

# **CLASSIFICAÇÃO E ANÁLISE DE DADOS**

## **Métodos e Aplicações II - CLADMAp II**

A large, stylized geometric logo in the background, composed of overlapping triangles and squares in shades of blue and purple. The logo features the letters 'CLAD' in a bold, blue, sans-serif font, with a stylized 'A' that is wider and more triangular. The logo is centered and occupies a significant portion of the cover.

**CLAD**

**Editores**

Helena Bacelar-Nicolau

Fernanda Sousa

Fátima Ferreira

Luís M. Grilo

A. Manuela Gonçalves

Carlos Marcelo

# **CLASSIFICAÇÃO E ANÁLISE DE DADOS**

## **MÉTODOS E APLICAÇÕES II - CLADMAp II**

**Editores**

Helena Bacelar-Nicolau

Fernanda Sousa

Fátima Ferreira

Luís M. Grilo

A. Manuela Gonçalves

Carlos Marcelo

**Título**

Classificação e Análise de Dados – Métodos e Aplicações II

**Editores**

Helena Bacelar-Nicolau (Universidade de Lisboa)

Fernanda Sousa (Universidade do Porto)

Fátima Ferreira (Universidade de Trás-os-Montes e Alto Douro)

Luís M. Grilo (Instituto Politécnico de Tomar)

A. Manuela Gonçalves (Universidade do Minho)

Carlos Marcelo (Instituto Nacional de Estatística)

**Impressão**

Instituto Nacional de Estatística

Av. António José de Almeida

1000-043 LISBOA

**1.ª Edição**

**Lisboa, Abril de 2017**

**ISSN 2183-8801**

**Depósito legal 411987/16**

Tiragem: 250 exemplares

Todos os direitos reservados. Nenhuma parte desta publicação pode ser reproduzida por processo mecânico, eletrónico ou outro sem autorização escrita dos editores.

# Índice

## CLASSIFICAÇÃO E ANÁLISE DE DADOS MÉTODOS E APLICAÇÕES II – CLADMAp II

Prefácio .....	v
Agradecimento aos revisores .....	vii
Modelos com trajetória latente no estudo da privação material em Portugal entre 2007 e 2010 .....	1
<i>Paula C. R. Vicente e Maria de Fátima Salgueiro (JOCLAD 2011)</i>	
Entropia relativa em misturas de regressões lineares .....	13
<i>Susana Faria e Gilda Soromenho (JOCLAD 2011)</i>	
Vítimas mortais em acidentes de viação em Portugal continental: componentes principais e previsão .....	21
<i>Fernando Sebastião e Irene Oliveira (JOCLAD 2011)</i>	
Seleção de atributos valorativos da habitação: uma aplicação ao mercado habitacional de Aveiro e Ílhavo .....	31
<i>Paulo Batista, Gladys Castillo, João Marques e Eduardo Castro (JOCLAD 2011)</i>	
Novas aplicações de métodos multivariados na análise da mortalidade: um estudo na região norte de Portugal, 2001-2005 .....	43
<i>Lara Teixeira, Vasco Machado, Manuela Felício e A. Manuela Gonçalves (JOCLAD 2011)</i>	
Estimação de um modelo com trajetória latente com dados omissos resultantes de um painel rotativo .....	55
<i>Paula C. R. Vicente e Maria de Fátima Salgueiro (JOCLAD 2012)</i>	

Análise dos perfis de consumo de <i>cannabis</i> pelos adolescentes de Ponta Delgada .....	65
<i>Áurea Sousa, Hélder Rocha Pereira, Sara Raposo, Osvaldo Silva e Helena Bacelar-Nicolau (JOCLAD 2012)</i>	
Seleção robusta em modelos de regressão linear com um grande número de preditores .....	77
<i>Shirin Shahriari, Susana Faria e A. Manuela Gonçalves (JOCLAD 2013)</i>	
A deeper glance on the percentage warping path distortion measure .....	85
<i>Joana Hora e Pedro Campos (JOCLAD 2013)</i>	
Elevados níveis de ferro nos doentes alcoólicos: um contributo para esclarecer esta correspondência .....	97
<i>Ana Matos, Carla Henriques, Luís Costa Matos, Nuno Monteiro e Paulo Batista (JOCLAD 2013)</i>	

## Prefácio

A série de publicações CLASSIFICAÇÃO E ANÁLISE DE DADOS - Métodos e Aplicações, CLADMap, tem, como principal objetivo, estimular e promover a difusão de informação técnica e científica no campo da Classificação e Análise de Dados.

Os trabalhos apresentados habitualmente nas JOCLAD – Jornadas de Classificação e Análise de Dados, cobrem e combinam um vasto número de métodos e aplicações da análise de dados multivariados (por exemplo, em Saúde, Biologia, Ambiente, Sociologia, Economia, Finança, Demografia e Estatística) e acreditamos que a sua publicação pela CLAD potencia o impacto da investigação científica que lhes deu origem.

Assim, este segundo volume da série, CLADMap II, vem dar continuidade à vontade expressa de associados e participantes nas sucessivas JOCLAD, de que a CLAD desse apoio à divulgação de trabalhos neles apresentados, deste modo os valorizando. Os artigos aqui incluídos, após processo de revisão interpares, são desenvolvimentos de trabalhos apresentados nas JOCLAD 2011-2013 e espelham a interdisciplinaridade e a diversidade de tópicos que integram estas jornadas.

Esta publicação inclui maioritariamente textos em português e alguns em inglês. Dos primeiros, uns seguem o mais recente acordo ortográfico, outros não, já que deixamos aos autores a liberdade dessa escolha.

Agradecemos aos autores, aos revisores e a todos os que direta ou indiretamente nos apoiaram na criação do CLADMap II.

Ao INE, parceiro privilegiado da CLAD, que, desde o início, vem cooperando nas suas atividades nacionais e internacionais, e particularmente na edição das publicações, o nosso agradecimento especial.

Finalmente, fica um convite aos investigadores teóricos e aplicados nesta área científica, de que submetam os seus trabalhos para publicação no próximo CLADMap III.

Lisboa, Abril de 2017

Helena Bacelar-Nicolau (Universidade de Lisboa)

Fernanda Sousa (Universidade do Porto)

Fátima Ferreira (Universidade de Trás-os-Montes e Alto Douro)

Luís M. Grilo (Instituto Politécnico de Tomar)

A. Manuela Gonçalves (Universidade do Minho)

Carlos Marcelo (Instituto Nacional de Estatística)

## **Agradecimento aos revisores**

Os revisores dos artigos submetidos a esta publicação CLADMAp II, vêm listados na tabela seguinte, por ordem alfabética dos seus apelidos. Aos Colegas que connosco colaboraram, generosa e arduamente, os editores agradecem.

Conceição Amado	Universidade de Lisboa
Helena Bacelar-Nicolau	Universidade de Lisboa
Jorge Cadima	Universidade de Lisboa
Paulo Canas Rodrigues	Universidade Federal da Baía, Brasil
Miguel de Carvalho	Universidade Nova de Lisboa
Marco Costa	Universidade de Aveiro
Pedro Duarte Silva	Universidade Católica Portuguesa
Fátima Ferreira	Universidade de Trás-os-Montes e Alto Douro
Adelaide Figueiredo	Universidade do Porto
A. Manuela Gonçalves	Universidade do Minho
Luzia Gonçalves	Universidade Nova de Lisboa
Luís M. Grilo	Instituto Politécnico de Tomar
Manuela Neves	Universidade de Lisboa
Irene Oliveira	Universidade de Trás-os-Montes e Alto Douro
Rosário Oliveira	Universidade de Lisboa
Ana Pires	Universidade de Lisboa
Ana Sousa Ferreira	Universidade de Lisboa
Fernanda Sousa	Universidade do Porto





# Modelos com trajetória latente no estudo da privação material em Portugal entre 2007 e 2010

Paula C. R. Vicente<sup>1</sup> · Maria de Fátima Salgueiro<sup>2</sup>

© The Author(s) 2017

**Resumo** Este estudo tem por objetivo modelar longitudinalmente o conceito de privação material em Portugal, utilizando dados do Inquérito às Condições de Vida e Rendimento (ICOR), participação portuguesa na base de dados europeia EU-SILC. É proposto um modelo com trajetória latente, condicionada, de segunda ordem, para explicar a trajetória da privação material nos anos de 2007 a 2010. É discutida a inclusão de variáveis explicativas no modelo quando consideradas como variantes e invariantes no tempo.

**Palavras-chave:** EU-SILC, ICOR, Modelos com Trajetória Latente de Segunda Ordem, Privação Material.

## 1 Introdução

A pobreza é um conceito multidimensional e bastante complexo de ser medido. De acordo com TOWNSEND (1979), os indivíduos ou agregados dizem-se em pobreza quando perdem os recursos para obter o tipo de alimentação, participar em atividades, ter as condições de vida e comodidades a que estão habituados ou pelo menos usuais nas comunidades em que se inserem.

---

<sup>1</sup>Universidade Lusófona de Humanidades e Tecnologias, Escola de Ciências Económicas e das Organizações, *p951@ulusofona.pt*

<sup>2</sup>Business Research Unit e Departamento de Métodos Quantitativos para Gestão e Economia, Instituto Universitário de Lisboa (ISCTE-IUL), *fatima.salgueiro@iscte.pt*

A privação material corresponde à ausência forçada de uma combinação de itens que contribuem para condições materiais de vida ou de qualidade de vida, tais como condições de habitação, posse de bens, capacidade para suportar necessidades básicas na sociedade a que os indivíduos pertencem, ou mesmo condições ambientais (MACK & LANSLEY, 1985).

O Inquérito às Condições de Vida e Rendimento das Famílias (ICOR) é um painel com uma periodicidade anual, o qual foi implementado com o objetivo de assegurar a participação da população portuguesa na base de dados europeia denominada EU-SILC (*European Statistics on Income and Living Conditions*).

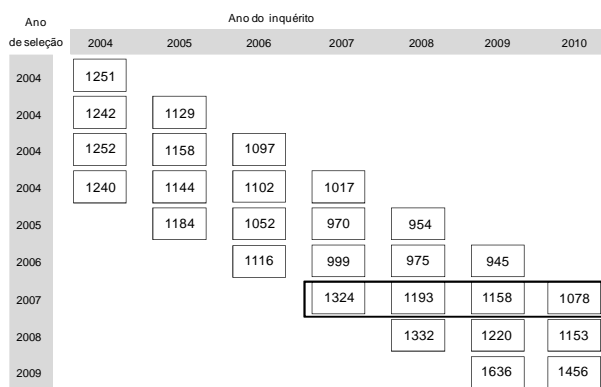
Neste trabalho são utilizados os dados do ICOR correspondentes a quatro momentos temporais, os anos de 2007 a 2010, com o objetivo de modelar longitudinalmente a privação material, considerando este conceito como medido em três dimensões: i) constrangimentos económicos; ii) posse de bens duradouros; e iii) condições da habitação (GUIO, 2009). Com este propósito são calculados *scores* para cada uma das três dimensões da privação, sendo estes utilizados como indicadores ordinais de medida do construto *privação material* em cada um dos momentos temporais. Esta modelação da privação material é feita recorrendo a um modelo com trajetória latente de segunda ordem. Uma vez descrita a trajetória da privação material no período em análise, são consideradas como variáveis explicativas desta trajetória, o rendimento, a dimensão e a área de urbanização do local de residência do agregado. É discutida a inclusão destas variáveis no modelo condicional quando consideradas como variantes e como invariantes no tempo. Os modelos com trajetória latente de segunda ordem, não condicionada e condicionada, propostos neste trabalho são estimados com recurso ao pacote estatístico Mplus 6 (MUTHÉN & MUTHÉN, 1998-2010). Os indicadores de medida utilizados são modelados como ordinais.

## 2 Amostra em estudo

A participação portuguesa no EU-SILC é assegurada pelo Instituto Nacional de Estatística, desde 2004, dispondo-se de dados longitudinais para o período de 2004 a 2010, quer ao nível dos agregados, quer ao nível dos indivíduos. Este painel apresenta ainda a particularidade de ser um painel rotativo, isto é, uma fração da amostra é renovada todos os anos (INE, 2009), conforme representado na Figura 1.

A privação material foi considerada como medida em três dimensões: i) constrangimentos económicos; ii) posse de bens duradouros; e iii) condições da habitação. Os *scores* para cada uma das dimensões da privação material foram calculados como uma soma de itens disponíveis no ICOR. A dimensão *constrangimentos económicos* inclui os cinco itens (correspondentes a ausência de capacidade financeira para): 1) pagar uma semana de férias fora de casa a todo o agregado, por ano; 2) ter uma refeição de peixe ou carne, pelo menos de dois em dois dias; 3) suportar despesas inesperadas, sem recorrer a crédito; 4) fazer face às

despesas e encargos usuais; e 5) ter a casa aquecida. Na dimensão *posse de bens duradouros* foram considerados os itens (correspondentes a ausência de disponibilidade económica para possuir): 1) TV a cores; 2) telefone fixo ou móvel; 3) máquina de lavar roupa; e 4) veículo ligeiro de passageiros ou misto. Na dimensão *condições da habitação* foram considerados: 1) ter telhado que deixa passar água, paredes/fundações/chão húmido, caixilhos ou chão apodrecido; 2) ausência de retrete; e 3) ausência de instalações de banho ou de duche no interior. Deste modo, a variável constrangimentos económicos toma valores numa escala de 0 a 5 itens em privação, a variável posse de bens toma valores entre 0 a 4 itens em falta e a variável condições da habitação numa escala de 0 a 3 itens em falta.



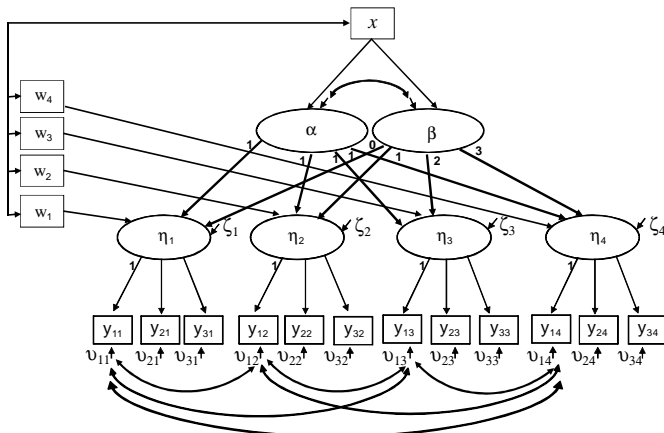
**Figura 1** - Desenho do painel rotativo ICOR. Dentro da caixa os agregados considerados neste estudo.

A amostra em estudo inclui 967 agregados com respostas válidas a todas as questões sobre constrangimentos económicos, posse de bens duradouros e condições de habitação, nos anos de 2007 a 2010 (ver Figura 1). Deste conjunto de agregados, 46.9% não têm capacidade financeira para assegurar mais de dois itens relativos a constrangimentos económicos, no ano de 2007, sendo os correspondentes valores de 43.7% em 2008, de 41.7% em 2009 e de 42.5% em 2010. Relativamente à posse de bens duradouros, 82.3%, 85.3%, 85.7% e 86.1% dos agregados, possuem todos os bens considerados respetivamente nos anos de 2007, 2008, 2009 e 2010. Quanto às condições da habitação, 77.2%, 77.6%, 78.9% e 78.6% dos agregados não apresentam qualquer um dos problemas considerados, nos anos de 2007 a 2010, respetivamente. As variáveis dimensão do agregado, rendimento disponível do agregado e área de urbanização foram consideradas como explicativas da trajetória da privação material. A dimensão média dos agregados, nos anos de 2007 a 2010, é de 2.74, 2.76, 2.71 e 2.67 elementos, com um desvio padrão de 1.298, 1.304, 1.274 e 1.265. O rendimento médio disponível das famílias aumentou entre 2007 e 2010, com valores de 16283.467, 17334.729, 17821.481 e 18619.823 euros, respetivamente. O desvio padrão do rendimento tomou valores de

12892.597, 15578.820, 17723.414 e 15728.051 euros, nos anos de 2007 a 2010, respetivamente. Quanto à área de urbanização, os agregados dividem-se em número aproximadamente igual pelas três categorias: área muito povoada, intermédia e pouco povoada.

### 3 Modelo com trajetória latente de segunda ordem condicionada

Os modelos com trajetória latente (BOLLEN & CURRAN, 2006) permitem, através da estimação de uma trajetória (linear ou não): i) estimar o nível médio inicial e a taxa média de mudança do conjunto dos indivíduos; ii) estimar a variabilidade entre indivíduos no valor inicial e na taxa de mudança; e iii) explicar a heterogeneidade observada nas trajetórias, recorrendo a modelos condicionais. A ideia subjacente a este tipo de modelação estatística é a de que a verdadeira trajetória não é observada, mas sim latente, sendo a mesma estimada a partir da estrutura de médias e de associações entre as medidas repetidas das variáveis observadas, isto é, não é modelada a evolução das variáveis, mas sim os processos que lhe estão subjacentes. Quando o objetivo não é modelar a trajetória das variáveis observadas, mas sim a trajetória de uma variável latente medida por múltiplos indicadores, deve ser considerado um modelo com trajetória latente de segunda ordem, sendo ainda possível incorporar no modelo variáveis que procuram explicar essa trajetória, recorrendo a modelos condicionais (como o apresentado na Figura 2).



**Figura 2** - Diagrama de um modelo com trajetória latente de segunda ordem, condicionada, com variáveis explicativas invariantes ( $x$ ) e variantes no tempo ( $w_t$ ). São consideradas quatro medidas repetidas de três indicadores ( $y_{jt}$ ). Por simplificação, apenas estão representadas as correlações entre termos residuais das medidas repetidas do primeiro indicador.

Num modelo com trajetória latente de segunda ordem o valor de  $y_{ijt}$ , o indicador  $j$ , para o elemento  $i$ , no momento temporal  $t$ , é dado por

$$y_{ijt} = v_{jt} + \Lambda_{jt}\eta_{it} + u_{ijt},$$

em que  $v_{jt}$  é o intercepto do indicador  $j$  no momento  $t$ ,  $\Lambda_{jt}$  corresponde ao peso fatorial do indicador  $j$  no momento  $t$  e  $u_{ijt}$  é a perturbação do elemento  $i$ , no momento  $t$ , para o indicador  $j$ . É pressuposto do modelo que o termo residual  $u_{ijt}$  tem média zero e não está correlacionado com os outros termos da equação. É também assumido que o termo residual tem distribuição normal e que a estrutura da matriz de variâncias-covariâncias depende dos termos que se permitirem correlacionar, sendo usual permitir a correlação entre os termos residuais das medidas repetidas ao longo do tempo de uma mesma variável observada. A equação da trajetória da variável latente  $\eta_t$  tem a seguinte expressão

$$\eta_{it} = \alpha_i + \lambda_t\beta_i + \zeta_{it},$$

em que,  $\eta_{it}$  representa a variável latente para o elemento  $i$  no momento  $t$ ,  $\lambda_t$  define a trajetória (é igual a  $t - 1$  se a trajetória considerada é linear), e  $\zeta_{it}$  representa a perturbação do elemento  $i$ , no momento  $t$ , sendo um pressuposto do modelo que  $\zeta_{it}$  tem média zero, variância  $\psi_t$  e que não está correlacionado com nenhum dos outros termos da equação. O intercepto e o declive da trajetória são dados, respetivamente, por

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i} \text{ e } \beta_i = \mu_\beta + \zeta_{\beta_i}$$

em que,  $\mu_\alpha$  e  $\mu_\beta$  são, respetivamente, a média do intercepto e a média do declive aleatórios para todos os elementos. É pressuposto do modelo que os termos residuais  $\zeta_{\alpha_i}$  e  $\zeta_{\beta_i}$  têm distribuição normal, com média zero e variâncias dadas, respetivamente, por  $\psi_{\alpha\alpha}$  e  $\psi_{\beta\beta}$ , e que não estão correlacionados com  $\zeta_{it}$  e  $\lambda_t$ .

Num modelo com trajetória latente de segunda ordem não condicionada os parâmetros de interesse são as médias e as variâncias dos efeitos aleatórios, bem como a covariância entre estes efeitos. Além destes, também a estrutura dos pesos fatoriais do modelo de medida (que relaciona a variável latente com os indicadores de medida) é considerada importante. Num modelo com trajetória latente não condicional apenas o efeito tempo descreve/explica o comportamento da trajetória.

Quando, para além de descrever, o objetivo é explicar a trajetória, recorre-se a um modelo condicional, com trajetória latente condicionada (como é o caso do apresentado na Figura 2). As variáveis explicativas podem ser consideradas como invariantes ou como variantes no tempo, sendo a sua incorporação no modelo realizada de forma distinta num e noutro caso. Assim, e sendo,  $x_i$  uma variável explicativa invariante no tempo, o intercepto e o declive aleatórios são dados por

$$\alpha_i = \mu_\alpha + \gamma_\alpha x_i + \zeta_{\alpha_i} \text{ e } \beta_i = \mu_\beta + \gamma_\beta x_i + \zeta_{\beta_i},$$

em que  $\gamma_\alpha$  e  $\gamma_\beta$  representam os seus coeficientes de regressão. Se a variável explicativa não se mantém constante com a passagem do tempo, então o interesse é agora o efeito das suas medidas repetidas nas medidas repetidas da variável latente

$\eta$  para a qual se modela a trajetória. Nesse caso, a equação da trajetória da variável latente  $\eta_t$  para o elemento  $i$  vem dada por

$$\eta_{it} = \alpha_i + \lambda_t \beta_i + \gamma_t w_{it} + \zeta_{it},$$

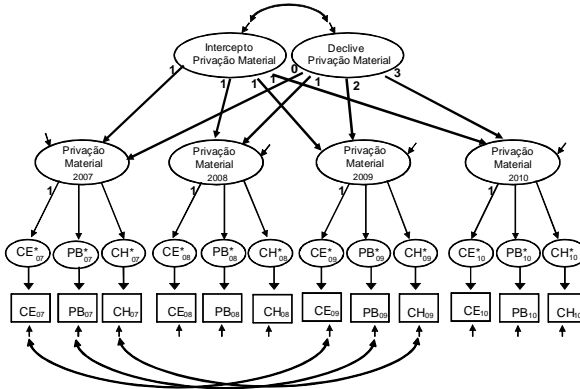
sendo  $w_t$  a variável explicativa variante no tempo e  $\gamma_t$  o seu coeficiente de regressão. Assim, os parâmetros de interesse no primeiro caso são os coeficientes de regressão dos efeitos aleatórios sobre a variável  $x$  ( $\gamma_\alpha$  e  $\gamma_\beta$ ), e no segundo caso os coeficientes de regressão das medidas repetidas de  $\eta_t$  sobre as medidas repetidas da variável explicativa  $w_t$  (os  $\gamma_t$ ). De salientar que, considerar as variáveis explicativas como invariantes no tempo resulta num modelo mais parcimonioso, todavia tratar uma variável explicativa como invariante no tempo, sendo ela variante, resulta em problemas de especificação do modelo e num provável enviesamento nos resultados (BOLLEN & CURRAN, 2006).

Se os indicadores de medida considerados são ordinais então  $y_{ijt}$  deve ser substituído pela variável latente contínua que lhe está subjacente  $y_{ijt}^*$ . A relação entre  $y_{jt}$  e  $y_{jt}^*$  é dada pela equação:  $y_{ijt} = c$  quando  $\tau_{c-1} < y_{ijt}^* \leq \tau_c$ , onde  $c = 1, 2, \dots, C$  é o número total de categorias ordenadas,  $\tau_{c-1}$  e  $\tau_c$  são os limites inferior e superior da categoria  $c$ , com  $\tau_0 = -\infty$  e  $\tau_C = +\infty$ , sendo os  $C-1$  valores de *threshold* ordenados do menor para o maior.

Para medir a qualidade do ajustamento modelo-dados são usualmente utilizadas na literatura as seguintes medidas: índice *Tucker-Lewis* (TLI); índice de ajustamento comparado (CFI); o *Root Mean Square Error of Approximation* (RMSEA) e o *Weighted Root Mean Square Residual* (WRMR). São considerados valores de ajustamento recomendáveis TLI>0.90, CFI>0.90, RMSEA<0.05 e WRMR<0.90, (SCHUMACKER & LOMAX, 2010; YU & MUTHÉN, 2002).

