.projavi.mec.pt/np4/179/, (acedido em 21 de fevereiro de 2014).

**ST III – Sessão 20 anos da CLAD – 6ª Feira, 11 de Abril, Salão Nobre (18h05)**

# Índice de Bem-estar em Portugal – Contributos para a interpretação dos resultados baseada em classificação de variáveis

**Paulo Gomes**[1]
[1]*ISEGI – Universidade Nova de Lisboa, paulo.pinhogomes@gmail.com*

**Sumário**

Apresentaremos os principais resultados do Índice de Bem-estar em Portugal, estudo recentemente divulgado pelo INE, complementando-os por recurso a uma análise estatística multivariada dos indicadores selecionados na caracterização do bem-estar, a qual proporciona uma tipologia desses indicadores que evidencia comportamentos diferenciadores na evolução do índice, nomeadamente no contraste da evolução em 2004-2008 e 2008-2011.

**Palavras-chave:** Análise em Componentes Principais, Análise Classificatória, Índice de Bem-estar, Seleção de variáveis.

## 1. Introdução

O Índice de Bem-estar (IBE), recentemente divulgado pelo INE, é um estudo estatístico de periodicidade anual e cujo âmbito geográfico é o país. As variáveis que integram a construção do IBE provêm de procedimentos administrativos e de operações estatísticas desenvolvidas no contexto do Sistema Estatístico Nacional, do Sistema Estatístico Europeu, do Banco Mundial e outros.

Do ponto de vista concetual, as condições materiais de vida das famílias e a qualidade de vida, foram identificadas como perspetivas essenciais na avaliação da evolução do bem-estar. Neste contexto, procurou-se que cada perspetiva fosse representada com indicadores, agrupados em domínios de análise, que correspondessem, tão fielmente quanto possível, à delimitação concetual definida.

Na perspetiva das condições materiais de vida pretende-se:

- Captar o domínio do bem-estar económico, através das possibilidades correntes e futuras de consumo, da realização do bem-estar material e da desigualdade de distribuição de rendimento;

- Avaliar a vulnerabilidade económica através da medição da pobreza monetária, da privação material, do endividamento e da vulnerabilidade da habitação;

- Avaliar a participação e inclusão social, a vulnerabilidade do trabalho e a disparidade salarial segundo o sexo, e a qualidade do trabalho.

Na perspetiva de qualidade de vida, foram considerados sete domínios de análise:

- Educação, conhecimento e competências – através da caracterização da educação formal, da aprendizagem ao longo da vida, da qualidade de educação e nível de competências adquiridas e da produção de conhecimento e inovação;

- Saúde – através dos indicadores-resultado na saúde, da avaliação da prestação de cuidados de saúde e dos indicadores relativos a fatores de risco;

- Balanço vida-trabalho – através da avaliação da conciliação do tempo afeto à família e ao trabalho e da avaliação subjetiva do balanço vida-trabalho;

- Segurança pessoal – através da avaliação da criminalidade e da avaliação subjetiva da segurança pessoal;

- Participação cívica e governação – através da avaliação da participação cívica e política e da confiança nas instituições;

- Relações sociais e bem-estar subjetivo – através da avaliação do bem-estar subjetivo social e do bem-estar subjetivo individual, dimensões que pela sua especificidade não serão objeto de análise conjunta;

- Ambiente – através da avaliação de qualidade da água e do ar, da intensidade apercebida de ruído, da análise do destino final dos resíduos, da medida da biodiversidade e da avaliação subjetiva da qualidade ambiental.

Em cada domínio foram previamente identificadas dimensões prioritárias de análise que evidenciaram as problemáticas a considerar em cada um deles, na ótica do bem-estar, as quais alicerçaram o processo de seleção de variáveis. O objetivo inerente à construção desta nova infraestrutura estatística é poder acrescentar à ênfase na medição da produção económica, a ênfase na medida do bem-estar das pessoas, num contexto de sustentabilidade. Com indicadores sintéticos ao nível de cada domínio, de cada perspetiva e a nível global, aprofunda-se o mecanismo de acompanhamento dos principais fatores críticos do desenvolvimento económico e social de Portugal, em termos de bem-estar das pessoas e das famílias.

As variáveis tomadas em cada domínio vêm expressas em diferentes unidades de medida, pelo que o recurso a números índice simples (baseados no rácio entre o valor da variável no ano j e o valor dessa variável no ano-base), e à função de agregação média dos índices associados aos indicadores referentes a cada domínio, proporciona uma escala unidimensional para a representação da construção multidimensional do Bem-estar. Independentemente da perda de informação subjacente à escolha desta escala, as vantagens desta opção situam-se ao nível da simplicidade e da transparência do método, da eliminação da heterogeneidade da medida, da comparabilidade entre indicadores, mas também da atenuação da sensibilidade dos valores finais dos índices à inclusão de indicadores com diferentes níveis de precisão estatística.

Apresentaremos os principais resultados do Índice de Bem-estar em Portugal, no período 2004-2011 e ainda dados preliminares relativos ao ano de 2012. Os referidos dados refletem a

evolução de 79 indicadores, repartidos pelos 10 domínios selecionados, tendo sido atribuída a mesma ponderação a todos os domínios, assim como a mesma ponderação às variáveis inseridas em cada domínio. Esta escolha não reflete qualquer apriorismo quanto à importância relativa de cada domínio na caracterização do bem-estar, nem tão pouco qualquer apriorismo quanto à importância relativa de cada variável no seio de um dado domínio. Isto é, não havendo uma justificação científica para a atribuição de pesos distintos aos domínios ou às variáveis, optou-se pela não diferenciação das ponderações.

Complementarmente procederemos a uma análise estatística multivariada dos dados recorrendo a métodos da análise fatorial e da análise classificatória com o objetivo de evidenciarmos os principais contrastes na evolução dos domínios em estudo e dos anos objeto de análise. Procederemos também à identificação de tipologias de indicadores face às características da respetiva evolução no período 2004-2012, onde colocaremos em evidência o efeito da crise na evolução do bem-estar em Portugal.

## 2. Principais resultados

O Índice de Bem-estar em Portugal evoluiu positivamente entre 2004 e 2011, atingindo o valor de 108,1 em 2011, estimando-se uma redução para 107,6 em 2012. Contudo, as duas perspetivas de análise do bem-estar – traduzidas através dos índices sintéticos de Condições materiais de vida e de Qualidade de vida – evoluíram em sentidos opostos: enquanto o índice que explica a evolução das Condições materiais de vida registou genericamente uma evolução negativa, atingindo o valor de 89,2 em 2011 (na comparação com o ano-base de 2004 = 100), o índice relativo à evolução da Qualidade de vida apresentou uma evolução continuamente positiva, atingindo em 2011 o valor de 116,2.

Os dados preliminares de 2012, também divulgados neste Destaque, reforçam esse contraste: o índice relativo às Condições materiais de vida teve novo agravamento com uma desvalorização de 13,2 pontos percentuais entre 2004 e 2012. Dada a forte associação existente entre muitas das variáveis que compõem este indicador sintético e o funcionamento do sistema económico, a sua evolução reflete o baixo crescimento da economia no período pré-crise e é particularmente sensível aos efeitos do aprofundamento da crise económica.

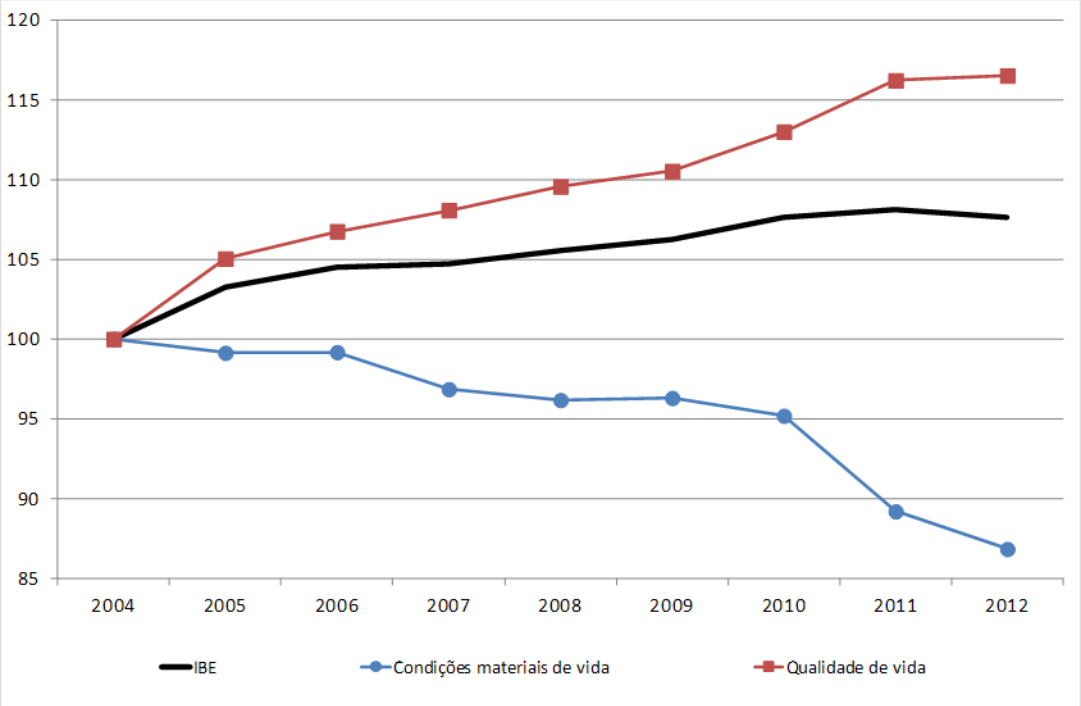



**Figura 1 - Índice de Bem-estar (IBE): global e por perspetiva (2004=100)**

Efetivamente, a análise da evolução nos períodos 2004-2008 e 2008-2012 evidenciou que à quebra registada na variação percentual do índice das Condições materiais de vida entre 2004 e 2008 (-3,8 pontos percentuais), se seguiu uma quebra mais acentuada no período 2008-2012, apontando-se para 2012 uma variação de -13,2 pontos percentuais. As taxas de variação média anual 2004-2008 (-1%) e 2008-2012 (-2,5%) sinalizam este contraste. Por sua vez, na perspetiva da Qualidade de vida, à evolução positiva entre 2004 e 2008 explicada por uma variação de 9,5 pontos percentuais, seguiu-se uma evolução também positiva no período 2008-2012, mas de menor intensidade (7,0 p.p.), estimando-se para 2012, uma variação de 16,5 pontos percentuais, face ao ano base de 2004.

## 3. Contributos para a interpretação dos resultados

A análise em componentes principais do quadro multivariado "domínios *versus* anos" sintetizou o contraste entre a evolução das condições materiais de vida dos portugueses, por comparação com o ano de 2004, e a evolução da respetiva qualidade de vida. Por outro lado, essa análise também evidenciou uma clivagem na evolução dos índices entre 2004-2008 e 2008-2012 e, por conseguinte, alguns sinais do efeito da crise no bem-estar dos portugueses. Complementarmente, a classificação hierárquica dos objetos "domínios do bem-estar" destacou o comportamento ímpar do domínio *educação, conhecimento e competências*, o qual se revelou o domínio com melhor desempenho, tendo a variação do índice sido de 53,9 pontos percentuais no período de 2004-2011. Por outro lado, essa classificação agrega 4 domínios transversais às

duas perspetivas de análise atrás referidas: *vulnerabilidade económica, trabalho e remunerações, participação cívica e governação* e *relações sociais e bem-estar subjetivo*.

Por último, procedeu-se a uma análise classificatória dos 79 indicadores retidos no estudo, identificando 6 *clusters* de indicadores, dos quais dois deles constituem pequenos grupos remanescentes de indicadores, com trajetórias singulares na evolução do respetivo índice. Relativamente a cada um dos 4 primeiros *clusters* identificados, recorremos à representação de *box-plots* associados à distribuição empírica da taxa de variação média dos índices nos períodos 2004-2008, 2008-2011 e 2004-2011.

**Referências**

INE (2013). Destaque: *Índice de Bem-estar 2004-2012*. 6 de Dezembro de 2013.

INE (2013). *Índice de Bem-estar: Documento metodológico*.

OECD (2011b). *How's Life?: Measuring Well-being*. OECD Publishing.

STIGLITZ, J. E., SEN, A., & FITOUSSI, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*.

# SESSÕES PARALELAS

**Classificação e Análise de Dados –5ª Feira, 10 de Abril, Salão Nobre (15h50)**

# Classifying a fairy tale: A case study

**Anna Carolina Finamore[1], M. Rosário Oliveira[2], Cláudia Pascoal[3], and António Pacheco[4]**

[1] *CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal,* anna.couto@tecnico.ulisboa.pt*;*

[2] *CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal,* rsilva@math.tecnico.ulisboa.pt*;*

[3] *CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal,* cpascoal@tecnico.ulisboa.pt*;*

[4] *CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal,* apacheco@math.tecnico.ulisboa.pt

**Abstract**

The enormous quantity of high-dimensional data available nowadays highlights the issue of how to transform data into useful knowledge. Many techniques and algorithms have been proposed for this purpose. In the paper, we explore variable selection algorithms for classification aiming at reducing data dimensionality by removing irrelevant and redundant input variables. We use a dataset containing measurements from books belonging to two exclusive classes (Multivariate Analysis and Children's books), concluding that the use of variable selection algorithms may lead to better performance than using all explanatory variables.

**Keywords:** Classification, Variable selection, Mutual information, Robustness

## 1. Introduction

Classification is the action of assigning a new instance to one of a given set of classes, on the basis of a training data set containing observations whose class membership is known. A set of quantifiable characteristics, known usually as explanatory or input variables, is chosen to characterize the objects to be classified and an algorithm that implements the classification procedure is constructed. In this respect, one should note that the easiness with which variables are nowadays measured and stored in databases leads frequently to assembling a relatively large number of input variables. It is common in practice that input variables are, exactly or approximately, functions of other variables, rendering them useless – or nearly useless - for classification purposes, which may lead to degrading the performance of the classifier. This fact calls for the introduction of variable selection procedures in classification problems, an approach that we follow in this work resorting to the maxMIFS algorithm, proposed in Pascoal (2014) and described in Section 2.

The aim of this study is to investigate if, in the context of a classification problem with highly redundant variables, the use of variable selection methods as a preprocessing step leads to as good classification performance measures as those obtained using the complete set of explanatory variables. Moreover, we also tackle the issue of whether the addition of irrelevant and not redundant variables is circumvented by maxMIFS. We address these issues using as

case study the classification of individual books as being either a “Children's book” or a “Multivariate Analysis book”.

## 2. Methods

In this section we present the main methods used in our work, namely: variable selection and classification methods.

### 2.1. Variable selection

Selecting the most interesting explanatory variables for a given classification problem is a preprocessing step that can be vital for solving real problems, when high quantities of data are involved. In this work, we use the maxMIFS algorithm (proposed by Pascoal, 2014), an alternative to the MIFS method, a pioneer work of Battiti (1994). The maxMIFS algorithm relies on Mutual Information (vide Cover and Thomas, 2006), an information-theoretic metric that captures both linear and nonlinear dependencies between random variables.

The maxMIFS algorithm can be described as follows. Let $X_1$, $X_2$,…, $X_p$, be the input variables, C be the correspondent class variable, S the set of variables already selected (initially S is an empty set), F the set of unselected variables, and $MI(X_i,C)$ the mutual information between an input variable $X_i$ and the class C. Our goal is to find a candidate variable $X_i$ that maximizes the triple mutual information between the class, the candidate variable and the set of selected variables, ie:

$$MI(C, X_i, S) = \max_{(X_s \in S)} MI(X_i, X_s) \text{ for } X_i \in F$$

This algorithm returns an order of the candidate input variables, balancing the relevance of a variable in explaining the class structure and its redundancy, given the subset of input variables previously selected. The involved indices are obtained using classical and robust estimators.

### 2.2. Classification methods

In this work, a classification problem is understood as the problem of assigning a new observation to one of possible populations (also denominated as classes), on the basis of a set of observations characterized by a vector of input variables and a categorical variable indicating the class each observation belongs to. If we consider that input variables have a joint multivariate normal distribution, with equal covariance matrix, the classification rule that minimizes the total probability of misclassification is a linear function of the input variables. By plugging the classical estimators of the parameters, one obtains the classifier associated with classical discriminant analysis (vide Johnson and Whichern, 2007). Pires and Branco (2010) proposed a robust estimator of the discriminant function based on projection pursuit that was implemented in the package *rrcov* by Todorov and Filzmoser (2009).

Among the data machine learning community, the Naïve Bayes classifier as well as the *knn* and C4.5 (a decision tree) classifiers (vide Tan et al., 2005) are common alternatives to classical discriminant analysis, and are also considered in this work. The Naïve Bayes estimator ignores the structure of dependence among the input variables in each class, considering that the input variables are independent normally distributed. The *knn* (with $k$=1) classifier relies on a simple, even though time consuming, (greedy) algorithm based on a nearest neighbor strategy. An object is assigned to the class of the nearest training object, where the used distance between two objects is the Euclidean distance. The C4.5 classifier constructs a tree where at each node the value of an input variable characterizing the object to be classified is compared with a threshold value. After a series of these tests, one reaches a leaf node of the tree, where each object is assigned to one of the classes. At the end, the trees are pruned in order to obtain simpler classifiers.

## 3. Results

In this section we introduce the dataset used and present and discuss the obtained classification results.

### 3.1. Dataset and input variables

In this study, we consider a dataset collected and gently provided to us by Professor João Branco. He measured 29 Multivariate Analysis books accordingly to 8 variables, described in the copy of his original diagram, reproduced in Figure 1, namely: height, width, thickness, diameter including height and width, diameter including height and thickness, diameter including thickness and width, superior perimeter, and lateral perimeter of a book. This dataset was part of a project on principal component analysis, in the first Multivariate Analysis Course taught at Instituto Superior Técnico, as part of the degree in Applied Mathematics and Computation.

Children's books were measured and the corresponding data was included by us in the study in order to have a second book class: Children's book class. Children's books were considered due to their shape variability. More precisely, the dataset used is composed of 29 Multivariate Analysis books and 41 Children's books. Note that the last 5 input variables are, apart from measuring errors, functions (not necessarily linear) of the height (A), width (L), and thickness (E) of books, thus being redundant variables.

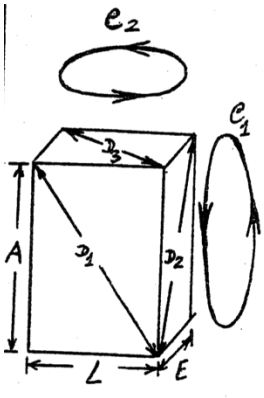



**Figure 1: Copy of Professor João Branco's diagram explaining the input variables, as given to students of the first Multivariate Analysis course taught at Instituto Superior Técnico.**

The dataset was enlarged with the inclusion of 3 irrelevant variables, generated independently from univariate normal distributions with expected value 0, 20, and 30, and standard deviations 1, 1, and 3, respectively. Thus, two sets of input variables were considered: the first one composed by the 8 input variables proposed by Professor João Branco, and a second set with 11 input variables obtained with the inclusion of the mentioned irrelevant variables. The variable selection algorithms, using classical and robust estimates of Mutual Information, were applied to these two sets of variables. The first three variables selected by the classical and robust procedure were: width, thickness, and height (in this order). The choice of the number of variables selected (three) was motivated by the problem itself and by the usual criteria to choose the number of principal component to be retained.

### 3.2. Classification results

The various classifiers under analysis are estimated based on the complete set of input variables (with and without irrelevant input variables) and on the three input variables selected by maxMIFS (classical and robust versions). The associated confusion matrices are summarized in Table 1 and the corresponding accuracy values are summarized in Table 2. We validated our classifier based on the Leave-One-Out strategy.

The use of a reduced set of three input variables lead generally to accuracy values that are better or equal than those obtained when using the complete set of input variables. The exceptions are the Classic LDA (Linear Discriminant Analysis) and Robust LDA without irrelevant variables, for which better classification results are obtained using the complete set of input variables. This could be explained by the fact that the 2 classes (MA: Multivariate Analysis books, and Child: Children's books) seems to have different covariance matrices, this violates the assumptions required to use linear discriminant analysis.

**Table 1: Confusion matrices obtained from several estimation methods under different sets of input variables. Off-diagonal numbers correspond to incorrectly classified books.**

| Estimation Method | Class | With Irrelevant Variables | | | | Without Irrelevant Variables | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All Variables | | Variable Selection | | All Variables | | Variable Selection | |
| | | MA | Child | MA | Child | MA | Child | MA | Child |
| Classic LDA | MA | 29 | 0 | 29 | 0 | 28 | 1 | 29 | 0 |
| | Child | 5 | 36 | 6 | 35 | 4 | 37 | 6 | 35 |
| Robust LDA | MA | 29 | 0 | 28 | 1 | 29 | 0 | 28 | 1 |
| | Child | 6 | 35 | 5 | 36 | 5 | 36 | 5 | 36 |
| Naïve Bayes | MA | 29 | 0 | 28 | 1 | 27 | 2 | 28 | 1 |
| | Child | 2 | 39 | 1 | 40 | 2 | 39 | 1 | 40 |
| *knn* | MA | 29 | 0 | 29 | 0 | 28 | 1 | 29 | 0 |
| | Child | 2 | 39 | 1 | 40 | 1 | 40 | 1 | 40 |
| C4.5 | MA | 28 | 1 | 28 | 1 | 28 | 1 | 28 | 1 |
| | Child | 2 | 39 | 2 | 39 | 2 | 39 | 2 | 39 |

**Table 2: Accuracies obtained from several estimation methods under different sets of input variables.**

| Estimation | With Irrelevant Variables | | Without Irrelevant Variables | |
|---|---|---|---|---|
| | All Variables | Variable Selection | All Variables | Variable Selection |
| Classic LDA | 0.929 | 0.914 | 0.929 | 0.914 |
| Robust LDA | 0.914 | 0.914 | 0.929 | 0.914 |
| Naïve Bayes | 0.971 | 0.971 | 0.943 | 0.971 |
| *knn* | 0.971 | 0.986 | 0.971 | 0.986 |
| C4.5 | 0.957 | 0.957 | 0.957 | 0.957 |

When adding irrelevant variables to the dataset, we observed that the order of the first three relevant variables selected by the classical and robust procedures was not altered. In general the results obtained with the original set of variables (excluding the irrelevant ones) lead to better results than the ones obtained with the original and irrelevant input variables, as expected. However, the same behavior was not observed when using the Naïve Bayes classifier, a fact that may be due to the assumptions of conditional independence and normality of the input variables, verified by the irrelevant input variables.

## 4. Conclusions

In the context of a classification problem with highly redundant variables, the use of variable selection methods as a preprocessing step may lead to better classification performance measures as those obtained using the complete set of input variables.

In the case study considered, we observed that the use of Linear Discriminant Analysis (Classic or Robust) lead to smaller accuracy values than the other methods (Naïve Bayes, *knn* and C4.5). Finally, one should note that the addition of irrelevant input variables was circumvented by the maxMIFS algorithm which did not select any of these variables.

**Acknowledgments:** This work was partially supported by FCT (Fundação para a Ciência e a Tecnologia) through project Pest-OE/MAT/UI0822/2014 and grant SFRH/BD/71780/2010. We express our most sincere gratitude to Prof. João Branco, for providing us the data he collected in the 1980's on measurements from Multivariate Analysis books, which served as inspiration to carry out this study. Finally, we would like to thank the referee for his insightful comments which helped to improve the paper.

## References

BATTITI, R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

COVER, T. & THOMAS, J. (2006). *Elements of Information Theory*. Wiley Sons, New York, NY, USA, 2nd edition.

JOHNSON, R. A. & WICHERN, D. W. (2007). *Applied Multivariate Statistical*, 6th Edition. Pearson Prentice Hall, New Jersey.

PASCOAL, C. (2014). Contributions to Variable Selection and Robust Anomaly Detection in Telecommunications, PhD Thesis, submitted for discussion.

PIRES, A. M. & BRANCO, J. A. (2010). Projection-Pursuit Approach to Robust Linear Discriminant Analysis. *Journal Multivariate Analysis,* **101**, 2464–2485.

TAN, P.-N., STEINBACH, M. & KUMAR, V. (2005). *Introduction to Data Mining*. Addison Wesley, Boston.

TODOROV, V. & FILZMOSER, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, **32**(3), 1-47.

**Classificação e Análise de Dados –5ª Feira, 10 de Abril, Salão Nobre (16h10)**

# Impact of input variables' stability on the classification of Internet applications

**M. Rosário Oliveira[1], Rui Valadas[2], Marcin Pietrzyk[3], and Denis Collange[3]**

[1]*CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Portugal, rsilva@math.tecnico.ulisboa.pt;*

[2]*Departamento de Eng. Electrotécnica e de Computadores and Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal, rui.valadas@tecnico.ulisboa.pt;*

[3]Orange Labs, France, *mj.pietrzyk@gmail.com and denis.collange@orange.com*

**Abstract**

The portability of a statistical classifier for Internet applications is its ability to operate correctly on a dataset it was not trained for. In this work we address the impact of the stability of input variables on the portability of classifiers. Our procedure, applied to 3 datasets, resorts to the notion of effect size and shows that the observed performance degradation for some Internet applications is explained by the lack of the input variables' stability between datasets captured at different locations and/or time periods.

**Key-words:** Internet traffic classification; Identification of Internet applications, Hypothesis testing, Effect size.

## 1. Introduction

The classification of Internet applications has deserved considerable attention in recent years. Deterministic approaches based on port numbers and Deep Packet Inspection (DPI) solutions turn to be unreliable or unrealistic given the increasing use of dynamic ports, encryption and obfuscation of packets content. To address these problems, recent studies have relied on statistical classification techniques to probabilistically map objects (traffic flows) to Internet applications (Dainotti et al. 2012, Kim et al. 2008, Li et al. 2009, Moore and Zuev 2005). In general, these proposals have been evaluated in the context of a single dataset and the reported performance is usually very high. However, there are two important aspects in the performance of classifiers that have deserved little attention so far.

First, the per-class performance of some applications may be poor even if the overall performance is very high. The overall performance follows the trend of the most numerous applications, and if their classification performance is good so it will the overall one. However, this may mask poor performance on less frequent applications.

Second, a classifier trained and tested on a particular dataset may perform poorly on a different dataset, collected on a different location or time period. We will refer to the ability of a classifier to operate correctly on datasets it was not trained for as its *portability*. Portability is an important requirement for network operators since it would be impractical to train every classifier that needs to be deployed in the numerous operator's Points-of-Presence (PoPs). For cost-effectiveness, large-scale classification infrastructures must rely on a small set of highly-

portable (pre-trained) classifiers. Moreover, it is important to highlight that the quest for portability is in the heart of the classification problem. Indeed, classifiers for Internet applications should seek capturing application specific behaviors and not dataset or site specific ones.

So far, the portability problem has only been addressed in a few works and mainly from an empirical perspective. Li et al. (2009) report low portability performance in P2P and INTERACTIVE traffic and Pietrzyk et al. (2009) in BITTORRENT and FTP traffic. In this work we go a step further, trying to understand the impact of the *stability* of input variables on the portability performance of classifiers. Informally, an input variable is stable if it keeps its main statistical properties across different datasets. An important question is then: what are the characteristics of input variables that must remain stable in order to assure portability? A first attempt to answer this question was given by Este et al. (2009). The authors used the mutual information measure to show that the amount of information carried by specific input variables remains constant irrespective of capture time or point of observation. However, preservation of the information content does not necessarily imply portability, since mutual information is invariant under scale or location transformations and under one-to-one transformations, that is, $MI(X;Y) = MI(U;V)$ where $u = g(x)$, $v = g(y)$, and $g$ is an invertible function (Dadkhah et al. 2010), and in general these properties do not hold for the classifiers under consideration.

In this work we propose a statistical procedure that relies on the notion of *effect size* (Thompson et al. 2007, Cohen 1988, Hedges and Olkin 1985) to identify stable input variables. To show the merits of our proposal we use three datasets collected at two different Internet locations in France, from the same Internet Service Provider. The dataset have some important spatial and temporal characteristics that are crucial to assess the portability of statistical classifiers: traces MS-I and R-III were captured at exactly the same time at two different locations; traces R-II and R-III were captured at the same location with an offset of seventeen days between them. Each object, which corresponds to an Internet traffic flow, is characterized by four input variables which enjoy wide acceptance as good discriminators for Internet applications (Bernaille et al. 2006, Pietrzyk et al. 2009, Este et al. 2009). The C4.5 algorithm is the chosen classifier since it has been widely employed in several works including (Kim et al. 2008, Li et al. 2009, Pietrzyk et al. 2009), and in our case provides the best trade-off between accuracy and model building time. We rely on the implementation of the C4.5 algorithm provided in the Weka suite (Hall et al. 2009).

We analyze the stability properties of the four input variables and show that our procedure helps explaining the poor portability performance of some applications when using these input variables as discriminators.

**References**

BERNAILLE L, TEIXEIRA R & SALAMATIAN K (2006). Early application identification. *Proc. of the 2006 ACM CoNEXT Conference (CoNEXT '06)*.

COHEN J (1988). *Statistical power analysis for the behavior sciences*. 2nd edn., Laurence Erlbaum Associates: New Jersey.

DADKHAH K, MIDI H & OLIMJON S (2010). The Performance of Mutual Information for Mixture of Bivariate Normal Distributions Based on Robust Kernel Estimation. *Applied Mathematical Sciences*, 4(29):1417–1436.

DAINOTTI A, PESCAPE A & CLAFFY K (2012). Issues and future directions in traffic classification. *IEEE Network*, 26(1):35–40.

ESTE A, GRINGOLI F & SALGARELLI L (2009). On the stability of the information carried by traffic flow features at the packet level. *SIGCOMM Comput. Commun. Rev.*, 39(3).

HALL M, FRANK E, HOLMES G, PFAHRINGER B, REUTEMANN P & WITTEN IH (2009). The Weka data mining software: An update. *SIGKDD Explorations*, 11(1).

HEDGES LV & OLKIN I (1985). *Statistical methods for meta-analysis*. Academic Press.

KIM H, CLAFFY K, FOMENKOV M, BARMAN D, FALOUTSOS M & LEE K (2008). Internet traffic classification demystified: myths, caveats, and the best practices. *Proc. of the 2008 ACM CoNEXT Conference (CONEXT '08)*.

LI W, CANINI M, MOORE AW & BOLLA R (2009). Efficient application identification and the temporal and spatial stability of classification schema. *Computer Networks*, 53(6):790–809.

MOORE AW & ZUEV D (2005). Internet traffic classification using bayesian analysis techniques. *Proc. of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'05)*, 50–60.

OLIVEIRA MR, VALADAS R, PIETRZYK M & COLLANGE D. (2014). Portability of statistical classifiers for the identification of Internet applications. *International Journal of Communication Systems*, submitted.

PIETRZYK M, COSTEUX JL, EN-NAJJARY T, URVOY-KELLER G (2009). Challenging statistical classification for operational usage: the ADSL case. *Proc. of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC '09)*.

THOMPSON B (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44(5):423–432.

**Classificação e Análise de Dados – 5ª Feira, 10 de Abril, Salão Nobre (16h30)**

# On circulant matrix approximation to correlation matrix: an application to sounds

**Eunice Carrasquinha[1], Conceição Amado[2], Ana M. Pires[3].**
[1]*CEMAT – Instituto Superior Técnico, eunice.trigueirao@ist.utl.pt;*
[2]*CEMAT – Instituto Superior Técnico, camado@math.ist.utl.pt;*
[3]*CEMAT – Instituto Superior Técnico, apires@math.ist.utl.pt;*

**Abstract**

Signal processing is an important task in our days that arises in various areas such as engineering and applied mathematics. A signal represents time-varying or spatially varying physical quantities. Signals of importance can include sound, electromagnetic radiation, images, telecommunication transmission signals, and many others. A signal carries information, and the objective of signal processing is to extract useful information carried by the signal. The received signal is usually disturbed by electrical, atmospheric or deliberate interferences. Due to the random nature of the signal, statistical techniques play an important role in analyzing the signal.

There are many techniques used to analyze these types of data, depending on the focus or research question of the study. Some of these techniques are Principal Component Analysis (PCA) and Fourier transform, in particular discrete Fourier transform (DFT). The main goal in this work is to exploit the relations between PCA and others mathematical transforms. To achieve that we propose a method, based on Toeplitz and circulant matrices, which the eigenvectors correspond to each column of the Fourier matrix. In this sense, the method presented relates the theory behind the Fourier transform through the Toeplitz and circulant matrices and the Principal Component Analysis. To illustrate the proposed methodology we will consider sound data.

**Keywords:** Circulant Matrix, Fourier Transform, Principal Component Analysis, Signal Processing, Toeplitz Matrix.

## 1. Introduction

Signal processing is an important task in our days that arise in various areas such as engineering and applied mathematics. Due to the constant presence of signs of interest, such as: sound, images, an analysis of these has become very important. There are several statistical techniques to analyze the signal processing, the most common are Principal Component Analysis (PCA) and the Fourier transform. Many signal processing algorithms rely on methods of decorrelation of the data. The advantages of techniques mentioned, are based on the principle that decorrelating the data serves to eliminate some of the redundant information. The PCA can transform a set of original, correlated, variables in a new set of uncorrelated variables, the principal components. The principal components are obtained by a linear combination of the variables, with weights chosen in a way that the principal components become mutually uncorrelated. Each component contains new information about the system, and is ordered so

that the firsts components account for most of the variability. PCA, often also known, particularly for continuous-time processing signals by Karhunen-Loève transform (KLT), or Hotelling transform. In this way KLT is closely associated with PCA, when we intend to analyze signals, including approximation, compression or classification. In the literature we can find various techniques for processing signals and images. The Fourier transform is one of the most used techniques, for processing signals due to its ability to decompose a signal into a sum of sinusoids. Fourier transform can be divided into four categories depending on the combination between: continuous or discrete signals; or periodic or aperiodic signals. One type is the discrete Fourier transform (DFT), which is used to perform Fourier analysis in signal processing. The method used to determine the DFT is called Fast Fourier transform (FFT). In general, the interpretation of the results based on DFT is not easy, and this study contributes to provide a methodology based on Toeplitz matrices and in particularly in circulant matrices, whose eigenvectors correspond to the columns of a Fourier matrix.

There has been a wide array of work on the problem to find the best circulant matrix associated to a Toeplitz matrix, more details can be found in Chan, T. (1988), Huckle. T. K. (1990), and Tismenetsky, M. (1991). In addition, there have been studies on approximation covariance matrix by Toeplitz, taking advantage of the properties of the latter. One of those studies was presented by Unser (1984) who used circulant and skew-circulant matrices to approximate covariance matrices. In 2001 Kouassi and Gouton presented a similar approach. However, both proposals are only valid for first-order Markov processes. Our proposed method generalizes those approximations for a correlation matrix associated with covariance matrix with any structure.

The new approach proposed is centered on Toeplitz and circulant matrices to approximate a correlation matrix based on least squares. Approximating a correlation matrix with a simple underlying structure can lead to faster computation times and the interpretation of the results becomes simpler. A brief introduction to the theory of Toeplitz and circulant matrices it will be present. Finally, the method is illustrated with sound data sets.

**Acknowledgement:** This work has been supported by Fundação para a Ciência e Tecnologia (FCT, Portugal).

## References

CHAN, T. (1988). An optimal preconditioner for Toeplitz systems. *SIAM J. Sci. Statist.Comput.,* 9(4)**,** 766-771.

HUCKLE, T. K. (1990). Circulant and skew-circulant matrices for solving Toeplitz matrix problems.*Copper Mountain Conference on Iterative Methods,* Cooper Mountain, Colorado.

KOUASSI, R.& GOUTON, P. (2001). Approximation of the Karhunen-Loève transformation and its applications to colour images.*Signal Processing: Image Communication,* 16**,** 541-551.

TISMENETSKY, M. (1991). A decomposition of Toeplitz matrices and optimal circulant preconditioning. *Linear Algebra and its Applications,* 154-156**,** 105-121.

UNSER, M. (1984). On the approximation of the Karhunen-Loève transform for stationary processes. *Signal Processing,* 7**,** 231-249.

**Classificação e Análise de Dados – 5ª Feira, 10 de Abril, Salão Nobre (16h50)**

# Metodologia STATIS em Controlo Estatístico da Qualidade

**Adelaide Maria Figueiredo[1], Fernanda Otília Figueiredo[2]**
[1]*Faculdade de Economia e LIAAD-INESC TEC, Universidade do Porto, adelaide@fep.up.pt;*
[2]*Faculdade de Economia, Universidade do Porto, e CEAUL, otília@fep.up.pt*

**Sumário**

Em situações reais a avaliação da qualidade global de um produto ou de um serviço depende de mais do que uma caraterística de qualidade, pelo que o desenvolvimento de cartas de controlo para dados multivariados é crucial. Com o objetivo de monitorizar processos multivariados e identificar as variáveis responsáveis por mudanças no processo, iremos utilizar a metodologia STATIS.

**Palavras-chave:** ACP, coeficiente RV, controlo estatístico da qualidade, monitorização de processos, STATIS**.**

## 1. Introdução

A avaliação da qualidade global de um produto ou de um serviço em situações reais depende de várias caraterísticas de qualidade. A utilização de técnicas univariadas para cada uma das caraterísticas é obviamente inapropriada, uma vez que estas caraterísticas de qualidade estão em geral correlacionadas, sendo necessário usar técnicas multivariadas de controlo de processos. A maior parte dos esquemas de controlo multivariados são baseados na estatística de Hotelling e são implementados sob a hipótese de dados normais multivariados. Estes esquemas consistem em monitorizar simultaneamente o valor médio e a matriz de covariâncias do processo ou então, monitorizar separadamente o vector médio e a matriz de covariâncias. Atendendo a que há várias variáveis correlacionadas em jogo para definir a qualidade de um produto, sempre que um esquema de controlo emite um sinal, é importante averiguar quais são as variáveis do processo responsáveis pela emissão desse sinal. Neste trabalho iremos aplicar a metodologia STATIS na monitorização de processos multivariados e, além disso, através desta metodologia, sempre que a carta emite sinal de fora de controlo identificaremos as variáveis responsáveis pela emissão desse sinal.

## 2. Metodologia STATIS

A metodologia STATIS (Structuration des Tableaux à Trois Indices de la Statistique) foi introduzida por L'Hermier des Plantes (1976) e L'Hermier des Plantes e Thiebaut (1977) e, posteriormente desenvolvida por Lavit (1988a, 1988b) e Lavit *et al.* (1994). Esta metodologia permite analisar simultaneamente vários quadros de dados quantitativos, recolhidos em ocasiões ou circunstâncias diferentes. Consoante o objetivo do estudo está centrado na análise dos indivíduos ou das variáveis, considera-se o método STATIS ou o método STATIS dual. Se

os indivíduos são os mesmos ao longo dos vários quadros de dados e as variáveis podem diferir ou não ao longo dos quadros, privilegia-se as proximidades entre indivíduos (método STATIS). Se os dados são recolhidos para as mesmas variáveis ao longo dos quadros de dados e os indivíduos podem diferir ou não ao longo desses quadros, evidencia-se neste caso as relações entre variáveis (método STATIS Dual). Estes dois métodos compreendem as seguintes etapas:

- Inter-estrutura: comparação global da estrutura das matrizes de dados. No método STATIS comparam-se as nuvens de indivíduos através das matrizes dos produtos escalares e no método SIATIS Dual comparam-se as nuvens de variáveis através das matrizes de correlações.
- Intra-estrutura: descrição da estrutura comum aos vários quadros de dados através da determinação do compromisso e da respetiva imagem euclidiana.
- Representação das trajetórias dos indivíduos/variáveis: estas representações permitem destacar os indivíduos (STATIS) ou variáveis (STATIS Dual) que mais contribuíram para as semelhanças ou diferenças encontradas entre os vários quadros. A partir da imagem compromisso traçam-se as trajetórias que descrevem o comportamento evolutivo de cada indivíduo (variável). Procede-se, ainda, à decomposição dos quadrados das distâncias entre quadros para averiguar quais os indivíduos que mais (ou menos) contribuíram para as diferenças entre os quadros.

Em Scepi (2002), Gourvénec *et al.* (2005) e Figueiredo *et al.* (2012) são apresentadas aplicações interessantes da metodologia STATIS a conjuntos de dados reais, as primeiras das quais em Controlo Estatístico da Qualidade.

## 3. Aplicação do STATIS ao Controlo Estatístico de Qualidade

As cartas de controlo são as ferramentas usualmente utilizadas para a monitorização de processos em Controlo Estatístico da Qualidade (CEQ). Inicialmente surgiram para a monitorização de processos industriais (introduzidas por Shewhart em 1924 nos *Bell Laboratories*) mas atualmente são aplicadas nas mais diversas áreas, entre elas, na Saúde, Medicina, Genética, Ambiente e Finanças. O objectivo destas representações é o de ajudar a tomar decisões sobre o estado do processo que está a ser monitorizado: sob controlo ou fora de controlo. Sempre que a carta emite um sinal de fora controlo, que eventualmente poderá ser um falso alarme, é necessário investigar as causas responsáveis pela emissão desse sinal para que sejam tomadas ações corretivas adequadas. Uma abordagem menos convencional para a monitorização de processos consiste em explorar melhor a teoria dos testes de hipóteses, definindo procedimentos de teste onde a tomada de decisão é feita com base na região de rejeição ou de aceitação, e em particular, no cálculo do p-value como medida quantitativa da verosimilhança de uma possível alteração no processo, tal como é sugerido por Li *et al.* (2013). Esta abordagem, no nosso entender, poderá impulsionar bastante o aparecimento de novas cartas de controlo não paramétricas, por adaptação dos variados testes não paramétricos existentes, e também o de novas cartas multivariadas baseadas em metodologias mais recentes de análise de dados.

Neste trabalho usaremos a metodologia STATIS com o objectivo de monitorizar processos normais multivariados. Iremos comparar as matrizes de correlações quando o processo está sob controlo com as matrizes de correlações quando o processo está fora de controlo. Assim, com base nas matrizes de correlações associadas a amostras de referência obtidas em diferentes instantes de tempo, quando o processo está sob controlo, construímos a matriz de correlações compromisso. Para amostras recolhidas em novos instantes de tempo, comparamos as respectivas matrizes de correlações com a matriz de correlações compromisso através do coeficiente RV (Escoufier, 1973). Se este coeficiente for significativamente elevado, concluímos que o processo está sob controlo e caso contrário, decidimos que o processo está fora de controlo. Para testar a significância estatística do coeficiente RV, utilizamos o teste proposto em Josse *et al.* (2008). No caso em que uma nova amostra leva a concluir que o processo está fora de controlo, identificamos as variáveis que são responsáveis por esta situação. Para esse efeito efetuamos a decomposição dos quadrados das distâncias entre as matrizes de correlações das amostras de referência e da nova amostra, em percentagem de contribuições das variáveis, tal como usualmente se faz na última fase do método Statis Dual.

Para ilustrar a estratégia que propomos para decidir se o processo normal multivariado está sob controlo ou fora de controlo, vamos considerar o seguinte exemplo, inspirado em Hawkins e Maboudou-Tchao (2008). Suponhamos um processo com quatro características de qualidade $X_1$, $X_2$, $X_3$, $X_4$, que seguem distribuição normal multivariada $N_4(\boldsymbol{\mu}, \Sigma)$. Quando o processo está sob controlo o vetor médio e a matriz de covariâncias são dadas por

$$\boldsymbol{\mu}_0 = \begin{Bmatrix} 126.61 \\ 77.48 \\ 80.95 \\ 97.97 \end{Bmatrix} \quad \text{e} \quad \Sigma_0 = \begin{pmatrix} 15.04 & & & \\ 8.66 & 5.83 & & \\ 10.51 & 5.56 & 15.17 & \\ 12.04 & 7.50 & 8.79 & 10.57 \end{pmatrix},$$

respetivamente.

Inicialmente, geramos 10 amostras de referência de dimensão n=10 com o processo sob controlo. Com base nestas amostras, e recorrendo ao método Statis dual, determinamos a matriz de correlações compromisso. Em seguida, geramos 3 amostras de dimensão 10, com o processo sob controlo e também geramos 3 amostras com o processo fora de controlo, considerando um acréscimo de 50% nas variâncias, um decréscimo de 50% nas covariâncias e não alterando o vetor médio. Com base nestas amostras, concluímos usando o critério definido anteriormente, se o processo está sob controlo ou fora de controlo. Nas situações de fora de controlo do processo, identificamos quais as variáveis que contribuem para essas situações.

**Agradecimentos:** Este trabalho é financiado por Fundos FEDER através do Programa Operacional Fatores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos FCOMP-01-0124-FEDER-037281 e PEst-OE/MAT/UI0006/2014.

**Referências**

ESCOUFIER, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, 751-760.

FIGUEIREDO, A., FIGUEIREDO, F., MONTEIRO, N. & STRAUME, O. (2012) Restructuring in privatised firms: a Statis approach. *Structural Change and Economic Dynamics*, 23, 108-116.

GOURVÉNEC, S., STANIMIROVA, I. & SABY, C. A. (2005) Monitoring batch process with the STATIS approach. *Journal of Chemometrics*, 19, 288-300.

HAWKINS, D. M. & MABOUDOU-TCHAO, E. M. (2008) Multivariate exponentially weighted moving covariance matrix. *Technometrics*, 50, 155-166.

JOSSE, J., PAGÈS, J. & HUSSON, F. (2008) Testing the significance of the RV coefficient. Computational Statistics and Data Analysis, 53, 82-91.

LAVIT, C. (1988a) *Analyse Conjointe de Tableaux Quantitatives*, Collection Méthodes+Programmes, Masson.

LAVIT, C. (1988b) Presentation de la méthode STATIS permettant l'analyse conjointe de plusieurs tableaux de données quantitatives. *Cachiers de la Recherche Developpement*, 18, 49–60.

LAVIT, C., ESCOUFIER, Y., SABATIER, R. & TRAISSAC, P. (1994) The ACT (Statis method). *Computational Statistics & Data Analysis*, 18, 97-119.

L'HERMIER DES PLANTES, H. (1976) *Structuration des Tableaux a Trois Indices de la Statistique*. Thése de 3 éme cycle. Université de Montpellier II.

L'HERMIER DES PLANTES, H., & THIEBAUT, B. (1977) Étude de la pluviosité au moyen de la methode S.T.A.T.I.S. *Révue de Statistique Appliquée*, 25 (2), 57–81.