## 4 Resultados

A modelação longitudinal do conceito de privação material foi realizada recorrendo a um modelo com trajetória latente de segunda ordem, sendo a variável latente *privação material* medida por três indicadores: constrangimentos económicos, posse de bens duradouros e condições da habitação. Estas variáveis observadas foram consideradas como ordinais, tendo sido definidos *thresholds*. Foi considerada a invariância do modelo de medida ao longo do tempo, assumindo os pesos fatoriais de cada indicador iguais nos diferentes momentos temporais. Permitiu-se ainda correlacionar os termos residuais das medidas repetidas da mesma variável manifesta, nos diferentes momentos temporais. Todavia, por forma a evitar sobrecarregar o diagrama, na Figura 3, apenas as correlações entre termos residuais dos indicadores no momento 1 com o momento 3, estão representadas.



**Figura 3** - Diagrama de um modelo com trajetória latente de segunda ordem da privação material, medida em 3 dimensões, por indicadores considerados como variáveis ordinais.

A estimação do modelo proposto foi realizada recorrendo ao pacote estatístico Mplus 6, tendo sido obtidos os seguintes valores para as medidas de ajustamento: CFI=TLI=0.998, RMSEA=0.022 e WRMR=0.689. Estes valores permitem concluir que existe um bom ajustamento modelo-dados (ver secção 3). Uma vez que os indicadores de medida são modelados como ordinais, o método de estimação mais adequado é o WLSMV (*Weighted Least Square Means and Variance Adjusted*). Este método de estimação resulta de uma adaptação apresentada por MUTHÉN, DUTOIT & SPISIC (1997) do estimador WLS (*Weighted Least Squares*), sendo um dos métodos de estimação robustos (com pressupostos mínimos acerca da distribuição das variáveis) disponibilizados pelo Mplus. De acordo com MUTHÉN & MUTHÉN (1998-2010, pag. 484) o método WLSMV utiliza “*weighted least square parameter estimates using a diagonal weight matrix with standard errors and mean and variance adjusted chi-square test statistic that use a full weight matrix*”.

Na Tabela 1 apresentam-se as estimativas da média, variâncias e covariância entre intercepto e declive, obtidas para o modelo com trajetória latente de segunda ordem. A média do intercepto toma o valor zero (porque a média de  $\eta_1$  é nula) e não é um parâmetro do modelo. A média do declive é significativa e negativa, sugerindo que a privação material das famílias diminui entre 2007 e 2010. A variância estimada para o intercepto é significativa, logo as famílias diferem na privação material que experimentam no momento inicial, ano de 2007. A variância estimada do declive não se mostrou significativa, o que quer dizer que não existem diferenças estatisticamente significativas na taxa média de mudança da privação material para o conjunto das famílias. O valor estimado para a covariância entre o intercepto e o declive não se mostrou significativo, não sendo portanto possível dizer que existe relação significativa entre o nível de privação material no ano de 2007 e a taxa média de mudança de 2007 a 2010.



**Tabela 1** - Estimativas (valores dos testes t) da média, variâncias e covariância do intercepto e declive para um modelo com trajetória latente de segunda ordem para a privação material.

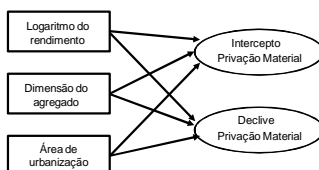
	<b>Privação Material</b>	
	intercepto	declive
<b>Média</b>	---	<b>-0.018</b> (-2.073)
<b>Variância</b>	<b>0.589</b> (8.607)	0.007 (1.300)
<b>Covariância</b>	-0.018 (-1.572)	

De seguida foram consideradas variáveis explicativas da trajetória de mudança da privação material, tais como, (logaritmo do) rendimento, dimensão do agregado e área de urbanização. Como a variável área de urbanização é definida em três categorias, foram criadas duas variáveis *dummy*, sendo a categoria de referência área pouco povoada. Procedimento idêntico foi seguido para a variável dimensão do agregado, para a qual foi calculada uma variável *dummy* com duas categorias, três ou menos indivíduos (categoria de referência) e mais de três indivíduos. A variável logaritmo do rendimento foi modelada como contínua. Foi utilizado o logaritmo do rendimento com o intuito de atenuar a heterogeneidade dos dados. Com o objetivo de estudar o efeito destas três variáveis na trajetória latente da privação material, estas foram introduzidas no modelo como invariantes no tempo (ver Figura 4), sendo apresentados na Tabela 2 os valores estimados para os coeficientes de regressão dos efeitos aleatórios nas três variáveis explicativas. A análise desta tabela permite observar que os impactos do (logaritmo do) rendimento, da dimensão do agregado e da área de urbanização do local de residência do agregado, no intercepto da trajetória latente da privação material se mostraram significativos. Assim, é possível dizer que agregados com rendimentos mais elevados experimentam uma menor privação material em 2007, isto porque o valor estimado é negativo. Por outro lado, um coeficiente de regressão estimado positivo para a variável dimensão do agregado, permite dizer que agregados com mais de três indivíduos têm maior privação em 2007, face a agregados com três ou menos indivíduos. Do mesmo modo, agregados que no ano de 2007 residem em áreas muito povoadas ou intermédias (grandes ou médias cidades) experimentam maior privação material face a agregados que habitam em áreas pouco povoadas (aldeias). Os efeitos das três variáveis explicativas na taxa média de mudança da privação material, ao longo do período entre 2007 a 2010, não se mostraram significativos. Alternativamente, foi considerado um outro modelo em que a variável (logaritmo do) rendimento foi modelada como variante no tempo (Figura 5) com o objetivo de explicar a variabilidade observada na privação material, no período de 2007 a 2010. As variáveis área de urbanização e dimensão do agregado foram introduzidas no modelo como invariantes, porque apresentam valores análogos ao longo do período em questão (2007 a 2010). Obteve-se os seguintes valores de ajustamento modelo-dados: TLI=0.997, CFI=0.998, RMSEA=0.018 e WRMR=0.721.

**Tabela 2** - Estimativas (valores dos testes t) dos coeficientes de regressão das variáveis explicativas do intercepto e do declive para um modelo com trajetória latente de segunda ordem condicionada.

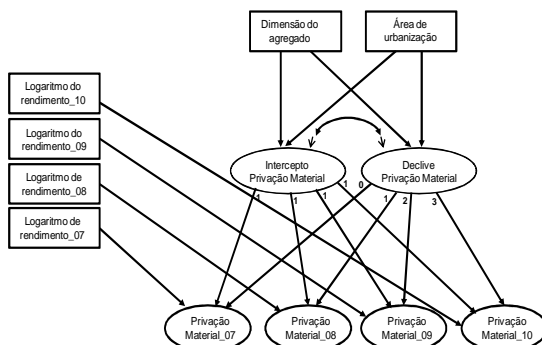
A negrito estão os valores que se mostraram significativos.

	<b>Privação Material</b>	
	intercepto	declive
<b>Logaritmo do rendimento</b>	<b>-0.831</b> (-17.802)	0.0140 (0.469)
<b>Dimensão do agregado</b> (três ou menos indivíduos)		
mais de três indivíduos	<b>0.355</b> (4.541)	0.037 (1.377)
<b>Área de urbanização</b> (pouco povoada)		
muito povoada	<b>0.339</b> (4.343)	-0.025 (-0.992)
intermédia	<b>0.190</b> (2.393)	0.0060 (0.217)

**Figura 4** - Diagrama de um modelo com trajetória latente de segunda ordem da privação material. As variáveis explicativas são consideradas como invariantes no tempo.

A análise da Tabela 3 permite concluir que o logaritmo do rendimento tem um impacto significativo e negativo na privação material nos anos 2007, 2008 e 2010, significando que valores mais elevados de rendimento implicam menor privação material nestes três anos. A dimensão do agregado e a área de urbanização apresentam um impacto positivo e significativo no intercepto da trajetória da privação material, embora esse efeito não se tenha mostrado significativo no declive da trajetória. Deste modo, pode concluir-se que agregados de maior dimensão (mais de três elementos) apresentam níveis de privação mais elevados em 2007, relativamente a agregados menores (três ou menos elementos). Agregados residentes em áreas muito povoadas ou intermédias apresentam, no ano de 2007, uma privação material superior face a agregados que habitam em áreas pouco povoadas. A comparação dos resultados obtidos na modelação da trajetória latente de segunda ordem da privação material considerando a variável explicativa (logaritmo do) rendimento como invariante ou variante no tempo permite conclusões distintas: no primeiro caso é possível dizer que agregados com rendimento mais elevado experimentam um nível de privação material mais baixo no ano de 2007, não sendo possível concluir que o (logaritmo do) rendimento influencie a taxa média de mudança da privação material, enquanto que no segundo caso se conclui que agregados com (logaritmo do) rendimento mais elevado apresentam valores de privação material mais baixos nos anos de 2007, 2008 e

2010. Foram obtidos os valores de ajustamento modelo-dados: TLI=0.987, CFI=0.988, RMSEA=0.032 e WRMR=1.141.



**Figura 5** - Diagrama de um modelo com trajetória latente de segunda ordem da privação material. A variável explicativa da trajetória, logaritmo do rendimento, é considerada como variante no tempo. As variáveis dimensão do agregado e área de urbanização são consideradas como invariantes.

**Tabela 3** - Estimativas (valores dos testes t) dos coeficientes de regressão da variável explicativa (logaritmo do) rendimento no fator latente privação material, e dos coeficientes de regressão das variáveis explicativas dimensão do agregado e área de urbanização do intercepto e do declive.

A negrito estão os valores que se mostraram estatisticamente significativos.

	Privação Material			
	2007	2008	2009	2010
<b>Logaritmo do rendimento</b>	<b>-0.526</b> (-6.288)	<b>-0.293</b> (-2.875)	-0.006 (-0.058)	<b>-0.642</b> (-5.383)
	<b>intercepto</b>		<b>declive</b>	
<b>dimensão do agregado</b> (três ou menos indivíduos)				
mais de três indivíduos	<b>0.394</b> (4.962)		0.046 (1.363)	
<b>Área de urbanização</b> (pouco povoada)				
muito povoada	<b>0.342</b> (4.369)		-0.027 (-1.057)	
intermédia	<b>0.188</b> (2.392)		0.001 (0.031)	

## 5 Discussão

Este trabalho propôs a modelação longitudinal do conceito de privação material medido nas suas três dimensões, constrangimentos económicos, posse de bens duradouros e condições da habitação, recorrendo a um modelo com trajetória latente de segunda ordem. Para cada uma das três dimensões foi calculado um *score* obtido como uma soma de itens, em cada momento temporal, e considerado na modelação como uma variável ordinal.

Os resultados obtidos sugerem que os agregados diferem na privação material que experimentam no ano de 2007. Por outro lado, a privação material que os agregados vivenciam diminui ao longo dos quatro momentos temporais, 2007 a 2010.

Um modelo com trajetória latente de segunda ordem condicionada foi ainda considerado por forma a determinar possíveis variáveis explicativas do comportamento da privação material. Estas variáveis foram consideradas como invariantes no tempo, tendo sido possível verificar que, no ano de 2007, os agregados com rendimentos mais elevados experimentam uma menor privação material. Por outro lado, agregados de maior dimensão e residentes em grandes ou médias cidades têm maior privação face a agregados menores e residentes em zonas rurais, respetivamente. Quando a variável explicativa rendimento é considerada como variante no tempo verifica-se que rendimentos mais elevados implicam menor privação material, nos anos de 2007, 2008 e 2010. O ajustamento modelo-dados é superior se as variáveis explicativas são consideradas invariantes.

## Referências

- BOLLEN, K.E. & CURRAN, P. (2006). *Latent Curve Models: A Structural Equation Perspective*, John Wiley & Sons, Inc., New Jersey.
- GUIO, A-C (2009). What can be learned from material deprivation indicators in Belgium and in its regions, *Institut Wallon de L'Évaluation de la Prospective et de la Statistique*, 901.
- INE (2009). Inquérito às Condições de Vida e Rendimento - ICOR, *Documento Metodológico*.
- MACK, J. & LANSLEY, S. (1985). *Poor Britain*, George Allen & Unwin, Ltd, London.
- MUTHÉN, B.O., DUTOIT, S.H. & SPISIC, D. (1997). *Robust Inference using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes*. Unpublished technical report, [www.statmodel.com](http://www.statmodel.com).
- MUTHÉN, L.K. & MUTHÉN, B.O. (1998-2010). *Mplus user's guide*, 6<sup>th</sup> edition, Los Angeles, CA: Muthén & Muthén.
- SCHUMACKER, R.E & LOMAX, R.G. (2010). *A beginner's guide to Structural Equation Modelling*, 2<sup>nd</sup> edition, Lawrence Erlbaum Associates, Inc.
- TOWNSEND, P. (1979). *Poverty in the United Kingdom*, Penguin Books, Hardmonsworth.
- YU, C.Y. & MUTHÉN, B.O. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. *Paper presented at the annual meeting of the American Educational Research Associations, New Orleans; LA*.



# Entropia relativa em misturas de regressões lineares

Susana Faria<sup>1</sup> · Gilda Soromenho<sup>2</sup>

© The Author(s) 2017

**Resumo** Os Modelos de Mistura são muito utilizados para modelar a heterogeneidade populacional. Medidas de entropia para analisar a heterogeneidade dos dados têm sido amplamente utilizadas. Nos modelos de mistura, critérios de entropia podem ser aplicados para estudar a sobreposição das componentes da mistura. Neste trabalho utilizámos uma medida de entropia baseada na distância de Kullback-Leibler para estudar a sobreposição das componentes de modelos de mistura de regressões lineares. Um estudo de simulação foi realizado com o objetivo de analisar a qualidade das estimativas dos parâmetros do modelo de mistura de regressões lineares em função da entropia dos modelos.

**Palavras-chave:** Distância de Kullback-Leibler, Entropia, Misturas de Regressões Lineares, Simulação.

## 1 Introdução

Atualmente, temos assistido a um crescente interesse pelos modelos de mistura finita uma vez que se têm revelado serem os modelos mais adequados em diversos campos de aplicações. Entre estes modelos, os modelos de mistura finita de regressões lineares têm sido amplamente estudados.

O interesse que se tem verificado pelos modelos de mistura de regressões lineares deve-se ao facto dos mesmos serem os mais adequados em aplicações onde a estimação de um único modelo de regressão não é eficiente. Estas aplicações

---

<sup>1</sup>Centro de Matemática, Departamento de Matemática e Aplicações, Universidade do Minho, [sfaria@math.uminho.pt](mailto:sfaria@math.uminho.pt)

<sup>2</sup>Instituto de Educação, Universidade de Lisboa, [gspereira@ie.ul.pt](mailto:gspereira@ie.ul.pt)

surtem, quando os dados são provenientes de uma população formada por vários grupos (aos quais se ajustam modelos de regressão com coeficientes distintos) e se desconhece quais as observações que pertencem a cada grupo.

O modelo de mistura de regressões lineares é dado por

$$y_i = \begin{cases} X_i^T \beta_1 + \varepsilon_{i1} & \text{com probabilidade} & \pi_1 \\ X_i^T \beta_2 + \varepsilon_{i2} & \text{com probabilidade} & \pi_2 \\ \vdots & & \\ X_i^T \beta_J + \varepsilon_{iJ} & \text{com probabilidade} & \pi_J \end{cases}$$

onde  $y_i, i = 1, \dots, n$ , é a variável resposta da  $i$ -ésima observação e  $X_i$  é o vetor de dimensão  $(p + 1)$  das variáveis explicativas da  $i$ -ésima observação,  $\beta_j = (\beta_{0j}, \beta_{1j}, \dots, \beta_{pj})^T$  é o vetor de dimensão  $(p + 1)$  dos coeficientes de regressão da  $j$ -ésima componente do modelo,  $\pi_j$  são as proporções de mistura com  $0 < \pi_j < 1, j = 1, \dots, J, \sum_{j=1}^J \pi_j = 1$  e  $\varepsilon_{ij}$  são os erros aleatórios com distribuição normal univariada de valor médio nulo e variância  $\sigma_j^2$ .

A função densidade de probabilidade da variável resposta condicional aos valores observados das variáveis explicativas é dado por

$$f(y|X) = \sum_{j=1}^J \pi_j f_j(y; X^T \beta_j, \sigma_j^2)$$

em que  $f_j(\cdot)$  é a função densidade de probabilidade da variável aleatória normal univariada de valor médio  $X^T \beta_j$  e variância  $\sigma_j^2$ .

O método da máxima verosimilhança recorrendo ao algoritmo *Expectation Maximization* (EM) (Dempster *et al.*, 1977) tem sido o método mais aplicado na estimação dos parâmetros de misturas de regressões lineares.

O objetivo deste trabalho é estudar o desempenho do algoritmo EM na estimação dos parâmetros de modelos de mistura de regressões lineares com duas componentes em diferentes configurações para as verdadeiras retas de regressão componentes da mistura e com diferentes níveis de sobreposição entre as componentes da mistura. Baseado na distância de Kullback-Leibler (Kullback-Leibler (1951)) é proposto um critério para avaliar a sobreposição das componentes do modelo de mistura de regressões lineares.

Segue-se a organização do artigo. Na Secção 2 descrevem-se algumas medidas de entropia e desenvolve-se uma medida de entropia para estudar a sobreposição das componentes de modelos de mistura de regressões lineares. Na Secção 3 é apresentado o estudo de simulação e os principais resultados. Finalmente, na Secção 4 são apresentadas as principais conclusões.

## 2 Medida de entropia

Em muitas situações, pretende-se estudar a distância entre as componentes de uma mistura. A distância de Kullback-Leibler, designada por entropia relativa, é uma das medidas mais usuais da distância entre duas distribuições,  $f(x)$  e  $g(x)$

$$KL(f: g) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

Esta distância não é simétrica e deve ser substituída pela distância simétrica de Kullback-Leibler (Fruhworth-Schnatter (2006)),

$$J(f: g) = KL(f: g) + KL(g: f) = \int f(x) \log \frac{f(x)}{g(x)} dx + \int g(x) \log \frac{g(x)}{f(x)} dx.$$

Algumas extensões da distância de Kullback-Leibler têm sido estudadas por alguns autores, como por exemplo, Barapour e Rad (2012) e Park *et al.* (2012). No contexto de modelos de mistura, Li e Wang (2010) propuseram um teste de homogeneidade baseado na distância de Kullback-Leibler. Leisch (2004) aplicou a distância de Kullback-Leibler para investigar a sobreposição das componentes de um modelo de mistura. Naik *et al.* (2007) desenvolveram um critério para selecionar componentes e variáveis em modelos de mistura de regressões, baseado na distância de Kullback-Leibler.

Considere-se um modelo de mistura de regressões lineares de duas componentes,

$$f(y|X) = \pi_1 f_1(y; X^T \beta_1, \sigma_1^2) + \pi_2 f_2(y; X^T \beta_2, \sigma_2^2).$$

Baseada na distância simétrica de Kullback-Leibler, propomos uma medida de entropia ( $EC$ ) para estudar a sobreposição de um modelo de mistura de regressões lineares de duas componentes,

$$\begin{aligned} EC(\pi_1 f_1: \pi_2 f_2) &= J(\pi_1 f_1: \pi_2 f_2) \\ &= KL(\pi_1 f_1: \pi_2 f_2) + KL(\pi_2 f_2: \pi_1 f_1) \\ &= 2\pi_1 \ln \left( \frac{\pi_1}{\pi_2} \right) + \ln \left( \frac{\pi_2}{\pi_1} \right) + \pi_1 KL(f_1: f_2) + \pi_2 KL(f_2: f_1), \end{aligned}$$

em que  $KL(f_i: f_j) = \frac{n}{2} \left( \ln \frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} \right) + \frac{1}{2\sigma_j^2} (X^T \beta_i - X^T \beta_j)^2 - \frac{n}{2}$ , é a distância de Kullback-Leibler entre duas distribuições de probabilidade normais.

A medida de entropia proposta tem como principal vantagem o facto de ter em consideração as proporções do modelo de mistura.



### 3 Estudo de simulação

De seguida, apresentamos um estudo de simulação para investigar o desempenho do algoritmo EM no cálculo das estimativas de máxima verosimilhança dos parâmetros de misturas de regressões lineares de duas componentes com diferentes configurações para as verdadeiras retas de regressão componentes da mistura e com diferentes níveis de sobreposição entre as componentes da mistura. Para avaliar a sobreposição entre as duas componentes aplicamos a medida de entropia *EC* proposta na Secção 2.

O estudo de simulação foi realizado utilizando o *software* R, versão 3.1.0 (R Development Core Team, 2014).

Considere-se os modelos de mistura de regressões lineares,

$$\text{- modelo 1: } f(y|X) = \pi_1 f_1(y; X^T \beta_{f1}, \sigma_{f1}^2) + \pi_2 f_2(y; X^T \beta_{f2}, \sigma_{f2}^2)$$

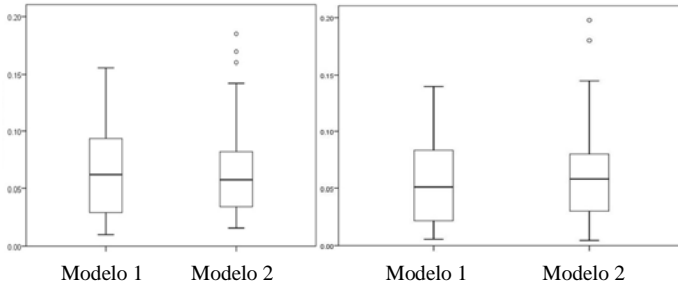
$$\text{- modelo 2: } g(y|X) = \pi_1 g_1(y; X^T \beta_{g1}, \sigma_{g1}^2) + \pi_2 g_2(y; X^T \beta_{g2}, \sigma_{g2}^2).$$

Gerámos amostras de dimensões  $n=50$ ,  $n=100$  e  $n=500$  de cada modelo de mistura variando os valores dos parâmetros de acordo com as situações descritas na Tabela 1 (realça-se que não foi considerado o caso de iguais proporções das misturas). O valor da variância ( $\sigma_{g2}^2$ ) foi escolhido de modo que os dois modelos de mistura (modelo 1 e modelo 2) tivessem o mesmo valor de *EC*. Para cada dimensão de amostra e para cada conjunto de valores dos parâmetros, gerámos 200 amostras. Considerámos duas diferentes configurações para as verdadeiras retas de regressão componentes da mistura: as retas de regressão eram paralelas ou concorrentes entre si. Foram utilizados 20 valores aleatórios como valores iniciais do algoritmo EM e a solução com o maior valor da função de log-verosimilhança foi a usada. O algoritmo EM terminava quando a diferença relativa da função de log-verosimilhança entre duas iterações consecutivas fosse menor que  $10^{-10}$ . Em cada uma das amostras geradas, calculámos a distância de *Mahalanobis* entre os valores verdadeiros dos parâmetros do modelo e os valores estimados dos parâmetros.

Resumidamente, o processo de simulação consiste nos seguintes passos:

- Gerar uma amostra de dimensão  $n$  do modelo de mistura 1. Estimar um modelo de mistura de regressões lineares usando o algoritmo EM e guardar as estimativas dos parâmetros desconhecidos do modelo e calcular a estimativa *EC*;
- Determinar ( $\sigma_{g2}^2$ ) de modo que os dois modelos de mistura (modelo de mistura 1 e modelo de mistura 2) tenham o mesmo valor estimado *EC*;
- Gerar uma amostra de dimensão  $n$  do modelo de mistura 2. Estimar um modelo de mistura de regressões lineares usando o algoritmo EM e guardar as estimativas dos parâmetros desconhecidos do modelo;
- Calcular a distância de *Mahalanobis* entre os valores verdadeiros e os valores estimados dos parâmetros no modelo de mistura 1 e no modelo de mistura 2;
- Aplicar o teste de Mann-Whitney para avaliar se existem diferenças significativas entre as distâncias de *Mahalanobis* dos dois modelos de mistura.

Na Figura 1 apresentam-se os *boxplots* da distância de *Mahalanobis* entre os valores verdadeiros dos parâmetros do modelo e os valores estimados dos parâmetros para dois dos casos estudados. Estes gráficos sugerem que não existem diferenças significativas nas distâncias de *Mahalanobis* entre os dois modelos de mistura.



**Figura 1** - *Boxplot* da distância de *Mahalanobis* entre os valores verdadeiros e os valores estimados dos parâmetros (Caso II e Caso VIII).

Na Tabela 1 mostra-se os valores estimados de *EC* e o valor-p do teste de Mann-Whitney para tamanho de amostra  $n=100$ . Resultados semelhantes foram obtidos para a dimensão da amostra  $n=50$  e  $n=500$ .

Observando a Tabela 1 podemos afirmar que:

- não há diferenças significativas na precisão das estimativas dos parâmetros entre misturas de regressões lineares com componentes paralelas e com a mesma medida *EC* (casos IV, V, VI, XIII, XIV e XV);
- não há diferenças significativas na precisão das estimativas dos parâmetros entre misturas de regressões lineares com componentes concorrentes e com a mesma distância *EC* (casos I, II, III, X, XI e XII);
- não há diferenças significativas na precisão das estimativas dos parâmetros entre misturas de regressões lineares com componentes concorrentes e misturas com componentes paralelas com a mesma distância *EC* (casos VII, VIII, IX, e XVI, XVII, XVIII).

**Tabela 1** – Valores verdadeiros dos parâmetros, valores estimados de EC e valor-p do teste Mann-Whitney para  $n=100$ .

Caso	Configuração	Modelo 1							Modelo 2							valor-p
		$\beta_{01}$	$\beta_{11}$	$\beta_{02}$	$\beta_{12}$	$\sigma_{f1}^2$	$\sigma_{f2}^2$	$\widehat{EC}$	$\pi_1$	$\beta_{01}$	$\beta_{11}$	$\beta_{02}$	$\beta_{12}$	$\sigma_{g1}^2$	$\sigma_{g2}^2$	
I	Concorrentes/Concorrentes	2	1	4	-1	0.35	0.5	824.2	0.2	0	1	2	-1	0.3491	0.5	0.352
II	Concorrentes/Concorrentes	2	1	4	-1	0.35	0.5	708.3	0.4	0	1	2	-1	0.3489	0.5	0.859
III	Concorrentes/Concorrentes	2	1	4	-1	0.35	0.5	655.9	0.8	0	1	2	-1	0.3462	0.5	0.231
IV	Paralelas/Paralelas	1	1	3	1	0.35	0.5	552.89	0.2	3	1	5	1	0.3488	0.5	0.588
V	Paralelas/Paralelas	1	1	3	1	0.35	0.5	516.32	0.4	3	1	5	1	0.3498	0.5	0.916
VI	Paralelas/Paralelas	1	1	3	1	0.35	0.5	450.03	0.8	3	1	5	1	0.3456	0.5	0.470
VII	Paralelas/Concorrentes	2	1	4	-1	0.35	0.5	721.84	0.2	3	1	5	1	0.2556	0.5	0.289
VIII	Paralelas/Concorrentes	2	1	4	-1	0.35	0.5	676.07	0.4	3	1	5	1	0.2505	0.5	0.468
IX	Paralelas/Concorrentes	2	1	4	-1	0.35	0.5	586.91	0.8	3	1	5	1	0.1902	0.5	0.269
X	Concorrentes/Concorrentes	0	1	2	-1	0.15	0.35	1596.49	0.2	4	1	6	-1	0.1497	0.5	0.451
XI	Concorrentes/Concorrentes	0	1	2	-1	0.15	0.35	1392.23	0.4	4	1	6	-1	0.1423	0.5	0.906
XII	Concorrentes/Concorrentes	0	1	2	-1	0.15	0.35	994.30	0.8	4	1	6	-1	0.1059	0.5	0.589
XIII	Paralelas/Paralelas	1	1	4	1	0.3	0.35	1465.49	0.2	3	1	6	1	0.2834	0.5	0.678
XIV	Paralelas/Paralelas	1	1	4	1	0.3	0.35	1430.14	0.4	3	1	6	1	0.2587	0.5	0.965
XV	Paralelas/Paralelas	1	1	4	1	0.3	0.35	1343.74	0.8	3	1	6	1	0.1599	0.5	0.559
XVI	Paralelas/Concorrentes	1	1	4	1	0.3	0.35	1458.79	0.2	2	1	4	-1	0.1667	0.5	0.089
XVII	Paralelas/Concorrentes	1	1	4	1	0.3	0.35	1418.81	0.4	2	1	4	-1	0.1376	0.5	0.309
XVIII	Paralelas/Concorrentes	1	1	4	1	0.3	0.35	1331.42	0.8	2	1	4	-1	0.0645	0.5	0.150

#### 4. Conclusão

Neste trabalho definimos uma medida de entropia para avaliar o grau de sobreposição das componentes de modelos de mistura de regressões lineares com duas componentes. Usando esta medida, investigamos o desempenho do algoritmo EM no cálculo das estimativas de máxima verosimilhança dos parâmetros de misturas de regressões lineares com duas componentes com diferentes configurações para as verdadeiras retas de regressão das componentes e com diferentes níveis de sobreposição entre as componentes da mistura. No nosso estudo de simulação concluímos que, quando a sobreposição das componentes do modelo de mistura de regressões lineares é a mesma, a posição relativa das componentes não influencia a precisão das estimativas dos parâmetros.

## Agradecimentos

S. Faria foi financiada pelo Centro de Matemática da Univ. do Minho com Fundos Nacionais através da FCT no âmbito do projeto PEst-OE/MAT/UI0013/2014. Os autores agradecem ainda os comentários dos revisores do artigo.

## Referências

- BARAPOUR, S. & RAD, A. H. (2012). Testing goodness-of-fit for exponential distribution based on cumulative residual entropy. *Communications in Statistics-Theory and Methods*, 41(8), 1387–1396.
- DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. B*, 39, 1-38.
- FRUHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, New York.
- KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.
- LEISCH, F. (2004). Exploring the structure of Mixture Model Components, In J. Antoch (Ed.), *COMPSTAT 2004. Proceedings in Computational Statistics*, 1405-1412. Heidelberg:Physica-Verlag/Springer.
- LI, Y. & WANG, L. (2010). Testing for Homogeneity in Mixture Using Weighted Relative Entropy, *Communications in Statistics – Simulation and Computation*, 37, 1981-1995.
- NAIK, P.A., SHI, P. & TSAI, C.L. (2007). Extending the Akaike information criterion to mixture regression models. *J. Amer. Statist. Assoc.*, 102, 244-254.
- PARK, S., RAO, M. & SHIN, D.W. (2012). On cumulative residual Kullback-Leibler information. *Statistics and Probability Letters*, 82, 2025–2032.



# Vítimas mortais em acidentes de viação em Portugal continental: componentes principais e previsão

Fernando Sebastião<sup>1</sup> · Irene Oliveira<sup>2</sup>

© The Author(s) 2017

**Resumo** Uma das preocupações da segurança rodoviária é a de reduzir os níveis de sinistralidade e em particular a mortalidade resultante dos acidentes de viação. Este trabalho tem como objetivo explorar os dados mensais do número de vítimas mortais em acidentes de viação ocorridos em Portugal Continental, entre os anos de 1998 e 2009, através da Análise Espectral Singular. Pretende-se extrair e interpretar as componentes dominantes e períodos associados, e ilustrar como o método pode ser utilizado na previsão da mortalidade.

**Palavras-chave:** Acidentes de Viação, Análise Espectral Singular, Componentes Principais, Mortalidade Rodoviária, Séries Temporais.

## 1 Introdução

As políticas de segurança rodoviária visam sobretudo contribuir para que a sinistralidade rodoviária em Portugal tenha cada vez menos impacto na sociedade. Analisar o comportamento do número de vítimas mortais que resultam dos acidentes de viação permite uma melhor perceção da influência das políticas adotadas nos últimos anos, e pode contribuir para um planeamento de novas políticas de sensibilização, de prevenção e de aplicação da legislação.

O assunto continua a suscitar uma grande motivação de estudo, uma vez que no ano de 2010, a Organização das Nações Unidas (ONU) veio proclamar o período entre 2011 e 2020 como a Década de Ação para a Segurança Rodoviária com o objetivo de reduzir os acidentes de viação a nível mundial.

---

<sup>1</sup>Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, [fsebast@ipleiria.pt](mailto:fsebast@ipleiria.pt)

<sup>2</sup>Universidade de Trás-os-Montes e Alto Douro, CITAB-UTAD, UTAD, [ioliveir@utad.pt](mailto:ioliveir@utad.pt)

Neste trabalho procedemos a uma breve descrição da técnica da Análise Espectral Singular (AES) e apresentamos o estudo da série mensal do número de vítimas mortais em acidentes de viação que ocorreram em Portugal Continental entre 1998 e 2009. Em particular, são analisadas as componentes essenciais para a reconstrução dos dados originais e para aplicação posterior na previsão da mortalidade rodoviária mensal.

## 2 Análise espectral singular

A AES é encarada como um método particular de aplicação da Análise em Componentes Principais, ACP, cujas variáveis iniciais são versões desfasadas da série temporal unidimensional  $X_t = (x_1, \dots, x_n)$ .

Segundo Elsner e Tsonis (1996) e Golyandina *et al.* (2001), o método básico da AES é constituído por duas etapas: a decomposição e a reconstrução.

A primeira etapa contempla a obtenção da matriz dos desfasamentos ao longo do tempo (ou matriz de trajetória),  $\mathbf{X} = [\mathbf{X}_1 : \dots : \mathbf{X}_k]^T$ , que é constituída por  $k$  vetores desfasados  $\mathbf{X}_j = [x_j : \dots : x_{j+m-1}]^T$ , com  $j = 1, \dots, k$ , para um determinado tamanho do vetor desfasado ou comprimento da janela,  $m$ . A decomposição da série temporal na soma de várias componentes é efetuada recorrendo à decomposição em valores singulares da matriz  $\mathbf{X}$ , de forma a obterem-se os  $m$  valores próprios  $\lambda_1 \geq \dots \geq \lambda_m \geq 0$  e os respetivos vetores próprios ortonormados  $U_1, \dots, U_m$  da matriz  $\mathbf{X}^T \mathbf{X}$ . As componentes principais (CPs) são obtidas através de  $V_i = \mathbf{X} U_i / \sqrt{\lambda_i}$ , com  $i = 1, \dots, d$ , onde  $d$  é a característica de  $\mathbf{X}$ .

A segunda etapa tem como finalidade reconstruir a série original através da soma das componentes retidas como principais, eliminando aquelas que são essencialmente constituídas por ruído, de modo a que essa nova série reconstruída possa ser usada na previsão de novos valores. Um dos métodos de previsão, aplicado em muitos dos casos, é o algoritmo de previsão recorrente descrito detalhadamente em Golyandina *et al.* (2001, Capítulo 2).

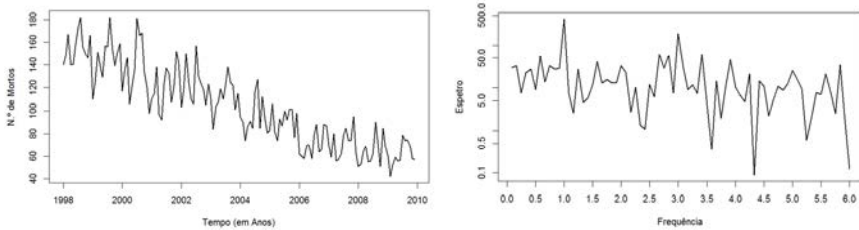
## 3 Caso de estudo

### 3.1 Breve comentário sobre os dados

Neste estudo é considerada a série mensal do número de vítimas mortais em acidentes de viação em Portugal Continental entre 1998 e 2009.

Os dados da Figura 1 recolhidos ao longo dos 12 anos em estudo ( $n = 144$  meses) foram provenientes da Direção Geral de Viação (DGV) que exerceu funções até 2006, dando origem à atual Autoridade Nacional de Segurança

Rodoviária (ANSR). A entidade responsável pelo tratamento e divulgação dos dados é o Instituto Nacional de Estatística (INE), segundo o qual, a definição de vítima de acidente, utilizada neste período em estudo, é aquela cujo óbito ocorra no local do evento ou no seu percurso até à unidade de saúde.



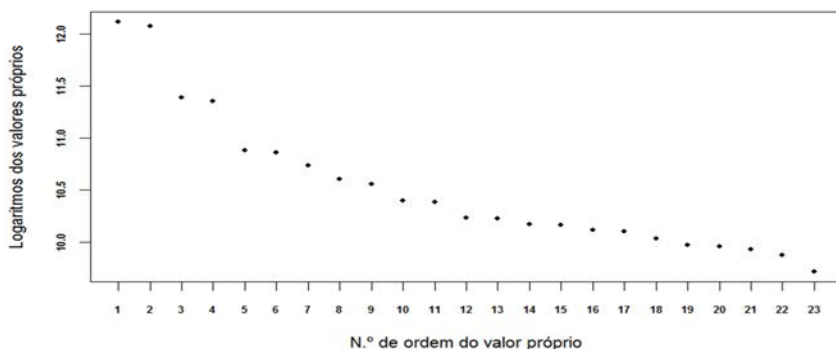
**Figura 1** – Número mensal de vítimas mortais em acidentes de viação em Portugal Continental entre 1998 e 2009 e respetivo periodograma.

### 3.2 Decomposição

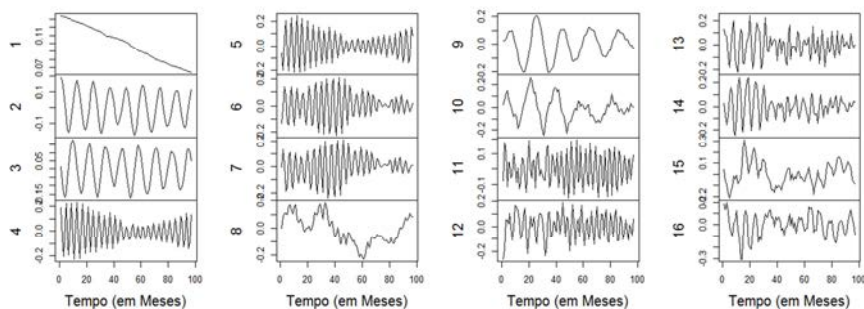
Uma questão importante na aplicação da AES é a decisão da escolha do comprimento da janela,  $m$ . Tendo em conta a especificidade do problema em análise e alguma informação prévia da natureza da série temporal, utilizou-se o comprimento da janela  $m = n/3 = 48$  meses, uma vez que era expetável que a série tivesse componentes sazonais de período inteiro, pelo que era aconselhável usar um comprimento da janela proporcional a esse período para obtermos uma melhor separação das componentes periódicas (Golyandina *et al.*, 2001; Hassani, 2007). Consequentemente, a matriz de trajetória é constituída por  $k = n - m + 1 = 97$  vetores desfasados correspondentes ao número de linhas e por  $m = 48$  colunas.

Com base na matriz de trajetória  $\mathbf{X}$  e através da decomposição em valores singulares, selecionaram-se as primeiras 16, de 48 componentes principais, com base nos respetivos valores próprios de  $\mathbf{X}^T\mathbf{X}$  ilustrados na Figura 2, associados às CPs apresentadas na Figura 3. Desta forma, as restantes componentes foram desprezadas por se admitir que continham a parte correspondente ao ruído. O primeiro valor próprio (não representado na Figura 2) é muito elevado em relação aos restantes, o que é natural uma vez que estamos a usar o modelo básico da AES sem qualquer tipo de centragem de dados, e os valores próprios obtêm-se da matriz  $\mathbf{X}^T\mathbf{X}$  em vez da matriz de variâncias covariâncias.





**Figura 2** – Logaritmos dos 24 primeiros valores próprios exceto o primeiro.



**Figura 3** – Primeiras 16 CPs com  $k = 97$  coordenadas.

Às 16 CPs que ficaram retidas associaram-se sete pares de componentes oscilatórias (CP2-CP3, CP4-CP5, CP6-CP7, CP9-CP10, CP11-CP12, CP13-CP14, CP15-CP16), por ostentarem um comportamento oscilatório semelhante assim como valores próprios muito similares, enquanto a CP1 corresponde à tendência decrescente do número de vítimas mortais (de 1998 para 2009 ocorreu uma redução de 60% no valor total anual) e a CP8 exibe uma mistura de tendência com componente oscilatória.

O total da percentagem de variância explicada pelas 16 primeiras CPs é 99.307%. A CP1 associada à tendência dos dados apresenta uma percentagem de variância muitíssimo elevada (97.569%) comparativamente com as restantes CPs. Podemos argumentar que tal ocorrência deve-se a uma nítida e predominante tendência decrescente do número de vítimas mortais em acidentes de viação ao longo dos anos. Situações similares a esta, em que a tendência se destaca

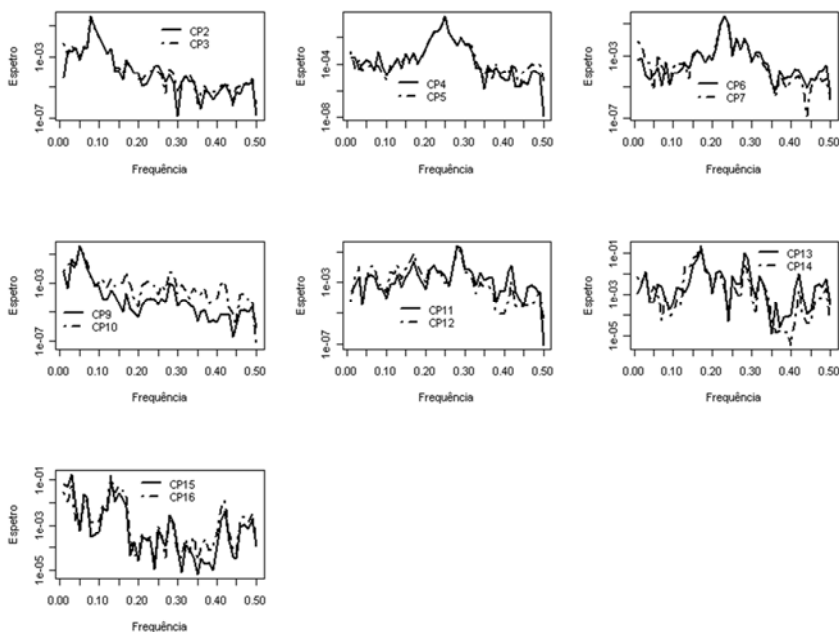
fortemente, ocorrem com alguma frequência em séries de valores relacionados com a produção industrial (Hassani *et al.*, 2009; Golyandina *et al.*, 2001, Capítulo 1).

### 3.3 Análise das periodicidades das CPs retidas

A Análise de Fourier considerada como um método tradicional de estimação espectral, pode ser utilizada para caraterizar picos espectrais em registos temporais que correspondem a oscilações regulares das funções trigonométricas seno e cosseno (Murteira *et al.*, 1993; von Storch e Zwiers, 1999, Capítulos 11 e 12).

Quando se analisa um espectro e, na presença de uma oscilação forte, se destaca um par de valores próprios consecutivos por serem quase iguais, então os vetores próprios e as CPs associados também surgem relacionados entre si aos pares para um mesmo período (Vautard e Ghil, 1989; Vautard *et al.*, 1992).

Para auxiliar na identificação dos períodos dos pares de componentes que são geradas por uma harmónica, recorreremos à visualização gráfica dos espectros (Figura 4) dos sete pares de CPs oscilatórias retidas, donde se concluiu que os períodos dominantes dos respetivos pares indicados são aproximadamente iguais a 12, 4, 4, 20, 4, 6 e 33 meses respetivamente.

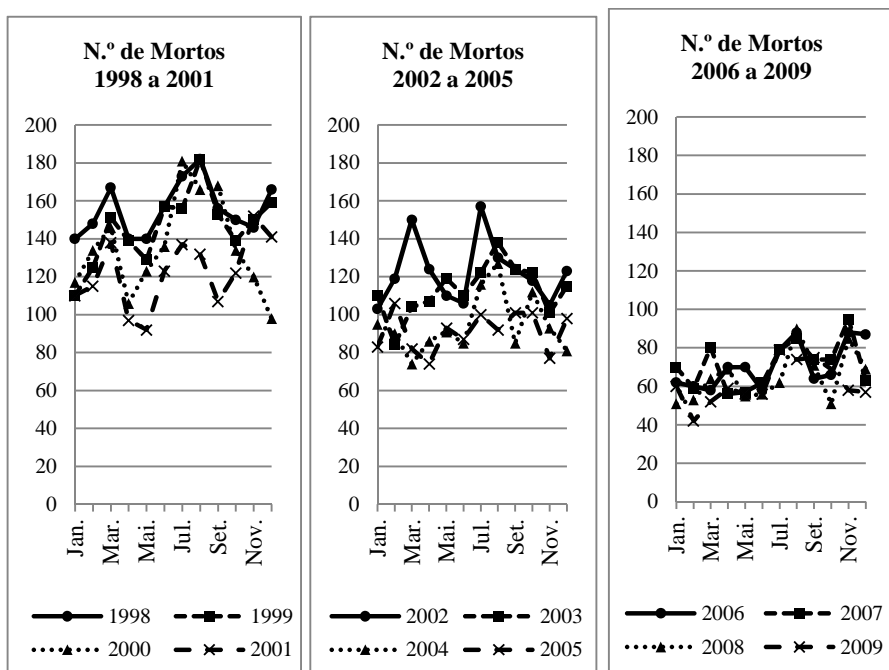


**Figura 4** – Espectros dos sete pares de CPs.

Note-se que embora em Vautard *et al.* (1992) se refira que o comprimento da janela deverá permitir analisar com sucesso os períodos pertencentes ao intervalo  $(m/5, m)$ , e que neste caso corresponde a (9.6, 48), tal recomendação não invalida que se possam considerar também os períodos de 4 e 6 meses explícitos em alguns pares de CPs.

Para complementar as análises anteriores e testar periodicidades escondidas de uma determinada frequência, consideremos a aplicação do teste de hipóteses de Fisher  $g$  (Brockwell e Davis, 1991, Capítulo 2; de Carvalho e Rua, 2014), sob a hipótese nula de que um pico espectral não é significativo contra a alternativa de que existe uma componente periódica. Todas as CPs dos sete pares retidos revelaram a existência de uma componente periódica estatisticamente significativa, para qualquer nível de significância usual.

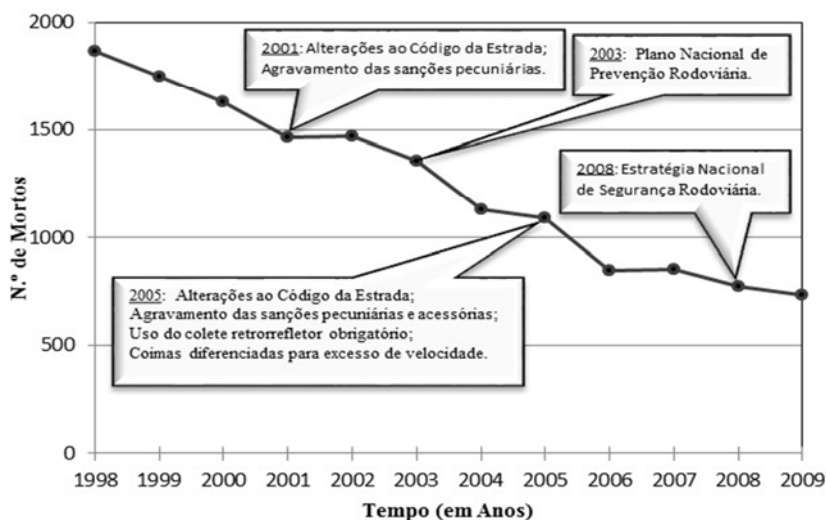
Tendo em conta a natureza da série em estudo, é possível efetuar uma interpretação dos ciclos temporais mais relevantes no contexto da sinistralidade rodoviária. Para além da compreensível sazonalidade anual, destaca-se o ciclo de 4 meses, uma vez que encontramos 3 principais picos de frequência em que o número de vítimas mortais é elevado, como podemos constatar nos cronogramas ao longo de cada ano na Figura 5.



**Figura 5** – Cronogramas do número de vítimas mortais mensais entre os anos de 1998 e 2009.

Estes 3 picos correspondem aos períodos de maior tráfego rodoviário em que muitas famílias gozam férias: férias escolares da Páscoa (março a maio), férias de verão (julho e agosto) e época natalícia. No período natalício, a falta de cuidados de alguns condutores na condução em condições climáticas adversas provoca um aumento de acidentes e naturalmente o número de mortos tem tendência a aumentar. Contudo há que ter em consideração que os dados em causa correspondem à totalidade dos mortos registados em todo o território nacional continental e não desagregados ao nível dos distritos, nos quais se verificam geralmente discrepâncias significativas entre as condições meteorológicas.

Há que realçar ainda o período com cerca de 33 meses (2 anos e 9 meses aproximadamente) na análise dos dados, o qual leva a pressupor uma relação com as alterações à legislação em vigor e com as campanhas de sensibilização introduzidas essencialmente na última década ilustradas na Figura 6.