LI, Z., QIU, P., CHATTERJEE, S. & WANG, Z. (2013) Using p-values to design statistical process control charts. *Statistical papers*, 54, 523–539.

SCEPI, G. (2002) Parametric and non parametric multivariate quality control charts. IN LAURO, C. et al. (Eds) *Multivariate Total Quality Control*,163–189. Heidelberg: Physica-Verlag.

**Análise de Dados em Medicina – 5ª Feira, 10 de Abril, Sala 316 (15h50)**

# Progression of carotid atherosclerotic plaques: speed and dependency from vascular risk factors

**A. Rita Gaio[1], Óscar Felgueiras[2], Rosa Santos[3], Elsa Azevedo[4]**

[1]*Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, argaio@fc.up.pt;*

[2]*Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, olfelgue@fc.up.pt;*

[3]*Departamento de Neurologia, CHSJ-Porto & Faculdade de Medicina da Universidade do Porto, rosampsantos2@gmail.com*

[4]*Departamento de Neurologia, CHSJ-Porto & Faculdade de Medicina da Universidade do Porto, elsaazevedo1@gmail.com*

**Summary**

We aimed to assess the progression rate of carotid atherosclerotic plaques, detected by ultrasonography (US), and its relation to vascular risk factors (VRF). We performed a retrospective study in 202 patients, who underwent a follow-up carotid US scan with an interval of 8-18 months. Stenosis progression and its relation to VRF were evaluated through multiple linear regression models, fitted by generalized least squares. Worse stenosis progression was significantly associated with higher initial stenosis. Significant cross-sectional VRF for stenosis severity were found.

**Keywords:** Carotid artery, Generalized least squares, Multiple linear regression, Progression of atherosclerosis, Vascular risk factors.

## 1. Introduction

Carotid stenosis progression reflects the atherosclerotic disease activity and is often associated with stroke. There have been several studies on the progression of atherosclerotic plaques either focusing on determining the number of patients who had evolved into a significant atherosclerotic stenosis and to vascular events [DELCKER ET AL., 1995] or identifying the VRF and their influence on the atherosclerosis progression [HERDER ET AL., 2012; SABETI ET AL., 2007]. Since Portugal presents a high stroke incidence, and there are no published studies concerning the evaluation of the carotid stenosis that could help explain the great prevalence of atherosclerosis, as well as its associated factors and progression velocity, the authors aimed to design a study in a hospital-based sample of Portuguese patients. The objective of the study was to assess the progression of carotid atherosclerotic plaques, detected by ultrasonography, and its relation to the presence and management of atherosclerosis' risk factors, identifying patients with a high risk of atherosclerosis progression. As secondary objective, the authors evaluated the variation of the effective control of VRF between the two examinations, thus studying the impact of atherosclerotic disease diagnosis into the clinical approach and its compliance by the patient.

## 2. Modelling and statistical analysis

We used a prospective database of the carotid Doppler ultrasound scans from the Neurosonology laboratory of the Neurology Department of São João Hospital Center. We retrospectively collected clinical data and information from 202 individuals related to the vascular risk factors, including data concerning demography, history of smoking habits, arterial hypertension (HT), diabetes mellitus (DM) and hypercholesterolemia (HC), previous stroke and coronary heart disease, as well as blood analytical parameters and reason for the examination. Included patients were those who had 1) undergone a minimum of two examinations with an interval of 8 to 18 months, and had at least one atherosclerotic stenosis in one of the carotid axes; 2) an obstruction percentage between ≥20% and ≤99% (measured by the ECTS method) in the common carotid artery (CCA), or located at its bifurcation or proximal internal carotid artery (bif/ICA); and 3) sufficient information on the clinical record related to the VRF, as well as the analytical blood values with a proximity inferior to 3 months to the Doppler ultrasound examination date.

Multiple linear regression models studied the stenosis progress throughout time. The generalized least squares method with normally distributed errors that were allowed to be correlated and/or heteroscedastic was applied [PINHEIRO & BATES, 2009].

Time was considered a dichotomous variable, reflecting the two evaluation periods. Hypertension was dichotomized depending on whether it was controlled (including inexistent) or not. A new variable, hereafter denoted by initial stenosis severity, classifying initial stenosis simply as being ≥50%, or <50%, was also defined. For time t, location l, age class a, status of controlled HT given by c, and status of initial severity given by s, the best fitted model was

$$\log(Stenosis)\,(t,l,a,c,s) = \beta_0(l,a,c,s,c*s) + \beta_1(s)t + \varepsilon \qquad (1)$$

with linear functions $\beta_0$and $\beta_1$, and errors $\varepsilon$ following a zero mean normal distribution. The errors variance-covariance matrix was found to be of compound symmetry at the 3-level grouping structure given by subject/axis/location, and with different variances according to the dichotomous value of the initial stenosis severity. Dummy variables were considered for the categorical variables time, location, age class, status of controlled hypertension at the second examination and dichotomous initial stenosis; reference categories were taken to be the 1st examination period, bif/ICA area, having less than 65 years-old, being either normotensive or hypertensive with controlled values at the second examination, and having an initial stenosis lower than 50%. Graphical analyses were used to assess normality and heteroscedasticity of model residuals and no compromising features were detected.

In order to comprehend which factors favor the greatest stenosis' progression, the authors selected, for each individual, the plaque presenting the highest stenosis at examination 2, from those with positive progression from examination period 1 to 2. Regarding those 112 plaques (from 112 different individuals and read at two different time points), and using the same terminology as above, the best fitted regression model was

$$\log(Stenosis)(t,l,c,s) = \beta_0(l,c,s) + \beta_1(s)t + \varepsilon. \quad (2)$$

Once again, the functions $\beta_0$ and $\beta_1$are linear and the errors ε follow a zero mean normal distribution. The structure of the errors variance-covariance matrix was found to be of compound symmetry at the individual level, and with different variances according to the dichotomous value of the initial stenosis severity. Dummy variables and reference categories were considered as in (1).

Final regression models were chosen on the basis of the lowest BIC (Bayesian Information Criterion) or of the likelihood ratio test, as appropriate. The obtained estimates for the regression coefficients are presented in Table 1 and Table 2.

**Table 1: Estimates obtained from the fitting of model 1.**

| Variable | Coefficient Estimate | 95% Confidence Interval |
|---|---|---|
| Intercept | 3.387 | (3.336, 3.438) |
| 2nd examination | 0.063 | (0.046, 0.081) |
| CCA location | -0.111 | (-0.177, -0.046) |
| ≥ 65 years-old | 0.072 | (0.031, 0.113) |
| Initial stenosis ≥ 50% | 0.691 | (0.628, 0.755) |
| Uncontrolled HT at 2nd examination | 0.103 | (0.044, 0.162) |
| 2nd examination*(Initial stenosis ≥ 50%) | -0.042 | (-0.067, -0.018) |
| (Uncontrolled HT at 2nd examination)*(Initial stenosis ≥ 50%) | -0.088 | (-0.171, -0.006) |
| Correlation parameter | 0.833 | (0.800, 0.862) |
| Error variance when initial stenosis ≥50%* | 0.691 | (0.621, 0.770) |
| Residual standard error | 0.243 | (0.225, 0.262) |

*: standardizing the variance of initial stenosis <50% to 1. LEGEND – CCA: common carotid artery; HT: arterial hypertension.

**Table 2: Estimates obtained from the fitting of model 2.**

| Variable | Coefficient Estimate | 95% Confidence Interval |
|---|---|---|
| Intercept | 3.505 | (3.420, 3.589) |
| 2nd examination | 0.170 | (0.136, 0.203) |
| CCA location | -0.305 | (-0.484, -0.126) |
| Initial stenosis ≥ 50% | 0.350 | (0.284, 0.417) |
| Uncontrolled HT at 2nd examination | 0.177 | (0.077, 0.277) |
| 2nd examination* *(Initial stenosis ≥ 50%) | -0.097 | (-0.146, -0.048) |
| Correlation parameter | 0.903 | (0.837, 0.943) |
| Error variance when initial stenosis ≥50%* | 0.698 | (0.568, 0.859) |
| Residual standard error | 0.318 | (0.268, 0.378) |

*: standardizing the variance of initial stenosis <50% to 1. LEGEND – CCA: common carotid artery; HT: arterial hypertension.

## 3. Conclusion

The two models gave consistent results with each other. Among the several VFR, only uncontrolled HT was shown to be statistically significant. Further studies with either larger sample sizes or longer time intervals between each US may be needed in order to identify other VFR with significant effects on stenosis progression.

**Acknowledgements:** The first and second authors were partially funded by the European Regional Development Fund through program COMPETE and by the Portuguese Government through FCT under the project PEst-C/MAT/UI0144/2013.

## References

DELCKER, A., DIENER, H.C. & WILHELM, H. (1995). Influence of vascular risk factors for atherosclerotic carotid artery plaque progression. *Stroke*, 26(11), 2016-22.

HERDER, M., JOHNSEN S.H., ARNTZEN K.A., MATHIESEN E.B. (2012). Risk factors for progression of carotid intima-media thickness and total plaque area: a 13-year follow-up study: the Tromso Study. *Stroke*, 43(7):, 1818-23.

PINHEIRO, J., BATES, D. (2009). *Mixed-Effects Models in S and S-Plus*, New York, USA, Springer Verlag.

SABETI, S., SCHLAGER, O., EXNER, M., MLEKUSCH, W., AMIGHI, J., DICK, P., MAURER, G., HUBER, K., KOPPENSTEINER, R., WAGNER, O., MINAR, E. & SCHILLINGER, M. (2007). Progression of carotid stenosis detected by duplex ultrasonography predicts adverse outcomes in cardiovascular high-risk patients. *Stroke*, 38(11) 2887-94.

**Análise de Dados em Medicina – 5ª Feira, 10 de Abril, Sala 316 (16h10)**

# Impacto das normas de orientação clínica na evolução do padrão de prescrição de antidiabéticos orais e antihipertensores em Portugal – Exemplo prático da análise de regressão segmentada a uma série temporal interrompida

**José Guerreiro[1], Carla Torre[2], Marta Gomes[3], Suzete Costa[4]**

*[1] Centro de Estudos e Avaliação em Saúde (CEFAR), Associação Nacional das Farmácias (ANF), jose.guerreiro@anf.pt;*

*[2] Centro de Estudos e Avaliação em Saúde (CEFAR), Associação Nacional das Farmácias (ANF), carla.torre@anf.pt;*

*[3] Centro de Estudos e Avaliação em Saúde (CEFAR), Associação Nacional das Farmácias (ANF), marta.gomes@anf.pt;*

*[4] Centro de Estudos e Avaliação em Saúde (CEFAR), Associação Nacional das Farmácias (ANF), suzete.costa@anf.pt*

**Palavras-chave:** Diabetes, Hipertensão, Intervenção, Regressão Linear, Série Temporal

## 1. Introdução

A análise de regressão segmentada a uma série temporal interrompida é uma técnica utilizada para identificar e quantificar o efeito de uma intervenção num *outcome* de interesse no tempo (Wagner, 2002). Este trabalho tem como objectivo, analisar a sua aplicação no contexto da investigação associada à utilização de medicamentos, com exemplos reais recolhidos a partir de uma base de dados Nacional sobre o consumo de medicamentos em ambulatório.

## 2. Metodologia

A informação sobre o consumo de medicamentos foi obtida a partir da base de dados hmR/SICMED, um sistema de informação representativo do consumo de medicamentos em ambulatório em Portugal, para o período de Janeiro de 2010 a Dezembro de 2013. Os medicamentos seleccionados fazem parte do grupo dos Antidiabéticos Orais (ADO) e dos Antihipertensores (ATH), uma vez que para estes grupos foram publicadas Normas de Orientação Clínica (NOC) (DGS, 2011a; DGS, 2011b; EURO MED STAT, 2004) por parte da Direcção-Geral da Saúde (DGS), no seguimento do Memorando de Entendimento (MoU, 2011). O momento de intervenção seleccionado foram as datas da publicação das NOC, Dezembro de 2011 para ADO e Setembro de 2011 para AHT. Os resultados foram analisados em função do número de DDD (Dose Diária Definida, OMS) consumidas. Foram ajustados modelos de regressão linear com variável dicotómica de intervenção (Normas DGS) e variável de tempo após intervenção à série da proporção do consumo de Metformina (medicamento de primeira linha) no total de ADO e a série do rácio de hipertensores ARA (Antagonistas dos Receptores da Angiotensina II) no total de ARA e IECA (Inibidores da Enzima de Conversão

da Angiotensina II), sendo que um rácio mais elevado indicava um pior padrão de consumo. A análise foi efectuada utilizando o software SAS Guide v4.1.

## 3. Resultados

Os resultados do modelo de regressão à série temporal da metformina ($p<0,0001$; $R^2=0,8692$) indicam ter havido uma alteração significativa na sua utilização após a publicação da Norma da DGS, verificando-se um ligeiro crescimento na tendência ($\beta=0,001$; $p<0,0001$) e no nível ($\beta=0,007$; $p<0,0001$) de consumo dentro dos ADO. O modelo relativo ao rácio ARA/(ARA + IECA) ($p<0,0001$; $R^2=0,9630$), tem vindo a aumentar de forma significativa ($\beta=0,001$; $p<0,0001$) mesmo após a publicação da NOC, ainda que de forma menos acentuada, mas sem alteração de nível ($\beta=-0,0005$; $p=0,4787$).

## 4. Conclusões

Os modelos de regressão segmentada a séries temporais interrompidas ajustados permitiram identificar alterações significativas no padrão de consumo dos medicamentos analisados. A publicação da Normas da DGS dos ADO parece ter contribuído ligeiramente para a inverter a diminuição do consumo de metformina. Contudo, a norma da DGS dos antihipertensores não parece ter tido impacto, pois não inverteu a tendência de crescimento do rácio ARA/(ARA + IECA).

## Referências

Direção-Geral da Saúde (2011a). Abordagem Farmacológica na Diabetes Mellitus tipo 2. https://www.dgs.pt/, (acedido em 7 de Janeiro de 2014)

Direção-Geral da Saúde (2011b). Abordagem Terapêutica da Hipertensão Arterial. https://www.dgs.pt/, (acedido em 7 de Janeiro de 2014)

MoU - Memorandum of understanding on specific economic policy conditionality (Mai-2011), http://www.portugal.gov.pt/media/371369/mou_20110517.pdf, (acedido em 10 Janeiro 2014)

EURO MED STAT (2004). The Library of European Union Pharmaceutical Indicators Expenditure and Utilization Indicators. Final version. http://ec.europa.eu/health/ph_projects/2001/monitoring/fp_monitoring_2001_frep_12_3_en.pdf (acedido em 25/09/2013).

WAGNER, A.K., SOUMERAI, S.B., ZHANG, F. & ROSS-DEGNAN, D. (2002) Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27, 299–309.

**Análise de Dados em Medicina – 5ª Feira, 10 de Abril, Sala 316 (16h30)**

# Equiparação das classificações dos cursos de Medicina

**A. Rita Gaio[1], Joaquim Costa[2], Milton Severo[3]**

[1]*Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, argaio@fc.up.pt;*

[2] *Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, jpcosta@fc.up.pt;*

[3] *Centro de Educação Médica e Departamento de Epidemiologia Clinica, Medicina Preditiva e Saúde Pública da Faculdade de Medicina da Universidade do Porto & ISPUP-Instituto de Saúde Pública da Universidade do Porto, milton@med.up.pt*

**Sumário**

Neste trabalho são apresentadas 3 propostas para definição de critérios para o concurso de especialidade médica, integrando a classificação média final do curso de Medicina (CMCM) e a classificação obtida na prova nacional de seriação. Os requisitos de equiparação e comparação entre as CMCM bem como de equivalência entre coortes de diferentes faculdades são avaliados e é efectuada uma análise crítica dos 3 critérios sugeridos.

**Palavras-chave:** Concurso de especialidade médica, Equiparação de classificações, scores T, scores Z.

## 1. Introdução

Neste trabalho são apresentadas e discutidas 3 propostas para definição de critérios para o concurso de especialidade médica, integrando a **c**lassificação **m**édia final do **c**urso de **M**edicina (CMCM) e a **c**lassificação obtida na **p**rova **n**acional de **s**eriação (CPNS). A CMCM varia entre 10 e 20 valores enquanto que a CPNS varia entre 0 e 100 pontos.

A integração da CMCM requer que os valores obtidos das várias Faculdades de Medicina portuguesas sejam comparáveis e equiparáveis. Para essa análise, precisamos de garantir que as classificações medem o mesmo conceito (*construto*), têm a mesma qualidade (*fiabilidade*), têm uma distribuição semelhante, e que a escolha da subpopulação (coorte de estudantes) não altera a equiparação. Será ainda necessário avaliar a equivalência entre coortes de estudantes de diferentes faculdades [HOLLAND e DORANS, 2007; LIVINGSTON, 2004].

## 2. Avaliação dos requisitos

Mostramos que a associação entre os z-scores da CMCM e da CPNS é independente da Faculdade de origem dos licenciados e que portanto as CMCM medem o mesmo constructo. Além disso, a variância total da CPNS explicada pela CMCM é de 42% pelo que ambas as classificações acrescentam informação relevante à classificação final.

A questão da igualdade das distribuições das CMCM das várias Faculdades é rejeitada, uma vez que existem diferenças significativas entre as suas médias e variâncias. Mais ainda, cerca de ¼ da variação total das CMCM pode ser atribuída à Faculdade frequentada. Verifica-se também que não existem diferenças significativas entre as médias das CMCM das duas coortes de estudantes analisadas, correspondentes aos anos de 2011 e 2012, pelo que a escolha da subpopulação não parece ter influência sobre a equiparação a efectuar. Podemos ainda afirmar que estamos na presença de coortes de estudantes (de diferentes Faculdades) que são equivalentes, na medida em que as diferenças existentes entre Faculdades explicam menos de 4% da variação total dos scores Z da CPNS.

## 3. Inclusão de candidatos estrangeiros

Não foi identificada qualquer associação entre os scores Z da CMCM e os scores Z da CPNS em candidatos estrangeiros, ao contrário do que se passa com os candidatos portugueses. Este facto sugere que o processo de candidatura seja diferenciado pela nacionalidade dos candidatos.

## 4. Propostas

### 4.1. Proposta 1

Esta proposta assume que as coortes de estudantes das diferentes Faculdades são equivalentes e que as CMCM são comparáveis. Sugere-se que a classificação final (CF) do candidato *i* da Faculdade *j* seja dada por

$$CF_{ij} = Y_{ij} * 0.25 + CPNS_{ij} * 0.75$$

onde $Y_{ij} = \left(X_{ij} - 10\right) * 10$ e $X_{ij}$ é a CMCM do candidato $i$ da Faculdade $j$. Pelo facto de as CMCM não serem comparáveis, esta proposta trará vantagens injustas a candidatos de determinadas Faculdades.

### 4.2. Proposta 2

Esta proposta assume que as coortes de estudantes das diferentes Faculdades são equivalentes e que as CMCM não são comparáveis. Sugere-se que a classificação final (CF) do candidato *i* da Faculdade *j* seja dada por

$$CF_{ij} = T_{ij} * 0.25 + CPNS_{ij} * 0.75$$

onde $T_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} * 10 + 50$, com $\bar{x}_j$ e $s_j$ correspondentes à média e desvio padrão das CMCM na coorte de estudantes da Faculdade $j$ que se estão a candidatar pela primeira vez, respectivamente.

A utilização do score T permite maximizar o efeito da CMCM na CF. Para um candidato de uma Faculdade estrangeira, a CF consistirá apenas da CPNS.

### 4.3. Proposta 3

Esta proposta assume que as coortes de estudantes das diferentes Faculdades não são equivalentes e que as CMCM não são comparáveis. Sugere-se que a classificação final (CF) do candidato *i* da Faculdade *j* seja dada por

$$CF_{ij} = T_{ij} * 0.25 + CPNS_{ij} * 0.75$$

onde $T_{ij} = \left(\frac{x_{ij}-\bar{x}_j}{s_j} + \frac{\bar{x}_{PNS,j}-\bar{x}_{PNS}}{s_{PNS}}\right) * 10 + 50$, com $\bar{x}_{PNS,j}$ correspondente à média das CPNS obtidas pelos candidatos da Faculdade $j$ e $s_{PNS}$ correspondente ao desvio padrão das CPNS entre as várias Faculdades. O termo $(\bar{x}_{PNS,j} - \bar{x}_{PNS})/s_{PNS}$ funciona como um indicador do desempenho dos alunos da Faculdade *j* relativamente à PNS.

Para um candidato de uma Faculdade estrangeira, a CF poderia consistir novamente apenas da CPNS. Esta proposta tem a desvantagem de depender da qualidade da PNS.

## 5. Conclusão

A curto prazo, e a partir dos dados disponíveis, a Proposta 2 parece-nos ser a mais indicada. De futuro, poder-se-á vir a adoptar a Proposta 3 caso se verifique entretanto: que existem diferenças significativas entre as CPNS das várias Faculdades; que a qualidade da PNS e a fiabilidade das CMCM sejam elevadas.

**Agradecimentos:** Os autores foram parcialmente financiados pelo Fundo Europeu de Desenvolvimento Regional através do programa COMPETE e pelo Governo Português através da FCT-Fundação para a Ciência e a Tecnologia através do projecto PEst-C/MAT/UI0144/2013.

## Referências

HOLLAND, P. W., DORANS, N. J. (2007) Linking and Equating Test Scores. IN BRENNAN, R.L. (Ed.), *Educational Measurement*,4, ed.,187–220. Westport, CT: Praeger Publishers.

LIVINGSTON, S. A. (2004). *Equating Test Scores (Without IRT)*. Educational Testing Service.

**Análise de Dados em Medicina –5ª Feira, 10 de Abril, Sala 316 (16h50)**

# M-regression, false discovery rates and outlier detection in genetic association studies

**Vanda M. Lourenço[1], Ana M. Pires[2]**
[1]*FCT/UNL and CMA-FCT/UNL, vmml@fct.unl.pt;*
[2] *IST/UTL and CEMAT-IST/UTL, apires@math.ist.utl.pt*

**Abstract**

Outlier detection is widely discussed in many areas of research one of which is the field of genetic association studies. Here, robust multiple linear regression methods have been shown to be a valuable asset, allowing us not to be concerned with the possibility of outliers disrupting the analysis results. Additionally these methods also provide an adequate setting to search for outliers. Therefore, we propose and discuss a robust outlier test to be used in the context of multiple M-regression illustrating the good performance of the test through a real data example application.

**Keywords:** Robust outlier test, Robust regression, Multiple testing, Single nucleotide polymorphism.

## 1. Introduction

Outlier detection is a widely discussed issue in many areas of research one of which is the field of genetic association studies where the main motivation is usually to clean the data so that classical models may be used in the analysis. Knowledge of these observations however, may be important in the assessment of the underlying mechanisms of that data, since outliers are not always a result of measurement errors. Robust multiple linear regression methods have been shown to be a valuable asset in genetic association studies (Lourenço *et al.*, 2011), allowing us not to be concerned with the eventuality of outliers in the data disrupting our analysis results. These methods also provide an adequate setting to search for outliers since the residuals obtained from robust fits are not usually affected by outlying observations. To this respect, we present a robust outlier test together with an appropriate robust estimate of scale and an adequate false discovery rate (FDR) correction measure, to be used in the context of multiple M-regression. We illustrate the good performance of the robust outlier test through an application that uses a real genetic data set from the literature (Weber *et al.*, 2008). Full results can be found in Lourenço and Pires (2014).

## 2. Methods

### 2.1. M-regression

We consider the general linear regression model

$$Y = X\beta + \varepsilon \qquad \textbf{(1)}$$

under the usual assumptions of independence, homoscedasticity and normality of the errors, $\varepsilon$, with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$. In our case: $Y$ is a continuous phenotypic trait; $X$ is the design matrix for categorical genotypic markers, the first column having 1's; $\beta$ is a vector of unknown parameters; the number of individuals in the sample is greater than the number of unknown parameters. In robust estimation the unknown parameters are estimated by:

$$\hat{\beta}_R = min_\beta \sum_i^n \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right)$$

where $\rho$ is the Huber function, $n$ the sample size, $\hat{\sigma}$ some robust estimate of scale and $r_i$ the residuals from the robust fit of model **(1)**. In this case the median absolute deviation (MAD) is considered. It can easily be seen that taking $\rho(x) = x^2$ allocates us to the least squares/maximum likelihood estimation setting.

### 2.2. Robust outlier test

To test the hypothesis

$$H_0: observation\ i\ is\ not\ an\ outlier,$$

the robust outlier test proceeds as follows:

1- take the standardized robust residuals $r_i^R = \frac{r_i}{\hat{\sigma}}$;
2- calculate $p_i = 2 \times \Phi\left(-\left|r_i^R\right|\right)$ for each observation $i$, where $\Phi$ is the normal p.d.f.;
3- adjust for multiple testing with an adequate FDR multiple testing correction procedure;
4- reject the null hypothesis at a given FDR threshold, e.g, 10%.

## 3. Application

We consider the maize data of Weber *et al.* (2008) on 493 plants, 1 quantitative trait (FERL – female hear length) and 61 candidate genetic markers (SNPs – single nucleotide polymorphisms) that was re-analyzed by Lourenço *et al.* (2011) via M-regression and Wald-type association tests. We perform the robust outlier test considering the residuals from the robust fit of the final adjusted multiple-SNP model. Results are then compared with the ones previously obtained where some rule of thumb was used to perform outlier detection. Finally,

some considerations are made to the adequacy of the proposed robust outlier test when the identified outliers are inspected further.

**Acknowledgements:** This work received financial support from Portuguese National Funds through FCT (Fundação para a Ciência e Tecnologia) under the scope of project PTDC/MAT-STA/0568/2012.

## References

LOURENÇO, V.M., PIRES, A.M. & KIRST, M. (2011). Robust linear regression methods in association studies. *Bioinformatics,* 27(6), 815-821.

LOURENÇO, V.M. & PIRES, A.M. (2014). M-Regression, false discovery rates and outlier detection with application to genetic association studies. *Submitted*.

WEBER, A.L. *et al.* (2008). The genetic architecture of complex traits in teosinte (Zea mays ssp. parviglumis): new evidence from association mapping. *Genetics,* 180, 1221-1232.

**Modelos Longitudinais – 6ª Feira, 11 de Abril, Salão Nobre (9h00)**

# Estimating bivariate integer-valued moving average models with the generalized method of moments

**Isabel Silva[1], Cristina Torres[2], Maria Eduarda Silva[3]**

[1]*Faculdade de Engenharia da Universidade do Porto, ims@fe.up.pt;*

[2] *ISCAP e Universidade do Porto, cmptorres@gmail.com;*

[3] *Faculdade de Economia da Universidade do Porto e CIDMA, mesilva@fep.up.pt*

**Abstract**

In this paper an estimation method for the parameters of Bivariate INteger-valued Moving Average (BINMA) model, based on Generalized Method of Moments (GMM) is proposed. The performance of the GMM estimator and its small sample properties will be investigated in a simulation study.

**Keywords:** BINMA models, Count time series, Generalized Method of Moments, Parameter estimation.

## 1. Introduction

Time series of counts arise when the interest lies on the number of certain events occurring during a specified time interval. In many situations the collected time series are multivariate in the sense that there are counts of several events observed over time and the counts at each time point are correlated. Several approaches and diversified models that explicitly account for the discreteness of the data have been considered, among which are the INteger-valued AutoRegressive Moving Average, INARMA, models (McKenzie, 2003). These models are constructed by replacing the multiplication in the conventional ARMA models by an appropriate random operator that preserves the discreteness of the counting process. The most popular of such operator is the **binomial thinning operator** (Steutal and Van Harn, 1979) defined as $\alpha \circ Y = \sum_{j=1}^{Y} B_j$, where $Y$ is a non-negative integer-valued random variable, $\alpha \in [0,1]$, and $B_j$, designated by counting series, is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables, independent of $Y$, such that $\mathrm{P}(B_j = 1) = 1 - \mathrm{P}(B_j = 0) = \alpha$.

Moving average (MA) models are widely used in econometric data. For example, inventories are often considered to be well described by MA processes. In fact, the periodic resetting of inventories to optimal levels implies an upper limit on the number of periods in which shocks can affect inventory levels, leading to autocorrelations functions characteristic of this type of processes. The **INteger-valued Moving Average** (INMA) process was introduced by McKenzie (1986) and Al-Osh and Alzaid (1988). A **Bivariate INMA** model that assumes independence between and within the thinning operations and that allows for both positive and negative correlation between the counts was proposed by Quoreshi (2006).

In this work, a BINMA model based on the same dependence structure of Al-Osh and Alzaid (1988) between thinning operations in the same equation is presented (Torres *et al.*, 2012), by generalizing the model proposed by Quoreshi (2006). In section 2, the model is described and emphasis is placed on models with Bivariate Poisson and Bivariate Negative Binomial innovations. The Method of Moments (MM) and the Generalized MM (GMM) are proposed in Section 3 in order to estimate the unknown parameters of these models. The methods will be compared and their performances will be investigated in a simulation study.

## 2. Bivariate integer-valued moving average models – BINMA($q_1$, $q_2$)

Let $\boldsymbol{X}_t = \left[X_{1,t} \quad X_{2,t}\right]'_{t\in Z}$ be a non-negative integer-valued random vector. Then $\boldsymbol{X}_t$ can be defined as a BINMA($q_1$, $q_2$) model, proposed by Torres *et al.* (2012), if satisfies the following equations

$$X_{1,t} = \varepsilon_{1,t} + \beta_{1,1} \circ \varepsilon_{1,t-1} + \cdots + \beta_{1,q_1} \circ \varepsilon_{1,t-q_1}$$
$$X_{2,t} = \varepsilon_{2,t} + \beta_{2,1} \circ \varepsilon_{2,t-1} + \cdots + \beta_{2,q_2} \circ \varepsilon_{2,t-q_2}$$

where $\circ$ denotes the binomial thinning operator, $\beta_{j,i} \in ]\,0,\,1\,[$ for $j=1,\ 2;\ i=1,\ldots,\ q_j$, and $\{\varepsilon_{j,t}\}$, $j=1,2;\ t\in \mathrm{Z}$, are i.i.d. sequences of non-negative integer-valued random variables, designated as innovations. Dependence between the two series that comprise the BINMA($q_1$, $q_2$) models is introduced by allowing for dependence between $\{\varepsilon_{1,t}\}$ and $\{\varepsilon_{2,t}\}$. It is assumed the same dependence structure proposed by Al-Osh and Alzaid (1988) between thinning operations in the same equation. In this way, each element $\{\varepsilon_{j,t}\}$ can be "*active*" in the system $q_j+1$ time units, the $\beta_{j,i}$ are the probabilities that an element of $\{\varepsilon_{j,t}\}$ will be "*active*" in the system at time $t+i$, independent of the other elements of the system, and then $\beta_{j,i} \circ \varepsilon_{j,t-i}$ represents the number of elements of generation $t-i$ which are "*active*" in the system at time *t*.

### 2.1. Poisson BINMA($q_1$, $q_2$) model

In the Poisson BINMA model the innovations of the two series follow jointly a bivariate Poisson distribution, denoted by $BP(\lambda_1, \lambda_2, \phi)$, for details see Kocherlakota and Kocherlakota (1992). Marginally each random variable follows a Poisson distribution with parameters $\lambda_1+\phi$ and $\lambda_2+\phi$, respectively. The parameter $\phi$ is the covariance between the two random variables.

### 2.2. Negative binomial BINMA($q_1$, $q_2$) model

In the negative binomial BINMA model the innovations of the two series follow jointly a bivariate negative binomial distribution, denoted as $BNB(\lambda_1, \lambda_2, \tau)$, for details see Cheon *et al.* (2009). Marginally each random variable follows a negative binomial distribution with mean $\lambda_1$

and $\lambda_2$ and variance $\lambda_1(1+\lambda_1\tau)$ and $\lambda_2(1+\lambda_2\tau)$, respectively. The covariance between the two random variables is $\lambda_1\lambda_2\tau$.

## 3. Parameter estimation

### 3.1. The method of moments

The MM estimates the population parameters by matching population (or theoretical) moments with corresponding sample moments. Suppose $\{X_{j,t}, j=1,2; t=1,\ldots,T\}$ is an observed sample from a BINMA($q_1$, $q_2$) model with true $q_j \times 1$ parameter vector $\boldsymbol{\theta}_0$ and let $\boldsymbol{\theta}$ be the parameter vector estimator. The **moment conditions** are defined by $\mathrm{E}\left[m\left(X_{j,t},\boldsymbol{\theta}_o\right)\right]=0$, where $m\left(X_{j,t},\boldsymbol{\theta}\right)$ is a continuous $p \times 1$ vector function of $\boldsymbol{\theta}$ and the expected value exists and is finite for all $t$ and $\boldsymbol{\theta}$ (Mátyás, 1999).

The estimator is obtained by solving the system of equations given by $\mathrm{E}\left[m\left(X_{j,t},\boldsymbol{\theta}\right)\right]=0$, with $p$ equations for $p=q_j$ unknowns. However, $\mathrm{E}\left[m\left(X_{j,t},\boldsymbol{\theta}\right)\right]$ is not observed and the analogous sample moment conditions defined by

$$m_T(\boldsymbol{\theta})=\frac{1}{T}\sum_{t=1}^{T} m\left(X_{j,t},\boldsymbol{\theta}\right)$$

is considered.

### 3.2. The generalized method of moments

The GMM estimation method was first introduced by Hansen (1982), into the econometrics literature and since then it has been widely applied to analyze economic and financial data. The GMM estimation is also based on population moment conditions, under the same conditions defined for the MM estimator. In this estimation method, the moment conditions are a set of $p$ equations with $q_j < p$ parameters and the estimators are obtained by minimizing the quadratic form

$$Q_T(\boldsymbol{\theta}) = m_T(\boldsymbol{\theta})'\, W_T\, m_T(\boldsymbol{\theta}),$$

where $W_T = \left(\mathrm{cov}\left(m_T(\boldsymbol{\theta})\right)\right)^{-1}$ is a positive definite weighting matrix. With any consistent estimator of $W_T$, the GMM estimator $\hat{\boldsymbol{\theta}}_W$ is consistent and

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_0\right) \rightarrow N\left(\mathbf{0}, \boldsymbol{\Sigma}_W\right),$$

where $\boldsymbol{\Sigma}_W = \left(\boldsymbol{D}'\,\boldsymbol{\Omega}^{-1}\boldsymbol{D}\right)^{-1}$, for $\boldsymbol{D} = \partial m_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ and $\boldsymbol{\Omega} = \lim_{T\rightarrow\infty} T W_T$.

### 3.3. Considerations on the simulation study

The aim of the simulation study will be to illustrate the small sample properties of the two estimators previously described and compare their behaviour, regarding bias and mean squared error. Furthermore, this study will address two issues on GMM estimation which are the choice of moment conditions and the estimation of the covariance matrix.

**Acknowledgments:** For the third author, this work was supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-OE/MAT/UI4106/2014.

## References

AL-OSH, M. A. & ALZAID, A. A. (1988). Integer-valued moving average (INMA) process. *Statistical Papers.* 29, 281-300.

CHEON, S., SONG, S. H. & JUNG, B. C. (2009). Tests for independence in a bivariate negative binomial model. *Journal of the Korean Statistical Society.* 38, 185-190.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica.* 50, 1029-1054.

KOCHERLAKOTA, S. & KOCHERLAKOTA, K. (1992). *Bivariate discrete distributions*. Markel Dekker, New York.

MÁTYÁS, L. [ed.] (1999). *Generalized method of moments estimation*. Cambridge University Press.

McKENZIE, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability.* 18, 679-705.

McKENZIE, E. (2003). Discrete variate time series. *Handbook of Statistics,* 21, 573-606.

QUORESHI, A.M.M.S. (2006). Bivariate time series modeling of financial count data. *Communications in Statistics-Theory and Methods.* 35, 1343-1358.

STEUTEL, F.W. & VAN HARN, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7, 893-899.

TORRES, C., SILVA, I. & SILVA, M. E. (2012). Modelos bivariados de médias móveis de valor inteiro. XX Congresso Anual da Sociedade Portuguesa de Estatística, 27-29 de Setembro, Porto, Portugal.

**Modelos Longitudinais – 6ª Feira, 11 de Abril, Salão Nobre (9h20)**

# Modelação e previsão da procura turística doméstica em Portugal numa conjuntura de crise económica e financeira

**Luís Nobre Pereira[1], Lara Noronha Ferreira[2]**

*[1]Escola Superior de Gestão, Hotelaria e Turismo & Centro de Investigação sobre o Espaço e as Organizações (CIEO)- Universidade do Algarve, lmper@ualg.pt*

*[2]Escola Superior de Gestão, Hotelaria e Turismo - Universidade do Algarve & Centro de Estudos e Investigação em Saúde da Universidade de Coimbra (CEISUC), lnferrei@ualg.pt*

**Sumário**

O objectivo deste estudo consiste em modelar e produzir previsões para a procura turística doméstica em Portugal. Foram usados dados trimestrais, desde 2009 a 2012, para um conjunto de variáveis económicas, como por exemplo o PIB *per capita*, o rendimento médio disponível das famílias e o indicador de confiança dos consumidores. Foram estimadas várias especificações de modelos econométricos considerando como variável dependente o número de dormidas de turistas residentes. As previsões da procura turística foram efectuadas com modelos de alisamento exponencial tradicionais.

**Palavras-chave:** Dados em painel, Elasticidades, Modelos econométricos, Previsões, Procura turística doméstica.

## 1. Introdução

Tendo em conta que, em 2012, quase um terço das dormidas em estabelecimentos hoteleiros em Portugal foi devida à procura doméstica, que mais de quatro quintos das dormidas dos turistas residentes ocorreram em Portugal (INE, 2013; WTTC, 2013), e que não existem estudos recentes sobre a procura turística doméstica em Portugal, em particular numa conjuntura de crise económica e financeira, considera-se relevante estudar quais são os factores que determinam essa procura turística. Assim, o objectivo deste estudo consiste em modelar e produzir previsões para a procura turística doméstica em Portugal.

## 2. Metodologia

### 2.1. Dados

Uma vez que só existem dados trimestrais da procura turística doméstica Portuguesa desde 2009, então foram usados dados desde o primeiro trimestre de 2009 até ao quarto trimestre de 2012.

De acordo com a literatura (Lim, 1997; Song *et al*, 2010), verifica-se que os determinantes da procura turística podem ser de natureza económica e de natureza não-económica. Apesar de existirem autores que defendem que existem factores não-económicos

que influenciam as escolhas dos turistas, e como consequência a procura turística (e.g. Eilat e Einav, 2004; Zhang e Jensen, 2007), verifica-se que a maioria dos estudos sobre a procura turística doméstica se limita a estudar os impactos de variáveis económicas. Assim, neste estudo decidiu delimitar-se os possíveis determinantes da procura turística doméstica a um conjunto de variáveis económicas, tendo em conta não só o contexto económico em que o país se encontra e o período em análise, mas também a informação disponível e com qualidade. Assim, tendo em conta a literatura e os dados disponíveis para Portugal, foi considerado o conjunto de variáveis explicativas apresentadas na tabela 1.

**Tabela 1: Variáveis explicativas usadas na modelação da procura turística doméstica em Portugal**

| Variável | Definição | Fonte |
|---|---|---|
| PIB | PIB *per capita* a preços constantes de 2006 | INE |
| RDF | Rendimento médio disponível das famílias | INE |
| ICC | Indicador de Confiança dos Consumidores | INE |
| TXD | Taxa de desemprego | INE |
| DES | Número de desempregados (em *stock*) | INE |
| TXA | Taxa de actividade | INE |
| IPT | IPC relativo à secção dos Transportes | INE |
| IPH | IPC relativo à secção de Restaurantes e Hotéis | INE |
| PET | Preço do barril de petróleo | BdP |
| TRK | Variável *dummy* que indica a presença da *Troika* em Portugal | --- |

Nota: INE – Instituto Nacional de Estatística; BdP – Banco de Portugal; IPC – Índice de Preços no Consumidor.

Tendo em conta os dados em painel disponíveis (n=7, T=16), considera-se que se trata de um pequeno painel e considerou-se tal facto ao nível da estimação econométrica (Cameron e Trivedi, 2010).

## 2.2. Modelos

Neste estudo é proposta a seguinte função da procura turística doméstica:

$$PT_{i,t} = f(PIB_t, RDF_t, ICC_t, TXD_t, DES_t, TXA_t, PET_t, IPT_t, IPH_t, TRK_t), \qquad (1)$$

onde $t$ é o período de tempo ($t$=1, …, 16); $i$ é a região NUTSII ($i$=Norte, Centro, Lisboa, Alentejo, Algarve, RA dos Açores, RA da Madeira) e $PT_{it}$ é o número de dormidas de residentes no período $t$, nos estabelecimentos hoteleiros da região $i$.

Tendo em consideração as especificidades do problema que está a ser investigado, então propõe-se que seja usado um modelo combinado, também conhecido como um modelo da média da população. Assim, o modelo proposto é o seguinte:

$$pt_{i,t} = \alpha + \mathbf{x'}_{it}\boldsymbol{\beta} + \mathbf{Z'}_{t}\boldsymbol{\gamma} + u_{it}, \qquad (2)$$

onde $\alpha$ é a constante do modelo, $\mathbf{x'}_{it}$ é um vector com as variáveis explicativas, $\mathbf{Z'}_{t}$ é um vector com variáveis *dummy*, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são vectores de parâmetros e $u_{it}$ é o termo de erro aleatório. O vector $\mathbf{Z'}_{t} = (T2, T3, T4)$ inclui três variáveis *dummy* sazonais que indicam o respectivo trimestre (segundo, terceiro e quarto trimestres, respectivamente), as quais foram incluídas no modelo de forma a evitarem a correlação seccional entre regiões. As variáveis logaritmizadas presentes no modelo (2) são representadas pelas respectivas letras minúsculas (genericamente para uma variável *X*, tem-se $x = \ln(X)$). Uma vez que a estimação do modelo (2) é mais eficiente pelo método dos mínimos quadrados generalizados (*pooled GLS*) do que pelo método dos mínimos quadrados ordinários (*pooled OLS*) quando existe algum tipo de correlação no termo de erro (Cameron e Trivedi, 2010, p. 254), então foram aplicados os testes de diagnóstico à heterocedasticidade e à correlação seccional dos resíduos (Baum, 2001).

A produção de previsões para a procura turística serão efectuadas com recurso a modelos de alisamento exponencial tradicionais. Estes modelos, desenvolvidos a partir dos trabalhos pioneiros de Holt (1957) e Winters (1960), são baseados em médias ponderadas de observações passadas, com pesos exponencialmente decrescentes para zero para observações mais antigas, e sendo o alisamento tanto mais acentuado quanto menor for o peso da observação relativa ao último período conhecido. Apesar de existirem modelos de séries temporais muito sofisticados que têm vindo a ser aplicados na produção de previsões para a procura turística internacional (e.g., Song e Li, 2008), não existe evidência clara que esses modelos produzam também previsões com melhor qualidade para a procura turística doméstica.

## 3. Conclusão

Os resultados obtidos indicam que em Portugal, numa conjuntura de crise económica e financeira, a procura turística doméstica é explicada pelo rendimento médio disponível das famílias, pelo número de desempregados e pelo índice de preços no consumidor relativo aos transportes. As estimativas obtidas indicam que a elasticidade do rendimento disponível na procura é aproximadamente +2,1%, do número de desempregados é cerca de -0,6% e do índice de preços dos transportes é aproximadamente +1,2%.

O modelo de Holt-Winters aditivo é o modelo que produz previsões com melhor qualidade para a procura turística doméstica nas regiões do Norte, Centro e Lisboa, enquanto o modelo de Holt-Winters multiplicativo é o modelo que produz previsões com melhor qualidade nas restantes regiões, bem como para a procura turística total de Portugal. Em termos globais, verifica-se que a qualidade das previsões nestes casos é boa, pois se for analisada a medida de qualidade mais popular - o Erro Percentual Absoluto Médio (EPAM), verifica-se que ela é sempre inferior a 6%. Em particular, o EPAM é inferior a 3% nas previsões da procura turística doméstica no Norte, Centro e em Portugal de forma agregada, e é inferior a 4% nas previsões relativas às regiões do Alentejo, RA dos Açores e Lisboa.

**Agradecimentos:** O CIEO e o CEISUC são financiados pela Fundação para a Ciência e a Tecnologia do Ministério da Educação e Ciência.

## Referências

CAMERON, A.C. & TRIVEDI, P.K. (2010). *Microeconometrics using Stata*. Stata Press.

EILAT, Y. & EINAV, L. (2004). Determinants of international tourism: a three-dimensional panel data analysis. *Applied Economics*, 36(12), 1315-1327.

INE (2013) *Estatísticas do Turismo 2012*. Lisboa: INE. Acedido em 30/09/2013 em www.ine.pt

HOLT, C.C. (1957). Forecasting seasonals and trends by exponentially weighted averages. O.N.R. Memorandum 52/1957, Carnegie Institute of Technology. Reimpresso com discussão em 2004 - *International Journal of Forecasting*, 20, 5-13.

LIM, C. (1997). Review of international tourism demand models. *Annals of Tourism Research*, 24, 835-849.

SONG, H. & LI, G. (2008). Tourism demand modeling and forecasting – A review of recent research. *Tourism Management*, 29(2), 203-220.

SONG, H., LI, G., WITT, S.F. & FEI, B. (2010). Tourism demand modeling and forecasting: how should demand be measured? *Tourism Economics*, 16(1), 63-81.

WINTERS, P.R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342.

WTTC (2013). *Travel & Tourism – Economic Impact 2013 Portugal*. London: World Travel & Tourism Council.

ZHANG, J. E JENSEN, C. (2007). Comparative advantage: Explaining Tourism Flows. *Annals of Tourism Research*, 34(1), 223-243.

**Modelos Longitudinais – 6ª Feira, 11 de Abril, Salão Nobre (9h40)**

# Modelo com trajetória latente com dados gerados a partir de um *planned missing design*: estudo de simulação

**Paula C.R. Vicente[1], Maria de Fátima Salgueiro[2]**

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

**Sumário**

Este estudo de simulação tem por objetivo analisar o efeito da existência de diferentes padrões de omissão (resultantes de um *planned missing design*) com diferentes percentagens de não resposta nos valores estimados dos parâmetros de um modelo com trajetória latente (*latent growth curve model*). São ainda considerados o efeito da dimensão da amostra e do número de momentos temporais em estudo no enviesamento dos parâmetros estimados pelo modelo de interesse.