**Figura 6** – Evolução anual do número de vítimas mortais e medidas de segurança rodoviária.

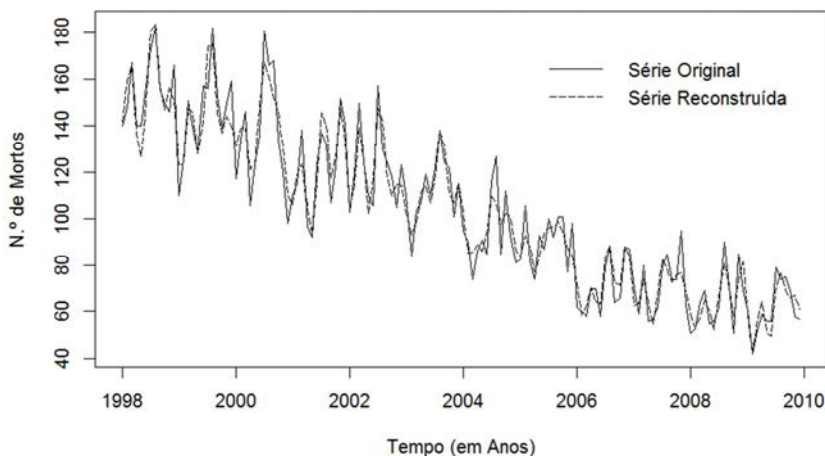
Simão (2010) descreve que entre os fatores mais significativos que têm contribuído para o decréscimo dos índices de sinistralidade em Portugal, onde se inclui a mortalidade rodoviária, destacam-se as sucessivas alterações à legislação em vigor e a eficiência na sua fiscalização.

Por outro lado, quando se compara Portugal com os países da União Europeia a 27 Estados Membros, este sobressai-se por ser um dos países com melhores resultados na diminuição dos valores de sinistralidade rodoviária, onde verificamos uma subida do 25.º lugar em 1998 para o 15.º lugar em 2009 dos países com menor número total de vítimas mortais por milhão de habitantes.

### 3.4 Reconstrução da série original do número de vítimas mortais e previsão

A Figura 7 mostra a reconstrução da série original à custa das primeiras 16 CPs retidas onde podemos constatar que a qualidade da reconstrução é bastante satisfatória.

As primeiras 16 CPs identificadas anteriormente foram utilizadas na reconstrução da série original sem ruído e permitiram aplicar o algoritmo de previsão recorrente em AES (Golyandina *et al.*, 2001, Capítulo 2) com o intuito de prever novos valores da série. Utilizámos o *software* de livre acesso *R*, no qual implementámos todos os passos da técnica assim como uma rotina para o algoritmo de previsão recorrente, cujas previsões foram também confirmadas no *software* *CaterpillarSSA*.

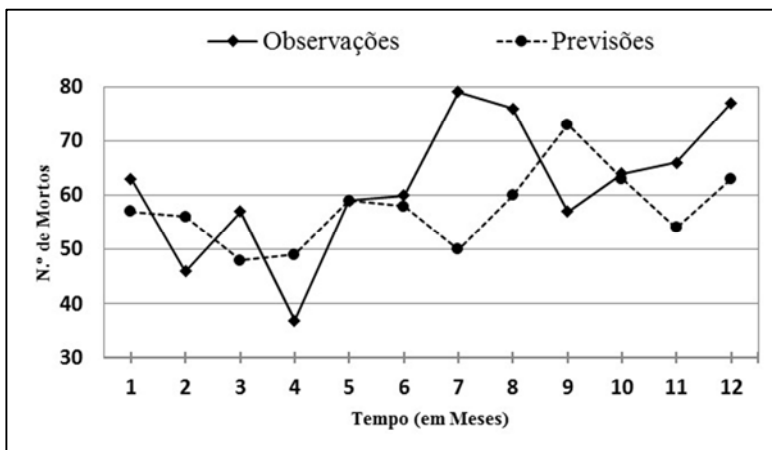


**Figura 7** – Reconstrução da série original com as 16 CPs retidas.

Constatamos que os valores previstos do número de vítimas mortais para os meses do ano de 2010 variam aproximadamente entre os 48 e os 73, os quais acompanham a tendência de uma certa estabilização suave que já era notória nos últimos anos desde 2006. Uma vez que atualmente estão divulgados os dados oficiais do número de vítimas mortais em acidentes de viação ocorridos no ano de 2010, podemos efetuar a comparação dos 12 valores mensais previstos para o ano de 2010 com os valores observados (Figura 8).

Verificamos que em julho de 2010, a diferença entre predição e observação é significativa e de aproximadamente 30 mortes, o que poderá estar a influenciar as

previsões posteriores e por isso admite-se a necessidade de um eventual reajuste da série.



**Figura 8** – Comparação do número de vítimas mortais previstas *versus* observadas para os 12 meses do ano de 2010.

Como medidas de precisão das previsões, verificamos que a raiz do erro quadrático médio é igual a 13.05 e que a precisão direcional média (percentagem de vezes em que a previsão e a série temporal variam no mesmo sentido entre dois momentos temporais consecutivos) é apenas cerca de 27.27%, o que nos leva a considerar que o ajustamento é fraco.

## 4 Discussão

Este trabalho pretende ilustrar como os métodos da AES permitem extrair as componentes oscilatórias e a tendência numa série temporal de dados relacionados com a sinistralidade rodoviária.

Para além do usual ciclo anual presente nos dados, detetou-se o ciclo de 4 meses cujos picos evidenciam uma relação com as férias escolares dos portugueses, nas quais aumenta a circulação automóvel, registando-se tendencialmente um maior número de acidentes envolvendo vítimas mortais. Também o período presente de 33 meses pressupõe uma ligação às alterações à legislação em vigor e às campanhas de sensibilização introduzidas na última década. Os valores obtidos resultantes da previsão mostram como a AES tem potencial para desempenhar um papel fundamental neste campo, nomeadamente no planeamento atempado para a introdução de novas políticas que se julguem mais eficazes no combate à mortalidade rodoviária.

## Referências

- BROCKWELL, P. & DAVIS, R. (1991). *Time Series: Theory and Methods*, 2nd Ed. Springer, New York.
- de CARVALHO, M. & RUA, A. (2014). Nowcasting the US Business Cycle: Singular Spectrum Analysis at Work. Working Paper 16, Banco de Portugal.
- ELSNER, J. B. & TSONIS, A. A. (1996). *Singular Spectrum Analysis. A New Tool in Time Series Analysis*, Plenum Press, New York.
- GOLYANDINA, N. E., NEKRUKTIN, V. V. & ZHIGLJAVSKY, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman & Hall, Boca Raton.
- HASSANI, H. (2007). Singular Spectrum Analysis: Methodology and Comparison, *Journal of Data Science*, 5, 239-257.
- HASSANI, H., HERAVI, S. & ZHIGLJAVSKY, A. (2009). Forecasting European Industrial Production with Singular Spectrum Analysis, *International Journal of Forecasting*, 25, 103-118.
- MURTEIRA, B., MÜLLER, D. & TURKMAN, K. (1993). *Análise de Sucessões Cronológicas*, McGraw-Hill, Lisboa.
- SIMÃO, N. (2010). *Análise Agregada da Eficácia das Políticas de Segurança Rodoviária em Portugal*, Dissertação de Mestrado, Instituto Superior Técnico da Universidade Técnica de Lisboa.
- VAUTARD, R. & GHIL, M. (1989). Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D*, 35, 395-424.
- VAUTARD, R., YIOU, P. & GHIL, M. (1992). Singular Spectrum Analysis: A Toolkit for Short, Noisy Chaotic Signals, *Physica D*, 58, 95-126.
- von STORCH, H. & ZWIERS, F. W. (1999). *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge.

## Seleção de atributos valorativos da habitação: uma aplicação ao mercado habitacional de Aveiro e Ílhavo

Paulo Batista<sup>1</sup> · Gladys Castillo<sup>2</sup> · João Marques<sup>3</sup> · Eduardo Castro<sup>4</sup>

© The Author(s) 2017

**Resumo** A construção de modelos de preços hedónicos, para a sua utilização em contextos de apoio à decisão, envolve, atualmente, a necessidade de manipular uma quantidade elevada de variáveis.

Este trabalho descreve uma abordagem *data mining* com o objetivo de testar a *performance* de um conjunto de técnicas de seleção de atributos, na definição de um conjunto representativo das múltiplas dimensões analíticas que descrevem o valor da habitação.

O trabalho permitiu identificar um conjunto reduzido de atributos, coerente com a teoria subjacente e mantendo uma aceitável capacidade explicativa para os objetivos analíticos propostos.

**Palavras-chave:** *Data mining*, Econometria, Mercado de habitação, Valor de mercado.

### 1 Introdução

A habitação é um elemento de grande relevância para os indivíduos, representando a sua posse o mais importante e valioso bem económico. Neste contexto, a identificação dos determinantes do seu valor constitui um elemento informativo fundamental para apoiar vários processos da organização social (Jones & Watkins, 2009).

Não obstante as diferentes correntes económicas inerentes à definição e determinação do valor (monetário) do bem habitação, a construção de modelos de

---

<sup>1</sup>Universidade de Aveiro, pauloricardolb@ua.pt

<sup>2</sup>Universidade de Aveiro, gladys@ua.pt

<sup>3</sup>Universidade de Aveiro, jjmarques@ua.pt

<sup>4</sup>Universidade de Aveiro, ecastro@ua.pt



preço de transação da habitação é, predominantemente, assente na teoria microeconómica neoclássica. Neste contexto, a teoria e técnica clássica utilizada é designada por modelo de preços hedónicos (Rosen, 1974), o qual se baseia na decomposição do valor da habitação num conjunto de atributos, mensuráveis, que a caracterizam. Não obstante a informação sobre o conjunto de atributos que descrevem a habitação constituir um aspecto crucial na modelação, a escassez de fontes de dados constitui um dos mais importantes desafios (Eurostat, 2013).

Os recentes sistemas de informação digitais vieram dar resposta ao problema de disponibilidade de informação, proporcionando um crescimento contínuo do volume e variedade de dados recolhidos. Assim, um dos emergentes desafios enfrentados na modelação do valor da habitação é a necessidade de construir procedimentos, eficientes, que permitam selecionar a informação útil dos cada vez maiores bancos de dados. Neste contexto a tarefa de seleção de atributos é consideravelmente complexa e, tal como defendem Caruana & Freitag (1994), as abordagens tradicionais (por exemplo, a seleção manual) tornam-se frequentemente ineficientes.

Metodologias de descoberta de conhecimento em bases de dados (KDD, na sigla inglesa) surgem como uma resposta a estes desafios. Fayyad (1996) define estas abordagens como guias metodológicos, de aplicação multidisciplinar. Estas facilitam a utilização e compreensão de um conjunto de diferentes algoritmos, com o intuito de desenvolver processos, não triviais, de extração de informação, útil e compreensível, previamente desconhecida, a partir de grandes volumes de dados digitalizados, não necessariamente recolhidos com o objetivo de análise e modelação pré-estabelecidos.

## 1.1 Objetivos

A seleção de atributos para a construção de modelos de preços hedónicos do valor da habitação destaca-se pelas exigências causadas pela complexidade do fenómeno e inerente multidisciplinaridade do processo. Num nível mais concreto, as teorias socioeconómicas subjacente destacam como um desafio importante a abstração concetual associada à mensuração de espaço / vizinhança / território (Marques (2012)) da habitação. Com efeito, este é um conceito difícil de traduzir, *à priori*, por um conjunto restrito de indicadores, uma vez que envolve elementos: geográficos (distâncias, ...), urbanísticos (morfologias físicas), sociais (vizinhança social), económicos (por exemplo, no que respeita à substituíbilidade e submercados espaciais) e estatísticos (como devem ser mensurados corretamente os atributos face às técnicas de modelação do preço a adotar e seus requisitos). Neste contexto o objetivo específico deste trabalho passa pelo desenvolvimento de uma abordagem analítica que permita coadjuvar o procedimento de seleção de atributos, com especial ênfase na seleção de atributos espaciais (territoriais), em contextos de modelação dos preços da habitação. Adotou-se como estudo de caso o mercado habitacional de venda, nos municípios de Aveiro e Ílhavo.

## 2 Metodologia

### 2.1 Visão geral

*Data mining* é uma expressão que tem vindo a adquirir um significado lato, relacionado com a aplicação de um conjunto alargado de algoritmos, em processos de análise de dados, que decorrem em ambientes de grande complexidade analítica (Tan, Steinbach, & Kumar, 2006). Nomeadamente, quando: i) estão envolvidas quantidades significativas de registos e variáveis, ii) os dados foram recolhidos indiretamente (reutilização de dados recolhidos para objetivos distintos daqueles que estão subjacentes ao processo analítico em construção) e ainda iii) quando a construção de modelos analíticos se reveste de especial complexidade ou ambiguidade. Numa definição mais concetual, tal como referido por Fayyad (1996), um conjunto alargado de algoritmos tem sido disponibilizado de forma acessível a analistas com diferentes bases científicas, enquadrados por guias metodológicos, que permitem enquadrar a sua utilização em contextos onde os analistas têm, por norma, um menor domínio científico das soluções técnicas (estatísticas e computacionais) eventualmente disponíveis. São exemplos destas propostas o CRISP-DM e o SEMMA (Azevedo, 2008) – sendo que o trabalho aqui apresentado enquadra-se na adoção de uma filosofia mais próxima da primeira.

A par da padronização de procedimentos atrás referida, registe-se a oferta de *software* que disponibiliza uma interface versátil e intuitiva para algoritmos previamente implementados, não exigindo assim capacidades técnicas muito avançadas aos analistas. Neste trabalho caso recorreu-se ao *software RapidMiner 5.0* e ao *RStudio 0.98*, com *R versão 2.5*, *package BMS versão 0.3.3* e *package GLMNET versão 2.0*; estas soluções permitem testar, de forma rápida, eficiente e versátil um número significativo de estratégias de seleção de atributos.

Para responder aos objetivos deste trabalho, a metodologia aqui descrita envolve a análise comparativa de um conjunto de algoritmos de seleção de atributos. A escolha recaiu em algoritmos de baixa complexidade, permitindo uma maior segurança do analista face aos requisitos técnicos necessários para a sua utilização transdisciplinar (secção 2.3). O objetivo final é a especificação do modelo econométrico (de preços hedónicos da habitação) (secção 2.2), procedimento que se irá basear numa análise comparativa dos resultados dos diferentes algoritmos implementados (secção 2.4), incluindo: a sua *performance*, coerência do modelo com a teoria subjacente (análise de cariz qualitativo) e simplicidade (baseada essencialmente no número de variáveis).

## 2.2 Determinação do valor da habitação

Os modelos de preços hedónicos, enquanto técnica que permite controlar a natureza heterogénea da habitação (Bourassa, Hoesli, & Sun, 2006), assentam na definição de bem compósito (Lancaster, 1966) e são usualmente modelados com base na regressão do preço de transação (variável dependente) com as características, previamente identificáveis pelo analista, que a descrevem (variáveis independentes). A abordagem econométrica procura obter estimativas do valor marginal de cada um dos atributos, descrevendo, em termos valorativos, os elementos diferenciadores de uma habitação. Desta forma, o modelo final permite posteriormente estimar o valor de mercado de uma dada habitação, para um período temporal (elemento não abordado neste trabalho) prévio à sua transação no mercado. O modelo de preços hedónicos é assim definido:

$$P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (1)$$

onde  $P$  é o preço da habitação,  $X$  é o vetor de atributos  $X_1, X_2, \dots, X_n$ ,  $\beta_1, \beta_2, \dots, \beta_n$  são os coeficientes de regressão a estimar e  $\beta_0$  é o parâmetro de interceção;  $\varepsilon$  corresponde à componente residual ou estocástica do modelo.

A utilização do método dos mínimos quadrados ordinários (OLS) para a obtenção de estimativas dos parâmetros da regressão linear múltipla, constitui a abordagem econométrica clássica (Wooldridge, 2008) e que aqui será seguida.

A teoria econométrica reconhece que a utilidade e adequabilidade de um modelo de preços hedónicos está diretamente relacionada com a correta definição dos atributos que descrevem a habitação (Ozanne & Malpezzi (1985)). Malpezzi (2003) sintetiza esta problemática, referindo que a análise exaustiva da literatura permite apenas concluir que existem três dimensões, fundamentais na definição destes atributos: **E** - atributos estruturais da habitação; **L** - características de localização e vizinhança; e **T** - atributos temporais que indiquem a época de transacção; conclui ainda o autor que é evidente a dificuldade em definir o conjunto exato de atributos.

## 2.3 Mecanismos de seleção de atributos

Dos diversos algoritmos de seleção de atributos, disponíveis nas soluções de *software* utilizadas, procurou-se testar um conjunto, representativo das opções mais usuais em problemas semelhantes, nomeadamente: i) no âmbito das ciências sociais (Hair, Black, Babin, Anderson, & Tatham, 2006), ii) da econometria (Wooldridge, (2008) and Koop, Poirier, & Tobias, (2007)), iii) bem como das abordagens generalistas de KDD (Blum & Langley, 1997) (Clarke, Fokoue, & Zhang, 2009).

No presente trabalho foram escolhidos 8 diferentes métodos, provenientes de quatro tipos de abordagens: i) métodos embutidos, ii) redução de dimensionalidade, iii) filtragem e iv) esquemas de pesagem. Note-se que enquanto as três últimas abordagens correspondem a um processo de modelação autónomo à construção do modelo hedónico (fase de pré-processamento da abordagem de *data mining*), no

primeiro caso, os algoritmos de seleção recorrem a critérios de seleção de variáveis baseados em medidas indiretas, obtidas pela aferição da eficiência do modelo (final) objeto de estimação (Guyon, 2006).

Os métodos de seleção automatizada de atributos analisados são:

- Métodos embutidos: foram comparados dois métodos em conjugação com o modelo de regressão linear tradicional:

**M1.** *Algoritmo M5prime* (Wang & Witten, 1996): o algoritmo é referenciado como um aperfeiçoamento do algoritmo *Quinlan's model-tree inducer* M5. Corresponde à combinação de um modelo de decisão em *árvore* com a aplicação de um modelo de regressão linear nas suas *folhas*, a que acresce uma funcionalidade de *poda* que permite uma redução significativa no tamanho da *árvore* (e, desta forma, nas variáveis selecionadas) baseado na aceitação de uma pequena penalidade na performance do modelo alvo.

**M2.** *Algoritmo de pesquisa "greedy forward"*: o algoritmo recorre a uma pesquisa exaustiva no conjunto de variáveis; a estratégia seguida inicia-se com um conjunto vazio, avaliando, de forma iterativa, a inclusão sucessiva de novos atributos, baseado na medida *Akaike Information Criteria* (Akaike, 1992).

- Análise em componentes principais: permite a construção de atributos de ordem superior (componentes principais), baseado na análise de variância dos atributos originais. Os novos atributos constituem combinações lineares do conjunto de atributos inicial e são ortogonais entre si. Este é um método "não supervisionado" uma vez que a variável dependente do modelo não tem qualquer papel nesta fase de seleção. Das várias abordagens, passíveis de serem testadas, para incorporar a informação obtida, optou-se aqui por implementar uma das mais simples.

**M3.** *Seleção de atributos representativos de cada uma das componentes*:

Como descrito em Hair et al. (2006), para manter o conjunto de variáveis original, pode-se optar por selecionar as variáveis iniciais, com os maiores *loadings* em cada uma das componentes, ordenadas por capacidade explicativa da variância original, até um máximo de 75% de variância cumulativa original explicada.

- Métodos de Filtragem Clássicos: foram utilizados dois métodos de filtragem, que implementam heurísticas de pesquisa exaustiva (Caruana & Freitag, 1994), utilizando como critério de avaliação, a medida CFS (*correlation-based feature selection* (Hall & Smith, 1998)). De acordo com a literatura, a utilização do critério CFS passa por considerar como subconjuntos ótimos aqueles em que os atributos são não-correlacionados entre si, mas fortemente correlacionados com a variável dependente do modelo hedónico. Foram implementadas duas estratégias de pesquisa do subconjunto ótimo (ou próximo do ótimo):

**M4.** *Eliminação inversa*: a pesquisa inicia-se com o conjunto completo de atributos disponíveis; o método prossegue com a remoção iterativa de

atributos, utilizando a medida CFS como critério de seleção do atributo a eliminar e de avaliação do subconjunto de atributos.

**M5.** *Seleção sequencial progressiva*: a pesquisa inicia-se com o conjunto vazio, prosseguindo com a adição iterativa e sequencial de atributos, mais uma vez utilizando a medida CFS como critério de seleção do atributo a adicionar e de avaliação do subconjunto de atributos.

- *Esquemas de pesagem*: no geral, estes métodos englobam a utilização de mecanismos para o cálculo de um peso, na modelação, associado a cada atributo do conjunto inicial. Este peso é utilizado como medida da sua relevância e, desta forma, permite uma hierarquização dos atributos e posterior processo de filtragem. Na filtragem os atributos são seleccionados em função de um limiar, usualmente estabelecido pela “sensibilidade” do investigador. Foram aplicados três abordagens de pesagem, recorrendo a filosofias de modelação distintas.

**M6.** *Pesagem através de uma máquina de suporte vetorial*: este método recorre aos coeficientes de uma máquina de suporte vetorial de forma a definir os pesos associados a cada atributo. A escolha deste esquema de seleção justifica-se por ser referido na literatura como uma abordagem que produz bons resultados, nomeadamente no domínio da bioinformática, especificamente na análise de *micro-arrays* (Guyon, 2006).

**M7.** *Mecanismo de regularização para selecção de variáveis - LASSO*: os métodos de regularização são mecanismos supervisionados que permitem melhorar o processo de modelação, especialmente indicados em situações em que a estimação clássica (neste caso o OLS) enfrenta variâncias altas. Esta família de métodos baseia-se, de forma genérica, na imposição de restrições que especificam um conjunto de soluções admissíveis (Fahrmeir, Kneib, Lang, & Marx, 2013). Dos vários métodos de regularização disponíveis, o LASSO (Least Absolute Shrinkage and Selection Operator) constitui uma abordagem clássica (Tibshirani, 1996), sendo inclusivé passível de utilização em exclusivo como mecanismo de seleção de variáveis (Belloni & Chernozhukov, 2009). O método recorre a um processo de estimação que introduz penalizações para a estimação dos coeficientes do modelo, a qual é governada pelo parâmetro  $\lambda$ . Este parâmetro determina, assim, o número de atributos significantes para o modelo e a sua escolha pode ser otimizada através da implementação de um esquema de “*cross-validation*”.

**M8.** *Mecanismo Bayesiano para selecção de variáveis - BMA*: a abordagem *Bayesiana* tem vindo a adquirir especial importância em processos de modelação nas ciências sociais (Lynch, 2007) (Koop et al., 2007) e em problemas de seleção de variáveis em particular (O’Hara & Sillanpää, 2009). Neste último caso, a seleção recorre a medidas comparativas das probabilidades do modelo (com base na distribuição *posterior*, resultado

da estimação). Especificamente, o método BMA - Bayesian Model Averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999) – é especialmente desenhado para as tarefas de selecção, sendo que se baseia na modelação Bayesiana exaustiva, de todos os modelos possíveis de serem construídos com um dado conjunto de variáveis (e para uma dada formulação), condicionais a um conjunto de *priors* pré-estabelecidos. Ao fornecer a probabilidade posterior de cada variável ser incluída nos modelos, esta medida pode constituir um mecanismo eficiente de selecção de variáveis a considerar para um modelo final.

## 2.4 Abordagem de análise de resultados

O coeficiente de determinação constitui uma medida clássica, de natureza genérica, para a avaliação da capacidade explicativa de modelos de regressão linear (ver por exemplo Hair et al. (2006)). Contudo, a utilização deste coeficiente deve ser alvo de alguns cuidados, visto que, entre outros aspetos, i) é sempre possível melhorar o coeficiente, adicionando um maior e correto número de atributos, ii) o coeficiente não é sensível à magnitude dos parâmetros, e iii) o coeficiente é suscetível de apresentar um erro sistemático positivo (estimativa otimista). No entanto, face à sua transdisciplinaridade (adoção alargada em diferentes domínios científicos) constitui uma escolha óbvia no contexto deste trabalho, sendo que para minimizar os problemas subjacentes à sua utilização foi implementado o cálculo do coeficiente de determinação ajustado, bem como um esquema de validação, usual em abordagens KDD, tendo sido escolhido o esquema *Hold-Out* – no qual o conjunto de dados é dividido de forma aleatória, tal que são definidos: i) um conjunto de treino, com 70% do número de casos, sob o qual é estimado o modelo; ii) um conjunto de avaliação, com os remanescentes 30%, que permite aferir a capacidade explicativa do modelo.

## 3 Estudo de caso: a construção de um modelo de preços no mercado de habitação de Aveiro - Ílhavo

O acesso aos dados armazenados pelo portal Casa Sapo, relativos a habitações publicadas no portal para efeitos de venda no período de 2000 a 2010, permitiu recolher 19 900 registos de potenciais habitações transacionadas nos municípios de Aveiro-Ílhavo.