A variação do parâmetro média do declive, mantendo constantes os restantes parâmetros do modelo a partir do qual os dados são gerados, não influencia a média das estimativas dos parâmetros para cada um dos modelos considerados, exceto para a média do declive, qualquer que seja a dimensão da amostra. Por outro lado, à medida que a variância do declive aumenta, o MSE aumenta.

**Palavras-chave:** *Latent growth curve model*, *Planned missing design*, Simulação de Monte Carlo.

## 1. Introdução

Nas últimas décadas tem aumentado o interesse pela recolha e análise de dados longitudinais, isto é, dados que descrevem a evolução dos acontecimentos durante um determinado período de tempo. Todavia, apesar das inúmeras vantagens que os dados longitudinais apresentam, a existência de não respostas resultantes do abandono do painel constitui uma problemática bastante frequente neste tipo de dados. Num inquérito longitudinal é usual fazer a distinção entre dados omissos intermitentes e *dropout*, mas as omissões podem também ser consequência do desenho do estudo. Num *planned missing design* há uma estrutura de omissões de dados que ocorre de forma intencional e de acordo com o planeado pelo investigador, sendo o objetivo deste planeamento minimizar o esforço de inquirição por parte dos respondentes e consequente abandono do painel (ENDERS, 2010). Assim, um inquérito que resulta de um painel rotativo pode ser considerado um exemplo de um *planned missing design*.

De acordo com RUBIN (1976) existem três diferentes mecanismos de omissão: 1) missing completely at random (MCAR), se a probabilidade de omissão de uma variável não está relacionada nem com outras variáveis, nem com o valor possível dessa mesma variável; 2) missing at random (MAR), quando a probabilidade de omissão de uma variável está

relacionada com uma outra variável presente na análise; e 3) missing not at random (MNAR), se a probabilidade de omissão de uma variável depende do valor que essa mesma variável assumiria.

Quando as omissões existentes num *planned mising design* podem ser consideradas como tendo um mecanismo de omissão MAR, a estimação por *full information maximum likelihood* é uma das abordagens estatísticas mais utilizadas para lidar com dados omissos (SCHAFFER & GRAHAM, 2002).

Neste trabalho pretendem apresentar-se os principais resultados de um estudo de simulação de Monte Carlo realizado com o objetivo de analisar o efeito da dimensão da amostra, de diferentes padrões de omissão (resultantes de um *planned missing design*), de diferentes percentagens de omissões e do número de momentos temporais considerados, nas estimativas dos parâmetros de um *Latent Growth Curve Model* (LGCM).

## 2. Estudo de simulação

A modelação de trajetórias longitudinais de mudança dos indivíduos é frequentemente objeto de estudo por parte de distintas áreas do conhecimento, tais como, ciências sociais e psicologia. Este procedimento requer grande parte das vezes uma base de dados longitudinais com uma dimensão considerável. A análise estatística das trajetórias de crescimento pode ser realizada usando técnicas de modelação como os LGCM, que permitem capturar informação sobre diferenças *interindividuais* nas mudanças *intraindivíduos* ao longo do tempo (NESSELROADE, 1991; BOLLEN e CURRAN, 2006).

O *software* estatístico Mplus 6 (MUTHÉN & MUTHÉN, 2010) permite de uma forma integrada, gerar *r* amostras de dados a partir da estrutura de um LGCM, cujos parâmetros populacionais são definidos a priori pelo investigador, com um determinado número de momentos temporais. Posteriormente, para cada uma das *r* amostras geradas, é estimado um LGCM, obtendo-se, deste modo, *r* estimativas para cada um dos parâmetros do modelo. Se nas amostras geradas existem omissões uma abordagem *full information maximum likelihood* é utilizada na estimação. Para cada um dos parâmetros do modelo o Mplus disponibiliza a média das estimativas, calculada a partir das *r* amostras independentes que foram geradas, e o desvio-padrão das estimativas, calculado para o conjunto das amostras geradas. Quando o número de amostras *r* é elevado este desvio-padrão pode ser considerado como o erro padrão do parâmetro populacional. São ainda disponibilizados o erro padrão médio de estimação, calculado para cada uma das *r* amostras, a *coverage*, que indica a proporção de amostras para as quais um intervalo a 95% contém o verdadeiro parâmetro, e o Mean Square Error (MSE) uma medida simultânea de *bias* e variância.

Num LGCM os parâmetros de interesse são as médias e as variâncias do intercepto e do declive aleatórios, bem como a covariância entre o intercepto e o declive. Para efeitos deste estudo de simulação a variância dos termos residuais assume-se como igual nos diferentes

momentos temporais. Neste estudo de simulação são apenas considerados dados com distribuição normal.

## 3. Resultados

Neste estudo de simulação foi considerada a geração de 1000 amostras de dados com distribuição normal, a partir de um LGCM com trajectória linear e um indicador (Y) medido em três, quatro ou cinco momentos temporais (a figura 1 representa um LGCM com quatro momentos temporais).

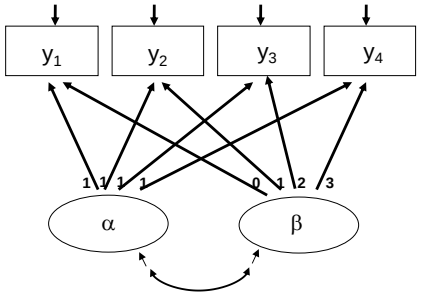



**Figura 1: Diagrama de um LGCM com quatro momentos temporais**

Foram geradas amostras de diferentes dimensões: N=50, N=250, N=500 e N=1000. Os valores considerados para os parâmetros populacionais que definem o modelo são a média do intercepto ($\mu_\alpha$=0) e a variância do intercepto ($\psi_{\alpha\alpha}$=1). Foi permitido variar os valores para a média e a variância do declive, respetivamente $\mu_\beta$=-4, -2, -1, 0, 1, 2, 4 e $\psi_{\beta\beta}$=0.2, 1, 2, 4, 16. Foram considerados os valores 0 e 0.5 para a covariância entre intercepto e declive ($\psi_{\alpha\beta}$) e a variância dos termos residuais foi fixada a valores que permitem obter uma fiabilidade de 0.5 para os indicadores, em cada um dos momentos temporais.

Para cada um dos modelos considerados foi calculado o *bias* na estimativa média dos parâmetros. Posteriormente, foram geradas amostras com padrões de omissão (que configuram um *planned missing design*) com percentagens de 10%, 20%, 30%, 40% e 50% de não respostas em cada momento temporal e avaliado o *bias* nas estimativas dos parâmetros de um LGCM, para amostras de diferentes dimensões.

Os resultados obtidos no estudo de simulação indiciam que fazer variar a média do declive, mantendo constantes os restantes parâmetros do modelo, não influencia a média das estimativas dos parâmetros para cada um dos modelos considerados, exceto para a média do declive, qualquer que seja a dimensão da amostra. Por outro lado, à medida que a variância do declive aumenta, o MSE aumenta.

**Referências**

BOLLEN, K.A. & Curran, P.J. (2006). *Latent Curve Models – A Structural Equation Perspective*, 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. (2010). *Mplus user´s guide*, Los Angeles, USA, Muthén & Muthén.

NESSELROAD, J.R. (1991). Interindividual differences in intraindividual change. IN COLLINS, L.M. & HORN, J.L.(Ed.) *Best Methods for the Analysis of Change.* New York, Am. Psychol. Association.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika,* 63, 581-592.

SCHAFFER, J.L. & GRAHAM, J.W. (2002). Missing Data: Our view of the state of the art. *Psychological Methods*, 7**,** 147-177.

**Modelos Longitudinais – 6ª Feira, 11 de Abril, Salão Nobre (10h00)**

# The effect of observed data deviations from normality on the parameter estimates of a latent growth curve model: a simulation study

**Maria de Fátima Salgueiro[1], Paula C.R. Vicente[2]**

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

[2]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Portugal, pvicente@netcabo.pt;*

**Abstract**

This simulation study aims at investigating the effect of deviations from normality of the data on the estimates of the parameters of a latent growth curve model. Varying number of time points and sample sizes are considered. The effects of various levels of data skewness and kurtosis on the bias, the mean square error and the coverage of the model parameter estimates are assessed. Preliminary results show that for models with a small mean intercept, skewness levels are low and kurtosis levels increase from low to moderate with the increase of the variability of the intercept and the magnitude of the slope. Consequently, the bias and the mean square error of the parameter estimates increase and coverage decreases, in particular for smaller sample sizes. Similar, but more severe, effects are obtained for the models with high mean intercept values.

**Keywords:** Deviations from normality, Latent growth curve model, Monte Carlo simulation study, Mplus.

## 1. Introduction

In recent years there has been a substantial increase in longitudinal data collection. Accompanying this growing availability of longitudinal data, there has been a considerable increase in the interest of developing statistical methods to analyze such data. Additionally, the implementation of the theoretical developments in statistical software has been taking place. Latent growth curve models (LGCM) are one of the very popular longitudinal statistical techniques, subject of intense interest both in terms of theoretical developments and more applied research. They are statistical models for longitudinal data allowing each individual under analysis to have distinct trajectories of change over time. These patterns of change are summarized in relatively few parameters which, in turn, are modeled as functions of other variables. For an overview see BOLLEN and CURRAN (2006).

Let $Y_{it} = \alpha_i + \lambda_t\beta_i + \varepsilon_{it}$ be the value for variable $Y$ for individual $i = 1, \dots, N$ and time $t = 1, \dots, T$. The random intercept of the LGCM is given by $\alpha_i = \mu_\alpha + \zeta_{\alpha_i}$ and the random slope by $\beta_i = \mu_\beta + \zeta_{\beta_i}$. The model assumes that $\varepsilon_i \sim N(0, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is a diagonal matrix with elements $\theta_t$. The model also assumes that $\varepsilon_{it}$, $\zeta_{\alpha_i}$ and $\zeta_{\beta_i}$ are mutually independent, and that $\zeta_{\alpha i} \sim N(0, \psi_\alpha)$; $\zeta_{\beta i} \sim N(0, \psi_\beta)$. Also, $cov(\zeta_{\alpha i}, \zeta_{\beta i}) = \psi_{\alpha\beta}$. The parameters of special

interest in this model are the means and the variances of the random effects, as well as the covariance between random effects.

SALGUEIRO (2012) has proposed using LGCM to describe and explain the growth trajectories of overall job satisfaction using longitudinal survey data. Four time points and a linear growth trajectory were considered. The proposed modeling strategy has accounted for missing data. However, the normality of the data under analysis had to be assumed, despite some evidence of skewness in the observed sample.

This talk presents the main results and conclusions from a Monte Carlo simulation study conducted in Mplus 6 in order to investigate the effect of observed data deviations from normality of the data on the estimates of the parameters of a latent growth curve model. Varying number of time points and sample sizes are considered.

## 2. Monte Carlo simulation study

The potentialities of Mplus 6 (MUTHÉN and MUTHÉN, 2010) to conduct Monte Carlo simulation studies are explored, using the mixture analysis procedure to generate non-normal data. First data are generated from a population with hypothesized parameter values: in our case the parameters of the LGCM specified by the researcher. The main parameters of interest are the means ($\mu_\alpha$ ; $\mu_\beta$) and the variances ($\psi_{\alpha\alpha}$ ;$\psi_{\beta\beta}$) of the random intercept and of the random slope, as well as the covariance between intercept and slope ($\psi_{\alpha\beta}$). In a second step, for each of the 1000 independent samples (or replications) drawn, the specified LGCM is estimated. In a third step Mplus computes and displays the average of the parameter estimates over the repeated draws of the independent samples. For each parameter, bias can be assessed by comparing the average of the 1000 obtained estimates to the hypothesized population value. Besides the standard deviation of the parameter estimates and the mean square error, the coverage (the proportion of the replications where a 95% confidence interval covers the hypothesized parameter value) is also computed and reported for each parameter (MUTHÉN and ASPAROUHOV, 2002).

In our study the hypothesized LGCM has a linear growth trajectory ($\lambda_t = t - 1$), three to six time points ($T = 3,4,5,6$) and sample sizes ($N$) ranging from 50 to 1000 observations. In Mplus non-normality of the observed data is obtained by generating data from two classes with different growth model parameter values, using the mixture analysis procedure (Muthén, 2002). The majority class includes 88% of the sample, whereas the minority class includes the remaining 12%. In this study the following values for the main parameters of interest of the majority class have been assumed: $\mu_\alpha = 0$ ; $\mu_\beta = 0$; $\psi_{\alpha\alpha} = 1$; $\psi_{\beta\beta}$ = 0.2. For the minority class different combinations of parameters have been assumed in order to induce different levels of skewness and kurtosis in the generated data. In particular the mean of the intercept has taken values 2.5 and 15. The variance of the intercept has assumed values 1 and 5. The mean of the slope has been fixed at zero and at four. The variance of the slope has remained unchanged

assuming a value of 0.2. In both classes $\psi_{\alpha\beta}$ has taken values 0 and 0.5. The variance of the residuals terms has been fixed to values that ensure a reliability of 0.5 for each observed indicator. For each of the combinations mentioned above normal distributed data were generated and for each resulting mixture a LGCM was estimated.

The effects of observed data deviations from non-normality on the bias, mean square root and coverage of the model parameter estimates was assessed. Preliminary results show that for models with a small mean intercept, skewness levels are low and kurtosis levels increase from low to moderate with the increase of the variability of the intercept and the magnitude of the slope. Consequently, the bias and the mean square error of the parameter estimates increase and coverage decreases, in particular for smaller sample sizes. Similar, but more severe, effects are obtained for the models with high mean intercept values.

## References

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

MUTHÉN, B.O. (2002). Using Mplus Monte Carlo Simulations In Practice: A Note On Assessing Estimation Quality and Power In Latent Variable Models. Mplus Web Notes: No. 1, Version 2.

MUTHÉN, B.O., & ASPAROUHOV, T. (2002). Using Mplus Monte Carlo Simulations In Practice: A Note On Non-Normal Missing Data In Latent Variable Models. Mplus Web Notes: No. 2, Version 2.

MUTHÉN, L.K., & MUTHÉN, B.O. (1998-2010). Mplus user's guide (6th Ed.). Los Angeles, CA: Muthén & Muthén.

SALGUEIRO, M.F. (2012). Modelling job satisfaction growth trajectories with missing data. IN Livro de Resumos das XIX Jornadas de Classificação e Análise de Dados, 74-76.

**Análise de Estruturas de Covariância – 6ª Feira, 11 de Abril, Sala 316 (9h00)**

# Depressão e risco de reincidência criminal face à delinquência juvenil

**Catarina Pral[1], Bruno Gonçalves[2], Catarina Marques[3]**

[1] *Faculdade de Psicologia da Universidade de Lisboa, catpral@hotmail.com;*

[2] *Faculdade de Psicologia da Universidade de Lisboa, bgoncalves@fp.ul.pt;*

[3] *Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, catarina.marques@iscte.pt*

**Sumário**

No presente trabalho procuraremos incidir sobre a associação entre a delinquência juvenil e a depressão na adolescência, bem como a coocorrência dos dois fenómenos e a sua relação com o risco de reincidência criminal. Nesta investigação participaram 283 jovens envolvidos com o sistema de justiça juvenil português, tendo-se estudado as variáveis em dois grupos: jovens institucionalizados (n=197) vs. jovens não institucionalizados (n=86).

**Palavras-chave:** Delinquência Juvenil; Depressão; Risco de Reincidência Criminal, Invariância do modelo fatorial.

## 1. Introdução

O efeito cumulativo de eventos de vida stressantes aumenta o risco de aparecimento, quer de problemas emocionais, quer de problemas comportamentais na adolescência (KIM *et al.*, 2003), sendo as experiências depressivas e o comportamento delinquente comuns durante o adolescer. Vários estudos apontam para a existência de uma inter-relação entre a depressão e a delinquência (e.g., ANGOLD & COSTELLO, 1993; WIESNER & KIM, 2006), sugerida pela frequência da coocorrência. Várias teorias explicativas da coocorrência têm sido defendidas por diferentes autores nomeadamente a **Teoria da falha**, segundo a qual os problemas externalizantes predizem os problemas internalizantes (CAPALDI, 1992); a **Teoria do acting-out** que defende que os problemas de internalização predizem os problemas externalizantes - *depressão mascarada* e a **Perspetiva da estabilidade** em que a coocorrência é explicada através de fatores de risco não específicos, como sejam a história familiar, a relação pais-filho e os acontecimentos de vida (KRUEGER *et al.*, 1998; OVERBEEK *et al.*, 2001).

De acordo com a investigação levada a cabo pelo *Northwestern Juvenile Project* (TEPLIN *et al.*, 2006), que avaliou a prevalência de doenças psiquiátricas nos jovens detidos no *Cook County Juvenile Temporary Detention Center*, no Illinois, quase dois terços dos indivíduos do sexo masculino e três quartos do sexo feminino preenchem os critérios de diagnóstico para uma ou mais doenças psiquiátricas, sendo que, mais de um quarto dos indivíduos do sexo feminino e quase um quinto dos indivíduos do sexo masculino, preenchiam critérios para uma ou mais doenças de foro afetivo (episódio depressivo major, distimia, episódio maníaco). AKSE *et al.* (2007) analisaram a coocorrência entre a delinquência e a depressão, tendo confirmado que o melhor modelo para explicar este fenómeno é o da perspetiva da estabilidade.

Na revisão bibliográfica realizada não se encontraram estudos comparativos entre grupos de delinquentes juvenis em situação de detenção e na comunidade. Este tipo de comparação parece-nos importante na medida em que a variável institucionalização poderá ter um papel ao nível da prevalência de sintomatologia depressiva e no risco de reincidência. Interessa-nos no presente estudo confirmar a coocorrência numa amostra de delinquentes Portugueses, pelo que iremos verificar a incidência de sintomatologia depressiva em jovens a cumprirem medidas tutelares educativas, introduzindo aqui a variável institucionalização no intuito de detetar se existem diferenças ao nível da depressão e risco de reincidência entre um grupo de delinquentes a cumprir medidas na comunidade e outro a cumprir a medida tutelar de internamento em Centro Educativo (CE), assim como avaliar se ambos estão relacionados.

## 2. Metodologia

Nesta investigação participaram 283 sujeitos, 39 raparigas e 244 rapazes, com idades compreendidas entre os 14 e os 19 anos de idade, sendo a média etária de 16,2 anos, 86 cumpriam medida de acompanhamento educativo, encontrando-se 197 a executar medida de internamento em centro educativo. A aplicação dos instrumentos selecionados, após autorização da Direção Geral de Reinserção e Serviços Prisionais e o consentimento informado dos sujeitos, ocorreu no período de março de 2011 e fevereiro de 2012. Decorreu, no que respeita aos jovens a cumprirem medida de acompanhamento educativo, nas equipas "Lisboa Tutelar Educativo 1" e "Lisboa Tutelar educativo 2" e nas equipas da então Delegação Regional do Algarve e Alentejo: Algarve 2 e Alentejo Interior. Relativamente aos adolescentes a cumprirem Medida Tutelar de Internamento em Centro Educativo, a recolha de dados ocorreu em todos os Centros Educativos do Continente. Participaram todos os adolescentes destes centros.

Foi utilizado um instrumento de medida psicológica - a Escala de Depressão do Centro de Estudos epidemiológicos (CES-D) (RADLOFF, 1977 - versão Portuguesa: FAGULHA & GONÇALVES, 2001) - e um instrumento de avaliação de necessidades e risco - Youth Level of Service/Case Management Inventory (YLS/CMI) (HOGE *et al.*, 2002 - Adaptação Portuguesa: DGRS, 2009). Cada um dos instrumentos foi sintetizado numa variável resumo que corresponde à pontuação total obtida como soma das respostas de todos os seus itens, podendo assumir valores entre 0 e 60 no caso do CES-D e entre 0 e 42 no caso de YLS/CMI.

Os dados obtidos foram analisados através de testes de hipóteses para comparação de médias de duas amostras independentes e de análise fatorial confirmatória. O modelo fatorial possui dois fatores, a depressão e o risco de reincidência, com 7 e 5 indicadores, respectivamente, que dependem apenas de um fator. Face ao nosso objectivo, pretendemos testar se os fatores se correlacionam e se há diferenças entre os dois grupos de jovens. As comparações nos modelos fatoriais foram realizadas com recurso a análise multigrupos. A invariância do modelo foi avaliada nos dois grupos por comparação do modelo livre (com pesos fatoriais, variâncias/covariâncias dos fatores e variâncias dos resíduos livres) com vários modelos onde foram aplicadas restrições de invariância de parâmetros. A escolha do modelo final recorreu ao

teste da diferença do qui-quadrado. Os modelos fatoriais foram estimados no AMOS 19, usando o método de máxima verosimilhança robusta.

## 3. Resultados

Comparando os resultados obtidos para a depressão, verifica-se a existência de diferenças significativas entre os dois grupos de jovens ($t=-3,878$; $p<0,001$), sendo a média da depressão maior no grupo de adolescentes a cumprirem medida de internamento em CE (média de 20,8) do que no grupo dos que cumprem medida de acompanhamento educativo (média de 16,6). Relativamente ao risco de reincidência, existem diferenças significativas entre os dois tipos de jovens ($t=-8,03$; $p<0,001$), sendo o risco de reincidência criminal maior em média entre os adolescentes a cumprirem medida de internamento em CE, dado que possuem média de 21,3, o que corresponde ao nível global de risco moderado de reincidência criminal, superior à dos adolescentes a cumprirem medida de acompanhamento educativo que possuem em média 14,1, o que corresponde ao nível global de risco moderado.

Os resultados da análise fatorial permitem afirmar que não existe relação entre a depressão e o risco de reincidência (correlação=0,148; p=0,067) se usarmos um nível de signficância de 5% (RMSEA=0,062,Pclose_fit=0,108,IC90%=(0,046;0,078); CFI=0,938; PCFI=0,783; GFI=0,918; PGFI=0,664). A análise multigrupos realizada permitiu a escolha do modelo com pesos fatoriais e covariâncias estruturais invariantes entre os dois grupos (RMSEA=0,051, Pclose_fit=0,448, IC90%=(0,031; 0,062); CFI=0,904; PCFI=0,830; GFI=0,863; PGFI=0,686), isto é, a importância de cada fator em cada um dos itens assim como a correlação entre a depressão e o risco de reincidência é a mesma nos dois grupos de delinquentes juvenis. Os resultados deste modelo confirmam a não existência de correlação entre a depressão e o risco de reincidência (correlação=0,037; p=0,651); contudo, dada a dimensão da amostra ao nível dos grupos ser pequena, esta afirmação é colocada com algumas reservas.

## 4. Discussão

Os resultados obtidos confirmam a coocorrência de sintomatologia depressiva e delinquência juvenil, verificando-se, através das diferenças significativas observadas nas médias nos dois grupos em estudo, que a institucionalização tem um papel relevante ao nível da sintomatologia depressiva.

Conforme seria expectável, dado tratar-se da medida tutelar educativa mais gravosa do sistema de justiça juvenil, os jovens que se encontram a cumprir medidas de internamento em CE, apresentam um risco médio de reincidência criminal superior à dos adolescentes a cumprirem medida de acompanhamento educativo.

Os jovens que se encontram a cumprir medida de internamento em CE apresentam níveis mais elevados de depressão e de risco de reincidência criminal. Contudo estes dois aspetos não se encontram relacionados, de acordo com os resultados obtidos através da análise fatorial, logo

teremos de considerar a interferência de outras variáveis, nomeadamente o facto dos jovens institucionalizados se encontrarem privados da liberdade e dos seus contextos de socialização habituais.

**Referências**

AKSE, J., HALE, B., ENGELS, R., RAAIJMAKERS & MEEUS, W. (2007). Co-Occurrence of Depression and Delinquency in Personality Types. *European Journal of Personality,* 21, 235-256.

ANGOLD, A. & COSTELLO, E.J. (1993). Depressive Comorbidity in children and adolescents: Empirical, Theoretical, and methodological issues. *American Journal of Psychiatry*, *150*, 1779- 1791.

CAPALDI, D.M. (1992). Co-occurrence of conduct problems and depressive symptoms in early adolescent boys: II. A 2-year follow-up at grade 8. *Development and Psychopathology*, 4, 125–144.

DGRS (2009). Manual do utilizador da YLS/CMI, versão experimental para uso interno.

GONÇALVES, B. & FAGULHA, T. (2000). *Estudo da Adaptação Portuguesa da Center for Epidemiologic Studies Depression Scale (CES-D).* Universidade de Lisboa.

HOGE,R., & ANDREWS, A. (2002). *Youth level of service/case management inventory (YLS/ CMI): User's manual*. Toronto: Multi-Health Systems (ed. 2008).

KIM, K., CONGER, R., ELDER, G., & LORENZ, F. (2003). Reciprocal influences between stressful life events and adolescent internalizing and externalizing problems. *Child Development*, 74(1), 127-143.

KRUEGER, R.F.; CASPI, A., MOFFITT, T.E., & SILVA, P.A. (1998). The structure and stability of common mental disorders (*DSM-III-R*): A longitudinal-epidemiological study. *Journal of Abnormal Psychology*, 107(2), 216-227.

OVERBEEK, G., VOLLERBERG, W., MEEUS, W., ENGELS, R. & LUIJPERS, E. (2001). Course, Co-Occurrence, and Longitudinal Associations of Emotional Disturbance and Delinquency from Adolescence to Young Adulthood: A Six-Year Three-Wave Study. *Journal of Youth and Adolescence*, 30 (4), 401-426.

RADLOFF, L (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Apply Psychology Measurement,* 1:385-401.

TEPLIN, L., ABRAM, K., MCCELLAND, G., MERICLE, A., DULCAN, M. & WASHBYRN, J. (2006). Doenças Psiquiátricas de Jovens em Situação de Detenção. *Infância e Juventude*, nº3/07.

WIESNER, M., & KIM, H.K. (2006). Co-occurring delinquency and depressive symptoms of adolescent boys and girls: A dual trajectory modeling approach. *Developmental Psychology*, 42, 1220-1235.

**Análise de Estruturas de Covariância – 6ª Feira, 11 de Abril, Sala 316 (9h20)**

# Visual fraud detection with Self Organizing Maps

**Paulo João[1], Victor Lobo[2]**

[1]*NOVA School of Statistics and Information Management, 1070-312 Lisbon – Portugal, Jolriao@gmail.com*

[2]*Portuguese Naval Academy, Alfeite, 2810-001 Almada – Portugal, Vlobo@isegi.unl.pt*

**Abstract**

Fraud has become a major focus of concern and even a political, social and economic issue in modern society. Fraud occurs in many fields, including healthcare, where the cost of meeting public demand for high quality and high technology services is very high. Fraud can be detected in many ways, and traditionally the State relies almost only on its internal control activities and internal auditors to prevent and detect fraud. These are general purpose solutions for fraud detection, but are not particularly adequate for specific data such as electronic prescription. The amount of data in healthcare repositories along with their high dimensional nature is too complex to be processed and analyzed by traditional methods. Self Organizing Maps (SOM), can improve the decision making with an approach based on dimensionality reduction and visualization of fraud. The goal of this paper is to propose a methodology to detect fraud in real time, where traditional human based approaches are not possible or are too complex to process.

**Keywords:** Outlier detection, data mining, SOM, fraud detection.

## 1. Introduction

A common major difficulty associated with fraud detection is that there is a large amount of data that needs to be analyzed and, simultaneously, there are only a few fraudulent samples to be used as the training data for the supervised methods (Olszewski, Kacprzyk and Zadrozny, 2013). Fraudulent behavior has become a major focus of concern and a political, social and economic issue in modern society. The requirement to meet public demand for quality and technology services is real and this is likely to become more widespread and intense (Yang and Hwang, 2006). The amount of data along with their high-dimensional nature requires sophisticated statistical methods to extract new and unexpected patterns embedded in that data. Therefore, effective, fraud detection is important for improving the quality and cut the costs of healthcare reimbursements (Li, Huang, Jin and Shi, 2008).

To detect fraud, first we analyze the available data and then we must detect the outlier data or unusual observations. It requires an understanding of the mathematical properties of data and relevant knowledge in the domain context where the outliers occur (Ilango, Subramanian and Vasudevan, 2012). Detection methods can be divided between: univariate methods, where each variable is explored separately; and multivariate methods, where many variables are explored together. We can also find two different approaches in each case: parametric methods, that assume characteristics or parameters on the data, *e.g.* Gaussian

distribution; and non-parametric methods, that don't assume such characteristics or parameters (Ben-Gal, 2005).

The problem of detection of multidimensional outliers is a fundamental and important problem in applied statistics. It involves a lot of computation time and dimension reduction. Despite this, most of known algorithms for detecting outliers are not fast enough when the underlying probability distribution is unknown, the size of the dataset is large and the number of dimensions in the space is high (Chaudhary, Szalay and Moore, 2002).

## 2. Outliers overview

Outliers are unusual observations on data. They are also known as: novelty; noise; extreme values; fraud, anomaly; defect data; skewed data; rare cases or noisy data. The difference is not methodological but related to the involved subject. For Ferdousi and Maeda (2006), Hodge and Austin (2004), Ahmed and Funk (2011), He, Xu, Huang and Deng (2004) outliers have been informally defined as rare observations in a dataset that seem to be inconsistent with the rest of that set of data, or deviate so much from other observations to arouse suspicions that they were generated by a different mechanism. "Fraud" refers to criminal activities or fraudulent behavior occurring in commercial organizations such as banks, insurance agencies, mobile companies, *etc*., and it is assumed that these activities are rare and generate data which differs from "normal" activities. Thus, outliers can be due to frauds, and it makes sense to look for frauds mainly amongst these outliers, even though they are not necessarily frauds.

### 2.1 Self Organizing Maps

A Self Organizing Map (SOM) is a neural network with feed-forward topology and an unsupervised training algorithm that uses a self-organizing process to configure its output neurons, according to the topological structure of the input data (Wasserman, 1989).

SOM were first proposed by Tuevo Kohonen in the beginning of the eighties (Kohonen, 1982). He draws some inspiration from the way we believe the human brain works. Research has shown that the cerebral cortex of the human brain is divided into functional subdivisions and that the neural activity decreases as the distance to the region of first activation increases (Kohonen, 2001).

To detect multiple outliers in multidimensional datasets, a SOM produces a topology preserving mapping of the multidimensional data onto lower dimensional visualizable plane (Chandola, Banerjee and Kumar, 2007). The objective is to extract the essential structures in a dataset, through a map resulting from an unsupervised learning process (Kaski and Kohonen, 1996; Kaski, Nikkilä and Kohonen, 1998). Each input data is mapped to one of a large (but fixed) number of nodes, and groups of these nodes (or even just a single one) represent a

cluster. The basic SOM training algorithm can be described as follows (Lobo, Cabral and Bação, 2007):

Let:
$\boldsymbol{x}$ be the set of $n$ training patterns $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n}$
$W$ be a $p_x q$ grid of units $\boldsymbol{w_{ij}}$ where $i$ and j are their coordinates on that grid
$\alpha$ be the learning rate, assuming values in the interval ]0,1[, initialized to a given first learning rate
$r$ be the radius of the neighborhood function $h(\boldsymbol{w_{ij}}, \boldsymbol{w_{mn}}, r)$, initialized to a given first radius

Do
1 Repeat
2 For $k$=1 to $n$
3 For all $\boldsymbol{w_{ij}} \in W$, calculate $d_{ij} = ||\boldsymbol{x_k} - \boldsymbol{w_{ij}}||$
4 Select the unit that minimizes $d_{ij}$ as the winner $\boldsymbol{w_{winner}}$
5 Update each unit $\boldsymbol{w_{ij}} \in$W: $\boldsymbol{w_{ij}} = \boldsymbol{w_{ij}} + \alpha h(\boldsymbol{w_{winner}}, \boldsymbol{w_{ij}}, r) ||\boldsymbol{x_k} - \boldsymbol{w_{ij}}||$
6 Decrease the value of $\alpha$ and $r$
7 Until $\alpha$ reaches 0

The main advantage of SOM is that it allows us to have some idea of data structure and outliers by observing a two dimensional map. This is possible mainly due to preservation of topological relations (*e.g.* patterns that are close in the input space will be mapped to units that are close in the output space). In most implementations this is a rectangular grid of units, but it can also be hexagonal (Kohonen, 2001). Single dimensional SOM is also common to solve problems where an ordering is necessary (*e.g*. travelling salesman problem).

The aim of this paper is to propose a method based on SOM to deal with outlier detection and give a useful tool to fight healthcare fraud. This method uses particular characteristics of healthcare data and can sometimes need adjustments in the general data mining methods, while in other cases minimal or no changes are necessary. These adjustments such as the size of the network, the learning rate and the neighborhood radius of the SOM, are dependent on the problem we are dealing with and on the expected results and analysis.

## 3. Conclusions and future work

SOM applications in healthcare to detect outliers can have tremendous potential and usefulness. However, the success of healthcare outlier detection hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare people consider how data can be better captured, stored, prepared, and mined. Possible directions of further work include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of this kind of applications. How to detect false positive errors and avoid them is a main issue to save computational time and deserves more attention from researchers.

**References**

AHMED, M. U., & FUNK, P. (2011). Mining rare cases in post-operative pain by means of outlier detection. *Signal Processing and Information Technology (ISSPIT), 2011 IEEE International Symposium on IEEE*, 035-041.

BEN-GAL, I. (2005). Outlier detection. *Data Mining and Knowledge Discovery Handbook*, 131-146.

CHANDOLA, V., BANERJEE, A., & KUMAR, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*, to appear.

CHAUDHARY, A., SZALAY, A. S., & MOORE, A. W. (2002). Very fast outlier detection in large multidimensional data sets. *Proceedings of the ACM SIGMOD Workshop in Research Issues in Data Mining and Knowledge Discovery (DMKD).*

FERDOUSI, Z., & MAEDA, A. (2006). Unsupervised outlier detection in time series data. *Proceedings of the 22nd International Conference on Data Engineering Workshops, IEEE, 51-56.*

HE, Z., XU, X., HUANG, J. Z., & DENG, S. (2004). A frequent pattern discovery method for outlier detection. *Advances in Web-Age Information Management*, Springer Berlin Heidelberg, 726-732.

HODGE, V., & AUSTIN, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 2 , 85-126.

ILANGO, V., SUBRAMANIAN, R., & VASUDEVAN, V. (2012). A Five Step Procedure for Outlier Analysis in Data Mining. *European Journal of Scientific Research*, 75, 3, 327-339.

KASKI, S. & T. KOHONEN (1996). Exploratory data analysis by the Self Organizing Maps: structures of welfare and poverty in the world. *Neural Networks in Financial Engineering*, N. Apostolos-Paul, Yaser Refenes, Yaser Abu-Mostafa, John Moody & A. Weigend. Singapore, World Scientific, 498-507.

KASKI, S., J. NIKKILÄ & T. KOHONEN (1998). Methods for interpreting a Self Organizing Maps in data analysis. *Proceedings of ESANN'98, 6th European Symposium on Artificial Neural Networks*, Bruges, Belgium, D-Facto.

KOHONEN, T. (1982). Self-organizing formation of topologically correct feature maps. *RecMap: rectangular map approximations*, 43, 1, 59-69.

KOHONEN, T. (2001). SOM, Vol. 30. Springer Verlag.

KOHONEN, T., HYNNINEN, J., KANGAS, J., LAAKSONEN, J., & TORKKOLA, K. (1996). *LVQ PAK: The learning vector quantization program package*. Technical report, Laboratory of Computer and Information Science Rakentajanaukio 2 C, 1991-1992.

LOBO, V., Cabral, P., & Bação, F. (2007). Self Organizing Maps for urban modelling. *Proceedings of the 9th International Conference on Geocomputation.*

OLSZEWSKI, D., Kacprzyk, J., & Zadrożny, S. (2013). Employing Self-Organizing Map for Fraud Detection. *Artificial Intelligence and Soft Computing* ). Springer Berlin Heidelberg, 150-161

WASSERMAN, P. D. (1989). *Neural computing: theory and practice*. Van Nostrand Reinhold Co..

YANG, W. S., & HWANG, S. Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31, 56-68.

**Análise de Estruturas de Covariância – 6ª Feira, 11 de Abril, Sala 316 (9h40)**

# Os ginásios da Cidade de Maputo: Os determinantes da satisfação e da lealdade dos clientes

**Edmundo Roque Ribeiro[1], Catarina Marques[2], Eduardo Correia[3]**
[1]*CIDAF- UP, Moçambique, edroquer@gmail.com;*
[2]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, catarina.marques@iscte.pt;*
[3]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, eduardo.correia@iscte.pt*

**Sumário**

O presente estudo pretende analisar os determinantes da satisfação e da lealdade do consumidor no contexto dos ginásios, recorrendo ao modelo ECSI. O modelo foi estimado por PLS-PM tendo por base uma amostra de 339 clientes de ginásios da Cidade de Maputo. Os resultados permitem concluir que a satisfação dos clientes é o principal factor responsável pela formação da lealdade dos clientes ao ginásio e a qualidade percebida é o factor com maior influência na satisfação dos clientes.

**Palavras-chave:** Qualidade**,** Satisfação, Lealdade, ECSI, PLS-Path Modeling.

## 1. Introdução

De acordo com SABA e PIMENTA (2008) um dos sectores com maior crescimento no mundo é o sector onde estão inseridos os ginásios. TSITSKARI e TSIOTRAS (2006) afirmam que uma das prioridades dos mesmos é a retenção dos clientes. COSTA (2006) advoga que os desafios da retenção são acrescidos pela existência actual de consumidores com maior consciência crítica, educação, conhecimento e informação. Nesse sentido, mais relevantes se tornam os esforços em identificar as dimensões da qualidade mais valorizadas pelos clientes e compreender a sua relação com a satisfação e a fidelidade ao ginásio. Este estudo pretende contribuir para analisar a temática da qualidade de serviços, da satisfação e da lealdade do consumidor no contexto dos ginásios da Cidade de Maputo.

A medição da satisfação e da qualidade de serviços tem ocupado um espaço bastante vasto na literatura do marketing de serviços existindo vários modelos para esse fim. O modelo *European Customer Loyalty Index* (ECSI) tem sido adoptado por muitos autores para a medição da satisfação e da lealdade do consumidor, com a consequente criação de um índice. O ECSI representa um modelo de equações estruturais constituído por variáveis latentes sendo a imagem, expectativas, valor e qualidade antecedentes da satisfação e as reclamações e a lealdade suas consequentes. Com o objectivo de testar a aplicabilidade do ECSI no contexto moçambicano, contexto social distinto daquele onde tem sido analisado e, em particular, no sector dos ginásios onde é rara a sua aplicação, o modelo foi adaptado dos usados por VILARES & COELHO (2005) e MARTENSEN et al. (2000). Para além disso, a qualidade percebida é fragmentada em qualidade técnica e qualidade funcional (GRONROOS, 2000) e, considerando as peculiaridades do sector dos ginásios, adicionou-se aos indicadores de aferição da qualidade alguns indicadores (os relativos a pessoal, programas, equipamento, instalações e

balneários) da *Service Quality Assessment Scale* (SQAS) de LAM et al. (2005), escala idealizada para a avaliação da qualidade de serviços em health clubs.

## 2. O modelo

A figura 1 apresenta o modelo conceptual e as 13 hipóteses a testar.

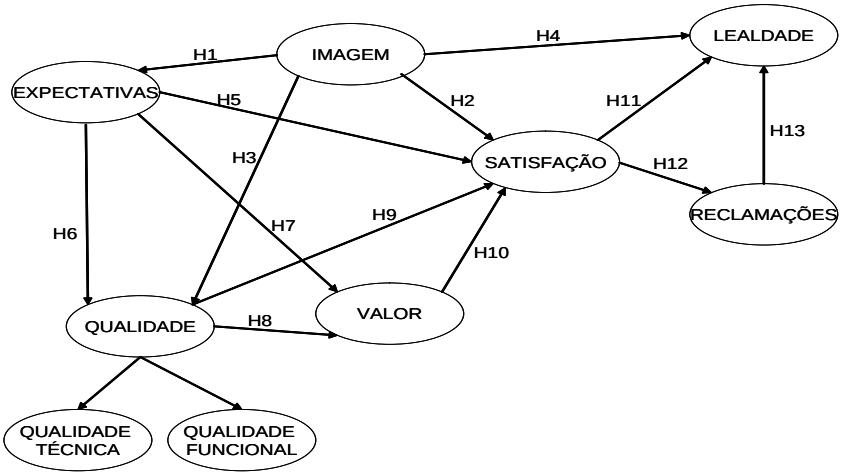



**Figura 1: Modelo Conceptual e Hipóteses**

## 3. Metodologia

Os dados foram obtidos através de um questionário auto-administrado constituído por 62 questões para medição dos construtos do modelo ECSI (medidas por uma escala de Likert de 7 pontos) e 12 questões dirigidas à identificação do perfil sócio demográfico dos respondentes. O questionário foi aplicado entre Maio e Agosto de 2012 a clientes de ginásios da Cidade de Maputo. Obtiveram-se 339 respostas de clientes que compareceram em dias e períodos horários escolhidos aleatoriamente em cada ginásio. O modelo foi estimado por PLS-PM (WOLD, 1985). Todas as variáveis de medida do modelo são reflectivas.

## 4. Resultados e discussão

Os resultados da análise revelam que 45,5% da variação da lealdade pode ser explicado pela variação da satisfação e da imagem, enquanto que 51% da variação da satisfação pode ser explicado pela variação da qualidade, das expectativas e do valor. Não obstante a existência de diferenças entre os modelos, resultados semelhantes foram encontrados por GONÇALVES (2012) que em estudo similar na área do fitness obteve que 51% da variação da lealdade é explicada pela variação da satisfação, da qualidade e das expectativas e 55% da variação da satisfação é explicada pela qualidade, pelas expectativas e pelo bem-estar. No que diz respeito

aos modelos de medida, a sua qualidade é assegurada uma vez que todas as medidas de fiabilidade e de validade do modelo estão dentro dos limites considerados recomendáveis.

A tabela 1 apresenta as estimativas standardizadas dos coeficientes das relações estruturais. No que diz respeito aos principais determinantes da satisfação e da lealdade, os resultados permitem concluir que (a) a qualidade percebida dos ginásios é o que mais influencia a formação da satisfação dos clientes (com um coeficiente estimado de 0,559); (b) o valor percebido e as expectativas formadas também têm um impacto positivo na formação da satisfação dos clientes; e (c) a satisfação dos clientes e a imagem dos ginásios são os factores principais responsáveis pela formação da lealdade dos clientes aos ginásios. As hipóteses H2 e H13 não se verificaram, pelo que se conclui que a imagem não afecta a satisfação do consumidor e que a resolução adequada das reclamações do consumidor não influencia a lealdade do consumidor.

**Tabela 1: Estimativas standardizadas dos coeficientes de regressão das relações estruturais**

| Hip. | Relação Estrutural | Estimativa | Erro Padrão | Estatística t |
|---|---|---|---|---|
| H1 | IMAGEM -> EXPECTATIVAS | 0,582 | 0,039 | 14,902 |
| H2 | IMAGEM -> SATISFACAO | *-0,017* | *0,059* | *0,287* |
| H3 | IMAGEM -> QUALIDADE | 0,390 | 0,052 | 7,489 |
| H4 | IMAGEM -> LEALDADE | 0,325 | 0,045 | 7,280 |
| H5 | EXPECTATIVAS -> SATISFAÇÃO | 0,128 | 0,062 | 2,054 |
| H6 | EXPECTATIVAS -> QUALIDADE | 0,332 | 0,058 | 5,752 |
| | QUALIDADE -> QUALIDADE TÉCNICA | 0,865 | 0,017 | 50,720 |
| | QUALIDADE -> QUALIDADE FUNCIONAL | 0,970 | 0,004 | 221,458 |
| H7 | EXPECTATIVAS -> VALOR | 0,152 | 0,062 | 2,451 |
| H8 | QUALIDADE -> VALOR | 0,432 | 0,061 | 7,044 |
| H9 | QUALIDADE -> SATISFAÇÃO | 0,559 | 0,063 | 8,842 |
| H10 | VALOR -> SATISFAÇÃO | 0,137 | 0,060 | 2,266 |
| H11 | SATISFAÇÃO -> LEALDADE | 0,434 | 0,062 | 6,970 |
| H12 | SATISFAÇÃO -> RECLAMAÇÕES | 0,517 | 0,062 | 8,304 |
| H13 | RECLAMAÇÕES -> LEALDADE | *0,048* | *0,065* | *0,743* |

Relativamente à imagem, é de realçar a sua influência nas expectativas criadas sobre o ginásio, na formação da qualidade percebida e na lealdade do cliente ao ginásio. Segundo SAÍAS (2007) as expectativas constituem o *standard* em relação ao qual é avaliada a qualidade percebida do serviço. Por outro lado, SABA e PIMENTA (2008) afirmam que uma boa imagem torna os clientes mais tolerantes aos erros cometidos pela empresa e exerce uma influência sobre o serviço. Os resultados do estudo estão de acordo com os autores citados, sugerindo que a construção de uma imagem robusta do ginásio deve constar das preocupações dos gestores dos ginásios.

## 5. Conclusão

A maioria das hipóteses foram comprovadas, desse modo o modelo ECSI utilizado revelou ser adequado para mensurar a satisfação dos clientes em relação aos serviços prestados nos ginásios da Cidade de Maputo. Conclui-se que a formação da satisfação dos clientes é influenciada pela qualidade percebida e que a satisfação dos clientes é o principal factor responsável pela formação da lealdade dos mesmos.

## Referências

COSTA, M. (2006) Construção da Marca Health & Fitness. IN A. Correia, A. Sacavém, C. Colaço (Eds.), *Manual de Fitness & Marketing*, 185-197. Lisboa: Visão e Contextos.

GONÇALVES, C. (2012) *Retenção de sócios no fitness- Estudo do posicionamento, expectativas, bem-estar e satisfação*. Dissertação de doutoramento em Motricidade Humana. Faculdade de Motricidade Humana – UTL.

GRÖNROOS, C. (2000) *Service Management and Marketing: a customer relationship management approach*. Chichester: John Wiley & Sons, Lta.

LAM, E., ZHANG , J. & JENSEN, B. (2005) Service Quality Assessment Scale (SQAS): An Instrument for Evaluating Service Quality of Health–Fitness Clubs. *Measurement in physical education and exercise science*, 9(2), 79–111.

MARTENSEN, A., GRØNHOLDT, L., & KRISTENSEN, K. (2000) The drivers of customer satisfaction and loyalty: cross-industry findings from Denmark. *Total Quality Management & Business Excellence, 11*(4,5,6), 544-553.

SABA, F. & PIMENTA, M. (2008). *Vendas e Retenção – 83 Lições para Academias e Clubes Esportivos*. São Paulo: Phorte Editora.

SAÍAS, L. (2007) . Marketing de Serviços – Qualidade e Fidelização de clientes. Lisboa: Universidade Católica Editora.

TSITSKARI, E. & TSIOTRAS, D. (2006) . Measuring Service Quality in Sport Services. *Total Quality Mangement*. Vol 17, 5, 623-631.

VILARES, M. & COELHO, P. (2005). *A Satisfação e a Lealdade do Cliente. Metodologias de Gestão, Avaliação e Análise*. Lisboa: Escolar Editora.

WOLD H. O. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopaedia of Statistical Sciences,* Volume 6 (pp. 581–591). New York, NY: John Wiley and Sons.

**Análise de Estruturas de Covariância – 6ª Feira, 11 de Abril, Sala 316 (10h00)**

# Near-exact distributions for the statistic used to test the reality of covariance matrix in a complex normal distribution

**Luís Miguel Grilo[1*], Carlos Agra Coelho[2*]**

[1]*Unidade Departamental de Matemática do Instituto Politécnico de Tomar, lgrilo@ipt.pt*;

[2] *Departamento de Matemática da Universidade Nova de Lisboa, cmac@fc.unl.pt*;

**Centro de Matemática e aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa*.

**Abstract**

To develop a family of near-exact distributions, for the likelihood statistic used to test the reality of the covariance matrix in a certain complex multivariate normal distribution, we start to show that the exact distribution of the negative logarithm of the likelihood statistic may be written as an infinite mixture of Generalized Near-Integer Gamma distribution. The near-exact distributions obtained are very close to the exact distribution but far more manageable and match, by construction, some of the first exact moments. The corresponding cumulative distribution functions allow us an easy computation of near-exact quantiles.

**Key-words:** Likelihood statistic, Characteristic function, Beta distribution**,** Gamma distribution, Mixtures.

## 1. Introduction

Let $\underline{X}$ be a random vector with a $p$ - variate complex normal distribution (Goodman, 1963), with variance-covariance matrix $\Sigma = \Sigma_1 + \mathrm{i}\Sigma_2$, which is a $p \times p$ positive Hermitian matrix, where $\Sigma_1$ is a $p \times p$ symmetric positive definite matrix, $\Sigma_2$ is a $p \times p$ skew-symmetric matrix and $\mathrm{i} = (-1)^{1/2}$ (Carter et al., 1976; Khatri, 1965a).

To test the reality of $\Sigma$, that is, to test

$$H_0 : \Sigma_2 = 0 \quad \textit{versus} \quad H_1 : \Sigma_2 \neq 0,$$

and assuming $\underline{X}$ with a non-null expected value, we may consider, for a sample of size $n + 1$, the power $2/(n + 1)$ of the likelihood ratio test statistic, obtained by Khatri (1965b),

$$\Lambda = \frac{|S_1 + \mathrm{i}S_2|}{|S_1|}, \tag{1}$$

where $S = S_1 + \mathrm{i}S_2$ is the maximum likelihood estimator of $\Sigma$.

When $\Sigma_2 = 0$, the statistic $\Lambda$, in (1), is distributed as a product of independent beta random variables (r.v.'s) with specific parameters. More precisely, if $p$ is even,

$$\Lambda \overset{st}{\sim} \prod_{j=1}^{\frac{p}{2}} Y_j \text{ where } Y_j \sim Beta\left(n - \frac{p}{2} - j + 1, \frac{p}{2} - \frac{1}{2}\right), \tag{2}$$

and, if $p$ is odd,

$$\Lambda \overset{st}{\sim} \prod_{j=1}^{\frac{p-1}{2}} Y_j \text{ where } Y_j \sim Beta\left(n - \frac{p+1}{2} - j + 1, \frac{p}{2}\right), \tag{3}$$

or for any $p$ , in (2) and (3), and taking $q^* = \left\lfloor \frac{p}{2} \right\rfloor$ and $q = \left\lceil \frac{p}{2} \right\rceil$ (where $\lfloor \cdot \rfloor$ denoting the floor of the argument, that is, the largest integer that does not exceed the argument and $\lceil \cdot \rceil$ denoting the ceiling of the argument, that is, the smallest integer not less than the argument), we may write

$\Lambda \overset{st}{\sim} \prod_{j=1}^{q^*} Y_j$ where $Y_j \sim Beta\left(n-q-j+1,\ q-\frac{1}{2}\right)$ are $q^*$ independent r.v.'s and '$\overset{st}{\sim}$' means 'stochastically equivalent to'. Thus, we may easily obtain (Grilo & Coelho, 2013)

$$E(\Lambda^h) = \prod_{j=1}^{q^*} E(Y_j^h) = \prod_{j=1}^{q^*} \frac{\Gamma\left(n-j+\frac{1}{2}\right)}{\Gamma(n-q-j+1)} \frac{\Gamma(n-q-j+1+h)}{\Gamma\left(n-j+\frac{1}{2}+h\right)}, \text{with } h > -(n-q-j+1). \tag{4}$$

Based in (4), we start to express the exact distribution of the negative logarithm of $\Lambda$ statistic as an infinite mixture of Generalized Near-Integer Gamma distribution (GNIG) and with this representation we develop near-exact distributions for $\Lambda$, which are finite mixtures of GNIG distributions.

In a number of papers we had already introduced the concept and explained the procedure used to develop near-exact distributions (Coelho, 1998, 2004; Coelho et al., 2006; Grilo, 2005; Grilo & Coelho, 2007, 2010a, 2010b, 2011) and we have also applied a similar procedure to obtain near-exact distributions for the product of an odd number of particular independent Beta r.v.'s (Grilo & Coelho, 2007). In the present case, based on the factorization of the exact characteristic function of the negative logarithm of $\Lambda$, we obtain near-exact distributions for this statistic which have better performance than the asymptotic distribution considered. The numerical studies developed, using a proximity measure, show the high closeness of these near-exact distributions to the exact distribution and also their excellent performance, namely for small samples sizes and for small values of $n - p$.

**Acknowledgments:** This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project PEst-OE/MAT/UI0297/2014 (Centro de Matemática e Aplicações).

## References

CARTER, E. M., KHATRI, C. G. & SRIVASTAVA, M. S. (1976) Nonnull distribution of likelihood ratio criterion for reality of covariance matrix. *Journal of Multivariate Analysis,* 6**,** 176-184.

COELHO, C. A. (2004) The generalized near-integer Gamma distribution: a basis for 'near-exact' approximations to the distribution of statistics which are the product of an odd number of independent Beta random variables. *Journal of Multivariate Analysis,* **89**, 191-218.

COELHO, C. A. (1998) The generalized integer Gamma distribution – a basis for distributions in Multivariate Statistics. *Journal of Multivariate Analysis,* **64**, 86-102.

COELHO, C. A., ALBERTO, R. P. & GRILO, L. M. (2006) A mixture of Generalized Integer Gamma distribution as the exact distribution of the product of an odd number of independent Beta random variables with first parameter evolving by 1/2. Applications. *Journal of Interdisciplinary Mathematics,* **9**, 2, 228-248.

GOODMAN, N. R. (1963) Statistical analysis based on a certain multivariate complex Gaussian distribution (An introduction). *Ann. Math. Statist.,* 34**,** 152-176.

GRILO, L. M. & COELHO, C. A. (2013) Near-exact distributions for the likelihood ratio statistic used to test the reality of a covariance matrix. *11th International Conference of Numerical Analysis and Applied Mathematic*s (*ICNAAM 2013*), AIP (American Institute of Physics) Conference Proceedings **1558**, 797.

GRILO, L. M. & COELHO, C. A. (2011) A family of near-exact distributions based on truncations of the exact distribution for the generalized Wilks Lambda statistic. *Journal of Communications in Statistics - Theory and Methods,* **41**: 1-21.

GRILO, L. M. & COELHO, C. A. (2010a) Near-exact distributions for the generalized Wilks Lambda statistic. *Discussiones Mathematicae Probability and Statistics,* **30** (1) 53-86.

GRILO, L. M. & COELHO, C. A. (2010b) The exact and near-exact distribution for the Wilks Lambda statistic used in the test of independence of two sets of variables. *American Journal of Mathematical and Management Sciences,* **30** (#1 & 2) 111-140.

GRILO, L. M. & COELHO, C. A. (2007) Development and Comparative Study of two Near-exact Approximations to the Distribution of the Product of an Odd Number of

Independent Beta Random Variables. *Journal of Statistical Planning and Inference*, **137**, 1560-1575.

GRILO, L. M. (2005) *Development of near-exact distributions for different scenarios of application of the Wilks Lambda statistic* (in Portuguese). Ph.D. thesis, Thecnical University of Lisbon, Portugal.

KHATRI, C. G. (1965a) Classical statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Statist.*, 36**,** 98-114.

KHATRI, C. G. (1965b) A test for reality of a covariance matrix in a certain complex Gaussian distribution. *Ann. Math. Statist.,* 36**,** 115-119.

**Análise de Dados em Economia e Gestão –Sábado, 12 de Abril, Salão Nobre (10h00)**

# Fatores chave de sucesso das equipas virtuais de tecnologias de informação em regime de outsourcing: do ponto de vista dos membros da equipa

**Fernando Santos[1], Ana Lorga da Silva[2], Isabel Duarte[3]**

*[1]Escola de Ciências Económicas e das Organizações - ULHT, Campo Grande, 376, 1749-024 Lisboa - PORTUGAL, fsantos.santos@gmail.com;*

*[2]CPES, Escola de Ciências Económicas e das Organizações - ULHT, Campo Grande, 376, 1749-024 Lisboa - PORTUGAL e CÉDRIC - CNAM, Paris – FRANCE, ana.lorga@ulusofona.pt;*

*[3]Escola de Ciências Económicas e das Organizações - ULHT, Campo Grande, 376, 1749-024 Lisboa - PORTUGAL 3, p789@ulusofona.pt*

**Sumário**

No presente trabalho, são identificados os principais fatores de sucesso das equipas virtuais das tecnologias de informação e de comunicação em ambiente de terceirização, do ponto de vista dos membros das próprias equipas. O questionário foi realizado através de um sítio *online* e aplicado aos membros das equipas das quatro regiões onde estes estão localizados, tendo em conta o seu sexo, idade e antiguidade na empresa.

As perguntas foram feitas utilizando uma Escala de Likert (1 a 5, onde 1 significa que não é importante e 5 muito importante). Os grupos principais de perguntas centram-se na equipa, condições de trabalho, comunicação e liderança. Os dados ordinais foram analisados utilizando métodos estatísticos paramétricos e não paramétricos.

**Palavras-chave:** Alfa de Cronbach**,** Análise multivariada de dados, CPCA, Escala de Likert, Fatores de sucesso em Equipas Virtuais.

## 1. Descrição da problemática

Num mundo globalizado em constante mutação e em procura acelerada pela redução de custos, as empresas tendem a adotar novas formas de organização, nomeadamente, recorrendo à constituição de Equipas Virtuais. A principal distinção entre uma equipa convencional e uma equipa virtual é o fator localização (Bell & Kozlowski, 2002).

Vários autores (Cvitovich, 2008; Ebrahim et al, 2012, entre outros) referem que o sucesso deste tipo de organizações de trabalho - pode levar a um aumento de eficiência e eficácia, **apenas e se**, estas forem geridas de forma a potenciar os seus benefícios e a minimizar as suas desvantagens.

A constante procura das empresas pela redução de custos levou a que, hoje em dia, as organizações de suporte sejam compostas por diversos grupos de indivíduos dispersos em termos geográficos e por diferentes fusos horários. As equipas de tecnologias de informação e comunicação são constituídas por um misto de grupos especializados localizados, por vezes,

em continentes diferentes daquele onde o cliente se encontra *Offshore*, próximos do cliente *Nearshore* e nas instalações do cliente *Onshore* (Carmel & Tija, 2005).

Como se pode observar, esta dinâmica leva à criação de equipas virtuais, as quais além de prestarem um serviço de qualidade ao cliente - interno e/ou externo – no caso de prestação de serviços em regime de *Outsourcing*; devem funcionar como se de uma equipa homogénea de trabalho se tratasse. Apesar de trabalharem no mesmo projeto, tendo em conta todas as variáveis, localizações geográficas dispersas, diferentes culturas, fusos horários, e por vezes, objetivos de carreira diferentes, estas equipas teriam à partida todos os ingredientes para não serem eficientes. Contudo, em algumas equipas verifica-se precisamente o contrário, pois, de acordo com os indicadores internos da empresa para a qual trabalham, são eficientes e eficazes. Quais serão então os ingredientes desse sucesso? O que os move e faz atuar como uma só equipa e serem altamente eficientes e eficazes? E porque é que esta eficácia se mantém ao longo do tempo? São estas e outras interrogações que justificam *per si*, o estudo desta temática.

### 1.1. Objetivo do Estudo

A presente investigação pretende aprofundar o conhecimento sobre os fatores chave de sucesso das equipas virtuais, em especial, das equipas de suporte às Tecnologias de Informação e Comunicação em regime de *Outsourcing* do ponto de vista das próprias equipas.

### 1.2. Significado da Pesquisa

Com o aumento do número de empresas a aderir à implementação de equipas virtuais, concomitantemente, nos últimos anos, os investigadores têm vindo a orientar as suas investigações para este tipo de organizações. A ênfase tem sido posta em dimensões específicas, colocando a equipa como elemento central das investigações, mas sem considerar o seu ponto de vista. Através de uma nova abordagem, relegando o papel do investigador para segundo plano e assumindo as equipas virtuais objeto do estudo um papel ativo; pretende-se identificar os fatores chaves de sucesso das equipas virtuais, de acordo com o ponto de vista das próprias equipas. Com este facto em mente, e, tendo em consideração que a investigação vista deste prisma poderia ser útil para o desenvolvimento do conhecimento, considerou-se oportuno avançar com este tipo de abordagem.

### 1.3. Hipóteses a Investigar

O sucesso das equipas objeto da presente investigação mede-se, confrontando as métricas dos níveis de serviço entregue mensalmente, com os níveis de serviço previamente definidos por contrato.

Reconhecendo que os processos utilizados pelas equipas objeto desta investigação, são *standards*, e que, os conhecimentos técnicos dos membros das equipas são os adequados para o desempenho cabal das tarefas, resta centrar a investigação na área motivacional. Sendo as equipas heterogéneas (idade, género, cultura, experiência, antiguidade na empresa/grupo de trabalho, nível de conhecimentos e/ou experiências profissionais e localização geográfica), até

que ponto, estas variáveis tem impacto no espírito de coesão da equipa e afetam a eficácia da mesma? De acordo com os indicadores internos da empresa, qual é a razão ou razões, para que a equipa se mantenha altamente eficiente e eficaz?

Nesse sentido formulam-se as seguintes hipóteses:

1. A motivação das equipas é considerada um fator essencial para o sucesso das Equipas Virtuais;
2. A existência de instrumentos e métodos para lidar com as diferenças individuais e culturais favorece a coesão e, é considerado pelos membros um fator de sucesso das equipas virtuais;
3. Os processos de comunicação formal e informal que promovem a participação ativa dos membros das Equipas Virtuais são percecionados pelos mesmos como essenciais para o seu sucesso;
4. A boa organização e coordenação do trabalho das Equipas Virtuais é considerado pelas mesmas um fator de sucesso;
5. Os membros das equipas consideram que a capacidade de prevenir e gerir conflitos e divergências potencia o sucesso das Equipas Virtuais.

## 2. Metodologia

### 2.1. Dados

Para a recolha dos dados recorreu-se à utilização de um sítio *online* onde se colocou um questionário com respostas fechadas, que incluem em relação a cada inquirido: género, região do mundo onde se encontra, idade (por intervalos), senioridade, tipo de cliente, e diversas questões com o objetivo de "responder" às hipóteses identificadas em 1.3., cujas respostas são dadas utilizando uma Escala de Likert (1 a 5, onde 1 significa que não é importante e 5 muito importante).

Foi aplicado um pré-teste que foi validado, tendo sido em seguida aplicado o questionário.

Do total da população de 1.380 elementos, recolheram-se 265 respostas, o que corresponde a uma taxa de participação de 19,2%. Numa primeira análise dos dados, verificou-se que 34 respondentes (12,8% do total), não tinham concluído o questionário, não se utilizou um método de imputação explícita, tendo sido as suas respostas eliminadas (método "*listwise*").

### 2.2. Análise dos dados

O questionário utilizado contém três dimensões - A equipa, Ambiente de Trabalho e Liderança, totalizando 39 itens. Observou-se a existência de uma diversidade de respondentes, quer em termos culturais, idade, género e senioridade.

A abordagem utilizada centrou-se na análise dos três grupos de questões que compõem o questionário; aplicou-se o teste de fiabilidade, α de Cronbach; e efetuou-se também uma análise global.

Para fundamentar as hipóteses descritas em 1.3, utilizou-se o método de Análise em Componentes Principais, adaptado a variáveis ordinais "*Categorical Principal Components Analysis / Nonlinear Principal Component Analysis*" (Gifi,1990; Blasius & Gower, 2005).

## 3. Conclusão

Com base no questionário elaborado e aplicado, conseguiu-se atingir o objetivo principal deste estudo – justificar as hipóteses que foram selecionadas para fundamentar os Fatores Chave de sucesso das Equipas Virtuais de Tecnologias de Informação em Regime de Outsourcing, do Ponto de Vista dos Membros da Equipa.

## Referências

BLASIUS, J. & GOWER, J. C. (2005) Multivariate Prediction with Nonlinear Principal Components Analysis: Application. *Quality & Quantity*, 39, 373–390.

BELL, B. S. & Kozlowski, S. W. J. (2002) A typology of virtual teams: Implications for effective leadership, http://digitalcommons.ilr.cornell.edu/hrpubs/8/ (acedido em 22 de Setembro de 2012).

CARMEL, E., TIJA, P. (2005) *Offshoring Information Technology: Sourcing and Outsourcing to a Global Workforce*. Cambridge University Press.

CVITKOVICH, K. (2008). Raising the bar: leading global virtual teams. *Mobility*, November issue, 1-6.

EBRAHIM, N., AHMED, S., RASHID, A., HANIM, S. & TAHA, Z. (2012). Technology Use in the Virtual R&D Teams. *American Journal of Engineering and Applied Sciences,* 5 (1) 9-14.

GIFI, A. (1990) *Nonlinear multivariate analysis*. Chichester, John Wiley & Sons.

**Análise de Dados em Economia e Gestão – Sábado, 12 de Abril, Salão Nobre (10h20)**

# Abordagem exploratória: análise hibrida de indicadores de sustentabilidade empresarial

**Winston Jerónimo[1] e Ana Amaro[2]**

[1]*CENSE, Center for Environmental and Sustainability Research. Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal. Is recipient of PhD fellowship SFRH/BD/35747/2007 from Fundação de Ciência e Tecnologia. winstonjeronimo@gmail.com*

[2]*Instituto Superior de Gestão, anaamaro@isg.pt.*

**Sumário**

A sustentabilidade empresarial é um tema de crescente importância sendo atualmente assumida pelos responsáveis das empresas como um processo de relevância estratégica para a sobrevivência da empresa no médio e longo prazo.Baseando-nos na divulgação realizada pelas empresas pretende-se, com este trabalho, identificar relações de interdependência entre três dimensões: Economia, Ambiente e Sociedade (*Triple Bottom Line* – TBL). Foram analisados os relatórios relativos a 2011 de 85 empresas (modelo de relato proposto pelo Global Reporting Initiative), distribuídas por 36 sectores económicos e com uma dispersão geográfica por 36 países pelos 5 continentes. Para analisar os dados utilizaram-se técnicas estatísticas descritivas, análise de contingência e de variância e ainda análise de correspondências. Devido ao carácter exploratório da análise, assim como a reduzida dimensão de amostra e grande variabilidade observada nos indicadores analisados, todas as técnicas (mesmo às análise de contingência e variância) foram utilizadas na ótica da exploração dos dados e não na da inferência estatística. Concluiu-se que as empresas centram a sua atenção em indicadores "ancora" e por esse motivo encontrámos baixa representatividade na integração e potenciais trade-off entre as dimensões. Encontrámos associações nos binómios Economia-Ambiente e Economia-Sociedade, no primeiro caso centrado no indicador económico Vendas e no segundo Remunerações.

**Palavras-chave:** Desenvolvimento sustentável, Relatórios de responsabilidade social, GRI, Triple Bottom Line, Análise exploratória de dados.

## 1. Introdução

Num mundo em que a dinâmica da população humana continua em crescendo, com a inevitável necessidade de obter mais recursos para fazer face às suas necessidades, cabe às empresas desenvolver e implementar estratégias que permitam a sustentabilidade dos sistemas onde operam *melhorando os seus processos de produção e consumo, assim como o seu relacionamento com as diferentes partes interessadas que com ela interatuam*.

Nos últimos trinta anos tem-se observado um movimento crescente que revindica através de grupos de pressão que as empresas alterem e se adaptem aos princípios propostos pelo desenvolvimento sustentável, que implementem praticas de responsabilidade social em investimento ético (Hubbard, 2011). Tem-se observado que esta mudança tem sido lenta mas

registam-se indícios de consolidação na modificação dos comportamentos, na gestão de riscos e na construção da imagem corporativa (Roca & Searcy, 2011; Holder-Webb *et al*., 2009).

Um dos normativos de relato com mais aceitação entre as empresas para demonstrarem e comunicarem a sua sustentabilidade é o Global Report Initiative (GRI) que surge em 1999 (GRI, 2006; 2011). A adesão a este tipo de relato tem tido um crescimento exponencial e começa a ser aceite como a tradução prática da sustentabilidade empresarial. (Moneva *et al*., 2006). O relato da sustentabilidade é um ato voluntário acarretando custos para a empresa (Branco & Rodrigues, 2006). Este normativo cobre através dos seus indicadores de progresso e avaliação as três dimensões da sustentabilidade "Economia, Ambiente, Sociedade" o denominado (TBL) sendo considerados os pilares da sustentabilidade (Elkington, 1999). Acreditamos que na interdependência das dimensões se potencia efeitos sinérgicos de sustentabilidade.

## 2. Dados e metodologia

As empresas que fazem parte da amostra foram selecionadas do total de empresas que registaram os seus relatórios de sustentabilidade na base de dados do GRI (GRI, 2011) em 2011. As variáveis que caracterizam as empresas são indicadores da sua atividade. Podem ser agrupados em económicos, ambientais e sociais, de acordo com a sua natureza, sendo alguns quantitativos e outros qualitativos.

Dada a grande variabilidade de sectores de atividade das empresas avaliadas, amplitude de variação de muitos dos indicadores utilizados, em associação à reduzida dimensão da amostra, as técnicas estatísticas adotadas para analisar os dados foram utilizadas com carácter exploratório e não de inferência estatística. Neste contexto para avaliar a possibilidade de existirem relações ou associações entre indicadores, dois a dois, efetuaram-se Análises de Contingência e de Variância. As hipóteses de base em análise (traduzidas pela hipótese nula) assumem a inexistência de associação entre os indicadores: Análise de Contingência (os dois indicadores (qualitativos) são independentes; e Análise de Variância (os valores médios do indicador quantitativo são iguais para todas as categorias). Com estas análises pretendeu-se detetar indícios da existência de associações entre dois indicadores de dimensões diferentes pelo que a verificação dos pressupostos destas duas análises não foi uma preocupação. Como critério de decisão consideraram-se interessantes as análises com p-value claramente inferior a um nível de significância de 0,05 (correspondente a um grau de confiança de 95%).

O resultado conduziu a um subconjunto de indicadores submetidos a uma Análise de Correspondências Múltipla para, em conjunto e para cada um dos binómios Economia-Ambiente e Economia-Sociedade, detetar associações de interesse (para albergar indicadores quantitativos efetuou-se a sua classificação). Para este efeito os indicadores quantitativos foram classificados em categorias garantindo a representatividade das mesmas.

O tratamento estatístico dos dados foi realizado utilizando o STATISTICA (2013).

## 3. Resultados

A amostra é constituída por um total de 85 empresas com uma distribuição equilibrada entre multinacionais, grandes empresas e PME, distribuindo-se por 36 sectores de atividades económicas. As empresas estão distribuídas por 36 países pelos 5 continentes.

Foram selecionados 84 indicadores do GRI (9 económicos, 30 ambientais e 45 de âmbito social), identificados os indicadores prevalentes para a análise de integração, assim como determinado o índice de relato das empresas: dos 84 indicadores selecionaram-se apenas 36 em função da frequência e variabilidade de resposta das 85 empresas.

Os resultados obtidos através da análise de contingência e de variância permitiram a identificação de indícios de relações entre indicadores de tipos diferentes (Quadro 1).

**Quadro 1 - Indícios de associação entre indicadores (Economia – Ambiente) e (Economia-Social) decorrentes de análises de contingência e variância**

| | | EC1-1 | EC1-7 | EC1-8 | EC8 |
|---|---|---|---|---|---|
| | | Vendas líquidas | Remunerações | Investimento na comunidade | Impacto dos investimentos em infraestruturas e serviços que visam o benefício público |
| **EN1** | Materiais utilizados | **Sim** | Não | Não | Não |
| **EN3** | Consumo direto de energia | **Sim** | **Sim** | Não | Não |
| **EN4** | Consumo indireto de energia | **Sim** | **Sim** | Não | Não |
| **EN8** | Consumo de água | Não | Não | Não | Não |
| **EN16** | Emissões diretas e indiretas de gases com efeito de estufa | Não | Não | Não | Não |
| **EN22** | Resíduos | Não | **Sim** | Não | Não |
| **EN23** | Presenta derrames | Não | Não | Não | Não |
| **EN28** | Apresenta sanções por incumprimento das leis e regulamentos ambientais | **Sim** | **Sim** | Não | Não |

| | | EC1-1 | EC1-7 | EC1-8 | EC8 |
|---|---|---|---|---|---|
| | | Vendas líquidas | Remunerações | Investimento na comunidade | Impacto dos investimentos em infraestruturas e serviços que visam o benefício público |
| **LA1-H** | Número de homens | **Sim** | **Sim** | **Sim** | **Sim** |
| **LA1-M** | Número de mulheres | **Sim** | **Sim** | **Sim** | **Sim** |
| **LA4** | Trabalhadores abrangidos por contratação coletiva | Não | **Sim** | Não | Não |
| **LA7** | Número de acidentes laborais | Não | **Sim** | Não | Não |
| **LA8** | Programas de formação diversificados | Não | Não | Não | Não |
| **LA12** | Análise de desempenho e desenvolvimento de carreira | Não | Não | Não | Não |
| **HR2** | Parceiros de negócio sujeitos a avaliação | Não | Não | Não | Não |
| **SO1** | Ações em estreita colaboração com a comunidade local, avaliação de impactos e programas de desenvolvimento | Não | Não | Não | Não |
| **SO8** | Coimas e sanções não monetárias por incumprimento de leis e regulamentos | **Sim** | **Sim** | Não | Não |
| **PR4** | Incidentes resultantes da não-conformidade com os regulamentos e códigos voluntários relativos à ção e rotulagem dos produtos e serviços | Não | **Sim** | **Sim** | Não |
| **PR9** | Coimas por incumprimento de leis e regulamentos relativos ao fornecimento e utilização de produtos e serviços | Não | **Sim** | Não | Não |

Estas associações/relações hibridas tem a ver com a porção de uma unidade coerente (sistema – dimensões do TBL) que por efeito de indução de alguns dos seus elementos (subsistema – indicadores) por outra unidade coerente diferente pode potenciar resultados

positivos, negativos, nulos de desempenho e que afetarão ambas as porções dos dois sistemas em grau e intensidade.

A análise de correspondências múltipla efetuada com os indicadores do sistema

- Economia-Ambiente revela que elevados valores para as Vendas líquidas (EC1-1) estão associados a grandes quantidades de materiais utilizados (EN-1), energia consumida (EN-3 e EN-4) e sanções ambientais (EN28).

- Economia-Sociedade revela que elevados valores para as Remunerações (Ec1-7) estão associados a maior proporção de trabalhadores com contratos de trabalho (LA4), mais acidentes laborais (LA7), mais incidentes resultantes de não conformidade (PR4), coimas (PR9) e multas (SO8).

A análise exploratória efetuada revelou-se muito útil tendo algumas limitações, nomeadamente na representatividade das categorias utilizadas para os indicadores o que justifica por ex. a associação aparente entre valores reduzidos para o indicador EN4 com o universo positivo do sistema Economia-Ambiente.

## 4. Conclusões e limitações

A ausência de referências na literatura que ajudem a validar o trabalho efetuado, a grande variabilidade dos indicadores, a sua menos equilibrada qualificação e a incorporação de inúmeros setores de atividade de apenas 85 empresas do mundo são alguns dos fatores que dificultaram a análise efetuada. Por esta razão consideramos estes resultados como preliminares.

A abordagem hibrida só terá lugar se não se verificar uma compartimentação de interesses quási estanque entre as dimensões do TBL que impeça a possibilidade da integração dos seus elementos constituintes. Isto é dever-se-á ver a sustentabilidade não como a materialidade de três dimensões, já que se assim for está-se a promover o interesse competitivos entre as dimensões ao invés de promover a interligação e o trade-off entre as suas dimensões.

## Referências

BRANCO, M & RODRIGUES, L. (2008). Factors influencing social responsibility disclosure by Portuguese companies. *Journal of Business Ethics* 83 (4), 685-701.

ELKINGTON J. (1999). Triple bottom line revolution: reporting for the third millennium, *Australian CPA*, 69: 75.

GRI (2006). Sustainability Reporting Guidelines Version 3.0 (G3). Global Reporting Initiative, Amsterdam. https://www.globalreporting.org/resourcelibrary/G3-Sustainability-Reporting-Guidelines.pdf (acedido em 10 de Outubro 2013).

GRI (2011). Global Reporting Initiative. What is GRI? Available from: http://www.globalreporting.org/AboutGRI/WhatIsGRI/. (acedido em 10 de Janeiro 2014).

STATISTICA (2013). Statsoft (2013), statsoft inc 2013 STATISTICA (data analysis software system), version 12. www.statsoft.com.

HOLDER-WEBB, L., COHEN, J., NATH, L., & WOOD, D. (2009). The Supply of Corporate Social Responsibility Disclosures Among U.S. Firms. *Journal of Business Ethics*, *84*(4), 497–527.

HUBBARD, G. (2011). The Quality of the Sustainability Reports of Large International Companies: An Analysis. *International Journal of Management*, *28*(3 - Part 2), 824–848.

MONEVA, J.M., ARCHEL, P., & CORREA, C., (2006). GRI and the camouflaging of corporate unsustainability. *Accounting Forum* 30, 121-137.

ROCA L.C., & SEARCY C. (2011). An analysis of indicators disclosed in corporate sustainability reports. *Journal of Cleaner Production*, doi: 10.1016/j.jclepro.2011.08.002.

**Análise de Dados em Economia e Gestão – Sábado, 12 de Abril, Salão Nobre (10h40)**

# Resultados de uma sondagem CATI móvel

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

**Sumário**

Na perspectiva da metodologia das sondagens o telemóvel é uma novidade e quer as suas potencialidades quer as suas fragilidades enquanto instrumento de sondagem são ainda insuficientemente compreendidas. A tecnologia de entrevistas telefónicas assistidas por computador (CATI) desenvolvida para os telefones fixos pode ser utilizada com telemóveis, mas importa avaliar a eficiência dos protocolos de marcação de contactos quando estes são aplicados a telemóveis. Esta apresentação descreve os resultados de uma sondagem CATI móvel e avalia a influência do horário e da repetição dos contactos naqueles resultados.

**Palavras-chave:** Entrevistas, Não contactos, Recusas, Sondagem CATI móvel

## 1. Introdução

O rápido crescimento da posse de telemóvel abriu novas oportunidades à recolha de dados. Os telemóveis são cada vez mais usados para realizar sondagens e é muito provável que a tendência de crescimento se mantenha (COUPER, 2011). Neste cenário torna-se relevante investigar os aspectos metodológicos associados ao telemóvel enquanto modo de sondagem.

Um dos problemas mais prementes das sondagens prende-se com a não resposta, causada quer pela dificuldade em contactar as pessoas quer pela dificuldade em convencê-las a participar nas sondagens. Este problema é transversal a qualquer modo de sondagem sendo provável que se verifique também nas sondagens por telemóvel. O utilizador do telemóvel, apesar de ter permanentemente o telemóvel consigo, pode desligá-lo ou colocá-lo em modo de silêncio, pode encontrar-se em zonas de fraca cobertura de rede ou pode não atender chamadas provenientes de números desconhecidos. Todas estas circunstâncias são desfavoráveis a um contacto bem sucedido que resulte na colaboração efectiva na sondagem.

A investigação sobre o problema da não resposta nas sondagens CATI (tradicionalmente realizadas por telefone fixo) revela que o protocolo de contacto com os números de telefone, i.e. o horário de contacto, o dia de contacto, o número de tentativas de contacto, tem implicações no resultado do contacto – entrevista, recusa ou não contacto (e.g. BRICK et al., 1991, HANSEN, 2008). Esta apresentação avalia se este mesmo efeito existe numa sondagem CATI móvel.

## 2. Dados

A base para analisar os protocolos de contacto é uma sondagem CATI móvel realizada em Portugal em 2012. Os números de telemóvel a contactar foram gerados aleatoriamente tendo a amostra sido estratificada por operador de serviço móvel.

Para concretizar o objectivo de obter 1500 entrevistas completas a utilizadores de telemóvel com 15 ou mais anos de idade, marcaram-se 11472 números de telemóvel. Verificou-se que 4110 desses números eram não elegíveis por serem não atribuídos ou desligados/desactivados.

Para cada número de telemóvel marcado conhece-se o número de tentativas de contacto, o resultado de cada tentativa de contacto, o dia e hora de cada tentativa. Para efeitos de análise os dias e horários de contacto são condensados em cinco categorias: 2ª a 6ª 17h-19h, 2ª a 6ª 19h-21h, 2ª a 6ª 21h-22h, fim de semana 10h-12h e fim de semana 12h-14h.

A análise é limitada às tentativas para estabelecer contacto com o número de telemóvel, i.e., às tentativas de contacto prévias a conseguir-se falar com o utilizador do telemóvel. Assim, os números não elegíveis ou relativos a indivíduos com menos de 15 anos são excluídos da análise. Os primeiros contactos são os mais críticos para o sucesso de qualquer sondagem e grande parte dos estudos sobre protocolos de contacto focaliza a atenção nas primeiras tentativas de contacto (e.g. BRICK et al., 1991, HANSEN, 2008). Neste seguimento, restringe-se a análise à primeira e segunda tentativa de contacto. Concretamente avalia-se o padrão horário das tentativas de contacto, a taxa de entrevista, de recusa e de não contacto por período horário e o padrão horário associado a cada resultado – entrevista, recusa e não contacto.

## 3. Resultados

Os resultados revelam que a distribuição dos números marcados por período horário é, em termos globais, semelhante na primeira e na segunda tentativa de contacto, sendo o período 2ª a 6ª 19h-21h aquele onde maior percentagem de números de telemóvel se marcaram, e o período fim de semana 10h-12h aquele com menor percentagem de números marcados. Apesar da semelhança entre a primeira e segunda tentativas na distribuição global dos números por horário verifica-se, entre os números que resultaram em "não contacto" após a primeira tentativa de contacto uma tendência de re-contacto em horário diferente do da primeira tentativa.

A taxa de não contactos foi superior à taxa de recusas na primeira e segunda tentativas de contacto, quer globalmente quer quando se faz a análise por período horário.

Uma análise de regressão logística revela um efeito significativo do período horário sobre a probabilidade de não contacto e sobre a probabilidade de recusa. Este efeito verifica-se quer

na 1ª tentativa de contacto quer na 2ª, embora com um padrão diferente. O efeito do período horário em cada um dos resultados é também diferente.

Na 1ª tentativa de contacto a probabilidade de não contacto é menor nos fins de semana entre as 10.00 e as 12.00 e é maior nos dias de semana entre as 17.00 e as 19.00. Quanto às recusas, acontecem com maior probabilidade nos fins de semana entre as 10.00 e as 12.00.

Na 2ª tentativa de contacto a probabilidade de obter um não contacto é menor nos fins de semana entre as 12.00 e as 14.00. A probabilidade de recusa, porém, é constante em todos os períodos horários.

**Agradecimentos:** Fundação para a Ciência e Tecnologia, PTDC/EGE-GES/116934/2010.

## Referências

BRICK, J., ALLEN, B., CUNNINHAM, P. & MAKLAN, D. (1991). Outcomes of a calling protocol in a telephone survey, *Proceedings of the American Association for Public Opinion Research Conference.* AAPOR, 142-149.

COUPER, M. (2011). The future of modes of data collection. *Public Opinion Quarterly*, 75, 889-908.

HANSEN, S. (2008). CATI sample management systems. IN LEPKOWSKI, J., TUCKER, C., BRICK, J, DE LEEUW, E., JAPEC, L., LAVRAKAS, P., LINK, M. & SANGSTER, R. (Eds) *Advances in telephone survey methodology*. New Jersey, Wiley.

**Análise de Dados em Economia e Gestão – Sábado, 12 de Abril, Salão Nobre (11h00)**

# Discriminant analysis of interval data: parametric versus distance-based approaches

**A. Pedro Duarte Silva[1], Paula Brito[2]**

[1]*Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa (Porto), Porto, Portugal, psilva@porto.ucp.pt*

[2]*Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Porto, Portugal, mpbrito@fep.up.pt*

**Abstract**

Building on probabilistic models for interval-valued variables, parametric classification rules, based on Normal or Skew-Normal distributions, are derived for interval data. The performance of such rules is then compared with distance-based methods previously investigated. The results show that parametric approaches generally outperform distance-based ones, and that restricted cases of the variance-covariance matrix which take into account the particular nature of interval data lead to parsimonious rules, which are sometimes quite effective in reducing expected error rates.

**Keywords** Discriminant analysis, Interval data, Parametric modelling of interval data

## 1. Discriminant approaches for interval data

In this paper, we are interested in the analysis of interval data, i.e., where elements are characterized by variables whose values are intervals of R, and investigate and compare different methods for discriminant analysis of such data.

Distance-based approaches to linear discriminant analysis of interval data are discussed in Duarte Silva & Brito (2006). Three fundamental approaches are considered. The first approach assumes an uniform distribution in each observed interval, derives the corresponding measures of dispersion and association, and appropriately defines linear combinations of interval variables that maximize the usual discriminant criterion; the second approach expands the original data set into the set of all interval description vertices, and proceeds with a classical analysis of the expanded set; finally, a third approach replaces each interval by a midpoint and range representation. These approaches lead to representations in the discriminant space in the form of intervals or single points, from which distance-based allocation rules are derived. In Brito & Duarte Silva (2012), a parametric modelling for interval data, assuming multivariate Normal or Skew-Normal (Azzalini, & Capitanio (1999)) distributions for the Midpoints and Log-Ranges of the interval variables, is proposed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, represented by different possible configurations. This approach is implemented in an R package, MAINT.DATA (Duarte Silva & Brito (2011)), available at the CRAN repository, which includes several tools for modelling and analysing interval data. In particular MAINT.DATA introduces a data class for representing interval data and provides methods and functions for parameter estimation,

statistical tests for the different covariance configurations, and parametric Discriminant Analysis.

Discriminant analysis of interval data has been investigated by other authors in different contexts. Ishibuchi, Tanaka and Noriko Fukuoka (1990) address discriminant analysis of interval data determining interval representations in a discriminant space using a mathematical programming formulation. Approaches of discriminant analysis of interval data based on imprecise probability theory may be found in Nivlet, Fournier & Royer (2001) and Utkin & Coolen (2011). In Lauro, Verde & Palumbo (2000), a generalization of classical Factorial Discriminant Analysis to symbolic data is proposed. This method is based on a numerical analysis of the transformed symbolic data, followed by a symbolic interpretation of the results; it allows considering quantitative, qualitative nominal or distributional variables; classification rules are then based on proximities in the factorial plane (see also Lauro, Verde & Irpino (2008)).

## 2. Comparison of classification methodologies

This paper evaluates the relative performance of different classification rules for interval data. It compares the distance-based classification rules considered in Duarte Silva & Brito (2006), the parametric classification rules derived from the models discussed in Brito & Duarte Silva (2012), and Factorial Discriminant Analysis (Lauro, Verde & Palumbo (2000)).

The comparisons rely on cross-validated classification rates of a real data set of temperatures in meteorological stations in China, with four interval variables and 899 observations, and on a controlled experiment with simulated data. As concerns the China dataset, the primary available data (obtained from the University Corporation for Atmospheric Research - UCAR) consists in monthly maxima and minima temperatures for the different stations, those have been aggregated for the four trimesters (Jan-Mar; Apr-Jun, Jul-Sep, Oct-Dec), leading in four interval variables. Twenty five different discriminant methods are applied to the resulting dataset, namely: nine distance-based approaches, Factorial Discriminant Analysis (FCA) considering single, average and complete linkage allocations, and sixteen parametric-based approaches, eight using the Gaussian model - Linear and Quadratic Discriminant Analysis, and eight using Skew-Normal Discriminant Analysis - Location and General model – both with 4 different configurations for the variance-covariance matrix (see Brito & Duarte Silva (2012)).

The simulation experiment uses a full factorial design for problems with two groups, three interval variables, and the following seven factors:

- Classification method (22 methods): all the methods compared in the China temperature data, except for the three Factorial Discriminant Analysis methods.
- Data Generating Process (2 levels): MidPoints generated by transformations using Gaussian and Skew-Normal variables.

- Separation (2 levels): Good and bad separation between group centroids.
- Range heterogeneity (2 levels): Same or group-specific distribution, used in the generation of Log-Ranges.
- Training sample size (4 levels): Total number of training sample observations, set at 30, 60, 100 and 150.
- Variance ratios (2 levels): Homocedastic and heterocedastic problems.
- True Variance-Covariance configuration (4 levels): Configuration of the population covariance used in the data generation. The same cases as assumed by the parametric methods under comparison.

The results show that parametric approaches generally outperform other approaches, and that restricted configurations of the variance-covariance matrix which take into account the particular nature of interval data lead to parsimonious rules, which can be quite effective in reducing expected error rates. Furthermore, the Gaussian parametric methods proved more reliable than the Skew-normal based methods, particularly in the cases with small training samples where the latter methods appear to suffer from estimation difficulties. With large samples, corresponding Gaussian and Skew-Normal based methods tend to produce similar results, but the Skew-Normal parametric methods are never clearly superior to corresponding Gaussian methods, even when the data generation is based on the Skew-Normal distribution.

**Acknowledgements:** This research is supported by the Project NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT) through the project PEst-OE/EGE/UI0731/2011.

## References

AZZALINI, A. & CAPITANIO, A. (1999). Statistical applications of the multivariate Skew-Normal distribution. *J. R. Statist. Soc. B*, 61 (3), 579–602.

BRITO, P. & DUARTE SILVA, A.P. (2012). Modelling interval data with Normal & Skew-Normal distributions. *Journal of Applied Statistics*, 39 (1), 3-20.

DUARTE SILVA, A.P. & BRITO, P. (2006). Linear discriminant analysis for interval data. *Computational Statistics*, 21 (2), 289-308.

DUARTE SILVA, A.P. & BRITO, P. (2011). MAINT.DATA: Model and Analyze Interval Data. R Package, version 0.2. Available at:
http://cran.r-project.org/web/packages/MAINT.Data/index.html.

ISHIBUCHI, H., TANAKA, H. & FUKUOKA, N. (1990). Discriminant analysis of multi-dimensional interval data & its application to chemical sensing. *International Journal of General Systems*, 16 (4), 311-329.

LAURO, N.C., VERDE, R. & PALUMBO, F. (2000). Factorial discriminant analysis on symbolic objects. IN: BOCK, H.-H. & DIDAY, E. (Eds.), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg, 212-233.

LAURO, N.C., VERDE, R. & IRPINO, A. (2008). Factorial discriminant analysis. IN DIDAY, E. & NOIRHOMME-FRAITURE, M. (Eds.), *Symbolic Data Analysis & the Sodas Software*. Wiley, Chichester, 341-358.

NIVLET, P., FOURNIER, F. & ROYER, J.J. (2001). Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. IN I*SIPTA'01*, 284-292.

UTKIN, L.V. & COOLEN, F.P.A. (2011). Interval-valued regression & classification models in the framework of machine learning. IN *Proc. 7th International Symposium on Imprecise Probability: Theories & Applications*. Innsbruck, Austria.

**Data Mining – Sábado, 12 de Abril, Sala 316 (10h00)**

# Agrupamento sobre uma matriz de distâncias UMAT – uma aplicação sobre dados financeiros

**Diogo Matos[1], Nuno C. Marques[2], Margarida G. M. S. Cardoso[3]**
[1]*DI-FCT/UNL, diogomatos38@gmail.com*
[2]*CITI & DI-FCT/UNL, nmm@fct.unl.pt (autor para envio de correspondência)*
[3]*ISCTE-IUL e UNIDE, margarida.cardoso@iscte.pt*

**Sumário**

Neste trabalho realiza-se uma Análise de Agrupamento sobre dados de uma matriz de distâncias UMAT, obtida mediante o algoritmo SOM - *Self Organizing Maps*. Os agrupamentos decorrem de vários níveis de *inundação* da UMAT, selecionando-se o que exibe melhor valor para o índice de Calinski e Harabasz. Após o ensaio do método proposto sobre dois conjuntos de dados do UCI *Machine Learning Repository*, este é depois aplicado para uma análise sistemática do comportamento conjunto de 11 empresas no índice S&P500, identificando-se distintos estados no mercado.

**Palavras-chave:** Análise de Agrupamento, Mapas Auto-Organizados, Mercados Financeiros.

## 1. Introdução

No âmbito das ciências da computação têm-se vindo a desenvolver modelos não paramétricos, com capacidade de descobrir padrões em grandes bases de dados, nomeadamente dados económicos e financeiros – v. (Corsetti et al., 2001), (Sarlin and Peltonen, 2013), por exemplo.

O algoritmo de (Kohonen, 1982) é um método não paramétrico para a obtenção de mapas auto-organizados (acrónimo em Inglês, SOM de *Self-Organizing Map*) que tem a capacidade de organizar dados multivariados, sendo capaz de diminuir a sua dimensionalidade, mantendo a representação das propriedades relevantes dos vetores de entrada e resultando num mapa que representa topograficamente as características do espaço de entrada. Este algoritmo tem sido aplicado no apoio à decisão em Economia e Finanças – v. (Sarlin and Peltonen, 2013) e (Chen et al., 2013), por exemplo.