Os municípios constituem uma área urbana portuguesa de média dimensão, composta por aproximadamente 117000 habitantes (em 2011), com um crescimento da população de 6% no período 2001 a 2011 (INE, 2013).

O portal *Casa Sapo* constitui um dos mais relevantes serviços de publicitação e pesquisa, de habitações para transação, em Portugal (Marktest, 2011). O serviço

gera receitas através da inserção de anúncios por parte das empresas de mediação imobiliária e dos proprietários. A informação acumulada no portal tem, desta forma, um objetivo puramente comercial o que, associado à ausência de exigências no tipo, quantidade e qualidade de informação inserida, constitui importantes limitações à utilização direta dos dados originais. Desta forma, o conjunto de dados iniciais sofreu um minucioso trabalho de depuração, que é descrito em Batista (2010). Após os processos de pré-processamento, a base de dados obtida contém 19900 observações e 55 atributos, distribuídos pelas categorias de atributos referidas na secção 2.2: 16 atributos estruturais, 39 características da localização e vizinhança e 11 variáveis *dummy* temporais.

De salientar que, tal como sugerido por Malpezzi (2003), as variáveis numéricas foram *logaritimizadas*, por forma a melhor adequar o modelo à teoria económica subjacente e permitir uma interpretação precisa à luz da teoria económica (os coeficientes poderão assim ser interpretados como *elasticidades-preço*). Uma nota ainda para os atributos temporais, que foram implementados como variáveis binárias, correspondentes ao ano em que o registo foi eliminado do portal. Desta forma eliminam-se os efeitos inflacionários gerais de considerar registos recolhidos em diferentes anos, tornando o modelo consistente para o ano de referência 2010. Por fim, foi ainda incluída uma variável TOM – *Time on Market* – a qual permite uma aproximação da variável preço utilizada como variável dependente (preço de oferta), para o eventual preço real de transacção – note-se, contudo, que esta abordagem correctiva dos preços de oferta (os preços disponíveis nos registos recolhidos) para os preços (reais) de transacção é alvo de algumas disputas (Sirmans, MacDonald, & Macpherson, 2010), mas considera-se que tal não afecta os objectivos concretos deste trabalho.

## 4 Análise de resultados

A seleção de variáveis é só uma das componentes do processo de construção do modelo de preços de habitação pelo que a análise de resultados deste trabalho procurou desenvolver uma metodologia, simples, de análise de resultados essencialmente útil para os objetivos propostos. Assim, consideraram-se dois indicadores: i) a capacidade explicativa do modelo final e ii) o número de variáveis envolvido no respectivo modelo de preços. Como facilmente podemos observar na Tabela 1, nestes indicadores-chave são as abordagens M4 e M5 aquelas que produzem melhores resultados na seleção automatizada de atributos. No caso do modelo M5 isto traduz-se numa redução de 96,5% do conjunto de variáveis em relação ao modelo de base, M0. A diferença entre ambos prende-se com dois atributos, relacionados com a dimensão territorial. No entanto, face à redução de capacidade explicativa inerente ao modelo M5, o modelo M4 reforça-se como aquele que tem melhor desempenho global.

Nos restantes modelos o seu desempenho caracteriza-se por, ou apresentam reduções muito grandes na capacidade explicativa, ou a capacidade de seleção do conjunto correto de atributos é mais reduzida.

Os resultados obtidos permitem ainda observar o conjunto de atributos sistematicamente selecionados. Neste aspeto, também a abordagem M4 apresenta uma melhor aproximação ao que a teoria sugere como um bom subconjunto de atributos (aquele que abarca as 3 grandes dimensões descritivas de uma habitação); para este modelo (M4) as estimativas, por ordem decrescente de importância, dão: área (-0,170), vizinhança de restaurantes (0,095), nível de conservação (-0,071), tipologia (0,035), proximidade a escolas primárias (0,033), vizinhança de comércio especializado (-0,029), vizinhança de divertimentos locais (0,144), vizinhança com densidade de edifícios superior a 80% (0,022), vizinhança com densidade de blocos multifuncionais – incluindo indústria, grandes infraestruturas, entre outros (-0,020), anúncio com referência a existência de lareira (-0,012) e proximidade a parques e jardins públicos (0,008). Em termos territoriais, os resultados deste modelo confirmam uma expetável (com base na teoria) valorização, seguindo um modelo centro (mais valorizado) – periferia (menos valorizado), do sistema urbano Aveiro – Ílhavo.

**Tabela 1** – Resultados globais, para o estudo de caso de Aveiro – Ílhavo.

Modelos	M0	M1	M2	M3	M4	M5	M6	M7	M8
Variáveis selecionadas	55	54	48	9	10	2	5	43	35
Variáveis significantes	30	30	46	9	10	2	4	39	35
Capacidade explicativa ajustada	0,618	0,617	0,617	0,310	0,563	0,456	0,469	0,616	0,618
Varição no nº de variáveis	-	-1,8%	-12,7%	-83,6%	-81,8%	-96,4%	-90,9%	-21,8%	-36,4%
Varição cap. exp. face a M0	-	-0,2%	-0,2%	-49,8%	-8,9%	-26,2%	-24,1%	-0,3%	0,0%

De referir que não se encontraram problemas assinaláveis ao nível da significância dos coeficientes estimados, bem como se regista a consistência no padrão geral de seleção (os diferentes subconjuntos de atributos apresentam semelhanças, quer ao conjunto de variáveis selecionadas, quer ao nível da magnitude dos coeficientes).

## 5 Conclusões

Com este trabalho, é possível concluir que a utilização de novas ferramentas estatísticas e computacionais, numa abordagem estruturada, com objectivos essencialmente instrumentais, é possível sem exigências de um conhecimento



muito avançado das áreas científicas de que provêm os referidos algoritmos. Desta forma, poderá facilita-se e melhor fundamentar as opções, exigidas aos profissionais que se debruçam sobre o valor da habitação e seus atributos. Neste caso concreto, note-se que os autores são, na sua maioria de profissionais que provêm da área de definição / estudo de políticas públicas – no caso concreto, especificamente do exercício de actividades no âmbito do planeamento urbano.

No caso de estudo aqui apresentado, mostra-se que a utilização e a comparação de técnicas (semi)automatizadas de seleção de atributos, permite uma maior consistência e segurança na execução da tarefa de selecção de atributos. Contudo, é importante salientar que: i) o processo de seleção estudado implica sempre uma diminuição da capacidade explicativa do modelo hedónico; ii) os atributos com uma reduzida relação com o preço da habitação apresentam uma maior variabilidade de seleção, pelas diferentes abordagens e algoritmos empregues; e iii) na utilização destas técnicas num ambiente multidisciplinar existem vantagens óbvias de não restringir os critérios de avaliação a medidas de análise quantitativa, pois a inclusão de aspetos de natureza qualitativa (por exemplo, uma análise da sucessiva seleção de determinado conjunto de atributos, nos vários métodos) permite um potencial maior controlo da abordagem metodológica adotada (tal como defendido nas metodologias de *data mining* aqui seguidas).

Por fim, regista-se a necessidade de uma avaliação crítica mais global que enquadre os resultados obtidos com as exigências e suporte teórico do modelo e respetivo domínio científico em que se aplica; contudo, essa análise extravasa os objetivos deste trabalho.

## Referências

- AKAIKE, H. (1992). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics* (pp. 610–624). Springer.
- AZEVEDO, A. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In *Proceedings from IADIS European Conference Data Mining* (pp. 182–185).
- BATISTA, P. (2010). *O data mining na identificação de atributos valorativos da habitação*. Dissertação de Mestrado. Universidade de Aveiro.
- BELLONI, A., & CHERNOZHUKOV, V. (2009). Least squares after model selection in high-dimensional sparse models.
- BLUM, A. L., & LANGLEY, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1), 245–271.
- BOURASSA, S. C., HOESLI, M., & SUN, J. (2006). A simple alternative house price index method. *Journal of Housing Economics*, 15(1), 80–97.
- CARUANA, R., & FREITAG, D. (1994). Greedy Attribute Selection. In *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 28–36).

- CLARKE, B., FOKOUE, E., & ZHANG, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science.
- FAHRMEIR, L., KNEIB, T., LANG, S., & MARX, B. (2013). *Regression: models, methods and applications*. Springer.
- FAYYAD, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems*, 11(5), 20–25.
- GUYON, I. (2006). *Feature extraction: foundations and applications* (Vol. 207). Springer.
- HAIR, J. F., BLACK, W. C., BABIN, B. J., ANDERSON, R. E., & TATHAM, R. L. (2006). *Multivariate data analysis* (Vol. 6). Pearson.
- HALL, M. A., & SMITH, L. A. (1998). Feature subset selection: a correlation based filter approach. In *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems* (pp. 855–858). Springer.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E., & VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382–401.
- INE, I. N. de E. (2013). INE - Base de dados dos censos da população portuguesa. Retrieved November 6, 2013, from [www.ine.pt](http://www.ine.pt)
- JONES, C., & WATKINS, C. (2009). *Housing markets and planning policy* (Vol. 40). John Wiley & Sons.
- KOOP, G., POIRIER, D. J., & TOBIAS, J. L. (2007). *Bayesian econometric methods*. Cambridge University Press.
- LANCASTER, K. J. (1966). A new approach to consumer theory. *The Journal of Political Economy*, 132–157.
- LYNCH, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer.
- MALPEZZI, S. (2003). Hedonic pricing models: a selective and applied review. In K. Gibb & A. O’Sullivan (Eds.), *Housing Economics and Public Policy: Essays in Honour of Duncan Maclellan* (pp. 67–89). Blackwell Science.
- MARKTEST. (2011). NetPanel.
- MARQUES, J. (2012). *The Notion of Space in Urban Housing Markets*. Universidade de Aveiro.
- O’HARA, R. B., & SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1), 85–117.
- OZANNE, L., & MALPEZZI, S. (1985). The efficacy of hedonic estimation with the annual housing survey. Evidence from the demand experiment. *Journal of Economic and Social Measurement*, 13(2), 153–172.
- ROSEN, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *The Journal of Political Economy*, 82(1), 35–55.
- SIRMANS, G. S., MACDONALD, L., & MACPHERSON, D. A. (2010). A meta-analysis of selling price and time-on-the-market. *Journal of Housing Research*, 19(2), 139–152.
- TAN, P.-N., STEINBACH, M., & KUMAR, V. (2006). *Introduction to data mining* (Vol. 1). Pearson Addison Wesley Boston.

- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- WANG, Y., & WITTEN, I. H. (1996). *Induction of model trees for predicting continuous classes* (No. 96/23).
- WOOLDRIDGE, J. (2008). *Introductory econometrics: A modern approach* (4th ed.). Cengage Learning.

## Novas aplicações de métodos multivariados na análise da mortalidade: um estudo na região norte de Portugal, 2001-2005

Lara Teixeira<sup>1</sup> · Vasco Machado<sup>2</sup> · Manuela Felício<sup>3</sup> · A. Manuela Gonçalves<sup>4</sup>

© The Author(s) 2017

**Resumo** A análise da mortalidade é fundamental no processo de planeamento da saúde e dos serviços de saúde, sendo a taxa de mortalidade padronizada pela idade um dos principais indicadores utilizados. A aplicação da Análise em Componentes Principais e da Análise Classificatória surge, neste trabalho, com o objectivo de identificar conjuntos de causas de morte que possam estar mais correlacionadas entre si e de reunir Agrupamentos de Centros de Saúde segundo perfis de mortalidade semelhantes.

**Palavras-chave:** Agrupamento de Centros de Saúde, Análise em Componentes Principais, Classificação Hierárquica Ascendente, Taxa de Mortalidade Padronizada.

### 1 Introdução

A mortalidade é considerada uma medida directa das necessidades em cuidados de saúde, reflectindo a carga global da doença na população, não só em termos da

---

<sup>1</sup>Departamento de Matemática e Aplicações (DMA), Universidade do Minho, [lara-teixeira@hotmail.com](mailto:lara-teixeira@hotmail.com)

<sup>2</sup>Departamento de Saúde Pública, Administração Regional de Saúde do Norte, I.P., [vmachado@arsnorte.min-saude.pt](mailto:vmachado@arsnorte.min-saude.pt)

<sup>3</sup>Departamento de Saúde Pública, Administração Regional de Saúde do Norte, I.P., [mfelicio@arsnorte.min-saude.pt](mailto:mfelicio@arsnorte.min-saude.pt)

<sup>4</sup>Centro de Matemática (CMAT) e Departamento de Matemática e Aplicações (DMA), Universidade do Minho, [mneves@math.uminho.pt](mailto:mneves@math.uminho.pt)

incidência da doença, como da capacidade de a tratar. Daí a importância dos indicadores de mortalidade no processo de planeamento da saúde e dos serviços de saúde.

A taxa de mortalidade padronizada pela idade é um dos indicadores usados na análise da mortalidade. A sua utilização, para um conjunto de causas de morte específicas, permite a comparação dos seus valores entre diferentes unidades territoriais. No entanto, os problemas de saúde não são balizados por fronteiras geográficas e, muito menos, administrativas. Um dos objectivos deste estudo é argumentar neste sentido. A Análise em Componentes Principais e vários métodos de Análise Classificatória, são usados neste estudo com o objectivo de identificar conjuntos de causas de morte que possam estar mais correlacionadas entre si e, também, conjuntos de Agrupamentos de Centros de Saúde (ACES<sup>1</sup>) da região Norte (respectivas áreas geodemográficas), cujo perfil de mortalidade seja mais semelhante.

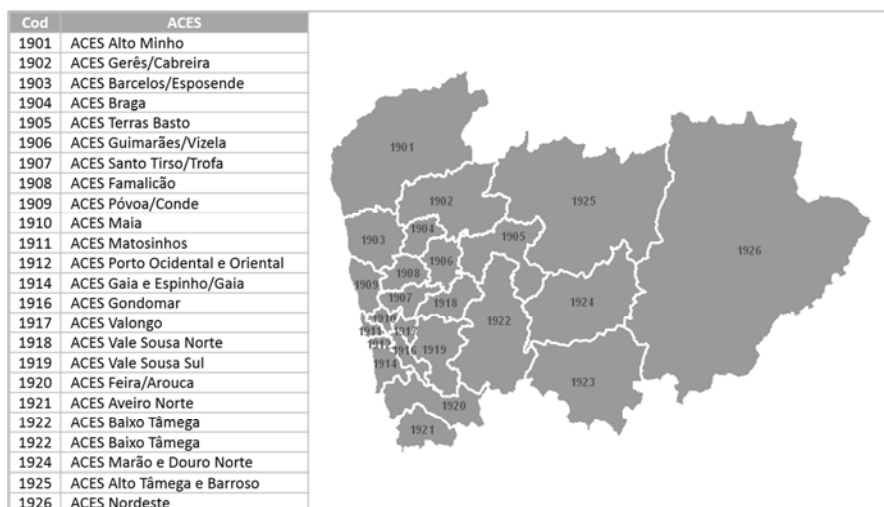
## 2 Base de dados e metodologia

É utilizada a base de dados regional dos óbitos para calcular a taxa de mortalidade padronizada pela idade de um conjunto de 13 causas de morte específicas, que se encontram classificadas de acordo com a 9<sup>a</sup> e 10<sup>a</sup> revisões da Classificação Internacional das Doenças (CID 9 e CID 10) (Tabela 1), para 24 ACES da região Norte (Figura 1). A selecção das 13 causas de morte específicas na análise para ambos os sexos resulta da magnitude destas causas no total dos óbitos e da sua prévia utilização em ferramentas de observação e monitorização do estado de saúde da população da região Norte (ARS Norte, 2008).

Calcula-se a taxa de mortalidade padronizada pela idade média anual no quinquénio 2001-2005, que será abreviada por TMP, apresentando-se aqui apenas as calculadas para ambos os sexos. Obtém-se uma matriz  $X=(x_{ij})$ , de dimensão 24 x 13, onde  $x_{ij}$  representa o valor da TMP na variável de ordem  $j$  (causa de morte específica) observada no indivíduo  $i$  (ACES).

---

<sup>1</sup> Os ACES foram criados em 2008 com a publicação em Diário da Republica do Decreto-Lei n.º 28/2008, que estabeleceu o regime da criação, estruturação e funcionamento dos agrupamentos de centros de saúde do Serviço Nacional de Saúde.



**Figura 1** – Códigos e designação dos ACES e sua distribuição geográfica.

**Tabela 1** – Códigos e designação das Causas de Morte.

Causa de Morte	Designação	CID 10	CID 9
Doenças Infecciosas e Parasitárias			
<b>Tuberculose</b>	<b>Tub</b>	A15-A19, B90	010-018, 137
<b>VIH / sida</b>	<b>VIH</b>	B20-B24	042-044
Tumores Malignos			
<b>Tumor Maligno da Traqueia, Brônquios e Pulmão</b>	<b>TMTBP</b>	C33-C34	162
<b>Tumor Maligno do Estômago</b>	<b>TME</b>	C16	151
Tumor Maligno da Próstata	TMP	C61	185
Tumor Maligno da Mama (Feminina)	TMM	C50	174
Tumor Maligno do Colo do Útero	TMCU	C53	180
<b>Tumor Maligno do Cólon e Recto</b>	<b>TMCR</b>	C18-C20	153-154.1
Doenças Endócrinas, Nutricionais e Metabólicas			
<b>Diabetes Mellitus</b>	<b>Diab</b>	E10-E14	250
Doenças do Aparelho Circulatório			
<b>Doença Isquémica do Coração</b>	<b>DIC</b>	I20-I25	410-414
<b>Doenças Cerebrovasculares</b>	<b>Dcer</b>	I60-I69	430-438
Doenças do Aparelho Respiratório			
<b>Pneumonia</b>	<b>Pneum</b>	J12-J18	480-486
<b>Doença Pulmonar Obstrutiva Crónica</b>	<b>DPOC</b>	J40-J47	490-496
Doenças do Aparelho Digestivo			
<b>Doença Crónica do Fígado e Cirrose</b>	<b>DCFC</b>	K70 e K73-K74	571
Causas Externas de Mortalidade			
<b>Acidentes de Transporte</b>	<b>AT</b>	V01-V99	e800-e848
<b>Lesões Autoprovocadas Intencionalmente (Suicídios)</b>	<b>Suic</b>	X60-X84	e950-e959

## 2.1 Taxa de mortalidade padronizada pela idade

O método directo de padronização consiste na aplicação das taxas específicas de mortalidade por idade a uma população padrão cuja composição etária é fixa, distribuindo-se pelos mesmos grupos etários das taxas específicas. O método consiste, portanto, em calcular as taxas de mortalidade esperadas na população padrão. De acordo com um procedimento corrente na literatura escolheu-se a população padrão europeia, com grupos etários decenais.

Considere-se a população padrão dividida em  $k$  classes ( $i = 1, 2, \dots, k$ ), as mesmas que estão na população em estudo. Seja  $w_i$  um factor de ponderação, isto

é, o peso para a classe  $i$  definido por  $w_i = \frac{m_i}{n_i \cdot m}$ , com  $m_i$  a representar o efectivo

populacional na classe  $i$  da população padrão,  $n_i$  o efectivo populacional na classe  $i$  da população em estudo e  $m$  o efectivo total da população padrão.

Sejam  $X_i$ ,  $i = 1, 2, \dots, k$ , variáveis aleatórias de Poisson de parâmetro  $\theta_i$  que representam o número de óbitos na classe  $i$  da população em estudo. A taxa de mortalidade padronizada pela idade (TMP) é uma variável aleatória  $Y$  que não é mais do que uma soma ponderada de variáveis aleatórias independentes de Poisson,

$$Y = \sum_{i=1}^k w_i \cdot X_i \quad \text{e} \quad E(Y) = \sum_{i=1}^k w_i \cdot \theta_i.$$

Para mais considerações ver Dobson *et al.* (1991) e Fay e Feur (1997).

Utiliza-se a comparação da TMP de duas áreas geográficas,  $A$  e  $B$ , para verificar se há diferenças estatisticamente significativas entre duas populações. Constrói-se um teste de hipóteses paramétrico sob a hipótese nula  $H_0 : Y^A = Y^B$ . Para cada teste foi identificado o valor- $p$  para perceber quais as diferenças mais significativas na TMP das populações em análise. É importante, muitas vezes, ter não só a identificação de que existem diferenças significativas, mas também poder observar se o valor da TMP é superior ou inferior ao valor esperado de uma população de referência. Assim, obtém-se uma classificação dos ACES, para cada causa de morte, em quatro classes: a TMP de um dado ACES é inferior ou superior à TMP da região, com a identificação dessa diferença ser estatisticamente significativa ou não. São produzidos mapas da região que permitem a visualização e identificação das diferenças testadas.

## 2.2 Análise em componentes principais

A Análise em Componentes Principais (ACP) tem por objectivo encontrar um conjunto de novas variáveis, as componentes principais, que são combinações lineares das variáveis originais (as causas de morte), não correlacionadas linearmente (Gordon, 1999). As componentes principais são determinadas por ordem decrescente das variâncias, isto é, da sua capacidade explicativa, e explicam a variância total dos dados. Geralmente é retido um número pequeno das primeiras componentes principais, que explicam uma percentagem considerável da variabilidade total.

O *biplot* (Gabriel, 1971) é uma representação gráfica de dados multivariados a duas dimensões que permite a representação simultânea dos indivíduos e das variáveis. A sobreposição das projecções das duas nuvens, dos indivíduos e das variáveis, no mesmo plano torna mais fácil a interpretação, desde que se compreenda que as nuvens têm significados diferentes.

## 2.3 Classificação hierárquica ascendente

A Classificação Hierárquica Ascendente tem como principal objectivo agrupar um conjunto de objectos a classificar (variáveis ou indivíduos) num número pequeno de classes, que reflitam as relações de semelhança e/ou dissemelhança entre esses objectos (Saporta, 2006). Aplicam-se métodos de classificação hierárquica, com recurso ao coeficiente de correlação de *Pearson* para a comparação de variáveis (causas de morte) e à distância Euclidiana para a comparação de pares de indivíduos (ACES). Como critérios de agregação foi utilizado o método do centróide para a classificação da variáveis e o método de Ward para os indivíduos.

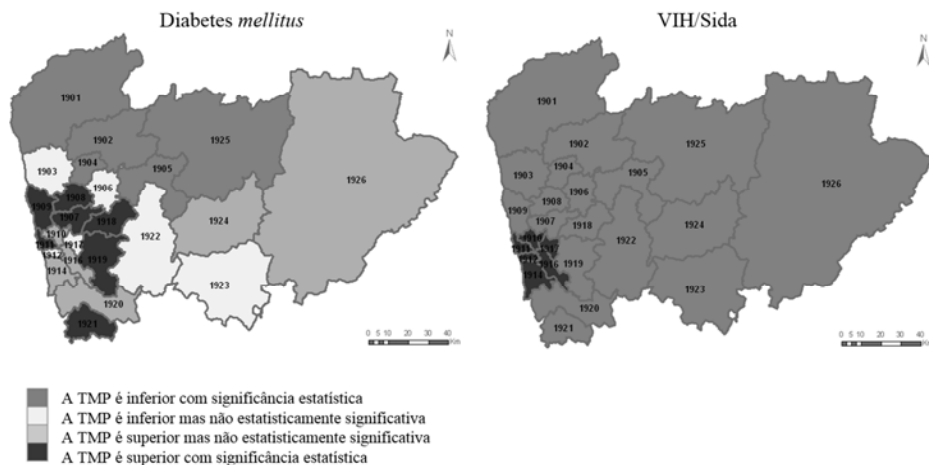
Foram encontrados grupos homogêneos de ACES da região Norte considerando todas as causas e, também, classes de causas de morte entretanto agrupadas. Uma vez constituídos os grupos de ACES, identificam-se os que diferem relativamente às variáveis que contribuem para o seu agrupamento, utilizando-se os testes não-paramétricos de *Kruskal-Wallis* e de *Tukey* (Higgins, 2004). Foi, ainda, realizada a representação com recurso a diagramas de caixa (*boxplot*) para auxiliar a visualização e interpretação dos resultados.

## 3 Resultados

O cálculo da TMP e o teste de hipóteses aplicado permitiram, para uma causa de morte específica, identificar os ACES que apresentam valores da TMP significativamente diferentes do valor da TMP da região Norte. Permitiu, também, para um determinado ACES, identificar quais as causas de morte em que esse



ACES apresenta valores da TMP significativamente diferentes dos valores da TMP da região Norte (ARS Norte, 2008). Na Figura 2 pode observar-se, por exemplo, a distribuição espacial das TMPs para as causas de morte Diabetes *mellitus* e VIH/sida. Para o VIH/sida e quando comparada com a TMP da região Norte pode observar-se valores significativamente superiores da TMP dos ACES da região do Grande Porto.