A nossa proposta parte dos resultados do SOM numa matriz de distâncias UMAT- acrónimo em Inglês *Unified Distance Matrix* (Ultsch, 1993, Ultsch et. al., 1990). O agrupamento recorre a modelos de *inundação* – v. (Bond, 2011), por exemplo – e para a determinação do número de grupos apoia-se na medida de (Caliński and Harabasz, 1974). A metodologia possibilita uma análise detalhada das tendências de evolução de um conjunto de ativos financeiros, agrupado com base na sua semelhança relativa de variação.

## 2. Metodologia proposta

Neste trabalho apresenta-se um método de agrupamento baseado no pós-processamento dos micro-grupos decorrentes da aplicação do SOM, tendo por base as correspondentes distâncias na UMAT (Ultsch, 1993). Viabiliza-se, nesta proposta, um agrupamento totalmente automático a partir dos resultados do SOM utilizando um algoritmo em três fases: **1.** identificação dos mínimos locais na UMAT; **2.** utilização de um processo de inundação para identificar um agrupamento (com critério de paragem baseado na distância média da UMAT); **3.** Afetação aos grupos de dados não agregados em **2.** utilizando o critério do vizinho mais próximo. O critério de paragem em **2.** é objeto de uma análise de sensibilidade: considerando o valor médio da UMAT como referência (valor reduzido à unidade) e testando valores acima e abaixo da unidade como valores possíveis de paragem. A medida CH de (Caliński and Harabasz, 1974) é utilizada para testar a bondade das sucessivas soluções de agrupamento: $CH(\Pi^K) = [\mathrm{B}(\Pi^K)/(K-1)]/[\mathrm{W}(\Pi^K)/(n-K)]$ em que $\Pi^K$ indica uma partição de $n$ observações em $K$ grupos e $B$ e $W$ se referem à variação entre (*between*) e intra (*Within*) grupos, respetivamente.

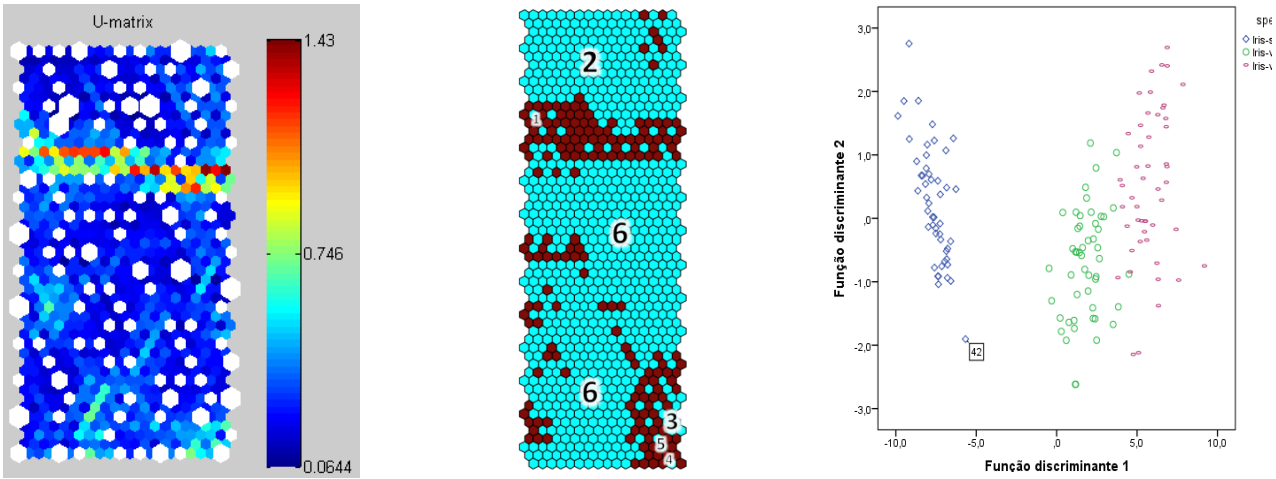
## 3. Análise e resultados

O método é primeiramente ensaiado nos dados *Iris* e *Wine* de UCI *Machine Learning Repository* - (Bache and Lichman, 2013). Os resultados obtidos sobre os dados *Iris* apresentam-se na Ilustração 1. Na análise de CH - v. Ilustração 2 - escolhe-se o valor máximo, que corresponde a um nível de inundação de 1.4 e à constituição de 6 grupos. Note-se que a solução *a priori* com 3 grupos (3 espécies de flores em Iris) corresponde ao nível 1.7 de inundação. De facto, o processo adotado agrupa as espécies *versicolor* e *virginica* num só grupo, excetuando 9 casos *virginica* que constituem 3 micro-grupos. O caso 42 isola-se num agrupamento singular o que decorre da sua especificidade face aos restantes – v. a representação clássica associada a duas funções discriminantes de Fisher na Ilustração 1 c).

Nos dados *Wine* o estudo aponta para a constituição de 27 grupos num nível de inundação 0.9 . Verificaram-se resultados pouco parcimoniosos, indo de encontro à especificidade dos dados.

A aplicação do método ilustra-se em dados reais referentes a 11 componentes do índice *S&P500* que se consideram mais correlacionados com a crise financeira de 2008.

A metodologia proposta consegue identificar zonas em que as variações de preços apresentam uma tendência semelhante e zonas onde esse comportamento é distinto (v. Ilustração 3). São propostos 47 grupos para um nível de inundação 1.2.




a) b) c)

**Ilustração 1 – Dados Iris: a) UMAT, b) 6 "lagos"/agrupamentos, c) funções discriminantes lineares**

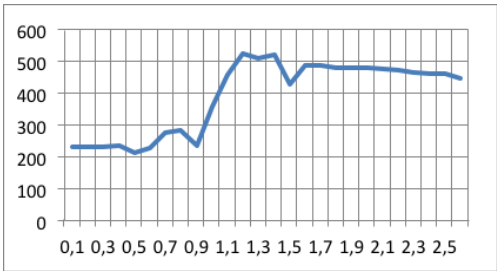
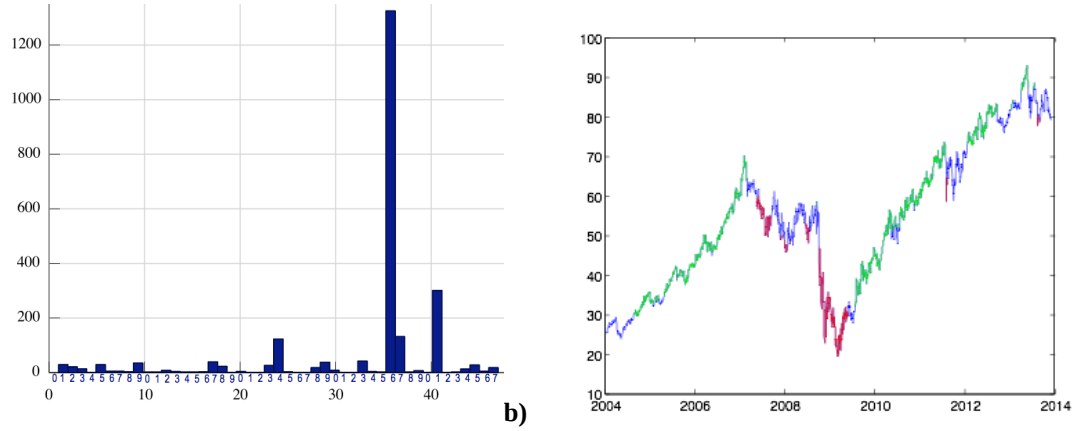



**Ilustração 2 – Dados Iris: Nível de inundação vs medida CH.**




**a)** **b)**

**Ilustração 3 – Dados financeiros: a) Gráfico de frequências por agrupamento. b) dois principais "lagos" projetados a vermelho e verde no preço médio (a azul).**

## 4. Discussão e perspetivas

A partir dos micro-grupos sumarizados na UMAT do SOM viabiliza-se a análise automática de agrupamento recorrendo a um indicador de qualidade de agrupamento - o índice

de Calinski e Harabasz. De acordo com os resultados obtidos as soluções são pouco parcimoniosas. Pode, contudo fixar-se um compromisso qualidade-complexidade que, em cada caso, guie a constituição de um número adequado de grupos.

A aplicação a dados reais confirma o interesse da técnica de agrupamento, nomeadamente por identificar um caso usual em que o mercado está estável e com padrão de subida constante com mais de 60% dos dados e um padrão comum de descida em perto de 15% das observações. Os restantes 25% constituem-se como momentos de forte instabilidade, devendo ser efectuada uma análise detalhada relativamente a comportamentos distintos entre as séries analisadas.

No futuro perspetiva-se o uso de novos índices de qualidade de agrupamento como guias de agrupamento automático a partir dos resultados do SOM. Realça-se que os micro-grupos iniciais do SOM se tornam particularmente relevantes num contexto de extração de dados sobre fluxos de dados contínuos (Silva, et. al., 2012), possibilitando-se a classificação imediata da variação de preços segundo classes como *normal* e *problemático*, sendo que os períodos normais apresentam maior segurança de investimento e os períodos problemáticos maior necessidade de atenção a componentes com variação complexa.

**Agradecimentos:** Os autores agradecem aos projectos *GoBusiness Finance* e *InspireBiz* (QREN: 18627/2011) pelos dados fornecidos e ajuda à aplicação do processo a dados reais.

**Referências**

BACHE, K. & LICHMAN, M. (2013). *UCI Machine Learning Repository* [Online]. Irvine, CA:University of California, School of Information and Computer Science. Available: http://archive.ics.uci.edu/ml.

BOND, C. (2011). An Efficient and Versatile Flood Fill Algorithm for Raster Scan Displays.

CALIŃSKI, T. & HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3**,** 1-27.

SILVA B. & MARQUES N.C. (2012). Applying Neural Networks For Concept Drift Detection In Financial Markets. *CEUR Workshop Proceedings*, 2012.

CHEN, N., RIBEIRO, B., VIEIRA, A. & CHEN, A. (2013). Clustering and visualization of bankruptcy trajectory using self-organizing map. *Expert Systems with Applications,* 40**,** 385-393.

CORSETTI, G., PERICOLI, M. & SBRACIA, M. (2001). Correlation analysis of financial contagion: what one should know before running a test. *Yale Economic Growth Center Discussion Paper*.

KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. 43.

SARLIN, P. & PELTONEN, T. A. (2013). Mapping the state of financial stability. *Journal of International Financial Markets, Institutions and Money,* 26**,** 46-76.

ULTSCH, A. & Semon, (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. *Proceedings of the International Neural Network Conference (INNC-90).*

ULTSCH, A. (1993). Self-organizing neural networks for visualisation and classification. *Information and classification.* Springer.

**Data Mining – Sábado, 12 de Abril, Sala 316 (10h20)**

# Análise de tendências políticas no Twitter para previsão de sondagens

**Luís Gomes[1], Pedro Saleiro[2], Carlos Soares[3]**

[1]*Labs Sapo UP / Faculdade de Engenharia da Universidade do Porto, lfc.gomes@fe.up.pt;*
[2]*Labs Sapo UP / Faculdade de Engenharia da Universidade do Porto, pssc@fe.up.pt;*
[3]*INESC TEC / Faculdade de Engenharia da Universidade do Porto, csoares@fe.up.pt*

**Sumário**

Neste projeto estudamos o problema de estimar o resultado de sondagens políticas com base na combinação de vários agregadores de *buzz* e sentimento obtidos de mensagens do Twitter. Tivemos acesso a mensagens do Twitter (*tweets*) de mais de 100 mil utilizadores classificados como sendo portugueses no período de Junho de 2011 a Dezembro de 2013. Os *tweets* foram filtrados de forma a utilizarmos na nossa análise apenas as mensagens que mencionam explicitamente o nome dos líderes dos cinco principais partidos portugueses. Estes *tweets* foram submetidos a uma classificação de polaridade (positivo, negativo, neutro) do sentimento em relação à respectiva menção ao líder partidário. Desta forma foi possível processar vários indicadores de frequência e polaridade que serviram como variáveis independentes dos nossos modelos de regressão. Utilizámos as sondagens da empresa Eurosondagem como variável dependente tanto para treinarmos o nosso modelo como para posterior avaliação. Os melhores resultados obtidos foram obtidos utilizando o algoritmo *Random Forests* e correspondem a um desvio médio de 0,50 em relação ao valor registado nas sondagens para o período de teste (Janeiro a Junho de 2013).

**Palavras-chave:** *Data Mining, Machine Learning, Regressão, Mídia Sociais.*

## 1. Introdução

Atualmente, as sondagens telefónicas tradicionais são o método mais utlizado para recolha de opinião pública. Uma das vantagens das sondagens é permitir uma escolha aleatória da amostra do público-alvo, evitando assim uma escolha enviesada. No entanto, são dispendiosas – cada chamada telefónica tem um custo associado – e demoram tempo a ser realizadas – tempo de recolha e processamento dos dados. Para além disso, tem-se verificado uma diminuição da taxa de resposta das sondagens, o que sugere que cada vez se torna mais difícil fazer com que as pessoas participem neste tipo de questionários.

Por outro lado, com o crescimento das redes sociais, nomeadamente Twitter e Facebook, o contacto das pessoas com as notícias modificou-se. Assim, torna-se possível a qualquer pessoa emitir livremente um parecer sobre determinada notícia em tempo real. Um dos desafios a que vários estudos tentam dar resposta é o de tentar perceber em que medida podem essas opiniões expressas nas redes sociais, e o sentimento a elas inerentes, serem um indicador de opinião pública. No entanto, em cada momento há simultaneamente opiniões negativas e positivas. Assim, e existindo a necessidade de obter um valor global que reflita a imagem de cada alvo

político nas redes sociais, num determinado período, utilizamos agregadores de sentimento. Resumidamente, um agregador de sentimento é uma fórmula matemática com base no número de menções positivas, negativas e neutras de cada um dos alvos políticos. Após um estudo exaustivo de vários trabalhos relacionados, foram recolhidos e implementados 25 agregadores de sentimento.

Assim, o objectivo deste trabalho é estudar e definir uma metodologia que nos permita estimar com sucesso o resultado das sondagens tradicionais, com base nas opiniões expressas nas redes sociais utilizando, representadas por agregadores de sentimento.

No entanto, e dada a periodicidade mensal das sondagens, surgiu a necessidade de agregar mensalmente os dados de forma a que o valor de cada agregador representasse um valor global mensal de cada alvo político. Transpusemos este problema para o caso de estudo português, utilizando dados políticos portugueses.

## 2. Conjunto de dados

Centrámos os nossos esforços nas opiniões expressas na rede social Twitter. O projeto POPSTAR, que recolhe e classifica mensagens do Twitter de acordo com a sua polaridade, forneceu-nos um conjunto de dados contendo o número diário de menções positivas, negativas e neutras para cada um dos cinco principais alvos políticos – líderes dos cinco principais partidos portugueses. Estes dados são referentes ao período compreendido entre Agosto de 2011 e Dezembro de 2013. As mensagens foram recolhidas de 100 000 utilizadores diferentes, classificados como portugueses.

No que diz respeito às sondagens, realizadas pela empresa Eurosondagem, foram-nos disponibilizados pelo Instituto de Ciências Sociais da Universidade de Lisboa os resultados mensais dos cinco principais partidos portugueses, desde Junho de 2011 a Dezembro de 2013.

## 3. Metodologia

Foi desenvolvida uma plataforma capaz de recolher e classificar as mensagens do Twitter. Para isso, foram escolhidos 1000 utilizadores portugueses. O próximo passo foi expandir a rede para os seus seguidores e para os utilizadores que os próprios seguiam, que escrevessem em português e que fossem utilizadores regulares da rede social. Esta rede foi expandida até um total de 100 000 utilizadores. Para além disso, a ferramenta desenvolvida também classifica as mensagens de acordo com a sua polaridade. Para isso, os modelos de classificação de sentimento utilizam como dados de treino uma amostra de 1500 *tweets* anotados manualmente por 3 utilizadores diferentes. Assim, após a recolha e classificação, foi possível termos acesso à contagem diária de menções positivas, negativas e neutras para cada um dos líderes cinco principais partidos políticos Portugueses. Para além disso, tivemos também acesso aos

respectivos resultados das sondagens. No entanto, e dada a periodicidade mensal das sondagens de opinião pública, surgiu a necessidade de agregar mensalmente essa contagem. Assim, seria possível aplicar os agregadores de sentimento às contagens mensais, de modos a que cada agregador representasse um valor global para cada mês, por alvo político. Utilizamos dois algoritmos de regressão: um algoritmo de regressão linear (*Ordinary Least Squares* – OLS) e um algoritmo de regressão não linear (*Random Forests* – RF). Os valores mensais absolutos dos agregadores de sentimento são usados como variáveis independentes do nosso modelo de regressão para prevermos o valor absoluto das sondagens. No entanto, e visto que existe pouca variação nos resultados das sondagens reais entre dois meses consecutivos, optámos por prever antes essa variação. Assim sendo, nesta abordagem em vez de utilizarmos os valores absolutos dos agregadores mensais de sentimento, utilizámos também as variações em relação ao mês anterior. Para além disso, replicamos cada uma das experiências incluindo e excluindo o resultado da sondagem do mês anterior como variável independente ($y_{t-1}$) no modelo de regressão.

No nosso caso de estudo tentamos prever o resultado das sondagens para o primeiro semestre de 2013. Para estimarmos o desempenho dos modelos de regressão, utilizámos uma técnica de *sliding window*:

- **Conjunto de treino** – contém o valor mensal dos agregadores dos 16 meses anteriores ao mês que pretendemos prever
- **Conjunto de teste** – contém o valor dos agregadores do mês que pretendemos prever.

Mantemos a janela temporal fixa de 16 meses anteriores ao mês que tentamos prever como conjunto de treino, e um período de teste de Janeiro a Junho de 2013. Após o cálculo da previsão de cada um dos meses, utilizamos como medida de avaliação o *Mean Absolute Error* (MAE), calculando o erro de previsão de cada mês. Posteriormente, fazemos a média dos seis MAE's, obtendo assim o erro de previsão global de cada modelo. Utilizámos como modelo de referência para a comparação uma *baseline* ingénua bastante utilizada em problemas de previsão de resultados: prever que o resultado da sondagem deste mês é igual ao do mês anterior, não havendo assim qualquer variação entre os dois meses consecutivos.

## 4. Resultados

Dada a pouca variação existente entre dois meses consecutivos, o modelo de referência mostrou ter um poder preditivo elevado, mostrando não ser tão ingénua quanto inicialmente se esperava. O erro global desse modelo de referência é de 0,56. Isto fez com que a tarefa de tentar estimar o resultado das sondagens com um erro global significativamente inferior à *baseline* se tornasse difícil.

Os modelos que usam os valores absolutos apresentam, de um modo geral, um erro global superior aos modelos que usam os valores das variações dos agregadores de sentimento como indicador para estimar o resultado das sondagens. O algoritmo de regressão linear (OLS)

apresenta um erro global de previsão maior do que o algoritmo de regressão não linear (RF). Os modelos que usam apenas os agregadores de *buzz* como variável independente têm um MAE inferior aos modelos que usam os agregadores de sentimento. O modelo com o erro mais baixo que conseguimos obter foi usando o algoritmo de regressão não linear (RF), utilizando como variável independente as variações dos agregadores de sentimento, e incluindo a variação da sondagem anterior de todos os candidatos. Este modelo obteve um erro global de 0,50.

Os resultados mostram que podemos estimar o resultado das sondagens com um erro de previsão reduzido. No entanto, é de salientar o facto de todas as melhorias referidas anteriormente serem muito pouco significativas, rondando a casa das centésimas.

**Referências**

ASUR, S. & HUBERMAN, B. (2010). Predicting the Future with Social Media, *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492-499.

BERMINGHAM, A. & SMEATON, A. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results, *Workshop at the International Joint Conference for Natural Language Processing (IJCNLP),* 13th November 2011, Chiang Mai, Thailand.

CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE B. & SMITH, N.(2010). From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series, *Proceedings of the Fourth INternational AAAI Conference on Weblogs and Social Media.*

HAN, J. & KAMBER, M. (2006) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

METAXAS, P., MUSTAFARAJ, E. & GAYO-AVELLO, D. (2011) How (not) to Predict Elections, *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 165-171.

TUMASJAN, A., SPRENGER, T., SANDNER, P. & WELPE, I. (2010) Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape, *2010 Social Science Computer Review, 29(4):*402-418.

WITTEN, I., FRANK, E. & HALL, M. (2011) *Data Mining. Practical Machine Learning Tools and Techniques.*

**Data Mining – Sábado, 12 de Abril, Sala 316 (10h40)**

# Análise de Afinidades entre Investigadores com Text Mining

**Luís Trigo, Pavel Brazdil**
*LIAAD INESCTec, {lptrigo,pbrazdil}@inescporto.pt*

**Sumário**

Localizar competências de pessoas dentro de um domínio pode ser um poderoso auxiliar na gestão de centros de investigação. A produção académica, embora num formato não estruturado, é facilmente acessível. Deste modo, recorre-se a fontes na web – instituições académicas e bases de dados bibliográficas – para visualizar as afinidades na produção académica como ponto de partida para a descoberta dos investigadores mais centrais em cada domínio e dos que ligam diferentes domínios, para além das afinidades não conhecidas entre os mesmos.

**Palavras-chave:** Agrupamento, Análise de Redes Sociais, Information Retrieval, Web Mining.

## 1. Introdução

Os investigadores procuram descobrir outros investigadores com interesses semelhantes para acompanharem o seu trabalho e perspectivarem futuras colaborações. Ao nível da gestão, permite perspectivar a localização dos investigadores num domínio de modo a auxiliar a implementação de medidas com impacto na implementação de parcerias com outras instituições e investigadores, identificando os indivíduos com mais afinidades. Outra das vantagens desta análise para o gestor é o facto de ir além da organização hierárquica formal dentro da própria organização, desvendando deste modo as suas ligações desconhecidas.

Como enunciam Goldstone e Rogosky (2002), no contexto de uma rede conceptual, o significado de um conceito - no caso deste trabalho, um investigador (mas também pode ser uma unidade terminológica) depende da sua relação com outros conceitos integrados no mesmo sistema, i.e. domínio de conhecimento. Deste modo, embora se utilizem as mesmas metodologias da análise de redes sociais, analisam-se redes de afinidades que ultrapassam a co-autoria.

O recurso ao cruzamento de fontes na *Web* coloca como primeiro problema a existência de vários investigadores com o mesmo nome e, com menos frequência, a identificação do mesmo investigador com vários nomes na base de dados bibliográfica. Trata-se de um clássico problema de reconhecimento de *entidades mencionadas*. Uma das técnicas utilizadas por Bugla (2009) parte da verificação de similaridades e agrupamento na instituição de origem. O trabalho pressupõe que o nome a desambiguar corresponda ao autor de publicações de maior semelhança e da mesma instituição.

No que concerne à descoberta de semelhanças entre investigadores, Price et al. (2010) desenvolveram uma metodologia para a web, denominada *SubSift*, que cria perfis para

investigadores e suas organizações a partir das suas publicações. Com base nestes perfis perspectivam uma tarefa típica de *Information Retrieval* que consiste em comparar os artigos submetidos a um congresso científico com o perfil dos avaliadores, de modo a optimizar a tarefa de distribuição dos artigos para avaliação.

## 2. Metodologia

Nesta secção procede-se à descrição do fluxo de tarefas empreendidas para completar a análise de afinidades. Tratando-se ainda de uma aplicação experimental, optou-se por escolher duas unidades de investigação do INESC-TEC relativamente próximas – o LIAAD e o CRACS. Por se ser uma base de dados bibliográfica aberta e abrangente na área da ciência da computação, o DBLP foi escolhido. Relativamente à ferramenta para extrair e tratar os dados, optou-se por utilizar o R com o seu pacote *'tm'* para a parte de *text mining*, o pacote *'XML'* para a parte de *web mining* e os pacotes *'igraph'* e *'sna'* para a parte de agrupamento e de análise de rede social.

### 2.1. Do web mining à matriz de semelhanças

A tarefa descrita nesta secção foi alvo de publicação numa edição anterior do JOCLAD (Trigo e Brazdil, 2010), mas teve entretanto a sua abordagem actualizada. As duas entidades de investigação possuem uma página com a listagem dos seus investigadores no seu website. As listas de investigadores serão facilmente extraídas, construindo uma expressão na linguagem de consulta *XPath* que permita extrair os nomes das tabelas de investigadores. De seguida, introduz-se o nome de cada um dos investigadores no URL de pesquisa do DBLP para aceder directamente à sua listagem das publicações. Ainda recorrendo ao *XPath*, extraem-se os títulos das publicações de cada um dos investigadores para um ficheiro de texto simples.

Depois do pré-processamento, onde se retiram números, *stop words*, pontuação e outros elementos irrelevantes, transforma-se a lista de documentos numa representação vectorial documento-termo com ponderação *tf-idf* (Feldman e Sanger, 2007). A partir da representação vectorial dos documentos obtém-se a matriz de semelhanças cosseno que servirá para gerar os agrupamentos e o grafo de ligações entre investigadores. Convém explicitar que as ligações podem não ser ligações efectivas ou co-autoria, reflectindo antes a abordagem de assuntos semelhantes.

### 2.2. Geração de agrupamentos

Partindo da matriz de semelhança para um formato de grafo, optamos por utilizar o algoritmo de procura de comunidades denominado *Walktrap* (Pons e Latapy, 2006). Esta técnica encontra subgrafos densamente ligados, também definidos como comunidades, através de passeios aleatórios. Parte do pressuposto que passeios aleatórios curtos tendem a ficar na mesma comunidade. Trata-se de uma abordagem aglomerativa hierárquica com base numa medida de distância entre vértices (nó a nó), cujo processo de agrupamento pode ser

visualizado na figura seguinte. Com base num nível óptimo de modularidade da rede baseado na ponderação entre as ligações internas e as externas da comunidade, o algoritmo encontra as comunidades organizadas de forma não hierárquica.

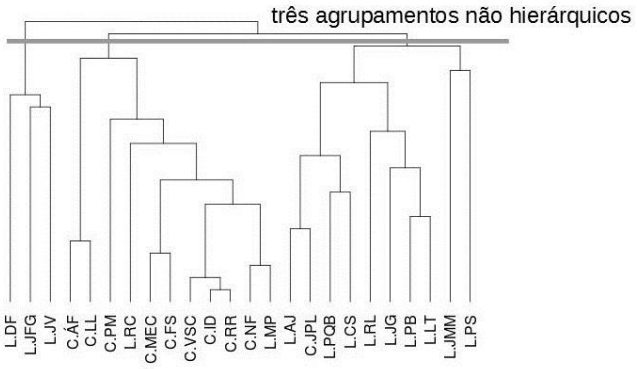



**Figura 1: Dendrograma gerado pelo algoritmo de agrupamento Walktrap**

### 2.3. Visualização e aAnálise da rRede de aAfinidades

O grafo de afinidades é obtido a partir da matriz discretizada de semelhanças – eliminou-se as ligações com valores cosseno inferiores a um limiar a partir do qual as relações são consideradas irrelevantes. A figura da rede de afinidades em baixo permite aferir a centralidade de alguns elementos dentro da sua comunidade (centralidade de grau) bem como os elementos que ligam as comunidades (centralidade de intermediação). As diferentes áreas densas do grafo são delimitadas pelas elipses originadas pelo algoritmo de agrupamento *Walktrap* e, em termos genéricos, vão de encontro à estrutura organizacional das entidades.

Adicionalmente aferiu-se algumas medidas de centralidade que levam em consideração as ponderações das diferentes ligações, como se pode ver na tabela em baixo. A *centralidade de grau* baseia-se no volume de ligações incidentes sobre um vértice. A *centralidade de intermediação* indica o número de vezes que um vértice liga outros dois vértices no caminho mais curto. A *centralidade de vector próprio* incorpora a importância dos vértices que ligam a um determinado vértice. As células destacadas a negro indicam que a maior centralidade de grau se encontra no CRACS, e que a maior centralidade de intermediação pertence a um membro do LIAAD (que foi agrupado junto dos investigadores do CRACS), como assinalado visualmente na figura anterior. Afigura-se ainda que, como assinala a centralidade de vector próprio, a influência dentro da própria comunidade das pessoas mais centrais no LIAAD é mais ténue do que no CRACS.

**Tabela 1: Medidas de centralidade para alguns dos investigadores mais relevantes**

| | .PB | RC | AJ | JG | .RR | C.FS | C.VSC |
|---|---|---|---|---|---|---|---|
| **Centralidade de grau (degree)** | 3,4 | 4,8 | 3,1 | 1,6 | | **5,6** | 4,7 |
| **Centralidade de intermediação (betweeness)** | 41 | **179** | 130 | 18 | 4 | 88 | 16 |
| **Centralidade de vector próprio** | 0,07 | 0,37 | 0,06 | 0,06 | 0,3 | **0,45** | 0,44 |

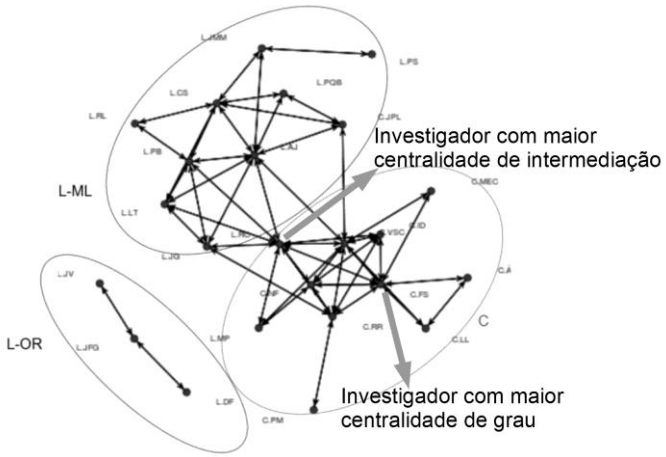



**Figura 2: Rede de Afinidades de Investigadores com visualização de agrupamentos não hierárquicos obtidos com algoritmo Walktrap**

## 3. Conclusões e trabalho futuro

O trabalho apresentado explorou Web/Text mining para obter nomes de investigadores e títulos das suas publicações. Isso permitiu gerar uma rede de afinidades, detetar comunidades e a sua análise. No trabalho futuro pretende-se processar os *abstracts* e considerar um número substancialmente maior de centros de investigação e/ou investigadores o que representa alguns desafios ao nível do agrupamento. Adoptar-se-á assim uma abordagem incremental de agrupamento e análise de afinidades.

## Referências

BUGLA, S. (2009), Name identification in scientific publications, Dissertação de Mestrado, Universidade do Porto, Porto

FELDMAN, R., SANGER, J. (2007), *Text Mining Textbook: Advanced Approaches in Analysing Unstructured Data*, Cambridge Univ. Press

GOLDSTONE, R., ROGOSKY, B. J. (2002), Using relations within conceptual systems to translate across conceptual systems, *Cognition*, 84. 295–320.

PONS, P., M. LATAPY (2006), Computing communities in large networks using random walks, *Journal of Graph Algorithms and Applications*, Vol. 10, no. 2, 191-218.

PRICE, S., FLACH, P. A., SPIEGLER, S., BAILEY, C., ROGERS, N. (2010), SubSift web services and workflows for profiling and comparing scientists and their published works, *Sixth IEEE International Conference on e–Science.*

TRIGO, L., BRAZDIL, P., (2010) Análise de Proximidade entre Investigadores de Alguns Centros de I&D da UP usando Text Mining sobre BD Bibliográficas, Livro de artigos JOCLAD 2010, ISCTE: Lisboa.

**Data Mining – Sábado, 12 de Abril, Sala 316 (11h00)**

# *Active learning* para análise de sentimento no Tweeter

**Vera Costa[1], Pedro Saleiro[2] Carlos Soares[3]**
[1]*Faculdade de Engenharia da Universidade do Porto, veracosta@fe.up.pt;*
[2]*Faculdade de Engenharia da Universidade do Porto, pssc@ fe.up.pt;*
[3]*INESC TEC, Faculdade de Engenharia da Universidade do Porto, csoares@fe.up.pt*

**Sumário**

A análise de sentimento estudada neste trabalho consiste em analisar a polaridade das mensagens de texto publicadas no Twitter sobre política em Portugal. Para usar abordagens de *data mining* para esse fim é necessário ter algumas mensagens previamente etiquetadas manualmente. Essa etiquetagem é cara pelo que se pretende minimizar o número de mensagens etiquetadas. Para tal, recorreu-se a *active learning*, cujo objetivo é precisamente a redução do custo de anotação manual de dados para métodos de classificação.

**Palavras-chave:** Análise de sentimento, *Active learning*, Classificação, *Data mining*

## 1. Introdução

A análise de sentimento baseado em mensagens do Twitter consiste em identificar a polaridade (que poderá ser positiva, negativa ou neutra) das mensagens de texto publicadas nessa rede social. Esta é uma tarefa difícil de trabalhar uma vez que a maioria dos *tweets* apresentam opiniões negativas, e apenas um reduzido número apresenta opiniões positivas. Isto leva a que o conjunto de dados respetivo tenha uma distribuição de classes muito desequilibrada.

O ideal seria uma anotação manual de uma grande quantidade de dados, o que permitiria melhorar os resultados dos algoritmos de classificação que podem ser usados para este fim. No entanto, na vida real isto nem sempre é possível uma vez que a anotação manual pode ser bastante dispendiosa e demorada (Li et al, 2012). Neste sentido, torna-se útil a abordagem de *active learning.*

*Active learning* é uma técnica de *machine learning* semi-supervisionada que consiste em seleccionar para anotação manual apenas os *tweets* considerados como mais incertos. Ou seja, aqueles que estejam mais próximos da "linha de fronteira" das várias classes de polaridade. Existem várias estratégias que podem ser usadas no método de *active learning* para seleção de dados incertos, como por exemplo: amostragem aleatória, amostragem por grupos, mudança de modelo esperado, redução do erro esperado, redução da variância e métodos de densidade ponderada (Settles, 2010).

A aplicação dos métodos referidos pode ser feita de duas formas: *stream-based* e *pool-based* (Olsson, 2009). Considerando um conjunto de dados não anotado, a primeira estratégia consiste em selecionar caso a caso os elementos do conjunto não anotado, analisá-los e decidir

se são ou não considerados casos incertos/duvidosos para anotação manual. Na segunda estratégia, a seleção é feita considerando um grupo de dados, aos quais é feita a análise para seleção de, eventualmente vários, casos duvidosos.

Apesar do *active learning* ser uma técnica bastante útil, enfrenta alguns desafios quando se trata de dados não balanceados (ou seja, com classes de tamanho bastante diferente). Assim, torna-se útil adaptar/ajustar a técnica de forma a ser possível obter os resultados mais optimizados quanto possível. (Li et al, 2010) sugerem os seguintes métodos: estratégia de amostragem aleatória, co-seleção com caraterísticas dos classificadores dos subspaços, co-seleção com dados rotulados da classe maioritária.

O presente estudo visa investigar o desempenho de vários métodos de *active learning* em combinação com vários algoritmos de classificação (por exemplo, árvores de decisão, SVM, Naive Bayes, regressão logística) com base num conjunto de *tweets* já rotulados.

## 2. Estudo Experimental

Este trabalho encontra-se a ser desenvolvido no âmbito do projeto POPSTAR (*Public Opinion and Sentiment Tracking, Analysis and Research*) e pretende otimizar os métodos de classificação usados na análise de sentimento político obtidos a partir do Twitter. Como já referido trata-se de uma análise de texto com dados não balanceados (cerca de 3% positivos, 43% neutros e 54% negativos).

Para esta análise estão a ser usadas técnicas de *active learning* numa amostra de 1030 *tweets* rotulados manualmente. Destes, 90% são considerados como conjunto de treino e os restantes 10% para conjunto de teste. Por sua vez, dentro do conjunto de treino, 20% são os dados rotulados inicialmente e os restantes 80% são os dados não rotulados que são analisados pelos métodos de *active learning*. O objetivo desta divisão é simular os dados reais e verificar que quanto maior for o número de dados rotulados, melhor é o desempenho do classificador a ser usado.

Uma vez tratar-se de dados não balanceados, os conjuntos atrás referidos foram efetuados de forma estratificada, garantindo a existência de *tweets* das três classes de polaridade em todos os conjuntos referidos.

Como método de referência para aferir o desempenho do método de *active learning*, é usado um método de amostragem aleatória para seleção de observações a serem rotuladas. Este método consiste simplesmente na seleção aleatória de uma observação para etiquetagem. Como algoritmo de classificação foi usada a regressão logística para classes múltiplas. A figura (1) ilustra o método de *active learning* utilizado, aplicado após a estratificação dos dados**.**

Entrada

- Dados classificados
- Dados não classificados
- Conjunto de teste

Saída

- Novo conjunto de dados classificado

Procedimento

- Para todos os dados não classificados
    1. Aplicar um algoritmo de classificação para os dados classificados
    2. Prever a incerteza de classificação para as observações não rotuladas (próximas da fronteira, possivelmente mal classificados)
    3. Escolher o caso com maior incerteza para etiquetar.

**Figura 1: *Active learning* usando a estratégia *stream-based***

Após identificado o método e algoritmo de classificação mais adequado para o conjunto de dados em estudo, pretende-se que estes sejam aplicados a um novo conjunto ainda não classificado, mas da mesma natureza. Os novos dados serão classificados e os casos duvidosos/incertos (próximos da linha de fronteira) serão identificados para anotação manual. O objetivo será obter o melhor conjunto de treino possível, com dados corretamente classificados. No entanto, como haverá sistematicamente novos *tweets*, serão sempre considerados novos casos para anotação manual. O ideal seria uma taxa de acerta de 100%, difícil de atingir.

## 3. Conclusões e trabalho futuro

Neste trabalho estamos a investigar a utilização de métodos de *active learning* para redução do custo de anotação manual de dados para métodos de classificação, num problema de análise de sentimento político em dados do Twitter.

Neste momento foi implementado um primeiro método simples e obtidos os primeiros resultados. Os próximos passos consistem na consolidação dos resultados, testando mais algoritmos de classificação e outros métodos de *active learning*.

Finalmente esperamos da análise dos resultados obtidos, desenvolver um novo método que seja particularmente adequado não só para este problema mas também para outros com características semelhantes.

## Referências

LI, S., JU, S., Zhou, G. (2012). Active Learning for Imbalanced Sentiment Classification. EMNLP-CoNLL'12 *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 139-148.

MELVILLE, P., Yang, S., TSECHANSKY, M., MOONEY, R. (2005). Active learning for probability estimation using Jensen-Shannon divergence. *Machine Learning: ECML 2005*, 268-279.

OLSSON, F. (2009). A literature survey of active machine learning in the context of natural language processing.

PRUDÊNCIO, R., SOARES, C., LUDERMIR, T. (2011) Uncertainty sampling-based active selection of datasetoids for meta-learning. *Artificial Neural Networks and Machine Learning–ICANN 2011*. 454-461.

SETTLES, B. (2010). Active learning Literature survey. *University of Wisconsin, Madison 52*, 55-66.

# POSTERS

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Metodologia estatística para a avaliação de um recurso natural (Minho e Galiza)

**Daniela Catalão[1], A. Manuela Gonçalves[2], Susana Faria[3], Jorge Oliveira[4]**

[1]*Departamento de Matemática e Aplicações, Universidade do Minho, danielacatalao@hotmail.com;*

[2]*Departamento de Matemática e Aplicações, CMAT – Centro de Matemática, Universidade do Minho, mneves@math.uminho.pt;*

[3]*Departamento de Matemática e Aplicações, CMAT – Centro de Matemática, Universidade do Minho, sfaria@math.uminho.pt;*

[4]*Snergeo, jorgeoliveira@snergeo.pt*

**Sumário**

Neste estudo pretendeu-se desenvolver uma metodologia geral, na área da Estatística Multivariada e da Inferência Estatística, com o objetivo de avaliar e interpretar a variabilidade de um alargado conjunto de informação sobre as características físico-químicas do solo que influenciam o desenvolvimento das videiras, a qualidade do mosto e das uvas e, por consequência, a qualidade dos vinhos. Este estudo incidiu sobre duas parcelas de cultivo de vinha, uma situada no Minho (Portugal) e outra localizada na Galiza (Espanha).

**Palavras-chave:** Análise fatorial, Análise de *clusters***,** Cultivo da vinha, Inferência estatística, Solo.

## 1. Introdução

Este trabalho resultou da necessidade de se identificarem quais as variáveis físico-químicas do solo que influenciam a produtividade das videiras e a qualidade das uvas, do mosto e, consequentemente, dos vinhos numa parcela da Estação Vitivinícola Amândio Galhano, no concelho de Arcos de Valdevez (Minho-Portugal) e numa parcela localizada na Bodega Santiago Ruiz, em Tomiño (Região da Galiza-Espanha). Pretende-se com este conhecimento, e num mercado cada vez mais competitivo, otimizar qualitativamente e quantitativamente a produção de vinho.

As vinhas em estudo do Minho eram constituídas apenas pela casta tinta Vinhão, enquanto a vinha da Galiza pela casta branca Alvarinho. Na parcela da Estação Vitivinícola Amândio Galhano, o solo é um solo residual granítico, enquanto na Bodega Santiago Ruiz em Tomiño (Galiza) o solo é um aluvião, constituído na sua maioria por acumulações de seixos, frequentemente cobertos por camadas arenoargilosas e matéria orgânica. Os dados analisados, recolhidos em campo e em laboratório, reportaram-se ao ano de 2011. O conhecimento adequado das características do solo e das exigências da vinha é fundamental para se conseguir atingir níveis de produção economicamente rentáveis e de qualidade do vinho (Jordão, 2007).

## 2. Metodologia

Na parcela do Minho foram georreferenciados 45 pontos com espaçamento regular, em 9 regiões, cada uma contendo 5 destes pontos onde foram recolhidas as amostras do solo, com o objetivo de se analisarem algumas variáveis referentes a características físico-químicas do solo. Nestas localizações também se avaliaram as variáveis referentes à produtividade das videiras. Foram recolhidas 9 amostras de mosto a partir das quais foram avaliadas algumas variáveis relativas às características físico-químicas do mosto; também foi feita uma análise do rendimento em sumo e da composição cromática e aromática do mosto. Por último, foram analisados os vinhos correspondentes às 9 regiões referidas anteriormente.

Na parcela da Galiza também foram georreferenciados 45 pontos de onde foram recolhidas as amostras do solo e analisadas algumas variáveis relativas às características físico-químicas do solo. Relativamente às variáveis alusivas à produção das videiras, estas foram avaliadas de acordo com 12 regiões. As variáveis analisadas referentes às uvas e ao mosto foram avaliadas apenas a partir de 6 amostras de uvas provenientes de 6 regiões. Por fim, foram analisados os 6 vinhos correspondentes a estas 6 regiões.

A título de exemplo na Tabela 1 apresentam-se as variáveis das características do solo que foram avaliadas em ambas as parcelas.

**Tabela 1: Variáveis do solo em estudo nas parcelas do Minho e da Galiza**

| | Variáveis | | Descrição | Existência | |
|---|---|---|---|---|---|
| | Designação | Unidades | | Minho | Galiza |
| **SOLO** | $DA$ | $(g/vol)$ | Densidade Aparente | × | × |
| | $MO$ | $(\%)$ | Matéria Orgânica | × | × |
| | $pH$ | | pH em Estrato Aquoso | × | × |
| | $FF$ | $(\%)$ | Fração Fina $(\%FF + \%FG = 100\%)$ | × | × |
| | $FG$ | $(\%)$ | Fração Grosseira $(\%FF + \%FG = 100\%)$ | × | × |
| | $P2O5$ | $(ug/g)$ | Fósforo Assimilável | × | × |
| | $K2O$ | $(ug/g)$ | Potássio Assimilável | × | × |
| | $Ca$ | $(ug/g)$ | Cálcio Assimilável | × | × |
| | $Mg$ | $(ug/g)$ | Magnésio Assimilável | × | × |
| | $AzT$ | $(\%)$ | Azoto Total | × | × |
| | $Ni$ | $(ug/g)$ | Níquel | × | × |
| | $Cr$ | $(ug/g)$ | Crómio | × | × |
| | $Cd$ | $(ug/g)$ | Cádmio | × | × |
| | $N$ | $(ug/g)$ | Nitratos | × | × |
| | $B$ | $(ug/g)$ | Boro | × | × |
| | $CTC$ | $(m.e./100g)$ | Capacidade de Troca Catiónica | × | × |

Os principais objetivos deste trabalho foram descrever o comportamento das diferentes variáveis em estudo e identificar regiões com semelhantes características físicas e químicas do solo, nas parcelas em estudo. Também se pretendeu verificar se existia um pequeno conjunto de variáveis que fosse responsável por explicar uma proporção elevada da variação total associada ao conjunto original das variáveis do solo que influenciam a qualidade do vinho e quais as regiões que apresentam grandes/pequenas concentrações dessas variáveis e, deste modo, reduzir a dimensionalidade do problema.

Para a concretização destes objetivos foram usadas metodologias da área da Estatística Multivariada (Análises de *Clusters* e Análise Fatorial) e de Inferência Estatística (em particular, testes de hipóteses não paramétricos de modo a lidar com o problema da não normalidade dos dados e do número reduzido de observações de algumas variáveis). A Análise de *Clusters* foi aplicada ao conjunto das variáveis do solo, a partir das variáveis originais do solo e a partir dos fatores retidos pela aplicação de uma Análise Fatorial a essas mesmas variáveis. A qualidade do vinho e o equilíbrio vegetativo da videira foram avaliados através de análise de correlações com as variáveis do solo.

Esta análise foi efetuada nas duas parcelas em estudo (Minho e Galiza) e, no final, os resultados de ambas foram analisados e comparados.

## 3. Resultados

A Análise de *Clusters* realizada a partir das variáveis originais do solo permitiu obter um agrupamento dos pontos de amostragem em três grupos homogéneos, relativamente às características do solo em ambas as parcelas (Minho e Galiza). Com a aplicação do teste não paramétrico de Kruskal-Wallis verificou-se que as variáveis do solo da parcela do Minho MO, FF, FG, P2O5, Mg, AzT, Ni, Cr, B e CTC apresentaram diferenças significativas entre os diferentes *clusters*. Estas diferenças foram ainda detetadas para as variáveis ácido málico do mosto e para a família de compostos em C6 do aroma do mosto na fração livre (FL1M). Relativamente à parcela da Galiza, as variáveis do solo MO, P2O5, K2O, Ca, Mg, AzT, Cd, N, B e CTC foram as que apresentaram diferenças significativas entre os três *clusters*. Foram, ainda, encontradas estas diferenças para a variável acidez total do mosto, para as famílias dos compostos em C6 do aroma das uvas na fração livre (FL1U), para as famílias compostos carbonilados do aroma das uvas na fração livre (FL5U), para as famílias de compostos em C6 do aroma das uvas na fração glicosilada (FG1U), para as famílias de compostos carbonilados do aroma das uvas na fração glicosilada (FG7U) e para a variável representativa da nota final atribuída ao vinho. O teste de comparações múltiplas não paramétrico LSD permitiu, ainda, avaliar em que *clusters* é que as variáveis do solo apresentavam diferenças significativas.

A aplicação de uma Análise de *Clusters* aos fatores resultantes das Análises Fatoriais efetuadas aos dados de ambas as parcelas, resultou em quatro fatores que explicaram aproximadamente 72% da variabilidade total dos dados originais da parcela do Minho e explicaram, aproximadamente, 85,5% da variabilidade total dos dados originais da quinta da Galiza. Os resultados obtidos a partir desta Análise de *Clusters* não expuseram uma clara estrutura dos grupos subjacentes aos dados da parcela da Galiza. Relativamente à parcela do Minho, a Análise de *Clusters* realizada a partir dos fatores, bem como os testes de comparações múltiplas, evidenciaram resultados semelhantes aos obtidos a partir das variáveis originais do solo.

A qualidade do vinho e o equilíbrio vegetativo da videira foram avaliados através de análise de correlações com as variáveis do solo e identificaram-se as zonas das parcelas que apresentaram défice/excesso nas concentrações dessas variáveis.

A análise dos resultados obtidos a partir das duas bases de dados que incorporavam um conjunto de dados, relativos ao ano 2011, pôs em evidência a influência do solo sobre o rendimento e qualidade das uvas e do vinho. Foi possível identificar algumas das características do solo que estão correlacionadas com o desenvolvimento das videiras e com a qualidade das uvas e os respetivos vinhos, nestas duas vinhas em estudo.

## 4. Conclusões

Este trabalho evidenciou um conjunto de indicadores que poderão ajudar no processo de tomada de decisão para a racionalização de utilização dos fatores de produção, como os nutrientes e os produtos fitofármacos, contribuindo, assim, para uma maior produtividade da videira e uma maior qualidade da uva e do vinho. A produtividade da videira, a qualidade das uvas, do mosto e do vinho foram avaliadas através das suas correlações com as variáveis do solo, permitindo avaliar e identificar as zonas das parcelas que apresentaram um défice/excesso nas concentrações destas variáveis e, assim, sempre que se justifique, efetuar correções diferenciadas para cada local da parcela de cultivo, efetuando-se deste modo, uma agricultura de uma forma mais sustentada.

**Agradecimentos:** Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT – Fundação para a Ciência e Tecnologia no âmbito do projecto PEstOE/MAT/UI0013/2014.

## Referências

BRANCO, J.A. (2004) *Uma Introdução à Análise de Clusters*, Évora, Sociedade Portuguesa de Estatística.

CHATEFIELD, C., COLLINS, A.J. (1980) *Introduction to Multivariate Analysis*, Cambridge, Chapman and Hall.

HIGGINS, J.J. (2004) *Introduction to Modern Nonparametric Statistics*, Thomson, Toronto, Duxbury Advanced Series.

JORDÃO, A.J. (2007) Gestão do Solo na Vinha. Texto elaborado no âmbito do Plano de Acção para a Vitivinicultura da Alta Estremadura.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Explaining contraceptive use by the wealth index in India: A latent variable approach

**José G. Dias[1], Isabel Tiago de Oliveira[2]**

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

[2]*Instituto Universitário de Lisboa (ISCTE-IUL), isabel.oliveira@iscte.pt*

**Abstract**

This analysis aims to measure the poverty-wealth effect on married Indian women contraceptive behavior. We formulate a probit regression model with a latent covariate – the wealth index – measured by a set of items and estimated by an Item Response Theory (IRT) model. Results confirm the positive impact of wealth on contraceptive adoption.

**Key-words:** Latent variables, IRT model, Probit regression model, Contraception, Wealth.

## 1. Introduction

This application addresses the poverty-wealth dimension impact on contraceptive adoption by Indian women. The purpose of this analysis is to estimate the impact of the poverty-wealth dimension on contraception using a latent variable approach. Section 2 describes the data set used in the empirical analysis and the statistical model. Section 3 gives a summary of the main results. The paper ends with concluding remarks.

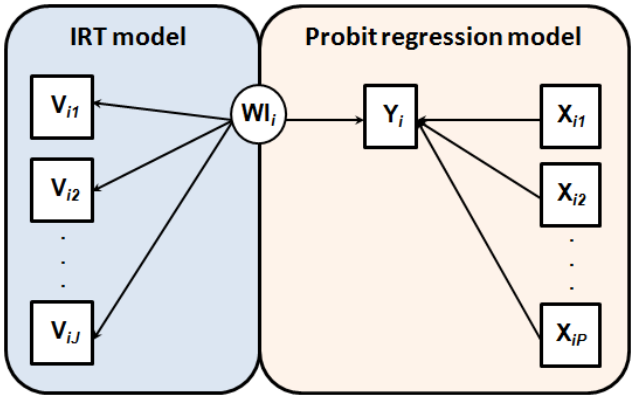



**Figure 1. Model diagram**

## 2. Data and Methods

This research uses the Indian National Family Health Survey from 2005-06 (IIPS and Macro International, 2007). The survey includes a large number of topics regarding women reproductive history, demographic, social and economic characteristics from women, and the household.

The proposed model takes the form of a probit regression model with a latent covariate. This covariate – the wealth index – is measured by a set of items (see Table 2) using an IRT model (VAN DER LINDEN e HAMBLETON, 1997; DEMARS, 2010). This model can be interpreted as a factorial model with a continuous latent variable and discrete manifest variables (see Figure 1). This model is formulated in a multilevel framework that takes the PSU (village) level into account (HOX, 2002).

**Table 1. Probit regression model estimates**

| **Variables** | Estimate | S.E. | p-value |
|---|---|---|---|
| **Level 1 - Fixed effects** | | | |
| **Female age** | -0.014 | 0.002 | 0.000 |
| **Caste** (ref: None of them) | | | |
| Caste_caste | -0.051 | 0.028 | 0.074 |
| Caste_tribe | -0.327 | 0.038 | 0.000 |
| Caste_obc | -0.091 | 0.023 | 0.000 |
| **Residence** (ref: Rural) | | | |
| Urban | 0.131 | 0.027 | 0.000 |
| **Religion** (ref: Hindu) | | | |
| Muslim | -0.336 | 0.034 | 0.000 |
| Other | -0.271 | 0.037 | 0.000 |
| **Houshold structure** (ref: Non-nuclear) | | | |
| Nuclear | 0.101 | 0.019 | 0.000 |
| **Female education** (ref: No education) | | | |
| Primary | 0.151 | 0.028 | 0.000 |
| Secondary | 0.337 | 0.026 | 0.000 |
| Higher | 0.576 | 0.041 | 0.000 |
| **Female occupation** (ref: Not working) | | | |
| Working | 0.095 | 0.020 | 0.000 |
| **Living children (reference: 0)** | | | |
| 1 | 1.105 | 0.037 | 0.000 |
| 2 | 1.757 | 0.043 | 0.000 |
| 3 | 1.908 | 0.048 | 0.000 |
| 4+ | 1.962 | 0.053 | 0.000 |
| **Living boys (reference: 0)** | | | |
| 1 | 0.267 | 0.025 | 0.000 |
| 2 | 0.522 | 0.033 | 0.000 |
| 3 | 0.445 | 0.044 | 0.000 |
| 4+ | 0.281 | 0.055 | 0.000 |
| **Wealth index** | 0.239 | 0.016 | 0.000 |
| **Thresholds** | 1.234 | 0.060 | 0.000 |
| **Level 2 - Random effects** | | | |
| Var(u1) | 0.231 | 0.014 | 0.000 |

Note: Residual variance equals 1.

## 3. Results

Results from the multilevel probit model for contraceptive adoption in India reveal the impact of the set of covariates plus wealth index on contraceptive use (Table 1). Covariates include the life cycle variables, residence factors, and the socioeconomic ones. These estimates reveal that as the women aged they tend to adopt contraception more often, and that the

offspring number and sex composition are important factors to the adoption of family planning methods. On the other hand, residence factors are significant: women living in urban areas and in nuclear households have comparatively more chances to use contraception than their counterparts. Regarding the Indian traditional socioeconomic and cultural differences, it is clear that the women in schedule tribes and in other backward classes had comparatively lower chances to adopt family planning than women that do not include themselves in none of these situations. On the other hand, both Muslim women and the women from other religious affiliations have lower odds to use contraception than Hindu women. Additionally, the contemporary differences associated with female work and education reveals that both factors increase the odds to adopt family planning methods. As expected the education gradient is very clear. Finally, the poverty-wealth dimension has significant impact on the contraceptive adoption: as wealth increases so it does the chances to adopt contraception. Finally, the upper level (PSU) explains 18.8% of the total variance.

**Table 2. IRT model estimates**

| Variables | Aggregate | Loadings | | | Thresholds | | |
|---|---|---|---|---|---|---|---|
| | | Estimate | S.E. | p-value | Estimate | S.E. | p-value |
| Household electrification (Yes) | 0.776 | 1.741 | 0.044 | 0.000 | -1.540 | 0.045 | 0.000 |
| House has windows with glass (Yes) | 0.244 | 1.253 | 0.027 | 0.000 | 1.118 | 0.030 | 0.000 |
| Type of toilet facility | | | | | | | |
| Flush Toilet | 0.518 | 1.551 | 0.036 | 0.000 | -0.085 | 0.034 | 0.011 |
| Pit Toilet/Latrine | 0.082 | -0.300 | 0.019 | 0.000 | 1.452 | 0.023 | 0.000 |
| None/Other | - | - | - | - | - | - | - |
| Type of flooring | | | | | | | |
| Natural | - | - | - | - | - | - | - |
| Rudimentary | 0.078 | -0.116 | 0.018 | 0.000 | 1.425 | 0.022 | 0.000 |
| Finished | 0.532 | 1.370 | 0.031 | 0.000 | -0.136 | 0.029 | 0.000 |
| Cooking fuel | | | | | | | |
| Good | 0.352 | 1.921 | 0.047 | 0.000 | 0.849 | 0.040 | 0.000 |
| Moderate | 0.045 | 0.041 | 0.012 | 0.000 | 1.696 | 0.023 | 0.000 |
| Poor | - | - | - | - | - | - | - |
| House ownership (Yes) | 0.848 | -0.136 | 0.014 | 0.000 | -1.037 | 0.016 | 0.000 |
| Ownership of a pressure cooker (Yes) | 0.551 | 1.881 | 0.040 | 0.000 | -0.281 | 0.036 | 0.000 |
| Ownership of a cot/bed (Yes) | 0.879 | 0.431 | 0.017 | 0.000 | -1.273 | 0.018 | 0.000 |
| Owership of a radio/transistor (Yes) | 0.363 | 0.394 | 0.012 | 0.000 | 0.377 | 0.012 | 0.000 |
| Ownership of a colour television (Yes) | 0.373 | 2.010 | 0.038 | 0.000 | 0.736 | 0.036 | 0.000 |
| Ownership of any telephone (Yes) | 0.189 | 1.457 | 0.032 | 0.000 | 1.566 | 0.033 | 0.000 |
| Ownership of a computer (Yes) | 0.043 | 1.393 | 0.047 | 0.000 | 2.950 | 0.064 | 0.000 |
| Ownership of a refrigerator (Yes) | 0.240 | 1.998 | 0.046 | 0.000 | 1.582 | 0.043 | 0.000 |
| Ownership a car (Yes) | 0.055 | 1.304 | 0.044 | 0.000 | 2.624 | 0.054 | 0.000 |
| Ownership a tractor (Yes) | 0.022 | 0.170 | 0.021 | 0.000 | 2.049 | 0.026 | 0.000 |
| Ownership a motorcycle/scooter (Yes) | 0.247 | 1.120 | 0.023 | 0.000 | 1.032 | 0.022 | 0.000 |

Table 2 gives the estimates of the Item Response Theory (IRT) model that accounts for the measurement of the latent variable: the wealth index. It measures the latent variable by a battery of items about the characteristics of the household. All indicators are significant in the measurement of the latent variable.

## 4. Conclusion

This paper provides an empirical approach to measuring the impact of wealth index on contraceptive use, when the wealth index is a latent indicator and is measured by a battery of items. Results show that the wealth index has a significant impact on contraceptive use.

**Acknowledgements:** This research was supported by the Fundação para a Ciência e Tecnologia (Portugal), Grant PTDC/CS-DEM/108033/2008.

## References

DEMARS, C. (2010). *Item Response Theory*. New York: Oxford University Press.

HOX, J. (2002). *Multilevel Analysis: Techniques and Applications*, Mahwah, Lawrence Erlbaum Associates.

INTERNATIONAL INSTITUTE FOR POPULATION SCIENCES (IIPS) & MACRO INTERNATIONAL (2007). *National Family Health Survey (NFHS-3), 2005–06: India: Volume I,* Mumbai, India, IIPS.

VAN DER LINDEN, W. J., & HAMBLETON, R. K. (Eds.). (1997). *Handbook of Modern Item Response Theory*, New York, Springer-Verlag.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# The impact of population heterogeneity on factor analysis estimation

**Catarina Marques[1], José G. Dias[2]**
[1]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, catarina.marques@iscte.pt;*
[2]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, jose.dias@iscte.pt*

**Abstract**

This paper discusses the impact of population heterogeneity on factor analysis under the homogeneity assumption and potential biases resulting from aggregation. An illustration using the Holzinger-Swineford Mental Abilities data shows that the factor mixture model relaxes factor analysis specification taking population heterogeneity and potential outliers into account.

**Keywords:** Mixture factor models, Unobserved heterogeneity, Outliers detection.

## 1. Introduction

Factor analysis assumes that observations are sampled from a homogeneous population with parameters invariant across individuals. This assumption tends to be strong and in most empirical applications not very realistic thus parameter estimates turn out to be biased (MUTHÉN, 1989; JEDIDI et al., 1997; YUNG, 1997; WALL et al., 2012). Unobserved heterogeneity may result from the existence of subgroups in population, which are characterised by distinct latent factors (WALL et al., 2012). For instance, YUNG (1997) argues that the presence of outliers can be viewed as a special case of unobserved heterogeneity.

Mixture factor models incorporate discrete unobserved heterogeneity, i.e., a clustering structure, into covariance structure models. Although these models have gained much popularity in the literature in recent years, most of the research has been focused on simulation studies and are not widely applied in empirical research. This study illustrates the impact of (discrete) unobserved heterogeneity on parameter estimates using an empirical dataset.

## 2. Data and models

This research uses the Holzinger-Swineford Mental Abilities Data (HOLZINGER & SWINEFORD, 1939), which has been defined as a benchmark data set for factor analysis studies. It contains 26 items hypothesised to measure a general factor - mental abilities - and five specific factors: spatial ability, verbal ability, mental speed, recognition/memory, and general deduction (or mathematical ability). This battery of tests was administered to 301 American seventh and eighth grade students in two different Illinois schools: the Grant-White School (n= 145) and the Pasteur School (n= 156). Parents of students in the first school were mostly American-born, whereas for the latter school they were mostly from working-class or foreign-born (speak their native language at home). In this study, we restrict ourselves to the 19 tests/items that intended to measure the first four abovementioned domains. In addition to the

19 items, the structural model contains four covariates: school (1-Grant-White; 2-Pasteur), gender (0-female; 1–male), age, and grade (0–7th grade; 1–8th grade). Table 2 presents the measurement model for these four factors in the first column: each item loads on only its hypothesised factor. Four items (CUBES, ADDITION, FIGURER, FIGUREW) were removed from the analysis as a result of low loadings.

We set up a two-component mixture factor model that accommodates unobserved heterogeneity with two distinct segments. The baseline model assumes population homogeneity with a single component (the traditional factor model).

## 3. Results

First, we estimated the baseline model that shows an overall good fit (Chi-square statistic value=257.828, df=106, p<0.001; CFI=0.915; TLI=0.891; RMSEA=0.069; SRMR=0.056). Then, model specification takes the unobserved heterogeneity into account. In order to minimise the convergence to local optima, 20 random starting values were set for each model. The following constraints were also imposed to achieve model identification: (1) as factor loadings were high for the aggregate model (items are appropriate for measuring constructs), the measurement model was kept invariant across components, i.e., loadings, error variances, and item intercepts were fixed across components; (2) the factor metric was achieved by fixing its first loading to be one; (3) the latent factor intercepts for components 1 and 2 were fixed to be zero; (4) the latent factor means and item intercepts of component 1 were fixed to be 0 (reference component); and (5) factor variance was fixed to be component-invariant.

Table 1 presents the information criteria statistics for model selection; all of them point to the solution of the two-component model, i.e., to the existence of unobserved heterogeneity. However, the first component contains almost all the students (96.1%).

**Table 1: Information Criteria Statistics**

| | **Aggregate model** | **Two-component model** |
|---|---|---|
| AIC | 30,865.274 | 30,099.882 |
| BIC | 31,083.994 | 30,366.794 |
| Sample-Size Adjusted BIC (aBIC) | 30,896.879 | 30,138.451 |

Table 2 provides maximum-likelihood estimates of the standardised loadings for the aggregate and two-component models. Results show that the component 2 has the highest values of standardised loadings on SPATIAL, SPEED and MEMORY factors, even when compared with the aggregate model. The loadings on the latter factor are higher than those of component 1 and those of the aggregate model. Component 1 presents loadings that are more similar to those obtained in the aggregate model. Therefore, the group 2 is composed by students for which test scores are highly influenced by their mental abilities; however, this group is very small, thus these observations may be outliers.

**Table 2: Standardised loading estimates for the aggregate and the two-component models**

| Factors | Aggregate model | | | Two-component model | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Component 1 | | | Component 2 | | |
| Items | Estimate | S.E. | P-Value | Estimate | S.E. | P-Value | Estimate | S.E. | P-Value |
| SPATIAL | | | | | | | | | |
| VISUAL | 0.725 | 0.047 | 0.000 | 0.712 | 0.053 | 0.000 | 0.820 | 0.103 | 0.000 |
| PAPER | 0.500 | 0.055 | 0.000 | 0.497 | 0.062 | 0.000 | 0.629 | 0.145 | 0.000 |
| FLAGS | 0.610 | 0.051 | 0.000 | 0.600 | 0.055 | 0.000 | 0.727 | 0.103 | 0.000 |
| VERBAL | | | | | | | | | |
| GENERAL | 0.844 | 0.020 | 0.000 | 0.843 | 0.022 | 0.000 | 0.751 | 0.044 | 0.000 |
| PARAGRAP | 0.817 | 0.022 | 0.000 | 0.814 | 0.025 | 0.000 | 0.714 | 0.043 | 0.000 |
| SENTENCE | 0.867 | 0.018 | 0.000 | 0.867 | 0.019 | 0.000 | 0.785 | 0.040 | 0.000 |
| WORDC | 0.744 | 0.029 | 0.000 | 0.740 | 0.031 | 0.000 | 0.624 | 0.051 | 0.000 |
| WORDM | 0.844 | 0.020 | 0.000 | 0.838 | 0.022 | 0.000 | 0.744 | 0.043 | 0.000 |
| SPEED | | | | | | | | | |
| CODE | 0.711 | 0.049 | 0.000 | 0.764 | 0.055 | 0.000 | 0.815 | 0.037 | 0.000 |
| COUNTING | 0.562 | 0.051 | 0.000 | 0.566 | 0.053 | 0.000 | 0.633 | 0.078 | 0.000 |
| STRAIGHT | 0.709 | 0.048 | 0.000 | 0.655 | 0.047 | 0.000 | 0.718 | 0.067 | 0.000 |
| MEMORY | | | | | | | | | |
| WORDR | 0.553 | 0.057 | 0.000 | 0.569 | 0.063 | 0.000 | 0.742 | 0.088 | 0.000 |
| NUMBERR | 0.523 | 0.059 | 0.000 | 0.505 | 0.068 | 0.000 | 0.684 | 0.088 | 0.000 |
| OBJECT | 0.648 | 0.054 | 0.000 | 0.604 | 0.064 | 0.000 | 0.772 | 0.049 | 0.000 |
| NUMBERF | 0.562 | 0.056 | 0.000 | 0.521 | 0.092 | 0.000 | 0.699 | 0.082 | 0.000 |

Examining the mimic model for the aggregate and the two-component models (results not shown), the effects of mental abilities on the test scores by gender and grade are different across groups. Age and school do not influence student mental abilities. Results can be summarised as follows: (1) boys in group 1 tend to have more spatial abilities, in contrast to the girls who have more speed and memory abilities. In group 2, boys have more verbal and speed abilities and girls more memory abilities; (2) students of the $8^{th}$ grade in group 1 tend to have more verbal and speed abilities, whereas those in group 2 have more verbal and memory abilities. Students of $7^{th}$ grade of group 2 are those who have more speed abilities; (3) the gender and grade effects of the students in group 2 in spatial abilities are not significant, because of high standard errors; and (4) coefficient estimates of gender and grade are similar in magnitude and signal for the aggregate model and the group 1. Regarding factor variance and covariance estimates (results not shown), the former are similar and close to 1 for all factors in the aggregate model and component 1 and the latter are also very similar for these sets of students (total and group 1). Factor variance estimates are low and not significant for component 2 due to the small number of observations. In contrast, covariance estimates between factors are high but mostly are not significant.

## 4. Conclusion

Although parameter estimates of aggregate model and component 1 are similar they differ. The aggregate model estimates are affected by the extremely high mental abilities of the students in component 2. This subgroup is structurally different and the source of unobserved heterogeneity is detected by the mixture model. Thus, this empirical example illustrates that factor mixture models can be applied to detect population heterogeneity in empirical research, namely outlier observations. Without taking unobserved heterogeneity into account parameter estimates are biased and hypothesis testing may not be trustworthy (MUTHÉN, 1989; JEDIDI et al., 1997; WALL et al., 2012).

## References

HOLZINGER, K. J. & SWINEFORD, F. A. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Education Monographs*, 48. University of Chicago.

JEDIDI, K., JAGPAL, H. S. & DESARBO, W. S. (1997). Finite-fixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1), 39-59.

MUTHÉN, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.

WALL, M.M., GUO, J. & AMEMIYA, Y. (2012). Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivariate Behavioral Research*, 47(2), 276-313.

YUNG, Y. F. (1997). Finite mixtures in confirmatory factor analysis models. *Psychometrika*, 62, 297–330.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Estudo empírico do índice de satisfação da procura dos candidatos aos cursos superiores de engenharia

**Raquel Oliveira[1], A. Manuela Gonçalves[2], Rosa M. Vasconcelos[3]**

*[1]Centro de Matemática, Universidade do Minho, Portugal, rmro_17@hotmail.com;*

*[2]Departamento de Matemática e Aplicações, CMAT – Centro de Matemática, Universidade do Minho, Portugal, mneves@math.uminho.pt;*

*[3]Departamento de Engenharia Têxtil, 2C2T – Centro de Ciência e Tecnologia Têxtil, Universidade do Minho, Portugal, rosa@det.uminho.pt*

**Sumário**

O objectivo deste trabalho é descrever e caracterizar a colocação dos estudantes no ensino superior português, nomeadamente nos cursos de engenharia. A aplicação de metodologias multivariadas ao índice de satisfação da procura dos estudantes permitiu detectar grupos dos cursos superiores e identificar alguns determinantes das escolhas dos estudantes no acesso ao ensino superior. Os dados utilizados referem-se ao ano lectivo 2010/2011, tendo sido disponibilizados pelo Ministério da Educação.

**Palavras-chave:** Análise de Agrupamentos, Ensino superior, Índice de satisfação da procura, Política de ensino superior.

## 1. Introdução

A implementação do Processo de Bolonha em Portugal foi liderada pelo Ministério da Ciência e Tecnologia e Ensino Superior (MCTES), fazendo parte do processo de organização e racionalização do sistema de ensino superior europeu, OCDE (2006). O MCTES determinou que as Instituições de Ensino Superior (IES) pudessem reestruturar os planos curriculares dos seus cursos de acordo com os princípios de Bolonha a partir do ano lectivo 2006/2007 tendo este processo de reestruturação de estar terminado no ano lectivo 2008/2009. Esta reestruturação provocou profundas alterações no sistema de ensino português. Este estudo pretende contribuir para a descrição e caracterização da colocação dos estudantes no ensino superior, por via da análise do índice de satisfação da procura (razão entre o número de alunos alocados na primeira escolha e o número de vagas disponíveis em cada programa) dos cursos superiores de engenharia oferecidos pelas IES portuguesas.

## 2. Metodologia

Os dados analisados estão disponíveis *on-line*, no site da Direcção Geral do Ensino Superior (DGES), sendo uma das suas atribuições a disseminação dos resultados dos concursos nacionais de acesso ao ensino superior. Os dados recolhidos respeitam ao ano lectivo de 2010/2011 e as variáveis consideradas são: procura global de cada curso (número total de alunos que incluíram o par instituição/programa entre suas preferências, independentemente da sua classificação), bem como o número de alunos que escolheram o curso como sua primeira

opção, como segunda opção e, assim por diante (até um máximo de seis opções), número de colocados por opção, número de vagas disponíveis para cada curso na primeira fase do processo de candidatura, número total de candidatos por sexo, número total de colocados por sexo, classificação do último estudante colocado, número de candidatos colocados nas seis escolhas, número de anos que compõem o curso (primeiro ciclo ou mestrado integrado), grau conferido pelo curso (licenciado ou mestre) e o índice de satisfação da procura (razão entre o número de alunos alocados na primeira escolha e o número de vagas disponíveis em cada programa). Foram consideradas 14 universidades (U) e 20 institutos politécnicos (IP), do sector público, oferecendo 275 cursos de engenharia dos quais 58 são mestrados integrados e 217 são primeiros ciclos.

## 3. Análise dos dados e resultados

Os cursos superiores de engenharia foram organizados de acordo com a Classificação Nacional de Áreas de Educação e Formação (CNAEF) (Figura 1).

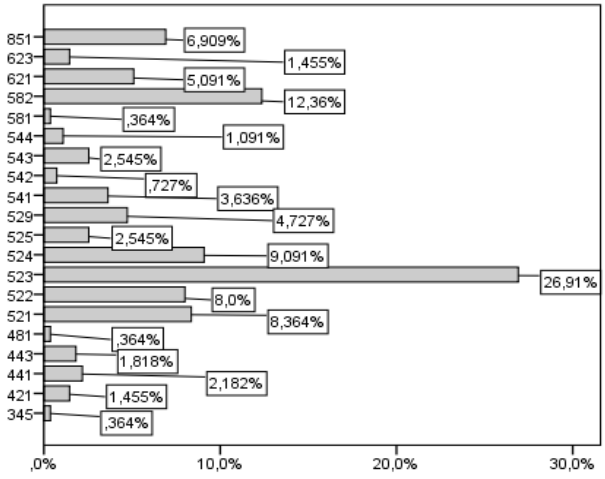



**Figura 1: Áreas de educação e formação**

Na Figura 2 apresenta-se a distribuição do índice de satisfação da procura de acordo com o tipo de IES. Foram aplicados alguns testes de hipóteses Mann-Whitney e Kruskal-Wallis (Conover, 1999), com o objectivo de verificar se o índice de satisfação da procura revela diferenças significativas por tipo de IES, do grau do curso superior ou das áreas de educação e formação. Os resultados do teste de Mann-Whitney indicam que o índide de satisfação da procura é significativamente diferente nos dois tipos de IES (estatística de teste=-8.367; p-value $\approx 0.00$), assim como é significativamente diferente nos dois tipos de tipos de cursos (estatística de teste=-8.069; p-value $\approx 0.00$). Situação idêntica é observada quando se compara as áreas de educação e formação. O teste de Kruskal-Wallis mostra-nos que o índice de satisfação da procura também revela diferenças significativas por áreas de educação e formação (estatística de teste= 34.852; p-value=0.015). Estes resultados reforçam o pressuposto de que o índice de

satisfação da procura está fortemente dependente das variáveis consideradas. Aplicou-se a Análise Classificatória Hierárquica pelo método aglomerativo (Gore, 2000) aos dados referidos, usando as variáveis áreas de formação e cursos. Foi usada a distância Euclidiana e como critério de agregação o método de Ward (Barnett, 1981). O dendrograma obtido e as classes definidas estão representados na Figura 3.

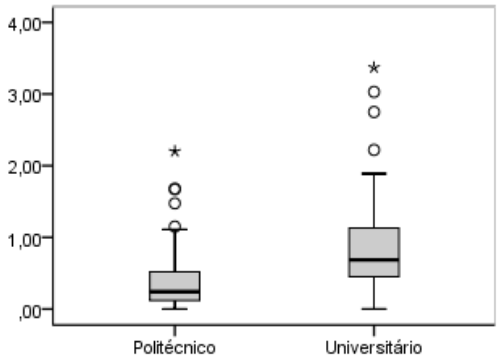



**Figura 2: Índice de satisfação da Procura por tipo de IES**

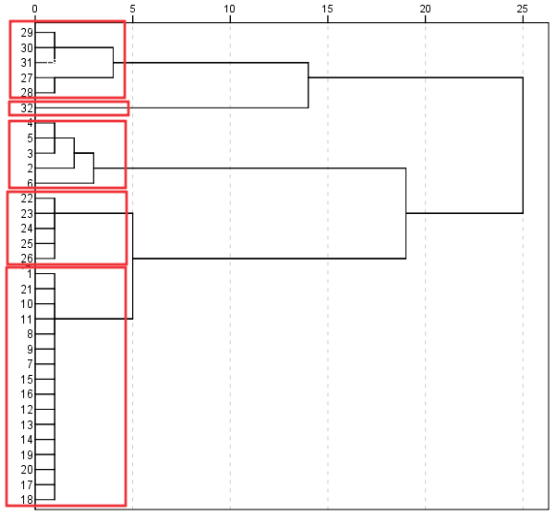



**Figura 3: Dendrograma obtido pelo modelo (distância euclidiana, método de Ward)**

A análise do dendrograma aponta para a formação de cinco classes definidas na Tabela 1, permitindo assim agrupar os cursos mais correlacionados no que respeita ao índice de satisfação da procura pelas áreas de formação, de acordo com as variáveis correlacionadas. Considerando as médias do índice de satisfação entre cada classe, resulta a seguinte ordenação das classes: Classe 2, Classe 5, Classe 1, Classe 3 e Classe 4. Este resultado confirma o conhecimento prévio sobre o rácio entre o número de vagas disponíveis e o número de estudantes colocados na 1ª opção, isto é, a classe com a média do índice de satisfação maior é aquela que corresponde aos cursos com menos candidatos.

**Tabela 1: Identificação de classes**

| | |
|---|---|
| **Classe1** | 27,28,28,30,31 |
| **Classe2** | 32 |
| **Classe3** | 2,3,4,5,6 |
| **Classe4** | 22,23,24,25,26 |
| **Classe5** | 1,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21 |

Este trabalho possibilitou um maior conhecimento da colocação dos estudantes nos cursos superiores de engenharia em Portugal e respectivo índice de satisfação. Este estudo empírico confirma que o índice de satisfação da procura é um indicador importante dos factores determinantes das escolhas dos estudantes no que respeita ao ensino superior, no entanto é um tema que merece ser objecto de um estudo mais detalhado e aprofundado em trabalho futuro, nomeadamente na aplicação de outras metodologias multivariadas, na análise de mais variáveis envolvidas no processo de acesso aos cursos de engenharia que estejam disponíveis.

**Agradecimentos:** Este trabalho foi parcialmente financiado pelo Centro de Matemática e pelo Centro de Ciência e Tecnologia Têxtil da Universidade do Minho por Fundos Nacionais através da FCT – Fundação para a Ciência e Tecnologia no âmbito do projecto PestOE/MAT/UI0013/2014 e do projecto Pest-C/CTM/UI0264/2013, respectivamente.

**Referências**

AMA(s.d.)http://www.dados.gov.pt/PT/CatalogoDados/Dados.aspx?name=ClassNacionaldeareasdeeducacaoeformacao (acedido em 25 de Fevereiro 2014).

BARNETT, V. (1981). *Interpreting Multivariate Data*. Sheffield, John Wiley & Sons.

CONOVER, W.J. (1999). *Practical Nonparametric Statistics*. Third Edition, New York, John Wiley & Sons.

DGES (s.d.)http://www.dges.mctes.pt/DGES/pt/Estudantes/Acesso (acedido 25 de Fevereiro 2014).

GORE, P.A., (2000). Cluster analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 297-321). New York: Academic Press.

OECD, ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2006). Reviews of National policies for education: Tertiary education in Portugal. Examiner's Report. Available at http://www.dges.mctes.pt/NR/rdonlyres/8B016D34-DAAB-4B50-ADBB-25AE105AEE88/2564/Backgoundreport.pdf (acedido em 25 de Fevereiro 2014).

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Estimação paramétrica e semi-paramétrica do índice de cauda utilizando o R

**Helena Penalva[1], Sandra Nunes[2], Manuela Neves[3]**

[1]*Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal, helena.penalva@esce.ips.pt;*

[2]*Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal e CMA/FCT/UNL, sandra.nunes@esce.ips.pt;*

[3]*Instituto Superior de Agronomia, Universidade de Lisboa e CEAUL, manela@isa.utl.pt*

**Sumário**

A Teoria de valores extremos tem como objetivo principal modelar eventos que ocorrem com probabilidade pequena. Na maioria das situações pretende-se a estimação de probabilidades de acontecimentos que são mais extremos do que aqueles que já foram observados. A estimação precisa do designado índice de cauda ou índice de valores extremos, $\gamma$, é de fundamental importância nesta teoria. Neste trabalho pretende-se utilizar a abordagem paramétrica e a semi-paramétrica na estimação de $\gamma$ e de outros parâmetros de interesse utilizando o software R.

**Palavras-chave:** Estimação paramétrica, Estimação semi-paramétrica, Índice de valores extremos, Software R, Teoria de valores extremos.

## 1. Introdução

Historicamente, as aplicações da Teoria de Valores Extremos (EVT) iniciaram-se em duas principais áreas: a área ambiental, com o estudo dos níveis do mar, velocidade do vento, caudal dos rios, entre outros; e a área da fiabilidade. Atualmente a Teoria de Valores Extremos tem surgido como uma das mais importantes áreas da Estatística em várias ciências aplicadas, tais como biologia, geologia e risco sísmico, hidrologia, finanças e telecomunicações.

A Teoria de Valores Extremos tem como um dos principais objetivos quantificar, descrever e inferir o comportamento dos valores extremos de um processo estocástico. A modelação dos máximos ou mínimos em amostras de dados de variáveis aleatórias ou da distribuição de excessos acima de certo limiar é uma peça chave desta teoria. Nestas distribuições, o parâmetro de forma, $\gamma$, designado por índice de cauda ou índice de valores extremos descreve o comportamento da cauda direita, $1-F$, do modelo subjacente aos dados. A sua estimação precisa é muito importante e de enorme influência na estimação de outros parâmetros, tais como quantis elevados, níveis e períodos de retorno, probabilidades de níveis elevados.

Neste trabalho pretende-se utilizar a abordagem paramétrica e a semi-paramétrica na estimação de $\gamma$ e de outros parâmetros de interesse utilizando o software R.

## 2. Conceitos básicos da Teoria de valores extremos

Numa primeira fase a teoria de extremos amostrais consistiu em determinar o comportamento limite do máximo $M_n := (X_1, \cdots, X_n)$ ou do mínimo $m_n := (X_1, \cdots, X_n)$, quando $n \to \infty$, de uma amostra $X_1, \cdots, X_n$ de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com função de distribuição (f.d.) F, desconhecida. Os primeiros resultados surgiram nos trabalhos de Fréchet (1927), Fisher and Tippet (1928), Gumbel (1935) e von Mises (1936), mas foi Gnedenko (1943) que definiu condições para a existência de sequências $\{a_n\} \in \mathbb{R}^+$ e $\{b_n\} \in \mathbb{R}$, tais que

$$\lim_{n\to\infty} P[(M_n - b_n)/a_n \leq x] = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \forall_{x\in\mathbb{R}}, \qquad (1)$$

sendo G uma função distribuição não degenerada. Esta função, chamada função de distribuição de valores extremos, geralmente denotada por $EV_\gamma$, é dada por

$$EV_\gamma(x) = \begin{cases} \exp\left[-\left(1+\gamma\dfrac{x-\mu}{\sigma}\right)^{-1/\gamma}\right], & 1+\gamma\dfrac{x-\mu}{\sigma} > 0, \ se\ \gamma \neq 0 \\ \exp\left[-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right], & x \in \mathbb{R}, \ \text{se } \gamma = 0, \end{cases}$$

onde $\gamma \in \mathbb{R}$, designado por índice de valores extremo, é o parâmetro de forma, e $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}^+$, são respetivamente os parâmetros de localização e escala. Quando o limite em (1) se verifica diz-se que F pertence ao domínio de atração (para máximos) de $EV_\gamma$ e escreve-se $F \in D_M(EV_\gamma)$ .

## 3. Abordagens paramétrica e semi-paramétrica

Estimar adequadamente vários parâmetros de interesse, tais como: os parâmetros de forma, $\gamma$; de localização, $\mu$; de escala, $\sigma$; quantis elevados ou níveis de retorno, período de retorno; o limite superior do suporte de uma distribuição F, continua a ser objeto de investigação.

A inferência estatística de extremos foi desenvolvida sob dois contextos: o paramétrico e o semi-paramétrico. A abordagem clássica, que remonta a Gumbel (1958), surgiu em contexto paramétrico, no qual a estimação é baseada em resultados assintóticos para o máximo. Pode-se referir a metodologia de máximos de blocos ou metodologia de Gumbel, a metodologia das *k*-maiores observações referentes a *m* blocos e o chamado método *Peaks Over Threshold* (POT), baseado em resultados de Pickands (1975). Neste contexto a inferência é feita baseada nos procedimentos habituais de estimação: máxima verosimilhança, máxima verosimilhança de perfil e método dos momentos ponderados de probabilidade.

A abordagem semi-paramétrica, iniciada pela escola holandesa de extremos (de Haan, 1970), considera as *k* maiores estatísticas ordinais associadas à totalidade das observações, não se definindo um modelo paramétrico, mas admitindo apenas que F satisfaz certas condições que

garantem que os máximos, convenientemente normalizados, convergem para a f.d. $EV_\gamma$, com parâmetro de forma $\gamma$. Nesta abordagem, e dado que não se define nenhum modelo paramétrico, os estimadores são definidos como funções das *k* maiores estatísticas ordinais. Tem surgido na literatura uma grande variedade de estimadores de $\gamma$, dos quais destacamos o estimador de Hill (1975), talvez o mais popular, o estimador de Pickands (1975), o estimador dos momentos, Dekkers et al (1989) e mais recentemente os estimadores de viés reduzido. Uma revisão muito interessante, apresentando ainda sugestões de trabalho futuro encontra-se em Beirlant et al (2012).

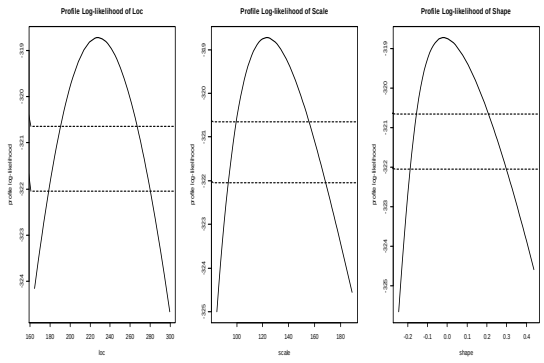
## 4. Uma aplicação a dados de valores extremos com o software R

O objetivo deste trabalho é ilustrar a estimação de $\gamma$ considerando as abordagens referidas, com recurso a um conjunto de dados referentes a níveis médios diários do caudal do rio Paiva medidos na estação hidrométrica de Fragas da Torre, entre 1946/47 a 1995/96. Os 50 anos observados correspondem exatamente ao período entre 1 de Outubro de 1946 e 30 de Setembro de 1996.

Da totalidade dos dados recolhidos, considerou-se apenas os períodos compreendidos entre Novembro de cada ano e Abril do ano seguinte, um pouco mais alargado que o período considerado em Gomes (1993). A estacionaridade deste conjunto de dados foi testada.

As estimativas dos parâmetros e respetivos intervalos de confiança podem ser obtidos usando os comandos:**library(evir)**; **gev(dados, block=181)**

As estimativas para $(\gamma, \mu, \sigma)$ são, respetivamamente: (-0.019, 227.875, 123.051).




**Gráficos da log-verosimilhança**

Para obter os intervalos a 95% confiança de Wald para os parâmetros da distribuição usámos:

**library(evd)**;**fgev(gev$data)**;**confint(fgev(gev$data),level=0.95).**

Para obter os intervalos de confiança de perfil da log-verosimilhança e o respectivo gráfico.

**x=fgev(gev$data); confint(profile(x), level=0.95); plot(profile(x), ci = c(0.95, 0.99)).**

| Parâmetros | Limite inferior | Limite superior |
|---|---|---|
| $\gamma$ | -0.16 | 0.21 |
| $\mu$ | 190.53 | 267.22 |
| $\sigma$ | 99.55 | 155.46 |

**I.C de perfil a 95% para os parâmetros**

Para visualizar alguns gráficos de diagnóstico:

**library(ismev)**; **gev.diag(gev.fit(gev$data)).**

Na abordagem semi-paramétrica, apresentamos aqui apenas a estimação de γ baseada no estimador de Hill que pode ser usado diretamente no R.

**library(evir); hill(gev$data, option="xi").**

Os outros estimadores, também usados neste trabalho, tais como os estimadores de viés reduzido, não estão ainda implementados no R, pelo que foram programados por nós.

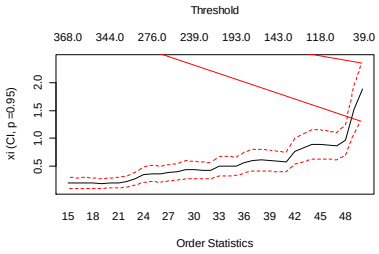



**Gráfico das estimativas de γ em função de *k***

**Agradecimentos:** Investigação parcialmente suportada por fundos nacionais através da FCT-Fundação para a Ciência e a Tecnologia, projetos PEst-OE/MAT/UI0006/2011, 2014 (CEAUL) e PEst-OE/MAT/UI0297/2011, 2014 (CMA/FCT/UNL).

## Referências

BEIRLANT, J., CAEIRO, F. & GOMES, M. I. (2012). An overview and open research topics in Statistics of Univariate Extremes, *Revstat*, 10:1, 1-31.

DEKKERS, A.L.M., EINMAHL, J.H.J.& de HAAN, L. (1989). A Moment Estimator for the Index of an Extreme-Value Distribution. *Ann. Statist*, 17(4), 1833-1855.

FISHER, R.A. & TIPPETT, L.H.C. (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180-190.

FRÉCHET, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math. (Cracovie)*, 6, 93-116.

GOMES, M.I. (1993). On the estimation of parameters of rare events in environmental time series. *Statistics for the Environment*, 226-241.

GNEDENKO, B.V.(1943). Sur la distribution limite d'une série aléatoire. *Annals of Mathematics*, 44, 423-453.

GUMBEL, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.

de HAAN, L. (1970). On regular variation and its Applications to the Week Convergence of Sample Extremes. *Mathematical Centre Tract 32*, Amsterdam.

HILL, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist*, 3, 1163-1174.

PICKANDS, J.(1975). Statistical inference using extreme order statistics. *Ann. Stat*., 3, 119-131.

VON MISES, R.(1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalkanique*, 1, 141-160.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Fatores determinantes na manutenção da relação de compromisso entre os *alumni* e a *alma mater*: aplicação de um modelo de equações estruturais

**Luís Nobre Pereira[1], Ilda Pedro[2], Helder Carrasqueira[3]**

[1]*Escola Superior de Gestão, Hotelaria e Turismo & Centro de Investigação sobre o Espaço e as Organizações (CIEO) - Universidade do Algarve, lmper@ualg.pt*

[2]*Escola Superior de Gestão, Hotelaria e Turismo-Universidade do Algarve, ipedro@ualg.pt*

[3]*Escola Superior de Gestão, Hotelaria e Turismo-Universidade do Algarve, hcarrasq@ualg.pt*

**Sumário**

Este trabalho pretende ser um contributo na obtenção de informações úteis para a definição de estratégias, tendentes à manutenção de relações duradouras entre uma Instituição de Ensino Superior (IES) e os seus *alumni,* por serem reconhecidamente parceiros que as IES não podem menosprezar. Tendo em vista esse objectivo, foi realizada uma sondagem, através de um questionário aplicado *online,* a uma amostra de 631 *alumni* de uma IES. Foi estimado um modelo concetual teórico que estabelece relações de dependência entre os fatores de satisfação e imagem em relação ao fator compromisso.

**Palavras-chave:** Sondagens, Modelo de Equações Estruturais, Compromisso, Imagem, Satisfação.

## 1. Introdução

No dealbar deste século têm sido significativas as mudanças a que as instituições de ensino superior (IES) portuguesas têm sido sujeitas. Esta conjuntura não é exclusiva da realidade do ensino superior português, tão pouco que tenha aparecido agora, apenas se agudizou. Países anglo-saxónicos, Estados Unidos, Austrália, testemunham também uma mudança significativa da realidade no ensino superior (Newman & Petrosko, 2011; Duarte, Alves, & Raposo, 2010; Çetin, 2004; McAlexander & Koenig, 2001; Kotler & Fox, 1994).

As IES vêm-se confrontadas com necessidades de ajustamento que, entre outras medidas, passam pela adoção de estratégias de marketing conducentes a um reposicionamento no mercado, formação da imagem institucional e foco na relação com os estudantes, de modo a ganharem vantagem competitiva. Através dos diferentes papéis que os *alumni* podem assumir nas instituições estes constituem parceiros com os quais há que manter uma relação duradoura, pelo que se impõe identificar os fatores determinantes na manutenção da relação de compromisso entre os *alumni* e a *alma mater,* sendo esta a nossa questão de partida.