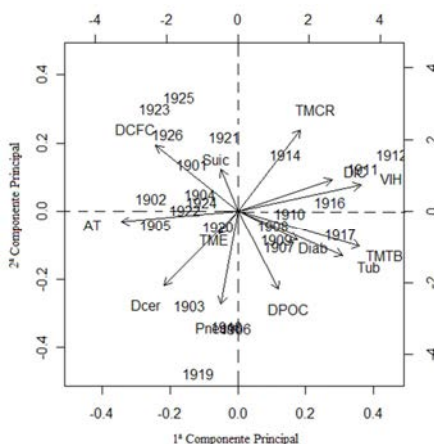


**Figura 2** – Mapas da distribuição espacial das TMPs para as causas de morte Diabetes *mellitus* e VIH/Sida.

Com a aplicação da ACP pode concluir-se quais as variáveis (causas de morte) que estão mais correlacionadas com cada uma das componentes principais retidas e quais os indivíduos (ACES) que mais contribuíram para a formação dessas componentes. Foram retidas quatro componentes (Tabela 2) que explicam 72,3% da variância total (a primeira explica 33,9%, a segunda 15,4%, a terceira 12,9% e a quarta 10,1% da variância total). Na Figura 3 pode observar-se o *biplot* da representação das variáveis e indivíduos no plano composto pela primeira e segunda componentes principais, que em conjunto explicam quase metade da variância total. Pode observar-se uma forte correlação positiva entre as variáveis VIH, TMTBP, Tub e DIC e a primeira componente. São os ACES do Porto, Matosinhos, Gondomar e Valongo que mais contribuem para a sua formação.

**Tabela 2** – Correlações entre as variáveis e as quatro primeiras componentes principais.

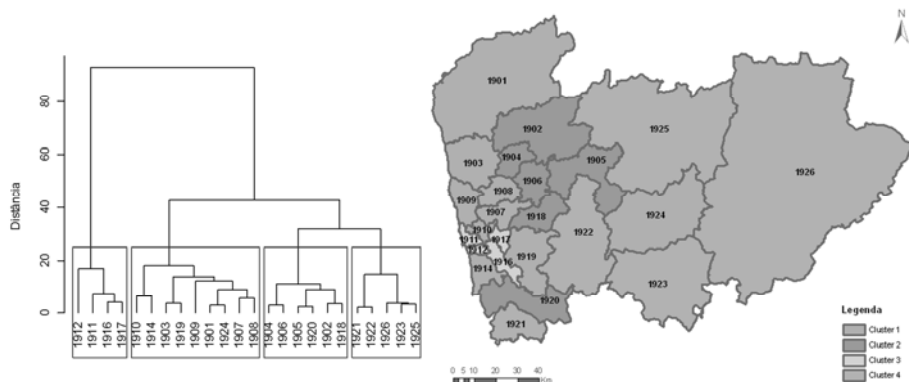
	Componentes Principais			
	1	2	3	4
Tub	<b>0,76</b>	0,30	0,36	0,18
VIH	<b>0,88</b>	-0,19	0,15	0,11
TMTBP	<b>0,87</b>	0,24	0,30	0,02
TME	-0,14	0,15	<b>0,71</b>	-0,17
TMCR	0,44	<b>-0,59</b>	0,20	-0,02
Diab	0,43	0,23	<b>-0,72</b>	0,26
DIC	<b>0,68</b>	-0,23	-0,10	0,49
Dcer	<b>-0,53</b>	<b>0,53</b>	-0,07	0,52
Pneum	-0,12	<b>0,65</b>	0,34	0,02
DPOC	0,29	<b>0,56</b>	0,14	0,07
DCFC	<b>-0,59</b>	-0,48	0,37	0,19
AT	<b>-0,83</b>	0,07	0,07	0,31
Suic	-0,12	-0,32	0,28	<b>0,72</b>



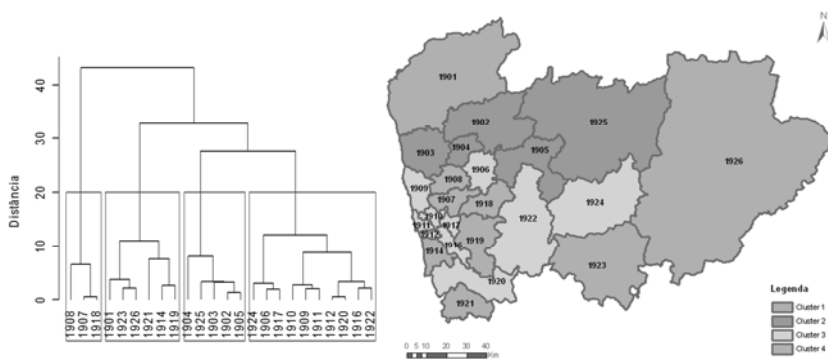
**Figura 3** – Biplot da 1ª e 2ª componentes.

Através da Análise de Classificação Hierárquica Ascendente, recorrendo ao coeficiente de correlação de *Pearson* como medida de semelhança e ao método do centróide como critério de agregação, classificaram-se as causas de morte em quatro classes (classe 1: Tub, TMTBP, VIH, DIC e TMCR, classe 2: Diab e DPOC, classe 3: Dcer, AT e Pneum, classe 4: DCFC, Suic e TME). Dentro de cada classe de variáveis é feita uma classificação hierárquica dos ACES da região Norte, usando a distância euclidiana e o critério de *Ward*. Para a primeira classe de variáveis agruparam-se os vinte e quatro ACES em quatro grupos (Figura 4), para a

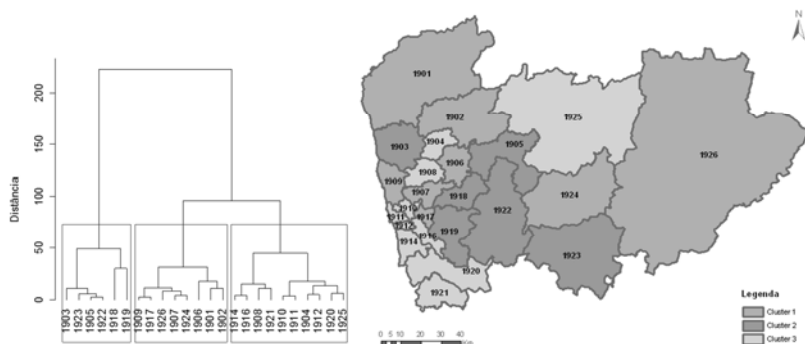
segunda classe também em quatro grupos (Figura 5), para a terceira classe em três (Figura 6) e para a quarta classe agruparam-se, novamente, os ACES em quatro grupos (Figura 7).



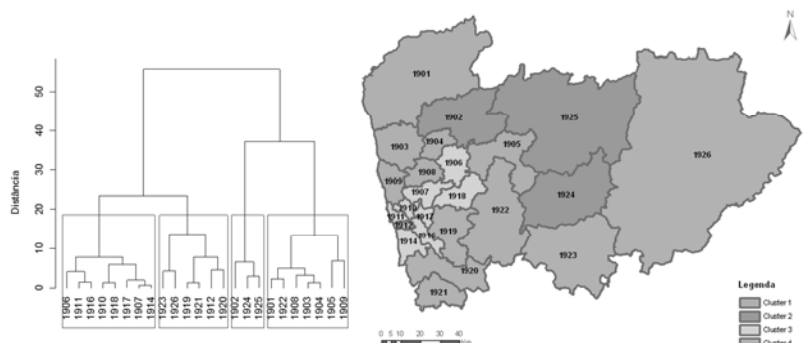
**Figura 4** – Dendrograma segundo o método de Ward com o número de grupos e a sua representação espacial para a classe 1.



**Figura 5** – Dendrograma segundo o método de Ward com o número de grupos e a sua representação espacial para a classe 2.



**Figura 6** – Dendrograma segundo o método de Ward com o número de grupos e a sua representação espacial para a classe 3.



**Figura 7** – Dendrograma segundo o método de Ward com o número de grupos e a sua representação espacial para a classe 4.

A aplicação dos testes não-paramétricos de *Kruskal-Wallis* e de *Tukey* foi realizada para as quatro classes de variáveis formadas, apresentando-se aqui apenas os resultados para a primeira classe. O teste de *Kruskal-Wallis* (Tabela 3) e o teste de *Tukey* (Tabela 4) permitiram detectar diferenças entre os grupos de ACES e evidenciar as causas de morte que mais contribuíram para a sua discriminação.

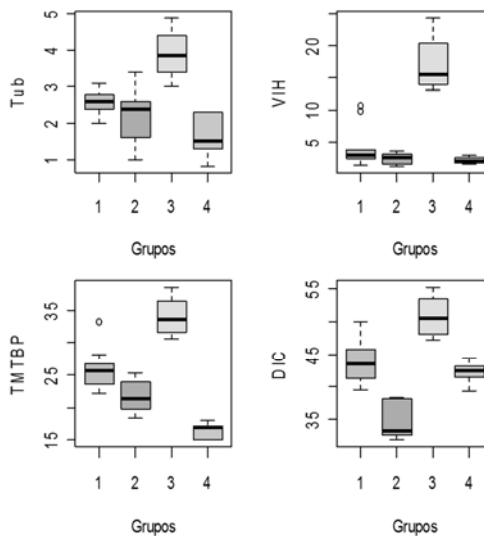
**Tabela 3** – Teste de *Kruskal-Wallis*.

	Estatística	Graus de liberdade	Valor-p
Tub	13,770	3	<b>0,003</b>
VIH	10,382	3	<b>0,016</b>
TMTBP	18,543	3	<b>&lt;0,001</b>
TMCR	6,336	3	0,096
DIC	17,391	3	<b>0,001</b>

**Tabela 4** – Teste não paramétrico de *Tukey*.

Variável	Grupo (I)	Grupo (J)	Diferença de médias (I-J)	Desvio-padrão	Valor-p	Intervalo de confiança	
						Limite inferior	Limite superior
Tub	1	2	3,000	2,569	0,653	-4,192	10,192
		3	<b>-8,333</b>	2,930	<b>0,046</b>	-16,533	-0,134
		4	<b>8,667</b>	2,719	<b>0,022</b>	1,056	16,278
	2	3	<b>-11,333</b>	3,147	<b>0,009</b>	-20,141	-2,525
		4	5,667	2,952	0,252	-2,596	13,929
	3	4	<b>17,000</b>	3,270	<b>0,000</b>	7,847	26,153
VIH	1	2	1,722	2,958	0,936	-6,557	10,001
		3	<b>-10,611</b>	3,372	<b>0,024</b>	-20,050	-1,172
		4	3,489	3,130	0,685	-5,273	12,250
	2	3	<b>-12,333</b>	3,623	<b>0,014</b>	-22,473	-2,194
		4	1,767	3,398	0,953	-7,745	11,278
	3	4	<b>14,100</b>	3,765	<b>0,006</b>	3,563	24,637
TMTBP	1	2	<b>5,500</b>	1,759	<b>0,025</b>	0,576	10,424
		3	<b>-6,667</b>	2,006	<b>0,016</b>	-12,281	-1,052
		4	<b>12,333</b>	1,862	<b>0,000</b>	7,122	17,544
	2	3	<b>-12,167</b>	2,155	<b>0,000</b>	-18,197	-6,136
		4	<b>6,833</b>	2,021	<b>0,014</b>	1,176	12,491
	3	4	<b>19,000</b>	2,239	<b>0,000</b>	12,733	25,267
DIC	1	2	<b>11,056</b>	1,984	<b>&lt;0,001</b>	5,504	16,607
		3	<b>-7,194</b>	2,262	<b>0,022</b>	-13,524	-0,864
		4	2,356	2,099	0,681	-3,520	8,231
	2	3	<b>-18,250</b>	2,429	<b>&lt;0,001</b>	-25,049	-11,451
		4	<b>-8,700</b>	2,279	<b>0,005</b>	-15,078	-2,322
	3	4	<b>9,550</b>	2,525	<b>0,006</b>	2,484	16,616

A classificação hierárquica permitiu identificar os ACES do Porto, Matosinhos, Gondomar e Valongo como sendo os que apresentam valores da TMP mais elevados nas causas de morte tuberculose, VIH/sida, tumor maligno da traqueia, brônquios e pulmão e doença isquémica do coração (Figura 8). Estes resultados vêm corroborar aqueles obtidos pela aplicação da ACP.



**Figura 8** – Representação das variáveis em função dos grupos.

## 4 Conclusão

Os procedimentos multivariados adoptados neste estudo foram muito úteis para identificar padrões homogêneos no comportamento das causas de mortalidade e dos ACES. Permitiram reduzir a dimensão do conjunto de informação e agrupar ACES segundo perfis de mortalidade semelhantes. A aplicação da Análise em Componentes Principais e dos métodos de Classificação Hierárquica Ascendente aos dados evidenciou a complementaridade dos métodos. A utilização destas metodologias em áreas aplicadas, como o Planeamento em Saúde e a Epidemiologia, assume cada vez maior importância.

## Agradecimentos

Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - “Fundação para a Ciência e a Tecnologia”, no âmbito do projecto PEstOE/MAT/UI0013/2014.

## Referências

- GORDON, A.D. (1999). *Classification*. 2nd ed, Chapman and Hall, London.
- HIGGINS, J.J. (2004). *Introduction to Modern Nonparametric Statistics*. Duxbury Advanced Series.
- SAPORTA, G. (2006). *Probabilités Analyse des Données et Statistique*. 2<sup>e</sup> edition, Technip, Paris.
- DOBSON, A., KUULASMAA, K., EBERL, E. & SCHERER, J. (1991). Confidence intervals for weighted sums of Poisson parameters. *Statistic in Medicine*, 10, 457-462.
- GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, vol.58, n.03, 453–467.
- FAY, M. P. & FEUER, E. J. (1997). Confidence intervals for directly adjusted rates: a method based on the gamma distribution. *Statistic in Medicine*, 16, 791-801.
- ARS Norte (2008). mort@lidades.geres – Mortalidade Geral e Específica, Região Norte 2001-2005, [http://portal.arsnorte.min-saude.pt/ARSNorte/dsp/AM\\_0105](http://portal.arsnorte.min-saude.pt/ARSNorte/dsp/AM_0105), (acedido em 1 de Janeiro 2011).

## Estimação de um modelo com trajetória latente com dados omissos resultantes de um painel rotativo

Paula C. R. Vicente<sup>1</sup> · Maria de Fátima Salgueiro<sup>2</sup>

© The Author(s) 2017

**Resumo** Este trabalho tem como objetivo avaliar o enviesamento na estimação dos parâmetros de um modelo com trajetória latente com dados omissos que configuram um *planned missing design*. São considerados dados do painel rotativo ICOR. As estimativas dos parâmetros obtidas por máxima verosimilhança são comparadas com as que resultam da estimação do modelo com dados completos. É realizado um estudo de simulação e fornecidas recomendações para efeitos de estimação de um modelo com trajetória latente com dados omissos.

**Palavras-chave:** Dados Omissos, *Full Information Maximum Likelihood*, ICOR, Modelo com Trajetória Latente, Monte Carlo, Painel Rotativo.

### 1 Introdução

A existência de dados omissos constitui uma problemática bastante frequente em estudos com dados longitudinais, sendo usual fazer a distinção entre dados omissos intermitentes e *dropout* (ou atrito). Ocorre *dropout* quando a partir de determinado momento temporal, um elemento deixa de responder às questões do inquérito. Se em determinado momento o elemento não responde mas mais tarde volta a

---

<sup>1</sup>Universidade Lusófona de Humanidades e Tecnologias, Escola de Ciências Económicas e das Organizações, [p951@ulusofona.pt](mailto:p951@ulusofona.pt)

<sup>2</sup>Business Research Unit e Departamento de Métodos Quantitativos para Gestão e Economia, Instituto Universitário de Lisboa (ISCTE-IUL), [fatima.salgueiro@iscte.pt](mailto:fatima.salgueiro@iscte.pt)



responder as omissões dizem-se intermitentes. Todavia, as omissões podem também resultar do desenho do estudo, como é o caso de um painel rotativo (ENDERS, 2010). Assim, num *planned missing design* há uma estrutura de omissão de dados que ocorre de forma intencional e de acordo com o planeado pelo investigador, sendo que o objetivo da utilização de um desenho deste tipo se prende com o propósito de minimizar o esforço de inquirição, e o consequente abandono do painel. O painel rotativo ICOR (Inquérito às Condições de Vida e Rendimento) pode ser considerado um exemplo de um *planned missing design*, (INE, 2009).

Uma das técnicas atualmente mais utilizadas para lidar com a problemática da omissão dos dados tem sido o método de estimação *full information maximum likelihood* (FIML) (SCHAFER & GRAHAM, 2002). Para efeitos de modelação estatística, ao escolher uma abordagem para lidar com dados omissos, é necessário ter em atenção a relação entre as variáveis observadas e a probabilidade de omissão, isto é, o mecanismo de omissão. O método FIML é adequado quando se assume um mecanismo de omissão de dados aleatório (MAR) ou completamente aleatório (MCAR).

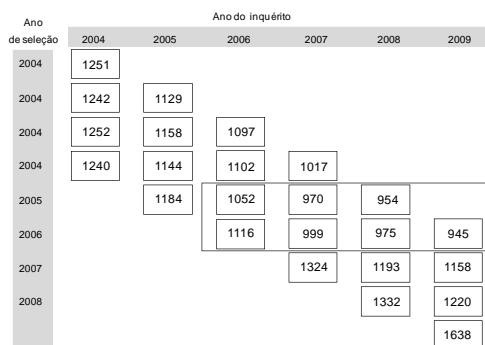
Neste trabalho é proposta a modelação de um indicador de privação material, para os anos de 2006 a 2009, com recurso a modelos com trajetória latente (BOLLEN & CURRAN, 2006). Ao longo dos quatro momentos temporais em análise permaneceram no painel um conjunto de agregados, para os quais existe informação completa. Existe ainda um outro conjunto de agregados que participaram no estudo entre 2006 e 2008 e que não estiveram em 2009 devido ao esquema de rotatividade do painel ICOR. A utilização de procedimentos de FIML permite estimar um modelo com trajetória latente para modelar a trajetória do *score* de privação material entre 2006 e 2009, considerando a totalidade dos agregados, ao invés do conjunto inicialmente considerado através da aplicação de uma técnica *listwise*, uma vez que no ano de 2009 existem omissões que resultam do desenho do estudo. Diferentes conjuntos de estimativas são então obtidas para os parâmetros do modelo, sendo discutido qual das duas opções metodológicas deve o analista tomar. Posteriormente, recorrendo a um estudo de simulação de Monte Carlo, e assumindo como verdadeiros os valores dos parâmetros estimados pelo modelo com trajetória latente para os dados completos, são geradas amostras com um padrão de omissão de 50% no último momento temporal. Para cada uma das diferentes amostras são estimados os parâmetros do modelo com trajetória latente, sendo o valor médio dessas estimativas comparado com o “verdadeiro” valor usado para gerar as réplicas. O pacote estatístico utilizado é o Mplus (MUTHÉN & MUTHÉN, 1998-2010).

## **2 Amostra em estudo**

O ICOR resulta da participação portuguesa na base de dados europeia EU-SILC (*European Statistics on Income and Living Conditions*), a qual é assegurada pelo

Instituto Nacional de Estatística (INE) desde 2004 com uma periodicidade anual. Este painel apresenta a particularidade de ser um painel rotativo com uma dinâmica de rotatividade de 1/4 da amostra (INE, 2009). Assim, um painel deste tipo configura um caso de dados omissos por desenho, uma vez que em cada ano sai uma fração da amostra (25%), não podendo nenhum agregado permanecer na amostra por mais de quatro anos, ver Figura 1.

No presente estudo foram considerados os 845 agregados que permaneceram no painel entre os anos de 2006 e 2009, constituindo 25% da amostra original, observados em quatro momentos temporais, e que se passam a designar por “caso completo”. Foram ainda considerados os 848 agregados que permaneceram no estudo de 2006 a 2008. O primeiro grupo de 845 observações, mais as 848 observações que correspondem aos agregados que saíram do painel em 2009, somam um total de 1693 agregados. Estes agregados constituem uma subamostra que se passa a designar por “caso com omissões”, e correspondem a aproximadamente 50% dos 4367 agregados que estavam no painel em 2006, conforme apresentado na Figura 1.



**Figura 1** - Desenho do painel rotativo ICOR. Os agregados que constituem o “caso com omissões” considerado no presente estudo encontram-se destacados na caixa que envolve os anos 2006 a 2009.

Em cada um dos quatro momentos temporais (2006 a 2009) foi calculado um indicador de privação material, obtido como uma soma de itens, que correspondem a perguntas do painel ICOR (GUIO, 2009). As questões consideradas para a construção deste *score* foram (ter capacidade financeira/ter disponibilidade económica para): i) ter uma refeição de carne ou de peixe, pelo menos de dois em dois dias; ii) pagar uma semana de férias, por ano, fora de casa, a todo o agregado; iii) suportar despesas inesperadas, sem recorrer a crédito; iv) ter a casa aquecida; v) fazer face às despesas e encargos usuais; vi) ter telefone móvel ou fixo; vii) ter TV a cores; viii) ter máquina lavar roupa; ix) ter veículo ligeiro de passageiros ou misto; e x) ter computador. O *score* calculado toma valores entre 0 e 10, indicando o valor 0 que o agregado não se encontra privado, isto é, tem capacidade financeira para assegurar todos os itens considerados, e 10 que não consegue assegurar nenhum dos itens considerados.

Uma análise às duas subamostras consideradas permite verificar que, na subamostra designada por “caso completo” a maior percentagem de agregados tem 3 itens em privação, nos anos de 2006 e 2007, enquanto que nos anos de 2008 e 2009, a maior percentagem de agregados tem 2 itens em privação. Na subamostra designada por “caso com omissões” a maior percentagem de agregados tem 3 itens em privação, nos anos de 2006, 2007 e 2009. No ano de 2008, a maior percentagem de agregados tem apenas 2 itens em privação.

### 3 Metodologia

Atualmente, os modelos com trajetória latente constituem uma técnica frequentemente utilizada no estudo da mudança usando dados longitudinais, dado que este tipo de modelação, através da estimação de uma trajetória latente ao longo do tempo, permite estudar a mudança quer individualmente, quer para o conjunto dos elementos (BOLLEN & CURRAN, 2006).

Um modelo com trajetória latente pode ser escrito como

$$y_{it} = \alpha_i + \lambda_t \beta_i + \varepsilon_{it},$$

em que  $y_{it}$  é o valor da variável  $y$  para o elemento  $i$ , no momento temporal  $t$ ,  $\alpha_i$  é o intercepto aleatório do elemento  $i$ ,  $\beta_i$  é o declive aleatório do elemento  $i$ ,  $\varepsilon_{it}$  representa o termo residual da trajetória traçada para o elemento  $i$  no momento  $t$ , e  $\lambda_t$  é uma constante que usualmente assume a forma  $(t-1)$  quando é considerada uma trajetória linear. Assume-se que o termo residual  $\varepsilon_t \sim N(0, \Theta_\varepsilon)$ , sendo  $\Theta_\varepsilon$  uma matriz diagonal com elementos  $\theta_{\varepsilon_t}$  na diagonal principal.

Os pesos fatoriais que ligam as variáveis observadas ( $y_{it}$ ) ao intercepto da trajetória ( $\alpha_i$ ) são fixos a um por forma a estabelecer o momento inicial. Os pesos fatoriais que ligam as variáveis observadas ao declive da trajetória ( $\beta_i$ ) estão habitualmente fixos a 0, 1, 2, 3, ... e refletem a passagem linear do tempo (i.e.  $\lambda_t = t - 1$ ), conforme apresentado na Figura 2.