Detemo-nos na satisfação e imagem, por serem duas variáveis frequentemente referidas como condicionantes de fatores como a lealdade, referências positivas, vontade de voltar a comprar, vontade de participar em atividades da instituição, reforço da confiança, entre outros aspetos (por nós entendidos como componentes do fator compromisso) (Rodrigues, 2012; Brown & Mazzarol, 2008; Helgesen & Nesset, 2007; Al-Alak, 2006; Çetin, 2004; Elliott &

Shin, 2002). Para melhor entender a influência destes fatores sobre o compromisso definimos como objetivos específicos a identificação dos componentes de ambos os fatores na manutenção da relação de compromisso, para além de um terceiro objetivo específico que se prende com a identificação de produtos específicos determinantes do relacionamento futuro.

## 2. Metodologia

### 2.1 População e amostra

A população-alvo deste estudo é formada por todos os *alumni* de uma IES, tendo-se definido como população em estudo todos os diplomados entre os anos letivos 2002/03 até 2011/12, num total de 2743 *alumni*. Todos os indivíduos desta população foram convidados para participarem no estudo através de um *email*, tendo sido possível observar uma amostra de 631 indivíduos. Estes indivíduos responderam a um questionário, com recolha de dados através de entrevista electrónica, cujo *link* de acesso foi colocado no referido *email*.

### 2.2 Hipóteses de investigação e conceção do modelo teórico

No estabelecimento de relações de dependência entre estas dimensões levantámos as hipóteses subjacentes no modelo apresentado na figura 1. Considerámos os resultados de estudos de idêntica natureza, e concomitantemente tivemos em conta a especificidade do objeto de estudo, pois na prática, queremos perceber as relações entre dimensões específicas. À semelhança da maior parte de outros estudos, entendemos considerar relações positivas entre estas dimensões.

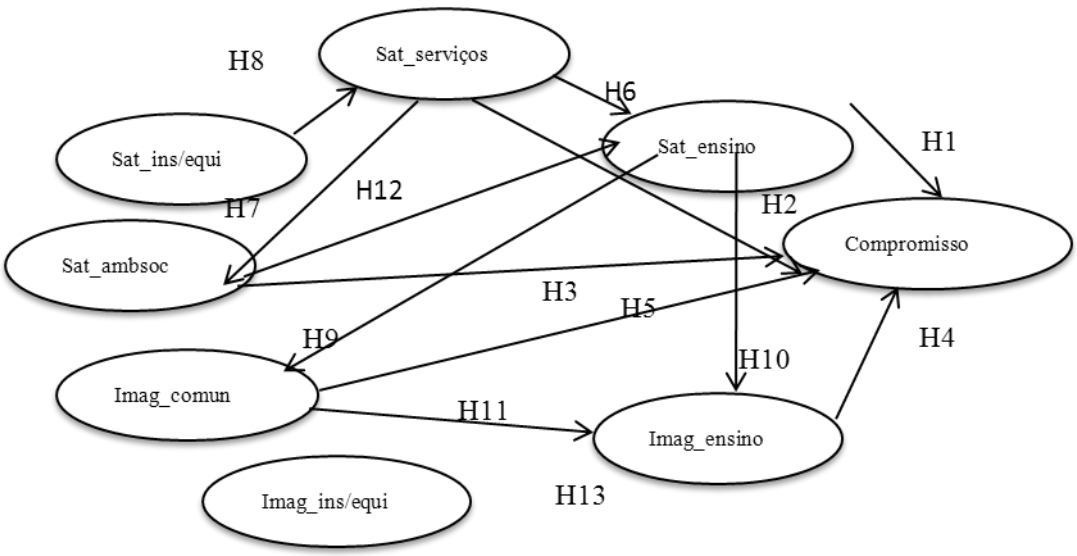



**Figura 1 - Modelo estrutural proposto**

## 3. Análise e interpretação dos resultados

Foi feita a análise da dimensionalidade através da análise fatorial confirmatória e avaliação da fiabilidade e consistência interna pelo *Alfa* de *Cronbach* e fiabilidade compósita. Foram analisadas também as validades, convergente e discriminante, no modelo de medida. Depois de terem sido removidos os *outliers*, o modelo apresentava as estatísticas de qualidade de ajustamento com índices considerados sofríveis: $\chi^2$=2098,468; $\chi^2$/gl=3.128; *p*-value <0,001; GFI=0,830; AGFI= 0,802 e os seguintes índices que permitiam considerar um bom ajustamento: NFI=0,916; CFI=0,941; RMSEA= 0,063. Para além disso, salienta-se que os valores do *Alfa* de *Cronbach* situaram-se na ordem dos 0,9, os pesos fatoriais estandardizados estavam todos muito acima de 0,5, os valores da variância extraída média encontravam-se todos acima de 0,5, os quais, por sua vez, eram superiores aos valores do quadrado da correlação entre os factores. Assegurada a qualidade do submodelo de medida procedemos à especificação e identificação do modelo estrutural global apresentado na figura 2. As variáveis *Q* representam as variáveis manifestas.

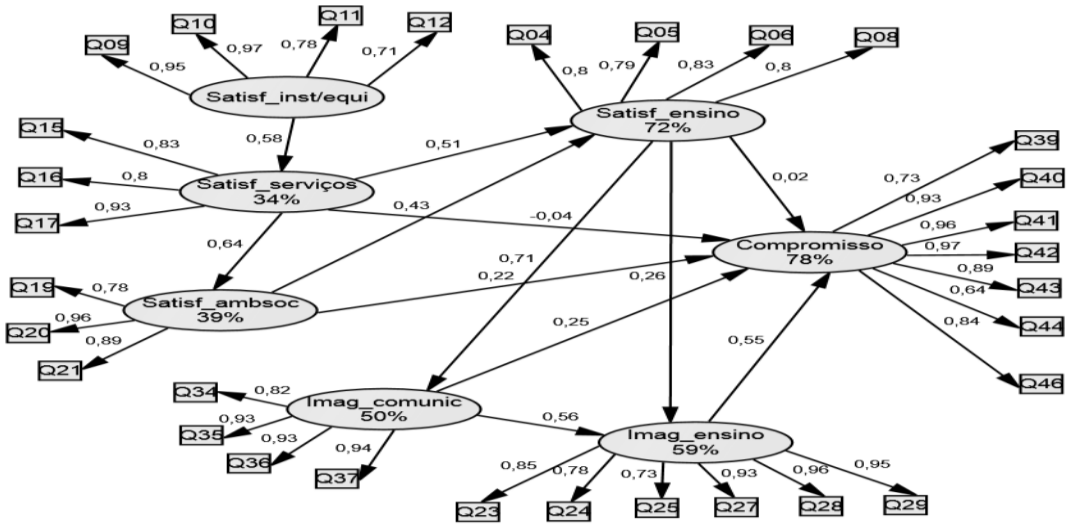



**Figura 2 - Modelo estrutural estimado**

A maioria dos *alumni* da amostra são mulheres. A idade média é de aproximadamente 34 anos com um desvio padrão de 7,6, sendo que a idade se encontra compreendida entre os 21 e os 66 anos. A grande maioria, 92,4%, trabalha em Portugal e é trabalhador por conta doutrem, 74,6%. A percentagem de inscritos na associação *alumni* é de 11,4%. No que se refere à estimação do modelo, verificámos que os índices de qualidade de ajustamento apresentavam os seguintes valores: $\chi^2$=1400,051; $\chi^2$/gl=3,357; p-value <0,001; GFI=0,863; AGFI= 0,837, índices que ainda assim mostravam um ajustamento sofrível. Por seu lado: NFI=0,925; CFI=0,946; RMSEA= 0,067, que revelavam um ajustamento superior ao modelo de medida. Em termos globais, verificámos que apenas as hipóteses H1 e H2 não se confirmam.

## 4. Conclusão

Pela análise dos resultados obtidos, concluímos que os determinantes para a manutenção da relação de compromisso entre os *alumni* e a *alma mater* são a imagem no ensino, a imagem da comunicação e a satisfação com o ambiente social e académico.

**Agradecimentos:** O CIEO é financiado pela Fundação para a Ciência e a Tecnologia.

## Referências

AL-ALAK, B. A. M. (2006). The Impact of Marketing Actions on Relationship Quality in the Higher Education Sector in Jordan. *Journal of Marketing for Higher Education*, *16*(2), 1–23.

BROWN, R. M., & MAZZAROL, T. W. (2008). The importance of institutional image to student satisfaction and loyalty within higher education. *Higher Education*, *58*(1), 81–95.

ÇETIN, R. (2004). Planning and Implementing Institutional Image and Promoting Academic Programs in Higher Education. *Journal of Marketing for Higher Education*, *13*(1-2), 57–75.

DUARTE, P. O., ALVES, H. B., & RAPOSO, M. B. (2010). Understanding university image: a structural equation model approach. *International Review on Public and Nonprofit Marketing*, *7*(1), 21–36.

ELLIOTT, K. M., & SHIN, D. (2002). Student Satisfaction : An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management*, *24*(2), 197–209.

HELGESEN, Ø., & NESSET, E. (2007). Images, Satisfaction and Antecedents: Drivers of Student Loyalty? A Case Study of a Norwegian University College. *Corporate Reputation Review*, *10*(1), 38–59.

KOTLER, P., & FOX, K. F. A. (1994). *Marketing Estratégico para Instituições Educacionais.* São Paulo: Editora Atlas.

MCALEXANDER, J. H., & KOENIG, H. F. (2001). University Experiences , the Student-College Relationship , and Alumni Support. *Journal of Marketing for Higher Education*, *10*(3), 21–44.

NEWMAN, M. D., & PETROSKO, J. M. (2011). Predictors of Alumni Association Membership. *Research in Higher Education*, *52*, 738–759.

RODRIGUES, S. L. (2012). *Os fatores que influenciam a formação da imagem das instituições de ensino superior: o caso do Instituto Politécnico de Leiria na perspetiva dos professores do ensino secundário do distrito de Leiria.* Dissertação de Mestrado, Instituto Politécnico de Leiria.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Modelação de contagens com excesso de zeros

**Jorge Santos[1], Susana Faria[2]**

[1]*Departamento de Matemática e Aplicações – Universidade do Minho, Campus de Azurém, 4800-058 Guimarães, jorge.mfd@sapo.pt;*

[2]*CMAT – Centro de Matemática, Departamento Matemática e Aplicações, Universidade do Minho, Campus de Azurém, sfaria@math.uminho.pt;*

**Sumário**

Os modelos de regressão para dados de contagem são muito utilizados nas mais variadas áreas de estudo para a modelação de fenómenos. Neste trabalho é feita uma abordagem aos modelos de regressão para variáveis resposta na forma de contagens, com problemas de excesso de zeros e/ou de sobredispersão. A aplicação desses modelos de regressão a dados bancários mostra que os modelos de zeros inflacionados apresentaram um melhor ajustamento, quando comparados com os modelos que não têm em consideração o excesso de zeros.

**Palavras-chave:** Dados de contagem, Modelos de regressão de Poisson e Binomial Negativa, Modelos de regressão de zeros inflacionados.

## 1. Introdução

Os dados de contagem são um tipo de dados muito frequentes nas mais diversas áreas deestudo. A natureza deste tipo de dados que assume apenas valores inteiros não negativos, correspondentes à ocorrência de um dado número de eventos durante um intervalo de tempo ou espaço, levaram ao aparecimento dos modelos de regressão para dados de contagem.

O modelo de regressão de Poisson é o modelo mais utilizado para este tipo de dados (Cameron e Trivedi, 1998), mas um problema comum neste modelo surge quando a variância da variável resposta é superior ao seu valor médio (no caso do modelo de Poisson a variância da variável resposta é igual ao seu valor médio). Este fenómeno designa-se por sobredispersão (Hinde e Demetrio, 1998). Uma possibilidade para ultrapassar este problema consiste em recorrer ao modelo de regressão binomial negativa (Hilbe, 2001). No entanto, nos dados de contagem é ainda muito comum existir excesso de zeros, e em muitas situações, este problema não é convenientemente modelado pelo modelo de regressão binomial negativa. Nestes casos, a modelação dos dados pode passar pela utilização de modelos de regressão de zeros inflacionados (Lambert, 1992, Cheung, 2002).

Os modelos de regressão de zeros inflacionados modelam as contagens como uma mistura de duas distribuições com dois processos subjacentes, um processo que trata do excesso de zeros, modelado por uma massa pontual em zero e assumindo que com probabilidade p a única observação possível é zero, e um outro que trata das contagens, modelado por uma distribuição de Poisson ou Binomial Negativa, com probabilidade 1−p (Zuur *et al*., 2009).

Neste trabalho estudou-se a relação do número de não pagamento da prestação do empréstimo de um cliente em função das características do cliente e do contrato. Em particular,

foram ajustados os modelos de regressão de Poisson, modelos de regressão Binomial Negativa, modelos de regressão de Poisson de zeros inflacionados e modelos de regressão binomial negativa de zeros inflacionados. Os resultados obtidos mediante a aplicação dos vários modelos foram discutidos e comparados.

A análise estatística foi realizada utilizando o *software* R, versão 2.14.1 (R Development Core Team, 2011). O modelo de regressão de Poisson encontra-se implementado no *package stats* recorrendo à função *glm( ).* O modelo de regressão Binomial Negativa encontra-se implementado no *package MASS* recorrendo à função *glm.nb( ).* Os modelos ZIP e ZINB encontram-se implementados no *package pscl* recorrendo à função *zeroinfl( ).*

## 2. Dados e Resultados

Os dados utilizados neste estudo dizem respeito a uma amostra aleatória de clientes de um banco a quem foi garantido crédito de consumo. Para cada cliente foi recolhida informação sobre várias características no início do contrato (idade, sexo, estado civil, profissão, montante contratado, prestação mensal do empréstimo, indicador se o cliente recebe o ordenado através do banco), assim como, o número total consecutivo de meses sem pagamento da prestação (variável resposta).

A análise da distribuição da variável resposta revelou claramente a existência de uma elevada incidência de zeros e de problemas de sobredispersão.

Vários modelos de regressão para dados de contagem foram ajustados aos dados. O modelo inicialmente estimado, o modelo de regressão de Poisson, revelou-se inapropriado, uma vez que apresentava sobredispersão. Neste modelo cerca de 30% dos valores ajustados são diferentes dos observados, e o valor ajustado pelo modelo para o número de não incumprimentos difere em 17% dos casos. Ajustou-se de seguida o modelo de regressão binomial negativa, verificando-se que este modelo é preferível ao modelo de regressão de Poisson. Para este modelo, a diferença entre os valores ajustados e os valores observados é bastante menor, com apenas 11% de diferenças. O valor ajustado pelo modelo para o número de incumprimentos igual a zero descresce para 2%, relativamente ao modelo de regressão de Poisson. No caso do modelo de regressão de Poisson de zeros inflacionados, os valores ajustados diferem 13% dos observados e o valor ajustado para o número de incumprimentos igual a zero é aproximadamente igual a 1%, revelando uma melhoria significativa em comparação com o modelo de regressão de Poisson. Por sua vez, o modelo de regressão binomial negativa de zeros inflacionados é o que melhor se ajusta aos dados, apresentando apenas 10% de diferenças entre os valores ajustados e os valores observados. O valor estimado para o número de não incumprimentos é de aproximadamente 2% dos casos.

A análise da qualidade de ajustamento dos diversos modelos indica os modelos de zeros inflacionados como sendo os modelos que melhor se ajustam aos dados (menores valores de AIC e BIC).

## 3. Conclusões

Os resultados obtidos mostraram que os modelos de regressão de zeros inflacionados apresentam um melhor ajustamento, quando comparados com os modelos que não têm em consideração o excesso de zeros. Mostraram ainda que os modelos baseados na distribuição Binomial Negativa, são os mais adequados para modelar estes dados, em vez dos modelos baseados na distribuição de Poisson.

Segundo o modelo de regressão binomial negativa de zeros inflacionados este modelo, o número esperado de incumprimentos aumenta com o aumento da idade do contrato e com o aumento do número de meses que ainda faltam para liquidar o empréstimo. Se o cliente for um cliente antigo e se receber o ordenado pelo banco, o número esperado de incumprimentos ao banco, diminui.

Para trabalho futuro, pretendemos aplicar os modelos de barreira *(Hurdle models*), os modelos de mistura de regressão de Poisson e os modelos de mistura de regressão binomial negativa a esta base de dados.

**Agradecimentos:** S. Faria foi financiada pelo CMAT da Universidade do Minho com Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto PEst-OE/MAT/UI0013/2014.

## Referências

CAMERON, A.C., TRIVEDI P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK.

CHEUNG,Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine* 21, 1461-1469.

HILBE J.M. (2001). *Negative Binomial Regression*, Second Edition, Cambridge University Press, New York.

HINDE, J., DEMETRIO, C. G. (1998). Overdispersion: models and estimation, *Computational Statistics and Data Analysis* 27(2), 151–170.

LAMBERT D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1-14.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011. http://www.R-project.org

ZUUR A.F., IENO N.E., WALKER N.J., SAVELIEV A.A., SMITH G.M. (2009). *Mixed effects models and extensions in ecology with R*. Springer.

**Sessão de Posters I – 6ª Feira, 11 de Abril (16:15)**

# Utilização de telemóveis entre a população sénior

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

**Sumário**

O telemóvel é um dispositivo de comunicação com grande penetração entre os jovens mas é na faixa etária 55-64 anos que se tem registado a maior taxa de crescimento de posse e utilização de telemóvel (Marktest, 2005). A população sénior tem valores, atitudes e percepções próprios e constitui-se como um grupo heterogéneo em termos de estilos de vida, *status* financeiro e outras características relacionadas com o comportamento do consumidor (Cedru, 2008). Este estudo avalia a utilização dos telemóveis pela população sénior e propõe uma segmentação baseada em atitudes e comportamentos face aos telemóveis relevante para o mercado das telecomunicações móveis.

**Palavras-chave:** Análise de Clusters, População sénior, Telemóveis.

## 1. Introdução

Antes da viragem do milénio já se previa que a população sénior seria responsável por mais de metade das decisões de compra na Europa, fazendo as empresas perceber quão o rápido crescimento do poder económico e mudança de estilos de vida dos mais velhos afectaria o seu desempenho comercial (Vuori & Holmund-Rytkönen, 2005). A investigação sobre comportamentos e atitudes da população sénior num contexto de marketing e comportamento do consumidor é no entanto minoritária se comparada com a que se faz nos segmentos mais jovens da população.

A utilização das tecnologias de informação e comunicação pelos séniores é uma temática ainda pouco investigada, pois os séniores são tradicionalmente retratados como relutantes em adoptar as novas tecnologias e em acompanhar os seus desenvolvimentos. De facto, segundo o Eurobarometer 241 os utilizadores mais intensivos dos telemóveis encontram-se na faixa etária 15-24 anos – 73,1% destes indivíduos usam o telemóvel várias vezes por dia; na faixa sénior (55 ou mais anos) essa percentagem não chega a 25% (European Commission, 2008).

No entanto, é de esperar que cada sucessiva geração de séniores se torne mais aderente e experiente em todos os tipos de tecnologia do que as gerações precedentes, o que se traduzirá a longo prazo numa utilização crescente dos telemóveis e outros dispositivos de comunicação por parte deste grupo da população.

Há portanto uma necessidade real de conhecer a população sénior de modo a construir novas perspectivas no desenvolvimento de estratégias de marketing direccionadas para este grupo. A segmentação é central para este propósito e o objectivo deste estudo prende-se justamente com uma proposta de segmentação da população sénior.

## 2. Dados e metodologia

Foi realizada uma sondagem CATI móvel aos utilizadores de telemóveis residentes em Portugal com 15 ou mais anos de idade. Os números de telemóvel a marcar foram seleccionados aleatoriamente tendo-se estratificado a amostra por operador do serviço móvel. Na faixa etária de 55 ou mais anos de idade inquiriram-se 363 indivíduos sobre os quais incidem as análises deste estudo.

O questionário incluía três secções: (1) posse e utilização de telemóvel, (2) atitudes face aos telemóveis e (3) caracterização demográfica. As atitudes foram avaliadas mediante uma bateria de 5 itens – o telemóvel ajuda-me no trabalho, a maioria das chamadas profissionais que recebo fora do período de trabalho são incómodas e invadem a minha privacidade, é importante para mim a opinião dos outros acerca do meu telemóvel, sem telemóvel sinto-me desligado do mundo, o telemóvel é apenas um equipamento técnico que permite fazer e receber chamadas - , sendo a resposta a cada item dada segundo uma escala ordinal tipo Likert de 1-concordo totalmente a 4-discordo totalmente.

Os 5 itens atitudinais foram utilizados numa Análise Hierárquica de Clusters com o objectivo de agrupar os utilizadores de telemóveis em função da partilha de atitudes semelhantes.

## 3. Resultados

A Análise de Clusters revelou três segmentos de utilizadores de telemóveis com os seguintes pesos: Grupo 1 – 48,2%, Grupo 2 – 28% e Grupo 3 – 23,8%. O Grupo 1 destaca-se dos restantes pela elevada discordância com a afirmação "o telemóvel ajuda-me no trabalho". O Grupo 2 destaca-se pela elevada concordância com as afirmações "sem telemóvel sinto-me desligado do mundo" e "o telemóvel ajuda-me no trabalho" e discordância com a afirmação "o telemóvel é apenas um equipamento técnico que permite fazer e receber chamadas". O grupo 3 destaca-se pela elevada discordância com a afirmação "sem telemóvel sinto-me desligado do mundo" e forte concordância com a afirmação "o telemóvel ajuda-me no trabalho". A distribuição de concordância com a afirmação "a maioria das chamadas profissionais que recebo fora do período de trabalho são incómodas e invadem a minha privacidade" não foi significativamente diferente entre os 3 grupos ($p>0,05$).

Os utilizadores do Grupo 1 são predominantemente mulheres, com mais idade do que os utilizadores classificados nos outros dois grupos e têm habilitações tendencialmente baixas. Este grupo tem uma menor percentagem de indivíduos profissionalmente activos face aos outros dois grupos; os utilizadores deste grupo pertencem sobretudo às classes sociais média baixa e baixa. O Grupo 2 e o Grupo 3 distinguem-se entre si sobretudo na situação profissional, nível de habilitações e classe social. Concretamente, o Grupo 3 é o grupo com menor percentagem de reformados/aposentados, tem a maior percentagem de indivíduos com habilitações de nível superior/universitário, e uma maior percentagem de utilizadores pertencentes às classes sociais alta e média alta.

Em termos de utilização do telemóvel o Grupo 3 destaca-se por os seus elementos utilizarem funcionalidades do telemóvel não somente ligadas à comunicação de voz ou de texto, mas outras que estão associadas eventualmente a uma atividade profissional como sejam consultar/editar agenda, ver e-mails ou aceder à Internet. O Grupo 1 regista as percentagens mais baixas de utilizadores das diversas funcionalidades avaliadas. Em termos de tarifário, o sistema pré-pago sem carregamentos obrigatórios é dominante entre os utilizadores do Grupo 1 enquanto que os sistemas de assinatura mensal ou o pré-pago com carregamentos obrigatórios são mais frequentes nos grupos 2 e 3.

Em síntese o Grupo 1 pode ser considerado como o segmento dos utilizadores "indiferentes", que vêem o telemóvel apenas como um aparelho que satisfaz as suas necessidades elementares de comunicação telefónica. O Grupo 3 agrega os utilizadores "funcionais" que vêem o telemóvel como um instrumento de trabalho mas que não valorizam os aspectos subjectivos e emotivos ligados ao telemóvel. Finalmente, o Grupo 2 agrega os utilizadores "móvel-dependentes" que além de valorizarem o telemóvel pela sua utilidade na actividade profissional valorizam também o telemóvel enquanto meio de contacto social e de status junto dos outros.

**Agradecimentos:** *Fundação para a Ciência e Tecnologia*, PTDC/EGE-GES/116934/2010.

## Referências

CEDRU (2008) Estudo de avaliação das necessidades dos séniores em Portugal, http://www.akdn.org/publications/2008_portugal_estudo%20seniores.pdf (acedido em 20 de fevereiro 2014)

EUROPEAN COMMISSION (2008) *Flash Eurobarometer 241:Information Society as seen by EU citizens.* Belgium, European Commission.

MARKTEST (2005). Barómetro das Comunicações , http://www.marktest.com/wap/a/n/id~946.aspx, (acedido em 20 de fevereiro 2014).

VUORI, S. & HOLMLUND-RYTKÖNEN, M. (2005). 55+ people as internet users, *Marketing Intelligence and Planning*, 23, 1, 58-76.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Análise fatorial confirmatória - Escala de integração comunitária de adultos com problemas psiquiátricos

**Joana Cabral[1], Célia Barreto Carvalho[2], Osvaldo Silva[3]**
[1]*Universidade dos Açores, joana.cabral@uac.pt*
[2]*Uiversidade dos Açores, ccarvalho@uac.pt*
[3]*Universidade dos Açores, CES-UA, osilva@uac.pt*

**Sumário**

A Escala de Integração Comunitária de Adultos com Problemas Psiquiátricos (EIC-APP) é um instrumento de medida que foi construído com base num modelo multidimensional de integração comunitária. Para verificar a validade da referida escala numa amostra de 183 indivíduos açorianos, com problemas psiquiátricos, foi feita uma análise fatorial confirmatória com o *software* AMOS. Os resultados obtidos levaram à escolha de um modelo de 2ª ordem com dois fatores, englobando 12 itens da escala inicial.

**Palavras-chave:** Análise fatorial confirmatória, Integração comunitária, Perturbações psiquiátricas.

## 1. Introdução

A Integração Comunitária tem-se revelado um fator importante na recuperação e bem-estar das pessoas com problemas psiquiátricos (Cabral & Barreto Carvalho, 2013), tornando-se cada vez mais relevante o estudo deste construto. Existem várias definições de Integração comunitária de doentes psiquiátricos, entre os quais, destaca-se o modelo multidimensional de integração comunitária referido por Wong e Solomon (2002) composto por três dimensões, que são: a física, a social e a psicológica. De acordo com este modelo, a **Integração Física** refere-se à forma como um indivíduo com problemas psiquiátricos passa o seu tempo fora de casa, participa em atividades comunitárias e utiliza os recursos da comunidade, por auto-iniciativa (Segal et al., 1980, *cit. in* Wong & Solomon, 2002). A **Integração Social** refere-se à quantidade, qualidade e diversidade das relações que estabelece com os membros da comunidade (Wolfensberger & Thomas, 1983, *cit. in* Wong & Solomon, 2002). Por fim, a **Integração Psicológica** diz respeito à forma como o indivíduo se vê como membro integrante da sua comunidade, expressa uma ligação emocional aos vizinhos, acredita na sua capacidade para satisfazer as suas necessidades a partir da vizinhança, bem como de exercer influência sobre a sua comunidade (Aubry & Myner, 1996; McMillan & Chavis, 1986, *cit. in* Wong & Solomon, 2002).

Um estudo realizado por Gulcur, Tsemberis, Stefancic e Greenwood, (2007) sugere que se acrescente a dimensão **Independência** ao modelo de Integração Comunitária de Wong e Solomon (2002), bem como, refere a necessidade de se criar um instrumento de medida que avalie as várias dimensões da integração comunitária.

## 2. Metodologia utilizada

A inexistência de instrumentos que avaliem o construto supracitado levou à construção da Escala de Integração Comunitária de Adultos com Problemas Psiquiátricos (EIC-APP), que foi desenvolvida tendo por base as dimensões (física, social e psicológica) definidas por Wong e Solomon (2002) e a dimensão independência sugerida por Gulcur (et al, 2007). A EIC-APP é um instrumento de auto-resposta, composto por um total de 34 itens. A resposta aos itens é realizada numa escala de *Likert* de 1 ("*Discordo totalmente*") a 5 ("*Concordo totalmente*"), sendo que as cotações mais elevadas indicam níveis mais altos de integração comunitária. A referida escala passou por vários procedimentos de construção e validação, detalhadamente descritos no estudo de Cabral e Barreto Carvalho (2013).

A amostra deste estudo é composta por 183 indivíduos com problemas psiquiátricos que vivem no Grupo Oriental da região Autónoma dos Açores (ilhas de Santa Maria e São Miguel); de ambos os sexos (72,7% ($n$=133) do sexo feminino e 27,3% ($n$=50) do sexo masculino) e idades compreendidas entre os 19 e os 78 anos (média=44,26; desvio padrão=13,5). Com o intuito de verificar a validade fatorial e a fiabilidade dos indicadores da escala da integração comunitária (EIC-APP), foi feita uma análise fatorial confirmatória com o software AMOS (versão 21, IBM SPSS). Para aferir a validade (fatorial, convergente e discriminante), foram utilizados, respetivamente, os pesos fatoriais estandardizados (assumindo que todos os itens são superiores a 0.5, o fator apresenta validade fatorial), a variância extraída média (VEM) (valores superiores a 0.5 indiciam validade convergente adequada) e a comparação das VEM dos fatores ($i$ e $j$) com o quadrado da correlação entre esses factores (assumindo a condição de que as VEM dos fatores $i$ e $j$ são superiores ao quadrado da correlação entre esses fatores ($r^2_{ij}$), pode ser admitida a validade discriminante). Para avaliar a fiabilidade dos indicadores em cada um dos fatores foi utilizada a fiabilidade compósita (FC), a qual estima a consistência interna dos itens reflexivos do fator (valores superiores a 0.7 indiciam uma fiabilidade apropriada).

Numa fase inicial, foram utilizados 34 itens subjacentes a quatro dimensões {integração física (IC_Fisica), integração social (IC_Social), integração psicológica (IC_Psicologica) e independência}, mas quando foi feita a estimação do modelo e avaliada a sua qualidade de ajustamento verificou-se que esta não se revelou satisfatória. Após terem sido retirados os itens que estavam a prejudicar o modelo, por apresentarem pesos fatoriais muito baixos (inferiores a 0.5), obteve-se um modelo final constituído por 12 itens e duas dimensões.

Com base neste modelo simplificado, foram avaliadas a fiabilidade compósita e a variância extraída média para cada fator, como descrito em Fornell e Larcker (1981). Foi avaliada a normalidade dos itens com base nos coeficientes de assimetria ($Sk$) e curtose ($ku$) uni e multivariada. Todas as variáveis apresentaram valores de $Sk$ ($|Sk|<3$) e $Ku$ ($|Ku|<10$), inferiores aos valores limites admissíveis para a distribuição normal (Kline, 2004). Foi, ainda, verificada a existência ou não de *outliers* com base no quadrado da distância de Mahalanobis.

A avaliação da qualidade do ajustamento do modelo apresentado, a qual tem como objetivo avaliar até que ponto este é capaz de reproduzir a estrutura correlacional dos itens

observados na amostra sob estudo, foi feita ainda com base em alguns índices (RMSEA; CFI; GFI; PCFI), considerando os respetivos valores de referência.

## 3. Resultados

Com base no modelo simplificado com dois fatores de 1ª ordem, foram obtidos para o "IC_Fisica" e para o "IC_Psico_Social", respetivamente, os valores 0.882 e 0.898, referentes à fiabilidade compósita. A variância extraída média (*VEM*), um indicador da validade convergente dos fatores, revelou-se também adequada, com os valores de 0.654 para o "IC_Fisica" e de 0.618 para o "IC_Psico_Social". Os valores de $VEM_{IC\_Fisica}$ (0.654) e $VEM_{IC\_Psico\_Social}$ (0.618) são superiores ao $r^2_{FP}$=0.218, pelo que podemos afirmar que os dois fatores têm validade discriminante.

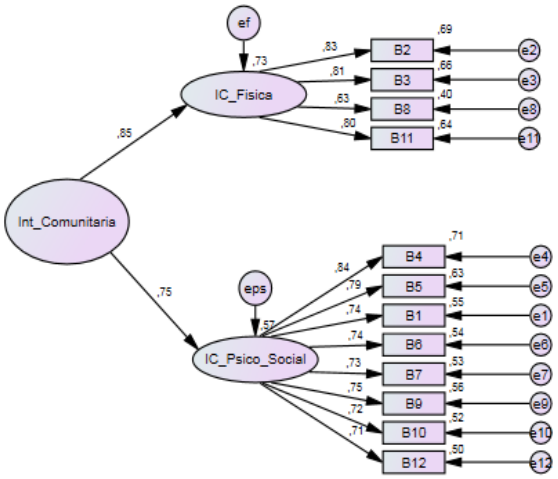



**Figura 1: Modelo das estimativas estandardizadas dos seus parâmetros**

| | Conteúdo dos itens |
|---|---|
| **B1** | Sinto que pertenço à minha comunidade |
| **B2** | Desloco-me sozinho(a) até aos sítios onde quero ir, a pé, em viatura própria, autocarro ou táxi. |
| **B3** | Vou sozinho(a) aos serviços sociais, à clinica, à farmácia ao centro de saúde / hospital, ou outros. |
| **B4** | As pessoas da minha comunidade pedem-me ajuda quando precisam. |
| **B6** | Sei que se precisar posso contar com o apoio das pessoas da minha comunidade. |
| **B7** | As pessoas da minha comunidade pedem-me ajuda quando precisam. |
| **B11** | Se necessário vou tratar de assuntos a locais públicos, tais como o banco, os correios, o supermercado, ou outros. |

**Figura 2: Itens que apresentam valores mais elevados da variância explicada pelos respetivos fatores.**

Depois de analisado e validado o modelo de 1ªordem com dois fatores e dada a existência de correlações significativas entre resíduos intra e interfatores (Gerbing e Anderson, 1984), foi testado um modelo de análise fatorial de 2ª ordem com dois fatores (Figura 1). A Figura 1 apresenta os valores dos pesos fatoriais estandardizados e a fiabilidade individual de cada um dos itens do modelo utilizado, enquanto a Figura 2 refere-se aos itens que apresentam valores mais elevados da variância explicada pelos respetivos fatores.O modelo com dois fatores associado à escala da integração comunitária ajustado à amostra considerada revelou uma qualidade de ajustamento global aceitável. Todos os itens da EIC-APP apresentam pesos fatoriais elevados ($\lambda_{ij} \geq 0.5$) e uma fiabilidade individual adequada ($(\lambda_{ij})^2 \geq 0.25$). Os valores obtidos para os índices RMSEA (0.06), CFI (0.87), GFI (0.85) e PCFI (0.71) são aceitáveis.

## 4. Considerais finais

A Analise Fatorial Confirmatória (AFC) revelou que vários itens da escala inicial não explicavam bem o modelo que visavam avaliar, tendo sido eliminados. A eliminação de alguns itens com base na AFC, incluindo os referentes à dimensão independência, vem refutar a sugestão de Gulgur et al. (2007) de se acrescentar a dimensão independência ao modelo. A Análise fatorial confirmatória não reproduziu na totalidade a estrutura fatorial que se esperava obter com base nos modelos teóricos existentes. Nos modelos inicialmente testados os índices de modificação indicavam que alguns dos seus itens e/ou os resíduos estavam também associado(s) a outro(s) fator(es), o que interfere na qualidade do ajustamento dos modelos.

Com base nos resultados obtidos, o modelo selecionado engloba dois fatores de 1ª ordem (Fisica e Psico_Social). No caso do fator Psico-Social existem itens pertencentes às dimensões Social e Psicológica, tendo-se constatado a existência de uma correlação relativamente elevada entre os itens dessas duas dimensões. Futuramente, será ainda relevante ter em consideração alguns aspetos que poderão ter influenciado os resultados, nomeadamente o número de doentes implicados, as características dos doentes (tipo e nível de gravidade da patologia dos participantes) e as especificidades do contexto comunitário onde os participantes estão inseridos.

## Referências

CABRAL, J.M. & BARRETO CARVALHO, C.M.O. (2013). *Estudo da integração comunitária de pessoas com problemas psiquiátri*cos. Dissertação de Mestrado. Ponta Delgada: Universidade dos Açores.

FORNELL, C., LARCKER, D. (1981). Evaluating SEM with Unobserved variables and measurement error. *Journal of Marketing Research*, 18(39), 39-50.

GERBING, D.W.&ANDERSON, J.C. (1984). On the Meaning of Within- Factor Correlated Measurement Errors. *Journal of Consumer Research: An Interdisciplinary Quarterly*, 11(1), 572-580.

GULCUR, L., TSEMBERIS, S., STEFANCIC, A. & GREENWOOD, R.M. (2007). Community integration of adults with psychiatric disabilities and historie of homelessness. *Community Mental Health Journal*, 43 (3), 211-228.

KLINE, R. (2004) *Principles and Pratice of Structural Equation Modelling*, 2nd Edition. New York, USA, Guilford Press.

WONG, Y.-L. I., & SOLOMON, P. L. (2002). Community integration of persons with psychiatric disabilities in supportive independent housing: A conceptual model and methodological considerations. *Mental Health Services Research*, 4, 13 – 28.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Clustering European industries using longitudinal data

**José G. Dias[1], Sofia B. Ramos[2]**
[1]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU, jose.dias@iscte.pt*
[2]*Instituto Universitário de Lisboa (ISCTE-IUL), BRU, sofia.ramos@iscte.pt*

**Abstract**

This paper analyzes the dynamics of 79 European industries from ten euro countries. We frame our research question into the dichotomy between industry vs. country effects. We take a two-step approach: first, time series indexes are filtered out using a (panel) regime switching model; then, based on the Kullback-Leibler distance between posterior probabilities, the hierarchical market structure is revealed. Results show that country factors tend to be more important in explaining industry indexes heterogeneity than industry effects, the exception being financial, technology, and telecommunications industries that show cross-country effects.

**Key-words:** regime-switching models, industry factors, cluster analysis, time series data.

## 1. Introduction

This paper discusses the identification of similarities between time series, i.e., the grouping of time series into homogeneous groups. We focus on the dynamic of euro zone stock markets at industry level. In particular, we explore and decompose country-industry indexes into country and industry effects by applying a new methodology that takes time series dynamics between regimes into account. Section 2 provides the clustering framework, Section 3 describes the data used in this paper, and Section 4 summarizes the results. The paper ends with concluding remarks.

## 2. Methodology

This paper uses regime switching models to filter out the dynamics of each time series (HAMILTON, 1989; MCLACHLAN and PEEL, 2000). We take a panel approach by modeling several time series simultaneously. We consider a multi-regime framework, where the selection of the number of regimes is based on the BIC. Indeed, regime-switching models have proven to be powerful tools for analyzing unobserved heterogeneity and asymmetry in time series data. Then, a matrix of distances between time series is obtained by computing the Kullback-Leibler distance between the posterior probabilities of each regime. That is, if two time series are very similar then the posterior probabilities of being in a given regime at a given time point should be very similar too. Finally, a hierarchical clustering algorithm reveals the structure of the panel time series.

## 3. Data

We use Datastream stock indexes for the ten original EMU countries: Austria, Belgium, Finland, France, Germany, Ireland, Italy, the Netherlands, Portugal, and Spain. To study country-industry indexes, we follow Industrial Classification Benchmark (ICB) and the industries analyzed are: Oil and Gas (OILGS), Basic materials (BMATR), Industrials (INDUS), Consumer Goods (CNSMG), Health Care (HLTCH), Consumer Services (CNSMS), Telecommunications (TELCM), Utilities (UTILS), Financials (FINAN) and Technology (TECNO). The panel is imbalanced because of different starting dates and because not all countries have listed companies for all industries. Because industry indexes have different starting dates and we need a common starting date we define the period from January 3, 1990 to December 30, 2012. The analysis is conducted in USD dollars to allow comparability of the stock markets and model estimation, before and after the euro launch.

## 4. Results

First, a multi-regime panel model is estimated. Based on BIC selection, six regimes provide the best fit to the panel data set. From the posterior probability estimates, a classification of the 79 industries into 5 clusters is given in Table 1.

**Table 1. Classification of industries into clusters**

| | OE | BG | FN | FR | BD | IR | IT | NL | PT | ES |
|---|---|---|---|---|---|---|---|---|---|---|
| OILGS | 4 | | | 3 | | 1 | 3 | 5 | | 3 |
| BMATR | 4 | 5 | 2 | 5 | 5 | 4 | 3 | 5 | 3 | 3 |
| INDUS | 5 | 5 | 3 | 5 | 5 | 2 | 3 | 2 | 3 | 3 |
| CNSMG | | | | 5 | 5 | 2 | 3 | 3 | 3 | 3 |
| HLTCH | 4 | 5 | 3 | 5 | 5 | 4 | 3 | 5 | 1 | 3 |
| CNSMS | 4 | 5 | 3 | 5 | 5 | 5 | 3 | 5 | 3 | 3 |
| TELCM | | 2 | | | 2 | | 2 | | | 2 |
| UTILS | 4 | 5 | | | 5 | | 3 | | | 3 |
| FIN | 4 | 5 | 3 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| TECNO | | 2 | 1 | 2 | 2 | | 2 | 2 | | |

Analyzing the clustering structure, we conclude that: Cluster 1 tends to be a very specific group, containing three distinct industries from 3 different countries. They show very specific cases containing the Finnish Technological industry (severely affected by the dotcom crisis), the Portuguese Health industry, and the Irish oil index. Cluster 2 contains telecommunications and technology industries and other specific industries from Finland, Ireland (2 industries), the Netherlands, and Spain. Cluster 3 contains most of the industries from Finland, Italy, Portugal, and Spain, indicating a peripheral cluster. Cluster 4 is country specific containing most of the Austrian industries and three Irish industries. Finally, Cluster 5 defines the core group, containing most of German, French, Dutch, and Belgium industries. Financial industry tends to be part of this core cluster.

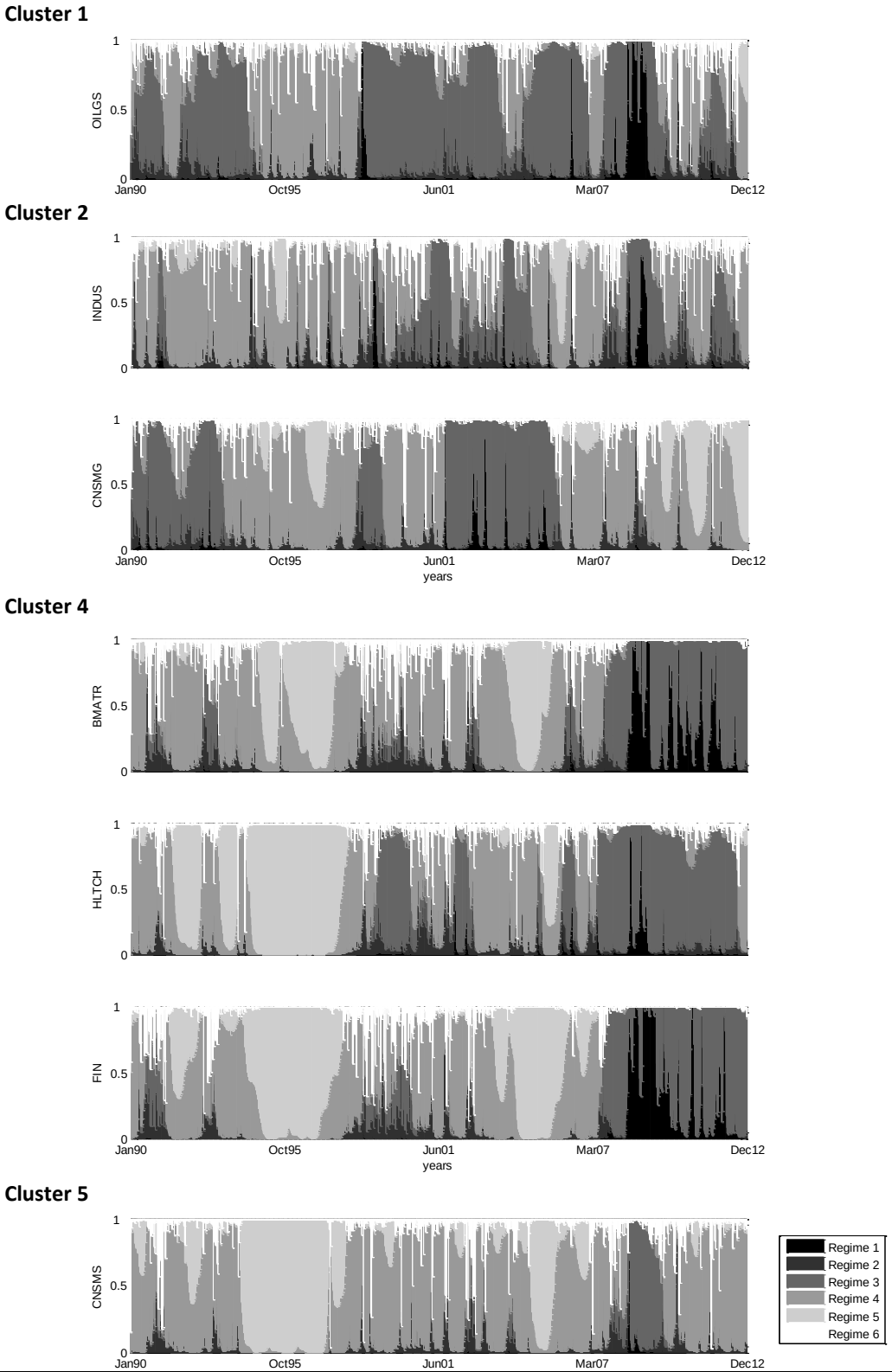



**Figure 1: Posterior probabilities for each industry in Ireland**

Figure 1 provides the longitudinal profile of Irish indexes. It shows that regime occupancies are quite distinct across groups. Regimes are ordered from negative expected returns (regime 1 the most negative one) to positive expected returns (regime 6 the most positive one). For example, in Ireland we observe that cluster 1 where Irish oil industry belongs

was severely affected from 2008 onwards, anticipating the recession that followed the subprime crisis then spread into a Global Financial crisis, as well as Cluster 4 that contains Irish Financial industry.

## 5. Conclusion

Our results show that the proposed method identifies groups of time series with different dynamics. A solution with five clusters shows that the country factor seems to be the most important one in the period. However, three industries tend to be homogeneous across countries: Financial, Technological, and Telecommunications.

**Acknowledgments:** Financial support from Fundação para a Ciência e Tecnologia is greatly acknowledged (PTDC/EGE-GES/103223/2008).