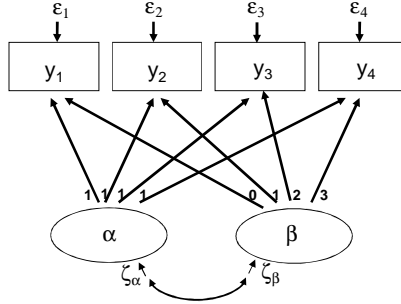
As variáveis latentes do modelo  $\alpha$  e  $\beta$  são, respetivamente, o interceto e o declive aleatórios da trajetória e são descritas pelas equações

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i} \text{ e } \beta_i = \mu_\beta + \zeta_{\beta_i},$$

em que,  $\mu_\alpha$  e  $\mu_\beta$  representam o valor esperado do intercepto e do declive, respetivamente. Os interceptos e os declives não estão correlacionados entre elementos e entre si, isto é,  $Cov(\alpha_i, \alpha_j) = 0$ ,  $Cov(\beta_i, \beta_j) = 0$  e  $Cov(\alpha_i, \beta_j) = 0$  com  $i \neq j$ . Assume-se que os termos residuais  $\zeta_{\alpha_i}$  e  $\zeta_{\beta_i}$ , que representam a variabilidade no momento inicial e em torno da taxa de crescimento, respetivamente, têm distribuição normal com valor esperado zero e não estão correlacionados com  $\varepsilon_{it}$ . As variâncias dos termos residuais  $\zeta_{\alpha_i}$  e  $\zeta_{\beta_i}$  são dadas por  $\psi_{\alpha\alpha}$  e  $\psi_{\beta\beta}$ , e a covariância por  $\psi_{\alpha\beta}$ . Assim,  $\zeta_{\alpha_i} \sim N(0, \psi_{\alpha\alpha})$ ,  $\zeta_{\beta_i} \sim N(0, \psi_{\beta\beta})$ ,  $Cov(\varepsilon_{it}, \zeta_{\alpha_i}) = 0$  e  $Cov(\varepsilon_{it}, \zeta_{\beta_i}) = 0$ , sendo a variância do intercepto ( $\alpha_i$ ) igual à

variância de  $\zeta_{\alpha_i}$ , isto é,  $\psi_{\alpha\alpha}$ , e a variância do declive ( $\beta_i$ ) igual à variância de  $\zeta_{\beta_i}$ , isto é,  $\psi_{\beta\beta}$ .

Os valores esperados e as variâncias do intercepto e do declive dessa trajetória permitem avaliar a média inicial e a taxa média de mudança para o conjunto dos elementos, assim como a variabilidade desses mesmos elementos em torno da média.



**Figura 2** - Diagrama de um modelo com trajetória latente linear, com quatro momentos temporais.

Quando existem observações omissas, a estimação de um processo de mudança usando um modelo com trajetória latente pode ser levada a efeito pelo método FIML, assumindo o pressuposto da normalidade multivariada da distribuição dos dados. Este método utiliza toda a informação disponível durante a estimação. Quando o mecanismo de omissão dos dados é aleatório, e os dados têm uma distribuição normal multivariada, o método FIML produz estimativas dos parâmetros, erros padrão e testes estatísticos que são consistentes e eficientes.

A função a maximizar, na presença de dados completos, é para a observação  $i$

$$\log L_i = -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})$$

onde  $k$  é o número de variáveis,  $\mathbf{Y}_i$  é o vetor para a observação  $i$ ,  $\boldsymbol{\mu}$  é o vetor das médias populacionais e  $\Sigma$  é a matriz das variâncias-covariâncias.

Com dados omissos, a função para a observação  $i$  toma a forma

$$\log L_i = -\frac{k_i}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

onde  $k_i$  representa o número de casos completos para aquela observação e os  $\boldsymbol{\mu}_i$  e  $\Sigma_i$  estão associados apenas aos dados disponíveis. Os cálculos para a função  $\log L_i$  para a observação  $i$  dependem apenas das variáveis e dos parâmetros para os quais essa observação tem dados completos (ENDERS, 2010).

O método da máxima verossimilhança para lidar com dados omissos não necessita de imputar ou substituir os valores omissos, usando, alternativamente, toda a informação disponível para estimar os parâmetros e os erros padrão (SCHAFER & GRAHAM, 2002).

Para medir a qualidade do ajustamento modelo-dados são usualmente utilizadas na literatura as seguintes medidas: índice *Tucker-Lewis* (TLI); índice de

ajustamento comparado (CFI); *Root Mean Square Error of Approximation* (RMSEA) e *Standardized Root Mean Square Residual* (SRMR). São considerados valores de ajustamento recomendáveis TLI>0.90, CFI>0.90, RMSEA<0.05, sendo valores até 0.08 considerados aceitáveis, e SRMR<0.08, (SCHUMACKER & LOMAX, 2010; YU & MUTHÉN, 2002).

Pode realizar-se um estudo de simulação utilizando procedimentos de Monte Carlo recorrendo ao pacote estatístico Mplus. Este *software* permite gerar  $m$  amostras de dados a partir da estrutura de um modelo com trajetória latente, cujos parâmetros populacionais são definidos a priori pelo investigador. Para cada uma das  $m$  amostras geradas, e de uma forma integrada, é estimado um modelo com trajetória latente, obtendo-se, deste modo,  $m$  estimativas para cada um dos parâmetros do modelo. Se nas amostras geradas existem omissões, é utilizada na estimação uma abordagem FIML. Para cada um dos parâmetros do modelo o *software* disponibiliza a média das estimativas, calculada a partir das  $m$  amostras independentes que foram geradas, e o desvio padrão das estimativas, calculado para o conjunto das amostras geradas. Quando o número de amostras  $m$  é elevado, este desvio-padrão pode ser considerado como o erro padrão do parâmetro populacional. São ainda disponibilizados a cobertura, que indica a proporção de amostras para as quais um intervalo a 95% contém o verdadeiro parâmetro, e a proporção de amostras nas quais o parâmetro se mostrou significativo. O enviesamento relativo na estimação de cada parâmetro,  $\theta$ , pode ser calculado utilizando o valor considerado como parâmetro populacional e a média das estimativas dos parâmetros obtidos nas várias amostras geradas, da seguinte forma

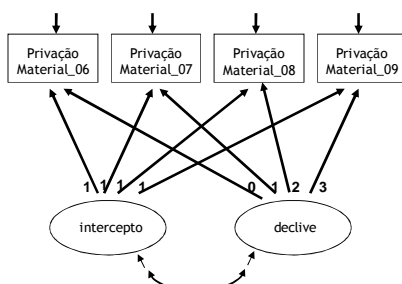
$$B(\hat{\theta}) = \frac{E(\hat{\theta}) - \theta}{\theta}.$$

## 4 Resultados

O presente trabalho propõe a modelação através de um modelo com trajetória latente de um *score* de privação material para os dados de 2006 a 2009. Apesar de a variável em estudo não ser contínua, para efeitos do presente trabalho foi assumida a normalidade da sua distribuição. Ao longo dos quatro momentos temporais em análise permaneceram no painel um total de 845 agregados, para os quais existe informação completa, designado “caso completo”. Mas há ainda um total de 848 agregados que participaram no estudo entre 2006 e 2008 e que não estiveram em 2009 devido ao esquema de rotatividade do painel. Ao conjunto destes dois grupos constituídos por um total de 1693 agregados designa-se por “caso com omissões”, no qual existem aproximadamente 50% de omissões no quarto momento temporal (2009).

A utilização de procedimentos de FIML permite estimar um modelo com trajetória latente (Figura 3) para modelar a trajetória do *score* de privação material entre 2006 e 2009, considerando a totalidade dos 1693 agregados, “caso com

omissões”. Também para os 845 agregados que constituem o “caso completo” é modelada a trajetória do *score* de privação material, no período considerado, recorrendo a um modelo com trajetória latente.



**Figura 3** - Diagrama de um modelo com trajetória latente linear de um *score* de privação material, medido em quatro momentos temporais: anos de 2006 a 2009.

Na Tabela 1 são apresentadas as medidas de qualidade do ajustamento modelados, para o “caso completo” e para o “caso com omissões”. Os valores obtidos para as várias medidas consideradas permitem concluir que existe um ajustamento razoável para as duas subamostras consideradas. Todavia, o valor do RMSEA para o “caso completo”, encontra-se no limite do razoável, enquanto que para o “caso com omissões” apresenta um valor superior a 0.08, valor recomendado na literatura da especialidade.

**Tabela 1** - Medidas de qualidade de ajustamento (“caso completo” e “caso com omissões”).

	<b>caso completo (845 agregados)</b>	<b>caso com omissões (1693 agregados)</b>
<b>RMSEA</b>	0.079	0.125
<b>CFI</b>	0.990	0.968
<b>TLI</b>	0.988	0.961
<b>SRMR</b>	0.025	0.046

**Tabela 2** - Estimativas dos parâmetros obtidas no ajustamento de um modelo com trajetória latente (“caso completo” e “caso com omissões”). A negrito estão os valores que se mostraram significativos.

	<b>caso completo (845 agregados)</b>	<b>caso com omissões (1693 agregados)</b>
<b>Média do intercepto (<math>\mu_{\alpha}</math>)</b>	<b>2.514</b>	<b>2.500</b>
<b>Média do declive (<math>\mu_{\beta}</math>)</b>	-0.025	<b>-0.106</b>
<b>Variância do intercepto (<math>\psi_{\alpha\alpha}</math>)</b>	<b>3.028</b>	<b>3.014</b>
<b>Variância do declive (<math>\psi_{\beta\beta}</math>)</b>	<b>0.123</b>	<b>0.168</b>
<b>Covariância intercepto/declive (<math>\psi_{\alpha\beta}</math>)</b>	<b>-0.293</b>	<b>-0.415</b>

As estimativas dos parâmetros do modelo com trajetória latente são apresentadas, para o “caso completo” e para “o caso com omissões”, na Tabela 2.

Para o conjunto dos 845 agregados que permaneceram no estudo de 2006 a 2009, é possível concluir que em 2006 o valor médio da privação é de 2.514 itens (valor da média do intercepto). O valor da média do declive não se mostrou significativo sugerindo que a taxa média de mudança na privação entre 2006 e 2009, embora negativa, não é significativa. As variâncias do intercepto e do declive mostraram-se ambas significativas, o que evidencia que nem todos os agregados têm o mesmo nível de privação material em 2006 e nem todos mudam da mesma forma, respetivamente. Para a covariância entre o intercepto e o declive obteve-se um valor negativo e significativo o que indicia que níveis mais elevados de privação em 2006 estão associados com uma taxa média de mudança mais lenta.

Para o conjunto dos 1693 agregados considerados obteve-se para a média do intercepto um valor de 2.500, o que permite dizer que em 2006 o valor médio de privação é de 2.5 itens. O valor obtido da média do declive é negativo e significativo, o que permite concluir que a taxa média de mudança na privação entre 2006 e 2009 diminui. As variâncias do intercepto e do declive revelaram-se significativas, o que permite concluir que nem todos os agregados têm o mesmo nível de privação material em 2006, e nem todos mudam da mesma forma, respetivamente. Para a covariância entre o intercepto e o declive o valor obtido é negativo e significativo, sendo possível dizer que níveis mais elevados de privação em 2006 estão associados com uma taxa média de mudança mais lenta. Embora, a estimativa da média do declive para o “caso com omissões” se tenha mostrado significativa, no “caso completo” isso não se verifica, permitindo conclusões distintas sobre a taxa média de mudança da privação material que os agregados experimentam entre 2006 e 2009, para as diferentes subamostras consideradas. Seguidamente, e para perceber qual das opções metodológicas deve o analista considerar, foi realizado um pequeno estudo de simulação em Mplus, recorrendo a procedimentos de Monte Carlo.

No estudo de simulação foram geradas 1000 amostras de dados, com 845 observações cada, com um padrão de 50% de omissões no último momento temporal, ano de 2009, recriando-se desta forma, uma amostra com um desenho semelhante à subamostra com todos os 1693 agregados. Os dados são gerados a partir de um modelo com trajetória latente com quatro momentos temporais, sendo utilizados como valores dos parâmetros populacionais que definem o modelo, os valores dos parâmetros que foram estimados pelo ajustamento de um modelo com trajetória latente para os dados que constituíam o “caso completo”.

Os resultados para a geração de 1000 amostras com 845 observações cada e com um padrão de omissão de 50% no quarto momento temporal são apresentados na Tabela 3. Os valores obtidos permitem concluir que numa percentagem de 28.8% das amostras geradas o parâmetro média do declive se mostrou significativo. Recorde-se que este é o parâmetro para o qual as conclusões se mostraram contraditórias, aquando da estimação dos modelos referentes ao grupo “caso completo” e “caso com omissões”. Quando calculado o enviesamento relativo na estimação dos parâmetros do modelo verifica-se que nenhum dos parâmetros

apresenta um valor de enviesamento superior a 0.05, ou 5% - valor a partir do qual o enviesamento é considerado não negligenciável, ver HOOGLAND & BOOMSMA (1998). Todavia, o parâmetro que apresenta um maior enviesamento é a média do declive (0.0320 ou 3.2%). Os valores de cobertura obtidos para todos os parâmetros são superiores a 0.90, valor considerado aceitável (COLLINS et al., 2001). Assim, é possível verificar que o parâmetro que, neste pequeno estudo de simulação, apresenta resultados menos consensuais é a média do declive.

**Tabela 3** - Resultados obtidos no estudo de simulação gerando 1000 amostras com 50% de omissões no quarto momento temporal.

	<b>estimativa média</b>	<b>desvio padrão médio</b>	<b>média dos erros padrão</b>	<b>cobertura</b>	<b>proporção de réplicas com coeficiente significativo</b>
$\mu_{\alpha}$	2.5160	0.0640	0.0628	0.938	1.000
$\mu_{\beta}$	-0.0258	0.0181	0.0183	0.945	0.288
$\psi_{\alpha\alpha}$	3.0326	0.1635	0.1679	0.955	1.000
$\psi_{\beta\beta}$	0.1220	0.0182	0.0187	0.950	1.000
$\psi_{\alpha\beta}$	-0.2929	0.0415	0.0429	0.961	1.000

## 5 Discussão

Este trabalho propôs a modelação longitudinal de um indicador de privação material, recorrendo a um modelo com trajetória latente, a dados de duas diferentes subamostras. Uma das subamostras correspondendo aos agregados que se mantiveram no estudo entre 2006 e 2009. A outra correspondendo a estes últimos mais os agregados que estiveram no estudo entre 2006 e 2008, mas não estiveram em 2009 devido ao esquema de rotatividade do painel ICOR. Assim, a primeira subamostra apresenta dados completos, e a segunda, dados com omissões.

Os resultados obtidos na estimação dos parâmetros de um modelo com trajetória latente permitem conclusões distintas, no que diz respeito à taxa média de mudança da privação material para as subamostras consideradas. Para o caso com omissões foi possível concluir que a taxa média de mudança na privação entre 2006 e 2009 diminui, isto porque o valor estimado para a média do declive se mostrou significativo. Todavia, para o caso completo não foi possível conclusão idêntica, porque apesar de a estimativa do parâmetro em questão assumir valor negativo não se mostrou significativa. Esta discrepância de conclusões mostrou a necessidade de um estudo de simulação que permita ao investigador saber qual das opções metodológicas deve considerar.

Assim, recorrendo a procedimentos de Monte Carlo foi implementado um estudo de simulação. Foram geradas amostras de dados com 50% de omissões no



quarto momento temporal, por forma a recriar um desenho amostral semelhante à subamostra “caso com omissões”. Os dados foram gerados a partir de um modelo com trajetória latente com quatro momentos temporais, sendo utilizados como valores populacionais dos parâmetros que definem o modelo, os valores estimados para esses mesmos parâmetros quando considerada a subamostra com dados completos. Com este estudo foi possível concluir que a média do declive é o parâmetro que apresenta maior enviesamento na estimação de um modelo com trajetória latente. Por outro lado, verificou-se que em 28.8% das amostras geradas este mesmo parâmetro se mostrou significativo. Deste modo, um estudo de simulação mais amplo revela-se necessário como trabalho futuro.

## Referências

- BOLLEN, K.E. & CURRAN, P. (2006). *Latent Curve Models: A Structural Equation Perspective*, John Wiley & Sons, Inc., New Jersey.
- COLLINS, L.M., SCHAFER, J.L. & KAM, C-M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330-351.
- ENDERS, C.K. (2010). *Applied Missing Data*, The Guilford Press, New York.
- GUIO, A-C (2009). What can be learned from material deprivation indicators in Belgium and in its regions, *Institut Wallon de L'Évaluation de la Prospective et de la Statistique*, 901.
- HOOGLAND, J.J. & BOOMSMA, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods and Research*, 26, 329-367.
- INE (2009). Inquérito às Condições de Vida e Rendimento - ICOR, *Documento Metodológico*.
- MUTHÉN, L.K. & MUTHÉN, B.O. (1998-2010). *Mplus user's guide*, 6<sup>th</sup> edition, Los Angeles, CA: Muthén & Muthén.
- SCHAFER, J.L. & GRAHAM, J. (2002). Missing Data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.
- SCHUMACKER, R.E & LOMAX, R.G. (2010). *A beginner's guide to Structural Equation Modelling*, 2<sup>nd</sup> edition, Lawrence Erlbaum Associates, Inc.
- YU, C.Y. & MUTHÉN, B.O. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. *Paper presented at the annual meeting of the American Educational Research Associations, New Orleans; LA*.

## Análise dos perfis de consumo de *cannabis* pelos adolescentes de Ponta Delgada

Áurea Sousa<sup>1</sup> · Helder Rocha Pereira<sup>2</sup> · Sara Raposo<sup>3</sup> · Osvaldo Silva<sup>4</sup>  
Helena Bacelar-Nicolau<sup>5</sup>

© The Author(s) 2017

**Resumo** A *cannabis* é a droga ilícita mais produzida e consumida na Europa e, embora seja uma “droga leve”, é reconhecido o seu impacto nas alterações de memória, nas sensações e nos comportamentos. Apresentam-se as principais conclusões obtidas, com base num questionário e em métodos de Análise de Dados (do univariado ao multivariado), com o objetivo de estimar a prevalência e determinar os perfis de consumo de *cannabis* por parte dos estudantes do Ensino Secundário do concelho de Ponta Delgada (Açores).

**Palavras-chave:** Análise de correspondências múltiplas, Análise classificatória hierárquica ascendente, *Cannabis*, Adolescentes, Ensino Secundário.

### 1 Introdução

A *cannabis* é a droga ilícita mais produzida e consumida em toda Europa, atingindo as maiores prevalências na faixa etária dos 15 aos 24 anos de idade (OEDT, 2010). O desenvolvimento da dependência da *cannabis* manifesta-se de uma forma mais gradual comparativamente ao de outras drogas (Wagner *et al.*, 2005; OEDT, 2010). No entanto, tem-se observado um crescente número de jovens que recorrem a instituições de saúde para desintoxicação de *cannabis*, sendo esta a segunda droga mais notificada, a seguir à heroína (OEDT, 2010). O consumo de *cannabis* está

---

<sup>1</sup>Universidade dos Açores, Departamento de Matemática, CEEAplA, aurea.st.sousa@uac.pt

<sup>2</sup>Universidade dos Açores, Escola Superior de Enfermagem de P. Delgada, helder.ja.pereira@uac.pt

<sup>3</sup>Hospital do Divino Espírito Santo de Ponta Delgada, srf\_raposo@hotmail.com

<sup>4</sup>Universidade dos Açores, Departamento de Matemática, CICS.NOVA.UAc, osvaldo.dl.silva@uac.pt

<sup>5</sup>Universidade de Lisboa, Faculdade de Psicologia; DataScience, hbacelar@psicologia.ulisboa.pt

relacionado com o surgimento de esquizofrenia, outras psicoses (Degenhardt *et al.*, 2009) e alterações do hipocampo, nomeadamente nas áreas da memória e das sensações (Schwilke *et al.*, 2009). Para além disso, o THC (delta-9-tetrahidrocannabinol) provoca alterações do humor, reações tóxicas, acentuada ansiedade e, possivelmente, efeitos nocivos mentais e fisiológicos nos consumidores mais jovens (Schukit, 1998). O consumo de *cannabis* é particularmente agressivo na adolescência, pelo facto da sua substância ativa (THC) se acumular em tecidos gordos de órgãos que ainda estão em formação nesta etapa da vida (Schuckit, 1998). O seu consumo regular na adolescência pode ter efeitos adversos na saúde mental dos jovens adultos, havendo dados que apontam para um maior risco de sintomas psicóticos e de perturbações que aumentam com a frequência do consumo (Hall e Degenhardt, 2009; Moore *et al.*, 2007).

O presente estudo visa estimar a prevalência e determinar os perfis de consumo de *cannabis* por parte dos estudantes do Ensino Secundário do concelho de Ponta Delgada (Açores). A Secção 2 aborda a metodologia e as variáveis do estudo, enquanto na Secção 3 são apresentados os principais resultados obtidos. Finalmente, a Secção 4 contém as principais conclusões relativas ao estudo desenvolvido.

## 2 Metodologias e variáveis do estudo

O presente estudo foi efetuado com base num questionário, apresentado em Raposo (2011), o qual foi submetido a um pré-teste e a uma validação de conteúdo. Ao longo de toda a investigação, salvaguardou-se o direito da livre participação dos intervenientes, assim como o anonimato e a confidencialidade dos dados (Raposo *et al.*, 2012).

**Tabela 1** – Repartição na população e composição da amostra.

	Feminino		Masculino		Total	
	População	Amostra	População	Amostra	População	Amostra
<b>10º Ano</b>	436	131	403	121	839	252
<b>11º Ano</b>	400	120	299	90	699	210
<b>12º Ano</b>	377	113	267	80	644	193
<b>Total</b>	<b>1213</b>	<b>364</b>	<b>969</b>	<b>291</b>	<b>2182</b>	<b>655</b>

Devido a alguns obstáculos no terreno, motivados essencialmente pela temática da investigação, e conhecendo-se a distribuição dos estudantes segundo as variáveis “Ano de escolaridade” e “Sexo”, relevantes para o estudo, optou-se por uma amostragem por quotas, considerando uma fração de amostragem exceçãoalmente elevada (0.3), de forma a se obter uma amostra representativa da população (ver Tabela 1) e não muito pequena (como os estudantes que já experimentaram *cannabis* são em menor número do que os que nunca experimentaram, pretendeu-se

evitar a sua sub-representação na amostra). Sempre que um estudante se recusou a responder ao questionário foi imediatamente substituído por outro com características equivalentes. Em conformidade com as quotas estabelecidas, a amostra é constituída por 655 estudantes, do Ensino Secundário do concelho de Ponta Delgada, abrangendo 30% da população, estratificada em função do ano de escolaridade (252 (38.5%) dos respondentes são do 10º ano, 210 (32%) são do 11º ano e 193 (29.5%) são do 12º ano) e do género (364 (55.6%) são do sexo feminino e 291 (44.4%) do sexo masculino). As idades dos inquiridos estão compreendidas entre os 16 e os 22 anos, sendo a média de idades de cerca de 17 ( $\pm 1.1$ ) anos. Do total de inquiridos, 30% residem em meio urbano e 70% em meio rural (geralmente, em freguesias próximas da cidade).

O questionário inclui, entre outras variáveis, a idade, o género, o ano de escolaridade, a área de residência, a frequência de consumo, a idade de início de consumo, o local de consumo, o(s) companheiro(s) de consumo, a profissão dos pais, a relação do adolescente com os pais, a prática de atividades extracurriculares, a autoperceção do comportamento escolar, a frequência de repreensão pelo diretor de turma/conselho executivo dos estudantes, o consumo de outras drogas e a opinião sobre a gravidade para a saúde decorrente do consumo de *cannabis*.

Para categorizar as profissões, recorreu-se ao Sistema Nacional de Profissões (SNP, 1995). O grupo que se denominou por A inclui os grandes grupos 1 (Quadros superiores de administração pública e empresas) e 2 (Especialistas das profissões intelectuais); o Grupo B refere-se aos grandes grupos 3 (Técnicos profissionais de nível intermédio), 4 (Administrativos e similares) e 5 (Pessoal dos serviços e vendedores); por fim, o grupo C diz respeito aos grandes grupos 6 (Agricultores), 7 (Operários e artífices), 8 (Operadores) e 9 (Trabalhadores não qualificados). De salientar que, para fins de tratamento de dados, optou-se pela consideração do progenitor com a categoria profissional mais elevada (Raposo, 2011).

Os dados recolhidos foram analisados usando métodos de Estatística Descritiva e de Estatística Inferencial (testes não paramétricos), a Análise de Correspondências Múltiplas (ACM) e a Análise Classificatória Hierárquica Ascendente (ACHA).

A ACM é uma técnica de análise da interdependência que tem como objetivo o estudo das associações entre as categorias de diversas variáveis qualitativas simultaneamente. A ACM é a generalização da Análise de Correspondências Simples (ACS) para matrizes com mais de dois fatores de classificação, sendo as associações entre as variáveis determinadas com base na distância do qui-quadrado entre as categorias das variáveis e entre os indivíduos. Os gráficos produzidos são interpretáveis com base nas contribuições de cada categoria para os eixos (dimensões) e nas proximidades e oposições entre as projeções das categorias nos eixos, permitindo avaliar visualmente se as variáveis de interesse se afastam do pressuposto de independência e sugerindo possíveis associações (e.g., Benzécri (1992); Greenacre and Blasius (2006)). O *output* da ACM apresentado na seguinte

secção, sob a forma gráfica (mapa percetual), foi obtido usando o procedimento “*Optimal Scaling*” do SPSS (*Statistical Package for Social Sciences*).