## References

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

MCLACHLAN, G., PEEL, D. (2000). *Finite Mixture Models*. New York, John Wiley & Sons.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Seleção de instâncias para algoritmos de aprendizagem não supervisionada: aplicação a dados de motores de aeronaves

**Leonor Fernandes[1], Roberto Henriques[2], Victor Lobo[3]**
[1]*ISEGI-UNL e EuroAtlantic Airways, mlfernandes@euroatlantic.pt;*
[2]*ISEGI-UNL, roberto@isegi.unl.pt;*
[3]*ISEGI-UNL e CINAV-Escola Naval, vlobo@isegi.unl.pt*

**Sumário**

Identificar o estado dos motores das aeronaves e prever a ocorrência de uma falha são informações relevantes na tomada de decisão da remoção de um motor de aeronave. Num voo a quantidade de dados disponíveis que carateriza o estado em que o motor se encontra é abundante e permite a utilização de técnicas de data mining. No entanto, numa abordagem exploratória, a abundância de dados torna o tempo de execução dos algoritmos muito lento ou é até um impedimento à aplicação de outros. O objectivo deste trabalho é utilizar diferentes métodos de seleção de observações para a aplicação de algoritmos de aprendizagem não supervisionada de clustering que permitam identificar grupos de voos com anomalias.

**Palavras-chave:** Seleção de observações (instâncias), Extração de conhecimento de dados (ECD), aprendizagem não supervisionada, seleção uniforme no tempo, Self Organizing Map (SOM).

## 1. Enquadramento

Identificar o estado de condição dos motores das aeronaves e prever a possibilidade de uma falha são informações importantes no apoio à tomada de decisão aquando da remoção de um motor de aeronave para manutenção. Usualmente o acompanhamento dos parâmetros do motor é realizado através da comparação entre os valores registados durante os voos e os limiares estabelecidos pelos fabricantes para cada um dos parâmetros. Quando aqueles se começam a aproximar destes, acções de manutenção devem ser tomadas. Este acompanhamento é feito individualmente, parâmetro a parâmetro.

Devido aos riscos de uma avaria inesperada, a tomada de decisão de remoção do motor é por vezes feita prematuramente. Este procedimento de segurança acarreta elevados custos.

Presentemente as empresas de aviação possuem o registo de um elevado número de dados que possibilita o desenvolvimento de novas formas de diagnóstico e prognóstico de avarias. A aplicação de técnicas de data mining aos registos existentes sobre motores e sobre as condições de realização dos voos, através do estudo das suas interações, permite outro tipo de informação acerca do desempenho dos motores.

Obter este tipo de informação é nos dias de hoje possível utilizando técnicas de extração de conhecimento de dados (ECD). Entende-se por ECD um processo de aproximar uma função

a partir dos dados da experiência passada (Gama J.; Carvalho A.; Faceli K.; Lorena A.; Oliveira M., 2012).

Durante a realização de um voo, a quantidade de dados disponíveis que permite caraterizar o estado em que o motor se encontra é abundante. Existem os registos dos diferentes parâmetros de desempenho e funcionamento dos motores segundo a segundo. O reconhecimento de registos anormais possibilitam identificar alterações no estado de funcionamento dos motores e possíveis falhas, mas durante um voo nem todos os registos são relevantes. Selecionar os registos significativos é uma tarefa importante na fase de pre-processamento dos dados em extração de conhecimento de dados (Fernández A.; Duarte A.; Hernández R.; Sánchez Á., 2010). Esta fase é uma das mais importantes e demoradas em todo o processo de ECD e as tarefas que a constituem condicionam o sucesso do processo (Reinartz T., 2002). Entende-se por seleção de observações ou instâncias a criação de um conjunto de treino S que está contido no conjunto de dados originais T. Não deverá existir perda de informação relevante na constituição do subconjunto S. Este tem que ser representativo do conjunto T. Pretende-se encontrar uma amostra com dimensão menor mas de forma a que as variáveis que a constituem apresentem os mesmos comportamentos dos dados originais (Gama J.; Carvalho A.; Faceli K.; Lorena A.; Oliveira M., 2012). Esta amostra permitirá a criação de grupos de voos com e sem anómalias e entender o ciclo de vida de um motor através da acção conjunta de todos os seus parâmetros de uma maneira mais rápida e menos complexa no processamento computacional.

## 2. Envolvente do problema

É vasta a literatura sobre a seleção de observações mas direcionada para problemas de aprendizagem supervisionada, salientando-se (Olvera-López J. A.; Carrasco-Ochoa J. A.; Martínez-Trinidad J. F.; Kittler J., 2010) que apresenta um resumo detalhado dos algoritmos mais utilizados com as suas características e compara os seus desempenhos. Em problemas de aprendizagem supervisionada estão disponíveis dados em que se conhecem os valores no espaço de input (ou seja variáveis independentes) e também os valores no espaço de output, ou seja a variável que se pretende prever, ou variável dependente.

Este trabalho, e uma vez que nos encontramos na fase exploratória dos dados, centra-se em aprendizagem não supervisionada, isto é, o conhecimento que existe é referente a variáveis no espaço de inputs (variáveis independentes) que permitirão identificar os grupos de voos. Neste contexto, não existe muita literatura. Os artigos encontrados reportam-se ao fim da década de 90 e inicio dos anos 2000 ( (John G.; Langley P., 1996), (Liu H.; Motoda H., 2002), (Palmer R.C.; Faloutsos C., 2000)). Estes artigos apontam como formas de seleção da amostra os processos de amostragem usuais (amostragem aleatória simples, amostragem uniforme ou sistemática, estratificada) e métodos de formação de clusters.

No entanto a quantidade de informação, mesmo na fase exploratória dos dados torna a sua análise complexa e pouco eficiente, tanto devido ao tempo de execução de alguns algoritmos

como ao impedimento de aplicação de outros algoritmos (Liu H.; Motoda H., 2002). É pois importante o desenvolvimento de novas formas de seleção de observações para a constituição do conjunto de treino.

O objectivo deste trabalho é avaliar e comparar diferentes métodos de seleção de observações (instâncias) para a construção de um conjunto de treino que deverá conter a informação relevante do conjunto de dados original.

## 3. Descrição dos dados e métodos de seleção das observações

Os dados originais são já uma amostra de registos de motores PW 4060 da Pratt & Whitney utilizados em cerca de 481voos realizados entre 2009 e 2013 em duas aeronaves B-767-300. Em média cada voo tem cerca de 17300 observações e 18 variáveis. As variáveis registadas descrevem: a) valores para 12 parâmetros de desempenho e funcionamento dos motores; b) valores para 6 características das condições de funcionamento do voo. O número de registos por voo difere com a duração do voo pelo que o nº de voos a serem selecionados para o conjunto de treino poderá ser proporcional ao do conjunto original. A dimensão da amostra final será obtida quando os indicadores estatísticos habituais (média, desvio padrão) para cada variável se aproximarem bastante dos valores dos dados originais em conjungação com o ganho de tempo utilizado no processo. Na tabela 1 apresenta-se a distribuição do nº de voos por ano e aeronave.

**Tabela 1 - Distribuição do nº de voos por ano e aeronave.**

| Aeronave 1 | nº de voos | Aeronave 2 | nº de voos |
|---|---|---|---|
| 2009 | 15 | 2009 | 17 |
| 2010 | 67 | 2010 | 93 |
| 2011 | 55 | 2011 | 27 |
| 2012 | 65 | 2012 | 76 |
| 2013 | 27 | 2013 | 39 |
| Total | 229 | Total | 252 |

Os métodos de seleção das observações utilizados serão:

1. Uniforme no tempo com intervalos constantes – Dos registos do ficheiro original, que se encontram ordenados, um elemento é selecionado para a amostra com um intervalo de amostragem de k=N/n, em que n será a dimensão da amostra e N a dimensão do ficheiro original.
2. Uniforme no tempo com médias de conjuntos de k observações – Selecionam-se de forma sistemática k observações e o valor selecionado para a amostra será a média de cada uma das variáveis.
3. Uniforme no espaço de carateristicas utilizando como critério o vizinho mais próximo – É selecionado para a amostra o primeiro registo do ficheiro original, define-se um limiar de

proximidade e de acordo com o critério do vizinho mais próximo seleciona-se o seguinte elemento, o terceiro elemento será o que comparado com os anteriores continua a obedecer ao limiar de proximidade e assim sucessivamente.

4. Som – Self organizing map – Constituem-se tantos clusters quanto a dimensão pretendida para a amostra, para cada um dos clusters encontrados seleciona-se a média dos registos e serão estes valores médios que farão parte do nosso ficheiro de treino.

## 4. Conclusão

O desenvolvimento deste trabalho está inserido na parte de pré processamento de informação e encontramo-nos neste momento na preparação, limpeza de todo o conjunto de dados para a aplicação dos métodos anteriormente referidos. O objetivo final desta etapa é a constituição do "melhor" conjunto de treino.

## References

GAMA, J., CARVALHO, A., FACELI, K., LORENA, A., OLIVEIRA, M. (2012). *Extração de Conhecimento de Dados*, Lisboa.

FERNÁNDEZ, A., DUARTE, A., HERNÁNDEZ, R., SÁNCHEZ, Á. (2010). GRASP for Instance Selection in Medical Data Sets. In: ROCHA, M., RIVEROLA, F., SHATKAY, H. & CORCHADO, J. (eds.) *Advances in Bioinformatics*. Springer Berlin Heidelberg.

LIU, H., MOTODA, H. (2002). On Issues of Instance Selection. *Data Mining and Knowledge Discovery*, 6, 115-130.

OLVERA-LÓPEZ, J. A., CARRASCO-OCHOA, J. A., MARTÍNEZ-TRINIDAD, J. F., KITTLER, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*, 34, 133-143.

PALMER, R.C., FALOUTSOS, C. (2000). Density biased sampling: an improved method for data mining and clustering. *SIGMOD Rec.*, 29, 82-92.

REINARTZ, T. (2002). A Unifying View on Instance Selection. *Data Mining and Knowledge Discovery*, 6, 191-210.

JOHN, G., LANGLEY, P. (1996). Static Versus Dynamic Sampling for Data Mining. In: Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA. AAAI Press, 367-370.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Avaliação do emprego da análise de agrupamentos nas revistas de saúde brasileiras no período de 1993 a 2011

**Fernando Frei[1], Franciele Karen Netto[2], Juliana Alves Pegoraro[3]**

*[1]Docente do Departamento de Ciências Biológicas – FCLAssis – UNESP - Brasil, ffrei@assis.unesp.br;*
*[2]Discente do curso de Estatística – FCT – UNESP – BRASIL, franciele.netto@gmail.com;*
*[3]Discente do curso de Engenharia Biotecnológica – FCLAssis – UNESP – Brasil, juliana_pegoraro@hotmail.com.*

**Sumário**

A Análise de Agrupamento é o nome genérico para um conjunto de métodos usados para identificar padrões. Diversos são os algoritmos com o mesmo objetivo. Pesquisadores tendem a adotar os métodos utilizados nos periódicos, que em grande parte não apresentam informações suficientes sobre os procedimentos adotados. O presente trabalho tem por objetivo identificar as deficiências no uso da Análise de Agrupamentos - Métodos Hierárquicos Aglomerativos - na área de Saúde, além de apresentar alternativas adequadas para a solução dessas deficiências. Pela análise efetuada pode-se concluir que nenhum dos vinte artigos avaliados contempla todos os quatro itens aferidos e apenas um artigo apresenta três itens. Este estudo revelou a necessidade de melhoria na forma de Análise de Agrupamentos e do relato na área de saúde.

**Palavras-chave:** Análise de agrupamentos, Métodos hierárquicos aglomerativos, Saúde.

## 1. Introdução

O reconhecimento de padrões é uma prática de grande importância na área da saúde. Esta prática permite a identificação de subgrupos da população que possuam determinados agravos ou comportamentos (Dini et al., 2011), ajuda a identificar grupos com maior grau de dependência e aumenta a eficácia das campanhas de saúde pública (Mackert e Walker, 2011). A análise de agrupamento é o nome genérico para um conjunto de métodos que podem ser usados para criar classificações e por conseqüência reconhecer padrões. Existem diversos métodos de agrupamentos e diversos algoritmos com o mesmo objetivo. As aplicações da Análise de Agrupamentos estão difundidas nas mais diversas áreas de estudo. O conjunto de técnicas hoje conhecidas é o produto da Era Digital, sem os computadores modernos atuais, o que conhecemos por Análise de Agrupamentos não existiria. Os principais programas computacionais, sejam comerciais ou livres, oferecem uma lista de aplicações crescentes na metodologia citada. A literatura científica tem apresentado um incremento no uso da metodologia e no desenvolvimento de novos algoritmos (Jain et al., 1996).

Os métodos Hierárquicos Aglomerativos caracterizam-se por sucessivas fusões de objetos em grupos. Neste método, o número de grupos não é conhecido *a priori*. Uma das maiores dificuldades da Análise de Agrupamentos é a determinação do número de grupos existentes no conjunto de dados analisados. Além disso, agrupamentos diferenciados surgem quando diferentes algoritmos são aplicados (Frei, 2006). Desta forma, uma das recomendações é a utilização de vários algoritmos e índices de homogeneidade, para que os agrupamentos

resultantes apresentem boas propriedades estatísticas, como homogeneidade e ótima separação (Brock et al., 2008). Se os resultados apresentam subestruturas semelhantes, após a aplicação de diferentes algoritmos, uma partição natural foi obtida, caso contrário, é pouco provável que os dados apresentem grupos naturais distintos.

## 2. Objetivo

O presente trabalho tem por objetivo identificar as deficiências no uso da Análise de Agrupamentos – Métodos Hierárquicos Aglomerativos em revistas brasileiras na área de Saúde, além de apresentar alternativas adequadas para a solução dessas deficiências.

## 3. Material e Método

Uma revisão dos procedimentos da Análise de Agrupamentos – Métodos Hierárquicos Aglomerativos foi realizada nos trabalhos publicados em revistas brasileiras da área de saúde disponíveis na base de dados Scielo no período de 1993 a 2011 de acordo com critérios

### 3.1. Seleção de Revistas

A área de saúde é ampla e as revistas pesquisadas das mais gerais, como por exemplo, revistas de saúde pública, até as mais específicas, como revistas de enfermagem e nutrição. As revistas de psicologia e psiquiatria não foram incluídas no estudo.

Artigos científicos que se referem à metodologia somente com a nomenclatura de Análise de Agrupamentos, Cluster Analysis ou Análise de Cluster, mas não indicavam outras características, não foram avaliados.

### 3.2. Avaliação

Diversos critérios devem ser observados para a melhor conduta na aplicação da Análise de Agrupamentos, entre eles o número de variáveis, padronização para variáveis métricas, análises de colinearidade e pontos atípicos. No entanto, para o estudo em foco, os artigos foram avaliados em relação a quatro quesitos necessários sugeridos por Aldenderfer e Blashfield (1984): Algoritmos, Medida de Similaridade, Índices Internos e Programa Computacional utilizado. Os índices internos avaliam a estrutura de agrupamento exclusivamente na amostra de dados estudada, sem levar em conta informação prévia. Para a construção do índice, o algoritmo de agrupamento é executado para várias soluções possíveis, onde k representa o número de grupos. Posteriormente, os valores dos índices obtidos a partir das partições geradas podem ser apresentados em gráficos em função de k. Nesse contexto, o melhor número de grupos é dado pelo mínimo ou o máximo dessa função, dependendo de como o índice foi definido.

## 4. Resultados

A tabela 1 apresenta 20 artigos científicos dos quais 11 não indicam o algoritmo ou algoritmos utilizados, o que representa mais da metade dos trabalhos. No quesito relacionado às medidas de similaridade, excluindo os trabalhos que mencionam o algoritmo de Ward, que pode ser utilizado diretamente na matriz de dados brutos, pode-se observar que do total de 15 artigos, apenas três apresentam informações. A adoção de índices internos não é relatada em nenhum dos artigos pesquisados. Finalmente, no quesito programa computacional metade dos artigos não cita o programa computacional empregado.

## 5. Conclusão

Este estudo revelou a necessidade de melhoria na forma de análise de cluster e do relato na área de saúde. Nenhum dos vinte artigos contempla todos os quatro itens avaliados e apenas um artigo apresenta três itens. A utilização de índices internos não é citada em nenhum dos artigos e o programa computacional SPSS é o mais empregado pelos pesquisadores.

O presente estudo seguramente não identificou todos os artigos da aplicação da Análise de Agrupamentos na área de saúde, visto que a pesquisa buscou palavras chaves associadas à metodologia, o que nem sempre é localizado na busca eletrônica. No entanto, fica evidente a fragilidade da aplicação da metodologia investigada. Orientações para conduzir e relatar a aplicação da Análise de Agrupamentos na área de saúde são necessários. Por essa razão, recomenda-se as seguintes etapas para o desenvolvimento da Análise de Agrupamentos: a) Medida de similaridade aplicada deve obedecer ao nível de mensuração dos dados; b) deve-se utilizar três algoritmos para buscar um padrão quanto ao número de grupos obtidos e estabilidade entre os objetos agrupados; c) Utilização de um ou mais índices internos que possam auxiliar a decisão do número de grupos baseados nos algoritmos empregados e d) Evitar o uso automático das configurações padrão em programas computacionais, visto que as mesmas podem não oferecer a análise mais adequada. Esse conjunto de métodos deve ser mencionado com clareza nos artigos para que outros pesquisadores possam replicar tais métodos, pressuposto básico do trabalho científico.

**Tabela1. Características dos artigos avaliados quanto a Análise de Agrupamentos.**

| Título do Artigo e Revista | Algoritmos utilizados | Medida de Similaridade | Utilização de índices | Programa Computacional |
|---|---|---|---|---|
| Mortalidade Precoce por Doenças ardiovasculares e Desigualdades Sociais em Porto Alegre: da Evidência à Ação. Arq Bras Cardiol. | Não citado | Não citado | Não | Não citado |
| Qualidade de vida, ponto de partida ou resultado final? Ciênc. saúde coletiva. | Complete Linkage e K-means | Distância euclidiana | Não | Não citado |
| Agrupamento de países segundo indicadores de padrão de vida. Rev. Saúde Pública. | UPGMA | Distância euclidiana | Não | Não citado |
| Distribuição espacial da dengue e determinantes socioeconômicos em localidade | Não citado | Não citado | Não | SPSS |

| | | | | |
|---|---|---|---|---|
| urbana no Sudeste do Brasil. Rev. Saúde Pública. | | | | |
| Perfil dos pacientes com perdas funcionais e dependência atendidos pelo PSF no município de São Paulo. Rev Esc Enferm USP. | Método de Ward | Distância euclidiana | Não | SPSS |
| Análise ecológica dos acidentes e da violência letal em Vitória, ES. Rev. Saúde Pública. | Não citado | Não citado | Não | Não citado |
| A qualidade das refeições de empresas cadastradas no Programa de Alimentação do Trabalhador na cidade de São Paulo. Rev. Nutr. | Não citado | Não citado | Não | SPSS |
| Controlar líquidos: uma intervenção de enfermagem para o paciente com excesso de volume de líquidos. Rev. Latino-Am. Enfermagem | Não citado | Não citado | Não | Não citado |
| Qualidade de vida na terceira idade: um conceito subjetivo. Rev. bras. Epidemiol. | Método de Ward e K-means | Não citado | Não | Statistica |
| Desigualdades na saúde reprodutiva das mulheres no Paraná. Rev. bras. epidemiol | Não citado | Não citado | Não | Não citado |
| Mortes por suicídio: diferenças de gênero e nível socioeconômico. Rev. Saúde Pública. | Método de Ward | Não citado | Não | SAS |
| Desigualdades sociais e gestão em saúde: metodologia de seleção de áreas urbanas visando à diminuição das desigualdades socioespaciais em regiões metropolitanas. Ciênc. saúde coletiva | Método de Ward | Distância euclidiana | Não | Não citado |
| Aspectos dietéticos das refeições oferecidas por empresas participantes do Programa de Alimentação do Trabalhador na Cidade de São Paulo, Brasil. Rev Panam Salud Publica. | Não citado | Não citado | Não | SPSS |
| Diferenciais intermunicipais de condições de vida e saúde: construção de um indicador composto. Rev. Saúde Pública | Não citado | Não citado | Não | Não citado |
| Doenças respiratórias agudas: um estudo das desigualdades em saúde. Cad. Saúde Pública | Não citado | Não citado | Não | Minitab |
| Perfil de produção do exame citopatológico para controle do câncer do colo do útero em Minas Gerais, Brasil, em 2002. Cad. Saúde Pública | Método de Ward e K-means | Não citado | Não | SPSS |
| Condição de vida e mortalidade infantil: diferenciais intra-urbanos no Recife, Pernambuco, Brasil. Cad. Saúde Pública, | Não citado | Não citado | Não | Não citado |
| Polimorfismo do sistema HLA em uma amostra de mestiços da população de Teresina, Piauí. Rev Assoc Med Brás. | Average Linkage | Não citado | Não | SPSS |
| Prevalências de sobrepeso, obesidade e hábitos de vida associados ao risco cardiovascular em alunos do ensino fundamental. Rev. Assoc. Med. Brás. | Não citado | Não citado | Não | SPAD |
| Critérios Ecocardiográficos para Definição de Graus de Disfunção Ventricular em Ratos Portadores de Estenose Aórtica. Arquivos Brasileiros de Cardiologia | Single Linkage | Distância euclidiana média | Não | Não citado |

**Agradecimentos:** FAPESP – Fundação de Amaparo à Pesquisa do Estado de São Paulo.

## Referências

ALDENDERFER, M.S. & BLASHFIELD, R.K. (1984). *Cluster analysis*. Newbury Park, CA: Sage Publishing.

BROCK G., PIHUR V., DATTA S. & DATTA S. (2008) clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, 25.

DINI A.P., FUGULIN F.M.T., VERÍSSIMO M.L.Ó.R. & GUIRARDELLO E.B. (2011) Sistema de Classificação de Pacientes Pediátricos: construção e validação de categorias de cuidados. *Rev Esc Enferm* USP, 45, No. 3, 575-80.

FREI, F. (2006) *Introdução à Análise de Agrupamentos: Teoria e Prática*. Editora Fundação UNESP, São Paulo.

JAIN, A.K., MURTY, M.N. & FLYNN, P.J. (1999) Data Clustering: A Review. *Journal ACM Computing Surveys*, 31 Issue 3, 264-323.

MACKERT M. & WALKER L.O. (2011) Cluster Analysis Identifies Subpopulations for Health Promotion Campaign Design. *Public Health Nursing*, 28, No. 5, 451–457.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Kidney-brain link in traumatic brain injury patients: A preliminary report

**A. Rita Gaio[1], Óscar Felgueiras[2], Celeste Dias[3], José-Artur Paiva[4], *Marek Czosnyka*[5]**

[1]*Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, argaio@fc.up.pt;*

[2]*Departamento de Matemática da Faculdade de Ciências da Universidade do Porto & CMUP-Centro de Matemática da Universidade do Porto, olfelgue@fc.up.pt;*

[3]*Serviço de Medicina Intensiva da Unidade de Cuidados Neurocríticos do CHSJ-Porto, mceleste.dias@gmail.com;*

[4]*Serviço de Medicina Intensiva do CHSJ-Porto & Departamento de Medicina da Faculdade de Medicina da Universidade do Porto, jarturpaiva@gmail.com;*

[5]*Neurosurgery Unit, Department of Clinical Neurosciences, University of Cambridge, mc141@medschl.cam.ac.uk*

**Summary**

The relationship between brain auto-regulation and kidney glomerular filtration rate in patients with traumatic brain injury remains to be investigated. We analyzed 194 daily observations collected from 18 patients. Multiple linear regression models estimated by generalized least squares evaluated the effect of clearance creatinine on cerebrovascular reactivity pressure index throughout time. Mixed-effects models were not shown to be adequate. Norepinephrine and fatal/nonfatal outcome presented significantly confounding effects.

**Key-words:** Augmented renal clearance, Cerebral auto-regulation, Generalized least squares, Kidney hyperfiltration**,** Multiple linear regression.

## 1. Introduction

Autoregulation of blood flow is the inherent capacity of vascular bed to maintain constant perfusion despite the variations of arterial blood pressure and intracranial pressure (brain) and is an important mechanism to maintain cerebral and kidney blood flow relatively constant. Cerebrovascular autoregulation and kidney function are frequently impaired in patients with acute brain lesion [LI ET AL, 2011]. Conversely, kidney hyperfiltration with augmented renal clearance and polyuria are also frequently observed in patients with traumatic brain injury (TBI), [MINVILLE ET AL., 2011; UDY ET AL., 2010]. The aim of this study is to report preliminary findings about the relationship between brain autoregulation impairment, estimated kidney glomerular filtration rate and fatal/nonfatal outcome in critically ill patients after severe traumatic brain injury.

## 2. Modelling and statistical analysis

We retrospectively performed an analysis of data collected for a prospective cohort of 18 consecutive patients admitted to Neurocritical Care Unit (NCCU) at Hospital S. João, Porto, with multiple trauma and severe traumatic brain injury and clinical indication for intracranial pressure (ICP) monitoring.

We analysed 194 complete daily observations from the 18 patients regarding cerebrovascular pressure reactivity index (PRx), creatinine (Cr), creatinine clearance (CrCl), noradrenaline dose, protein intake, water balance, sodium, and osmolarity. For one of the individuals, 38 daily observations were available but only the first 23 complete daily observations were retrieved. This cut-off value was essentially identified by two criteria: proximity to the highest number of complete daily observations from the remaining individuals in the sample, and representativeness of the mean of its remaining observations. For the remaining patients the number of complete daily values ranged from 3 to 18.

Multivariate linear regression models studied the crude and adjusted effect of creatinine clearance on PRx throughout time. The generalized least squares method with normally distributed errors that were allowed to be correlated and heteroscedastic was applied [PINHEIRO e BATES, 2009]. Amongst possible clinical confounders, the crude effect of clearance creatinine on PRx was only found to be significantly altered by noradrenaline. Denoting by *PRx(s,c,d)* the PRx value read for creatinine clearance *c* and noradrenaline *d* in a subject with dichotomous outcome *s* (fatal or nonfatal), the best model fitted by maximum likelihood was given by the equation

$$PRx(s,c,d) = \beta_0 + \beta_1 s + \beta_2 c + \beta_3 d + \varepsilon \qquad (1)$$

Here, the residuals ε were assumed to follow a normal distribution with zero mean and were found to have a variance-covariance matrix with an intra-individual time autocorrelation structure of order 1. The reference category for the fatal/nonfatal outcome corresponded to the nonfatal result. Graphical analyses were used to assess normality and homoscedasticity of model residuals and no compromising features were detected. The obtained estimates for the regression coefficients are presented in Table 1A. Crude effects were studied with the same structure of the variance-covariance matrix as that from model (1). The obtained 95% confidence intervals for the correlation parameter and residual standard error of the adjusted model were (0.2546, 0.5376) and (0.2161, 0.2733) respectively.

Intra-individual and inter-individual variations were also considered by linear mixed effects models, with observations grouped at the individual level [3]. The restricted maximum likelihood estimation provided a model with a single random effect at the intercept level. That model was discarded as only 1.9% of the response variation was explained by the random effect.

**Table 1: Estimates, and 95% confidence intervals, for the crude and adjusted effects of: A. creatinine clearance, fatal/nonfatal status and norepinephrine on PRx, from all patients; B. creatinine clearance and norepinephrine on PRx, from patients with normal kidney function.**

| A. Variables for all patients | Crude Effects | 95% CI | Adjusted Effects | 95% CI |
|---|---|---|---|---|
| Intercept | -------- | -------- | 0.1080 | (-0.0006, 0.2730) |
| Creatinine Clearance | -0.0014 | (-0.0019, -0.0008) | -0.0010 | (-0.0016, -0.0004) |
| Fatal/NonFatal Status | 0.2430 | (0.0985, 0.3874) | 0.1543 | (0.0220, 0.2866) |
| Norepinephrine | 0.0028 | (0.0011, 0.0044) | 0.0017 | (0.0001, 0.0033) |
| Sodium | 0.0052 | (0.0002, 0.0101) | 0.0012 | (-0.0040, 0.0065) |
| Protein Intake | 0.0001 | (-0.0018, 0.0020) | 0.0006 | (-0.0012, 0.0025) |
| Water Balance/100* | -0.0009 | (-0.0033, 0.0016) | -0.0017 | (-0.0042, 0.0007) |

*Water balance was divided by 100 only for scaling purposes

| B. Variables for patients with normal kidney function | Crude Effects | 95% CI | Adjusted Effects | 95% CI |
|---|---|---|---|---|
| Intercept | ----- | ----- | 0.1658 | (0.0147, 0.3169) |
| Creatinine Clearance | -0.0012 | (-0.0018, -0.0006) | -0.0011 | (-0.0017, -0.0005) |
| Norepinephrine | 0.0012 | (0.0004, 0.0020) | 0.0010 | (0.0002, 0.0018) |

Generalized least squares was also shown to be the most adequate method for an analogous study carried out only for individuals with preserved renal function with creatinine <1.3 mg/dl (2 male subjects who developed kidney failure with fatal outcome having 11 complete observations were removed). The restricted maximum likelihood estimation provided again a model with a single random effect at the intercept level which explained only 3.9% of the response variation. For the above notation, the chosen model was

$$PRx(c,d) = \beta_0 + \beta_1 c + \beta_2 d + \varepsilon \qquad (2)$$

with the Gaussian distributed residuals ε having a variance-covariance matrix as in (1). Model estimates are presented in Table 1B. The obtained 95% confidence intervals for the correlation structure parameter and residual standard error of the adjusted model were (0.2606, 0.5384) and (0.2148, 0.2729) respectively.

## 3. Conclusion

In TBI patients, better cerebral auto-regulation evaluated with cerebrovascular pressure reactivity index is significantly associated with augmented renal clearance and favorable outcome. Generalized least squares proved to be more adequate than mixed-effects regression in the modelling of this clinical situation.

**Acknowledgments:** The first and second authors were partially funded by the European Regional Development Fund through program COMPETE and by the Portuguese Government through FCT - Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0144/2013.

## References

LI, N., ZHAO W-G., ZHANG W-F. (2011). Acute Kidney Injury in Patients with Severe Traumatic Brain Injury: Implementation of the Acute Kidney Injury Network Stage System. *Neurocritical Care*, 14, 377-81.

MINVILLE, V., ASEHNOUNE, K., RUIZ S., ET AL. (2011). Increased creatinine clearance in polytrauma patients with normal serum creatinine: a retrospective observational study. *Critical Care*, 15, R49.

PINHEIRO, J., BATES, D. (2009). *Mixed-Effects Models in S and S-Plus*, New York, USA, Springer Verlag.

UDY, A., BOOTS, R., SENTHURAN, S., ET AL. (2010). Augmented creatinine clearance in traumatic brain injury. *Anesthesia & Analgesia*, 111, 1505-10.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Root resorption risk modeling

**S. Pereira[1], N. Lavado[2], L. Nogueira[3], M. Lopez[4], J. Abreu[1], H. Silva[5]**

[1]*Dep. of Orthodontics, Faculty of Medicine, University of Coimbra, soalves1@gmail.com;*
[2]*Polytechnic Institute of Coimbra (IPC-ISEC) and BRU-IUL, nlavado@isec.pt;*
[3]*Medical Genetics Department, Faculty of Medicine, University of Coimbra;*
[4]*Faculty of Engineering, University of Porto;*
[5]*Medical Genetics Department, Faculty of Medicine, University of Coimbra and CIMAGO;*

**Abstract**

We proposed to study nine clinical and treatment-related factors and polymorphisms of four genes in order to construct a model to predict orthodontic-induced external apical root resorption (EARR). We concluded that the main factors contributing to the EARR process are gender, treatment duration, use of a Hyrax appliance, premolar extractions and rs1718119 polymorphism of the P2RX7 gene. These five variables explained 27% of EARR variability, suggesting the existence of other aetiologic factors. This study has been recently published in the journal *Oral Diseases* (Pereira et al, 2013).

**Keywords:** Apical Root Resorption, Gene Polymorphism, Orthodontics, P2RX7 Gene

## 1. Introduction

External apical root resorption (EARR) is a frequent complication observed in association with orthodontic tooth movement and has been of great concern to clinical researchers. Although forces applied and tooth movement are the trigger for EARR, studies have shown that biomechanical factors may not account for more than one-tenth to one-third of the variation observed in root resorption. Many factors have been described including medication, endodontic treatment and patients' intrinsic factors such as gender, age, tongue thrust, the existence of anterior open bite, type of malocclusion and systemic diseases.

In this study we also included four genes, extensively explored as susceptible markers, as candidates for root resorption modeling. The polymorphisms chosen for this study have previously been associated with EARR or bone remodeling and might be functionally relevant: rs1143634 from IL-1B, rs3102735 from OPG gene, rs1805034 from RANK gene and rs1718119 from P2RX7.

### 1.1. Materials and Methods

For this retrospective study, patients followed by the same orthodontist were randomly selected from the archives of two orthodontic clinics and from the Department of Orthodontics, Dentistry, Faculty of Medicine of Coimbra. The 195 selected patients included 72 males and 123 females, with an average age of 17.24 years (s.d. 6.8 years).

Before and after orthodontic treatment panoramic radiographs were used to evaluate the percentage of EARR for each patient based on measurements on six teeth, the four maxillary

incisors (ISO System codes 11, 12, 21 and 22) and the two maxillary canines (ISO System codes 13 and 23). Measurements were conducted using a software prototype developed by the authors using Matlab, that included image preprocessing, point selection (manually marking 4 points on each selected tooth) and feature extraction to produce a set of linear measurements for parts (crown and root) of the teeth under analysis. The corrected final root measurement results from the application of a correction or enlargement factor corresponding to the ratio between the initial and final crown lengths, because it is accepted that during orthodontic treatment, the crown length does not change. In order to analyse the EARR phenomena for the individual and not only for each separate tooth, we propose a metric based on the maximum observed EARR (EARRmax) on the six selected teeth.

Eighteen explanatory variables were included in a stepwise regression (mixed forward/backward option) having EARRmax as dependent variable and the following explanatory variables: gender (binary), age (years), treatment duration (months), anterior open bite (binary), premolar extractions (binary), Hyrax appliance (binary), tongue thrust (binary), overjet (continuous orthodontic measure), skeletal class (3 categories, thus 2 binary variables), polymorphisms from TNFRSF11A, TNFRSF11B, IL-1B and P2RX7 genes (3 categories for each, thus 8 binary variables).

**Table 1. External apical root resorption (%) statistics for each tooth (n=195).**

| Statistics | | Tooth | | | | | |
|---|---|---|---|---|---|---|---|
| | | 13 | 12 | 11 | 21 | 22 | 23 |
| Average | | 8.0 | 11.5 | 9.5 | 9.9 | 10.2 | 8.3 |
| Standard deviation | | 6.9 | 8.6 | 8.2 | 9.6 | 8.1 | 7.2 |
| Maximum | | 47.7 | 37.0 | 42.1 | 49.7 | 37.0 | 35.4 |
| Percentile | 25 | 2.8 | 4.8 | 3.5 | 2.7 | 3.5 | 2.9 |
| | 50 | 6.2 | 9.5 | 7.2 | 6.3 | 8.2 | 6.6 |
| | 75 | 11.1 | 17.5 | 13.9 | 13.8 | 15.0 | 11.3 |
| | 95 | 22.4 | 29.0 | 26.4 | 31.6 | 26.6 | 23.7 |

## 2. Results

Characterization of the teeth's EARR is depicted in Table 1. On average, EARR ranged from 8% (tooth 13) to 11.5% (tooth 12). The difference between canines' and incisors' EARR average was significant, while there were no significant differences within incisors or between symmetrical teeth (repeated-measures ANOVA: Wilks' lambda = 0.826, F = 7.998, P < 0.0005; post hoc tests: P < 0.0005 only for differences between canines' and incisors' EARR average).

The EARR median ranged from 6.2% (tooth 13) to 9.5% (tooth 12). According to the value of percentile 95, five percent of patients showed EARR higher than 22.4% on tooth 13 and higher than 31.6% on tooth 21.

The stepwise model at the final stage showed that the selected variables explained 27% of the EARRmax variability (ANOVA: F = 15.315, P = 0.000; adjusted determination coefficient = 0.270, n = 195) and that the selected variables were (Table 2): gender (P < 0.05), treatment duration (P < 0.001), premolar extractions (P < 0.01), Hyrax appliance (P < 0.001) and GG genotype for the P2RX7 gene (P < 0.01).

**Table 2 - Results of stepwise regression model**

| Parameters of statitical model | N (%) | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | SD | Beta | t | *P value* |
| (Constant) | | 7.6 | 2.4 | | 3.145 | .002 |
| Female | 123 (63%) | -3.2 | 1.2 | -0.162 | -2.580 | .011 |
| Duration of treatment (months) | | 0.3 | 0.1 | 0.275 | 4.270 | .000 |
| Pre-molar extraction | 58 (30%) | 4.0 | 1.3 | 0.193 | 2.990 | .003 |
| Hyrax appliance | 14 (7%) | 8.4 | 2.3 | 0.230 | 3.599 | .000 |
| *P2RX7*, GG genotype (baseline AA genotype) | 84 (43%) | 3.1 | 1.2 | 0.162 | 2.595 | .010 |

The generalizability of our findings is supported by a 75/25% cross-validation. We verified that the model based on a training sample (n = 159) included the same five variables as the model based on the full data set (n=195) and also that the stepwise algorithm reaches convergence also after five iterations. Finally, we compared the accuracy of the model for the validation sample (n = 36) to the accuracy of the model for the training sample, resulting in a value of .0033 for shrinkage. Thus the validation was successful.

Table 2 also describes the contributions of factors included in the multiple linear regression model, assuming that all other variables remain constant: on average, female values on EARR were 3.2% below male values; each additional month of treatment represented an average increase of 0.3% ; patients with premolar extractions had about 4% higher values than those without; use of Hyrax appliance increased the amount of EARRmax by 8%; the presence of GG genotype on the P2RX7 gene was associated with an average increase of 3%.

**Acknowledgements:** The authors are grateful to the *Center of Investigation on Environmental, Genetics and Oncobiology* (CIMAGO), University of Coimbra for funding this study.

## References

PEREIRA, S., LAVADO, N., NOGUEIRA, L. LOPEZ, M., ABREU, J. & SILVA, H. (2013) Polymorphisms of genes encoding P2X7R, IL-1B, OPG and RANK in orthodontic-induced apical root resorption. *Oral Diseases*, Article first published online: 31 OCT 2013, DOI: 10.1111/odi.12185.

**Sessão de Posters II – Sábado, 12 de Abril (11:20)**

# Motivação e satisfação na função pública: um exemplo dos Açores

**Áurea Sousa[1], Maria Graça Batista[2], Deanna Medeiros[3]**

[1]*Departamento de Matemática; CEEApla; CMATI, Universidade dos Açores, aurea@uac.pt;*
[2]*Departamento de Economia e Gestão; CEEApla, Universidade dos Açores, mbatista@uac.pt;*
[3]*Departamento de Economia e Gestão, Universidade dos Açores, deanna_medeiros@hotmail.com*

**Sumário**

O objetivo geral deste trabalho é o de aferir os níveis e os fatores de motivação e de satisfação profissional dos trabalhadores da função pública na Região Autónoma dos Açores (R.A.A.). Apresentam-se aqui as principais conclusões obtidas com base na análise dos dados recolhidos, a partir de um questionário composto por três secções (caracterização geral, níveis de satisfação e níveis de motivação), previamente testado e validado. Conclui-se, em termos gerais, que a faixa etária é um fator de influência na motivação e na satisfação.

**Palavras-chave:** Função pública, Motivação, Recursos humanos, Satisfação

## 1. Nota introdutória e objetivos gerais do estudo

Atualmente, os recursos humanos são considerados a base da estrutura de uma organização. Os colaboradores de uma organização, através do seu conhecimento e capacidades técnicas, permitem a criação de valor para os *stakeholders* e a consecução dos objetivos da organização. A estratégia ao nível dos recursos humanos, com os seus efeitos na motivação e satisfação dos colaboradores é um elemento fundamental da estratégia de uma organização (Dinham e Scott, 1998; Perry e Wise, 1990; Santhapparaj e Alam, 2005; Sledge et al., 2008). No entanto, ainda existe, em muitas organizações, uma aposta incipiente tanto no recrutamento, como na formação, na motivação e na liderança dos seus membros.

A satisfação profissional tem sido associada a resultados organizacionais positivos, tais como maior inovação, menor stress, maior *empowerment*, maior produtividade, crescimento organizacional, baixos níveis de absentismo e rotatividade, e elevados níveis de motivação. Herzberg et al. (1959) sublinharam a necessidade de reforçar os fatores motivacionais para gerar satisfação, sendo de referir que Dinham e Scott (1998) salientam que a satisfação e a motivação estão correlacionadas. Colaboradores motivados e satisfeitos geram: acréscimo dos níveis de eficiência e eficácia; resultados organizacionais positivos; clima e cultura organizacionais como propulsionadores de criatividade e cooperação entre equipas; inovação e redução de absentismo e rotatividade. O objetivo geral deste trabalho é o de aferir os níveis e os principais fatores de motivação e de satisfação profissional dos trabalhadores da função pública na R.A.A. As conclusões obtidas são confrontadas, sempre que possível, com as de outros autores.

## 2. Metodologia e variáveis de estudo

Os dados, relativos a 120 funcionários de diversas instituições públicas da R.A.A, foram recolhidos através de um questionário devidamente testado e validado. Este questionário, além das variáveis de caracterização geral da amostra, contém duas escalas, uma para avaliar os níveis de satisfação, constituída por 19 itens, e outra para avaliar os níveis de motivação, constituída por 18 itens. Cada funcionário selecionou uma e uma só de 5 modalidades possíveis (1 = *Muito Insatisfeito*/ *Desmotivado*, 2 = *Insatisfeito/ Desmotivado*, 3 = *Pouco Satisfeito/ Motivado*, 4 = *Satisfeito/ Motivado* e 5 = *Muito Satisfeito/ Motivado*) em relação a cada um dos itens que avaliam as duas variáveis latentes em estudo ("*Satisfação*" e "*Motivação"*)

Entre os diversos métodos estatísticos aplicados, destacam-se a Análise Classificatória Hierárquica Ascendente (*ACHA*), a Análise em Componentes Principais Categórica (*CatPCA*) e a Regressão Ordinal.

A *ACHA*, baseada no coeficiente de afinidade (e. g. Bacelar-Nicolau et al., 1980, 1985, 1988), e em cinco critérios de agregação, dois clássicos e três probabilísticos no âmbito da Metodologia *VL* (e.g. Lerman, 1981; Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998), foi efetuada com o software *CLASSIF* (Nicolau, Bacelar-Nicolau e Colaboradores, 2009) e permitiu identificar duas tipologias de variáveis ligadas, respetivamente, à satisfação e à motivação.

Dada a natureza das variáveis, utilizou-se a Análise em Componentes Principais Categórica (*CatPCA*), para os dois conjuntos de itens (variáveis componentes) que avaliam as duas variáveis latentes em estudo, de forma a possibilitar a representação gráfica das proximidades desses itens num espaço bidimensional (representação dos itens nos dois primeiros eixos de uma *CatPCA*).

A Regressão Ordinal foi utilizada para relacionar as variáveis dependentes (alguns dos 19 itens que avaliam a satisfação e alguns dos 18 itens que avaliam a motivação) com as variáveis independentes, "*Sexo*", "*Habilitações literárias*", "*Idade*" e "*Estado civil*".

## 3. Conclusões e desenvolvimentos futuros

As principais conclusões relativas aos níveis e aos principais fatores de motivação e de satisfação profissional dos trabalhadores da função pública na R.A.A são sintetizadas em tabelas. Conclui-se, em termos gerais, que a faixa etária é um fator de influência na motivação e na satisfação, tendo-se comprovado também que os totais das duas escalas estão positivamente correlacionados, tal como era de esperar a partir da literatura.

A dimensão da amostra utilizada, condicionada pela dispersão geográfica das diferentes direções e serviços regionais da função pública na R.A.A. e pela falta de adesão de algumas instituições públicas (relativamente ao preenchimento dos questionários), constitui uma amostra limitada do cenário de investigação. No entanto, os resultados deste e de um eventual estudo

abrangendo uma amostra de maior dimensão, complementada pela realização de entrevistas, poderão ser úteis a nível da implementação de futuras reestruturações nas políticas de recursos humanos, com o intuito de incrementar os níveis de satisfação e motivação na função pública nos Açores.

**Referências**

BACELAR-NICOLAU, H. (1980). *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória,* Tese de Doutoramento, FCL, Universidade de Lisboa.

BACELAR-NICOLAU, H. (1985). The Affinity Coefficient in Cluster Analysis. IN BEKMANN, M.J. (Ed.) *Methods of Operations Research.* Munchen, Verlag Anton Hain, 507-512.

BACELAR-NICOLAU, H. (1988). Two Probabilistic Models for Classification of Variables in Frequency Tables. IN BOCK, H.-H. (Ed.) *Classification and Related Methods of Data Analysis.* North Holland, Elsevier Sciences Publishers B.V., 181-186.

DINHAM, S. & SCOTT, C. (1998). A Three Domain Model of Teacher and School Executive Career Satisfaction. *Journal of Educational Administration,* 36(4)**,** 362-378.

HERZBERG, F., MAUSNER, B. & SNYDERMAN, B. (1959). *The Motivation to Work,* New York, John Wiley & Sons.

LERMAN, I.C. (1981). *Classification et Analyse Ordinale des Données,* Paris, Dunod.

NICOLAU, F.C. (1983). Cluster Analysis and Distribution Function. *Methods of Operations Research,* 45**,** 431-433.

NICOLAU, F.C. & BACELAR-NICOLAU, H. (1998). Some Trends in the Classification of Variables. IN Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Ed.) *Data Science, Classification, and Related Methods.* Springer-Verlag, 89-98.

NICOLAU, F.C., BACELAR-NICOLAU, H., SOUSA, F., SOUSA, Á. & SILVA, O. (2009). *CLASSIF:Software de Análise Classificatória - Abordagens Clássica e Probabilística VL*. Estudos e Aplicações do LEAD, FP-UL.

PERRY, J. L. & WISE, L. R. (1990). The Motivational Bases of Public Service. *Public Administration Review,* 50, 367-373.

SANTHAPPARAJ, A. S. & ALAM, S. S. (2005). Job Satisfaction among Academic Staff in Private Universities in Malaysia. *Journal of Social Sciences*, 1 (2), 72-76.

SLEDGE, S., MILES, A. & COPPAGE, S. (2008). What Role does Culture Play? A Look at Motivation and Job Satisfaction among Hotel Workers in Brazil. *International Journal of Human Resource Management,* 19(9), 1667-1682.

# Índice de Autores