Em Análise Classificatória (*Cluster Analysis*), pretende-se identificar grupos (classes/*clusters*) relativamente homogéneos, com base nas (dis)semelhanças entre as entidades. Os métodos hierárquicos produzem famílias de classificações encaixadas, geralmente hierarquias de partições, e podem ser subdivididos em aglomerativos ou ascendentes e divisivos ou descendentes. Um caso frequente diz respeito a dados em que  $n$  objetos /indivíduos são descritos pela presença ou ausência de  $p$  características, em que  $x_{ij}=1$  se  $i$  possui o atributo  $j$  e  $x_{ij}=0$ , caso contrário. Diversos coeficientes de proximidade para dados binários são definidos com base numa tabela de contingência  $2 \times 2$  associada a um par de elementos do conjunto a classificar (Gordon, 1999). Neste trabalho, é efetuada a ACHA de variáveis relativas ao consumo de drogas, com base no coeficiente de Ochiai, caso particular do coeficiente de afinidade (e.g., Bacelar-Nicolau, 1980, 1987; Sousa, 2005) quando os dados são binários, e em três critérios de agregação clássicos (o da ligação simples, o da ligação completa e o da ligação média).

### 3 Apresentação de resultados

Verificou-se que 36.8% dos adolescentes da amostra já consumiram *cannabis* alguma vez na vida, constatando-se a maior prevalência nos adolescentes do sexo masculino (rácio de consumo entre os géneros de 1.24 rapazes para cada rapariga). A média de idades de início ao consumo nas raparigas foi de 15.05 ( $\pm 1.51$ ) anos, enquanto nos rapazes foi de 14.98 ( $\pm 1.49$ ) anos. Constatou-se, ainda, que 51.8% dos estudantes do 12º ano já experimentaram o consumo de *cannabis* comparativamente a 27.8% do 10º ano. Dos adolescentes que experimentaram o consumo de *cannabis*, a maior parte (92.4%) consumiu um charro pela primeira vez com amigos, 3.8% com familiares, 2.5% sozinhos e 1.3% com outros. Estes resultados eram expectáveis, já que, como defende Morel *et al.* (1998), é raro que a iniciação a uma droga se faça fora de um grupo. Diversos estudos (e.g., Kuntshe *et al.*, 2006; Chabrol *et al.*, 2006; Stephens *et al.*, 2009; Duarte *et al.*, 2006; Pérez *et al.*, 2010; Shehu *et al.*, 2008; Chen *et al.*, 2006) são unânimes quanto ao facto do consumo de *cannabis* ser grandemente influenciado pelas atitudes e opiniões do grupo de pares, constituindo-se este como o principal fator preditivo.

De acordo com os inquiridos que já experimentaram a *cannabis*, o principal motivo que os levou a consumir pela primeira vez foi a curiosidade (79.6%), a que se seguem a influência de pares (10.2%) e os problemas (6.6%). Os outros motivos referidos (por acharem “*bom*”, para “*celebrar*”) tiveram percentagens residuais (respetivamente, 1.8% e 0.9%). Quanto ao local de consumo, 21% consumiu, pela primeira vez, em casa (própria ou de amigos) e 20.5% fê-lo, pela primeira vez, em ambientes noturnos. Os restantes locais de iniciação ao consumo foram a rua

(19.7%), a escola (17.5%), o jardim (16.6%), a praia (3.5%) e os acampamentos (1.3%).

**Tabela 2** – Variáveis cuja associação com o consumo de *cannabis* foi significativa.

Variáveis	$\chi^2$	<i>p-value</i>
Gênero ( <i>Feminino, Masculino</i> )	4.44	0.035
Maior qualificação profissional dos pais ( <i>Grupo A, Grupo B, Grupo C</i> )	9.73	0.008
Relação familiar [ <i>Má</i> (R. M.), <i>Razoável</i> (R. R.); <i>Boa</i> (R. B.)]	31.34	0.000
Já reprovaste em algum ano? [ <i>Não</i> (N. Rep.), <i>Sim</i> (Rep.)]	5.83	0.016
Autoperceção do comportamento [ <i>Mau</i> (C. M.), <i>Razoável</i> (C. R.); <i>Bom</i> (C. B.)]	23.92	0.000
Consumo de álcool [ <i>Não</i> (N. C. Álcool), <i>Sim</i> (C. Álcool)]	32.82	0.000
Consumo de tabaco [ <i>Não</i> (N. C. Tabaco), <i>Sim</i> (C. Tabaco)]	173.28	0.000
Consumo de heroína [ <i>Não</i> (N. C. Heroína), <i>Sim</i> (C. Heroína)]	48.80	0.000
Consumo de cocaína [ <i>Não</i> (N. C. Cocaína), <i>Sim</i> (C. Cocaína)]	58.24	0.000
Consumo de anfetaminas (“ <i>speed, crystal e pastilhas</i> ”)? [ <i>Não</i> (N. C. Anfetaminas), <i>Sim</i> (C. Anfetaminas)]	83.52	0.000
Opinião sobre o consumo de charros ( <i>Muito Prejudicial</i> (M. Prej.), <i>Prejudicial</i> (Prej.), <i>Nem Prejudicial Nem Benéfico</i> (NPNB), <i>Benéfico</i> (Benef.), <i>Muito Benéfico</i> (M. Benef.))	170.74	0.000
Já foste chamado ao gabinete do diretor de turma ou do conselho executivo? ( <i>Não Fui chamado</i> (N.F. Cham.), <i>Fui chamado</i> (F. Cham.))	63.65	0.000

Usando o teste de independência do qui-quadrado, não se observou qualquer associação significativa entre a prática de atividades extracurriculares e o consumo de *cannabis* ( $\chi^2=0.009$ ; *p-value*=0.926). Quanto à área de residência, observou-se que de entre os 238 adolescentes que já consumiram *cannabis*, alguma vez na vida, 34.5% residem no meio urbano e 65.5% no meio rural. A aplicação do teste das probabilidades exatas de Fisher permite-nos concluir que a proporção de consumidores residentes no meio rural é significativamente superior à do meio urbano (*p-value* =0.037<0.05). Estes resultados estão de acordo com os obtidos por Chen e Killea-Jones (2006), que, no seu estudo, observaram que o consumo de *cannabis* era significativamente superior nos estudantes do meio rural (39.1%) comparativamente aos do meio urbano (28.5%).

Usando o teste de independência do qui-quadrado, foram observadas associações estatisticamente significativas entre o “consumo de charros” (*cannabis* e seus derivados) e as variáveis apresentadas na Tabela 2. Neste contexto, é de salientar, por exemplo, associações significativas entre o consumo de *cannabis* e uma mais elevada categoria profissional dos pais; a autoperceção de uma relação familiar de fraca qualidade; a autoperceção de comportamentos escolares desadequados e a reprovação do aluno em algum ano. Para os adolescentes cuja categoria profissional mais elevada dos pais se insere no Grupo A (respetivamente, Grupo B e C) a prevalência de consumo foi de 44.2% (respetivamente, 35.7% e 29.8%). É especialmente notória a diferença percentual entre a prevalência de consumo nos adolescentes cuja categoria profissional mais elevada dos pais pertence ao Grupo A e nos adolescentes cuja categoria profissional mais elevada dos pais pertence ao Grupo C. O facto da prevalência do consumo variar em função

da categoria profissional dos pais era expectável, já que maiores recursos financeiros podem facilitar o acesso a esta substância, embora existam diversas variáveis associadas ao “consumo de charros” (e.g., culturais, entre outras, algumas das quais referidas nesta secção).

Dos inquiridos que, alguma vez na vida, consumiram um charro apenas 13.4% admitiram que o seu consumo é muito prejudicial, enquanto dos que nunca consumiram esta droga 45.6% admitiram que o seu consumo é muito prejudicial. É de referir, ainda, que dos inquiridos que consideram o consumo de *cannabis* benéfico ou muito benéfico, 91.5% afirmaram ser consumidores ou já terem experimentado o consumo de charros. Dos adolescentes que consideram o consumo de *cannabis* prejudicial ou muito prejudicial, a maioria (78.2%) nunca consumiu charros. Este facto apoia a tese de que os adolescentes que têm uma opinião positiva acerca do consumo de *cannabis* e seus derivados estão mais predispostos ao seu consumo. Em consonância com estes resultados estão os obtidos por Chabrol *et al.* (2006), Duarte *et al.* (2006) e Stephens *et al.* (2009), que, de igual modo, observaram existir relação entre a percepção do adolescente acerca do consumo de *cannabis* e o seu consumo, sendo que opiniões positivas acerca do mesmo funcionam como fatores preditivos. Neste sentido, torna-se fundamental uma intervenção primária junto dos adolescentes, de forma a informá-los e esclarecê-los acerca dos efeitos e consequências do consumo desta droga ilícita.

Cerca de 44.7% dos adolescentes que já consumiram *cannabis*, pelo menos uma vez na vida, foram chamados ao gabinete do diretor de turma ou do conselho executivo (versus 16% dos que nunca consumiram). Foram, ainda, identificadas associações significativas entre o consumo de *cannabis* e outras drogas, tais como o álcool ( $p\text{-value}=0.000$ ), o tabaco, a heroína, a cocaína e as anfetaminas (“*speed*, *crystal* e *pastilhas*”), que são as mais comuns nesta faixa etária.

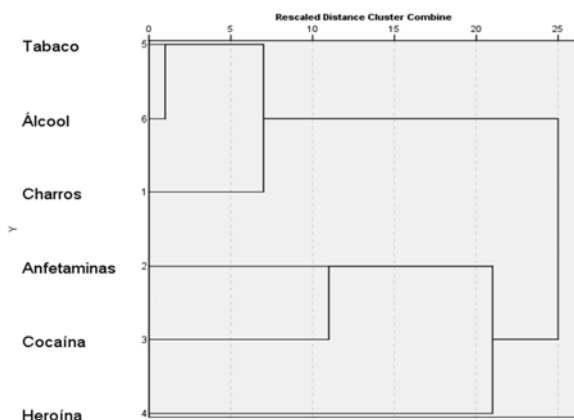
Foi efetuada a Análise de Correspondências Múltiplas (ACM), considerando como variáveis ativas as variáveis referentes ao consumo de charros, álcool, tabaco, heroína, cocaína e anfetaminas e a opinião dos adolescentes sobre o consumo de charros. Foram, ainda, consideradas as restantes variáveis da Tabela 2, à exceção do género, como variáveis suplementares, com o intuito de relacionar o consumo de charros com essas variáveis.





cham.)” estão relativamente próximas da categoria “*Consumo de charros*”, o que está de acordo com o referido previamente. É de salientar, ainda, que os alunos que não consomem charros tendem a ter uma relação familiar boa (R.B.) e a autopercecionarem o seu comportamento escolar como bom (C.B.). Relativamente à dimensão 2, as categorias que se revelaram mais dominantes foram, por ordem decrescente, “C. Heroína”, “C. Cocaína”, “M. Benef.”, “N. C. Álcool”, “C. Anfetaminas” e “C. M.”.

A ACHA das variáveis relativas ao consumo de charros, álcool, tabaco, heroína, cocaína e anfetaminas, com base no coeficiente de Ochiai e nos critérios de agregação da ligação simples, da ligação completa e da ligação média, forneceu dendrogramas muito similares, os quais apontaram, no nível três, para uma partição em três classes: C1: [Álcool, Tabaco, Charro]; C2: {Cocaína, Anfetaminas}; C3: {Heroína}. A classe 1 vai ao encontro da ideia de que o consumo de *cannabis* e seus derivados está associado ao consumo de outras drogas, tais como o tabaco e o álcool, sendo de salientar que o dendrograma associa a *cannabis* a estas drogas consideradas legais e apenas muito mais tarde às ilícitas.



**Figura 2** – Dendrograma obtido usando o critério da ligação média.

A Figura 2 mostra o dendrograma obtido com o critério da ligação média, onde podemos verificar que a droga que mais se distingue das restantes, em termos do seu consumo (*Sim/Não*), é a Heroína. Não são apresentadas, no presente trabalho, a análise classificatória dos indivíduos (articulação entre a ACM e o método não hierárquico das *k*-médias, seguida da análise classificatória hierárquica dos grupos resultantes – análise de dados complexos/simbólicos) e a análise detalhada da sua relação com a das variáveis.

## 4 Conclusão

Os métodos de análise de dados multivariados utilizados permitiram, já nesta primeira abordagem, obter ou confirmar alguns resultados relevantes sobre os dados. A aplicação e interpretação mais desenvolvidas destes e de outros métodos de classificação, tais como os métodos probabilísticos da Validade da Ligação (e.g., Bacelar-Nicolau, 1987), quer sobre o conjunto das variáveis quer sobre o dos indivíduos, poderão fornecer uma mais rica informação sobre a amostra estudada.

Os resultados obtidos permitiram, além de uma estimativa da prevalência do consumo de *cannabis* na população em estudo, no ano letivo de 2009-2010, identificar o perfil dos adolescentes mais predispostos ao consumo, definido nomeadamente, por características como o ano de escolaridade, contexto familiar e género, entre outras variáveis.

A maioria dos adolescentes que experimentaram outras drogas, já havia experimentado ou mantinha o consumo de *cannabis*, mostrando que apesar de não haver relação direta de causa-efeito a *cannabis* poderá funcionar como porta de entrada para o consumo de outras drogas ilícitas. A maior parte dos indivíduos que experimentaram o consumo de *cannabis*, fizeram-no na companhia de amigos, pelo que o planeamento de atividades e estratégias para a prevenção do consumo e promoção da saúde se deve realizar em contexto de grupo de amigos, dirigida ao grupo como um todo e não em particular. Um conhecimento desta natureza é crucial para o desenho de intervenções direcionadas ao foco dos problemas, aumentando a possibilidade de sucesso e evitando custos desnecessários.

Segundo o Inquérito Nacional em Meio Escolar de 2011 (3º Ciclo), o consumo de bebidas alcoólicas, de tabaco e, em menor grau, o de *cannabis* aumentou entre os alunos das escolas públicas, tendo-se mantido (com tendência para descida) as prevalências das “outras drogas”. A dimensão e a gravidade do problema tornam premente uma atuação em saúde pública direcionada para a prevenção do consumo da *cannabis*. Assim, torna-se fundamental uma intervenção primária junto dos adolescentes, no sentido de informá-los e esclarecê-los acerca dos efeitos e consequências a curto e longo prazo do consumo desta e de outras drogas.

## Referências

- BACELAR-NICOLAU, H. (1980). *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória*, Tese de Doutoramento, Universidade de Lisboa, Portugal.
- BACELAR-NICOLAU, H. (1987). *On the distribution equivalence in cluster analysis*, in DEVIJVER, P.A. & KITTLER, J. (Eds.) *Pattern Recognition Theory and Applications*, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, New York, Springer-Verlag, pp. 73-79.

- BENZÉCRI, J.P. (1992). *CORRESPONDENCE ANALYSIS HANDBOOK*. NEW YORK: MARCELL DEKKER.
- CHABROL, H., CHAUCHARD, E., MABILA, J.D., MANTOULAN, R., ADÈLE, A. & ROUSSEAU, A. (2006). Contributions of social influences and expectations of use to cannabis use in high-school students, *Journal of Intelligent Information Systems*, 31, 2116-2119.
- CHEN, K.W. & KILLEA-JONES, L.A. (2006). Understanding Differences in Marijuana Use Among Urban Black and Suburban White High School Students from Two U.S. Community Samples, *Journal of Ethnicity in Substance Abuse*, 5(2), 51-73.
- DEGENHARDT, L., HALL, D., LYNSKEY, M., MCGRATH, J., MCLAREN, J., CALABRIA, B. & WHITEFORD, H. (2009). Should burden of disease estimates include cannabis use as a risk factor for psychosis? *Plos Medicine*, 6(9), 100-133.
- DUARTE, R., ESCARIO, J. J. & MOLINA, J. A. (2006). MARIJUANA CONSUMPTION AND SCHOOL FAILURE AMONG SPANISH STUDENTS, *ECONOMICS OF EDUCATION REVIEW*, 25, 472-481.
- GORDON, A.D. (1999). *CLASSIFICATION*, 2<sup>ND</sup> ED: LONDON: CHAPMAN & HALL.
- GREENACRE, M.; BLASIUS, J. (2006). *MULTIPLE CORRESPONDENCE ANALYSIS*. CHAPMAN AND HALL/CRC.
- HALL, W. & DEGENHARDT, L. (2009). Adverse health effects of non-medical cannabis use. *The Lancet*, 374 (9698), 1383-1391.
- KUNTSCHKE, E. & JORDAN, M.D. (2006). Adolescent alcohol and cannabis use in relation to peer and school factors. *Drug and Alcohol Dependence*, 84, 167-174.
- MOORE, T., ZAMMIT, S., LINGFORD-HUGHES, A., BARNES, T., JONES, P., BURKE, M. & LEWIS, G. (2007). Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. *The Lancet*, 370 (9584), 319-328.
- OEDT (2010). *Relatório Anual do Observatório Europeu das Drogas e da Toxicodependência (OEDT): A EVOLUÇÃO DO FENÓMENO DA DROGA NA EUROPA*. European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), Lisboa, <http://www.emcdda.europa.eu/publications/annual-report/2010>, (acedido em 12 de fevereiro 2012).
- PÉREZ A., ARIZA C., SÁNCHEZ-MARTÍNEZ F. & NEBOT M. (2010). Cannabis consumption initiation among adolescents: a longitudinal study, *Addictive Behaviors (addict behav)*, 35(2): 129-34.
- RAPOSO, S. (2011). *Consumo de derivados de Cannabis Sativa pelos jovens do ensino secundário da área escolar de Ponta Delgada*, Tese de Mestrado, Universidade dos Açores, Portugal.
- RAPOSO, S., SOUSA, Á. & PEREIRA, H. (2012). Um Estudo sobre Consumo de Derivados de Cannabis Sativa pelos Adolescentes do Ensino Secundário de Ponta Delgada, IN *Livro de Resumos das XIX Jornadas de Classificação e Análise de Dados (JOCLAD 2012)*, 197-200, LISBOA, INE.
- SCHUKIT, M. A. (1998). *Abuso de álcool e drogas*, Lisboa, Climepsi Editores.
- SCHWILKE, E., SCHWOPE, D., KARSCHNER, E., LOWER, R., DARWIN, W., KELLY, D., GOODWIN, R., GORELICK, D. & HUESTIS, M. (2009). Delta-9-

- Tetrahydrocannabinol (THC), 11-Hydroxy-THC, and 11-Nor-9-carboxy-THC Plasma Pharmacokinetics during and after Continuous High-Dose Oral THC. *Clinical Chemistry*, 55 (12), 2180-2189.
- SHEHU, A.U. & IDRIS, S.H. (2008). Marijuana smoking among secondary school students in Zaria, Nigeria: factors responsible and effects on academic performance, *Annals of African Medicine [Ann Afr Med]*, 7 (4), 175-179.
- SOUSA, Á. (2005). *Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa*, Tese de Doutoramento, Universidade dos Açores, Portugal.
- STEPHENS, P.C., SLOBODA, Z., STEPHENS, R.C., TEASDALE, B., GREY, S. F., HAWTHORNE, R.D. & WILLIAMS, J. (2009). Universal school-based substance abuse prevention programs: Modeling targeted mediators and outcomes for adolescent cigarette, alcohol and marijuana use, *Drug Alcohol Depend*, 1; 102 (1-3), 19-29.
- WAGNER F., VELASCO-MONDRAGÓN, H., HERRERA-VÁZQUEZ, M., BORGES, G. & LAZCANO-PONCE, E. (2005). Early Alcohol or Tobacco Onset and Transition to Other Drug Use Among Students in the State of Morelos, MEXICO, *Drug and Alcohol Dependence*, 77 (1), 93-96.



# Seleção robusta em modelos de regressão linear com um grande número de preditores

Shirin Shahriari<sup>1</sup> · Susana Faria<sup>2</sup> · A. Manuela Gonçalves<sup>3</sup>

© The Author(s) 2017

**Resumo** Neste trabalho discute-se o problema de selecção de variáveis em modelos de regressão linear que envolvem um grande número de preditores, contaminados por *outliers* e observações atípicas. Como os métodos clássicos de selecção de variáveis não são resistentes à presença de *outliers* e outros tipos de contaminação, neste estudo são analisados métodos robustos de selecção de variáveis em modelos de regressão linear que envolvem um grande número de preditores. Estudos de simulação são realizados para avaliar e comparar o desempenho dos métodos de selecção de variáveis apresentados.

**Palavras-chave:** Regressão Linear Robusta, Selecção Robusta de Variáveis, *Outliers*.

## 1 Introdução

Os modelos de regressão têm vindo a ser aplicados em diferentes áreas do conhecimento: Medicina, Engenharia, Ambiente, etc. Cada vez é mais fácil recolher/monitorizar dados nas várias áreas. Assim, o estudo de modelos de regressão linear envolvendo um grande número de preditores é cada vez mais importante, em particular estudar métodos para seleccionar/identificar os preditores (quando há um grande número de preditores observados, muitas vezes

---

<sup>1</sup>CMAT-Centro de Matemática, Universidade do Minho, Portugal, [shirin.shahriari22@gmail.com](mailto:shirin.shahriari22@gmail.com)

<sup>2</sup>CMAT-Centro de Matemática, DMA-Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, [sfaria@math.uminho.pt](mailto:sfaria@math.uminho.pt)

<sup>3</sup>CMAT-Centro de Matemática, DMA-Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, [mneves@math.uminho.pt](mailto:mneves@math.uminho.pt)

correlacionados) que mais contribuem para a explicação da variável resposta, por forma a obter-se um modelo mais parcimonioso.

Na literatura, o estudo de inferência nos modelos de regressão linear tem sido considerado fundamentalmente sob a distribuição Normal. Contudo, em muitas situações, a inferência sob a normalidade é imprópria, por exemplo quando os dados provêm de uma distribuição com caudas mais ou menos pesadas que a distribuição Normal. Quando os dados contêm observações atípicas é também necessário a aplicação de métodos robustos de selecção de variáveis que são resistentes à presença de *outliers*, i.e., métodos em que a selecção correcta dos preditores não é influenciada por estas observações discordantes ou atípicas. Sabe-se que uma pequena proporção de *outliers* nos dados pode afectar os resultados dos critérios de selecção de modelos como o critério de informação de Akaike (AIC) (Akaike (1970)), o critério  $C_p$  de Mallows (Mallows (1973)) ou o critério de informação Bayesiana (BIC) (Schwarz (1978)).

Recentemente métodos alternativos robustos para a selecção de variáveis têm sido propostos na literatura: abordagens com base em critérios de selecção clássicos nomeadamente o critério AIC,  $C_p$  de Mallows, etc. (Ronchetti, 1985, Ronchetti e Staudte, 1994, Muller e Welsh, 2005). Adicionalmente, existem outros métodos que se baseiam em técnicas robustas de selecção de modelos como a validação cruzada ou em técnicas de selecção que diagnosticam simultaneamente o modelo correto e as observações atípicas (ver Ronchetti *et al.* (1997) e Morgenthaler *et al.* (2003)).

Salibian-Barrera e Van Aelst (2008) usam *Bootstrap* robusto para obterem um método rápido de selecção de modelos baseado na metodologia de *Bootstrap* que é exequível, em termos computacionais, para um grande número de preditores. Quando se tem um grande número de preditores é mais exequível e computacionalmente mais eficiente a utilização de métodos de selecção de variáveis baseados na introdução/eliminação de variáveis progressivamente no modelo tendo em conta a sua importância, seleccionando o melhor subconjunto de variáveis sequencialmente, adicionando ou removendo variáveis em cada passo, como acontece nos métodos de selecção *forward* e de eliminação *backward* (Weisberg (1985) ou Miller (2002)).

O método *Least Angle Regression* (LARS), uma versão modificada do método *forward stagewise*, proposto por Efron *et al.* (2004) é uma poderosa e eficiente, em termos computacionais, técnica de selecção sequencial de variáveis. Khan *et al.* (2007a) propuseram um método robusto de selecção de modelos baseado no método LARS, usando estimativas robustas de correlação entre os preditores e a variável resposta.

Neste trabalho apresenta-se um método robusto de selecção de variáveis em modelos de regressão que envolvem um grande número de preditores, e o seu desempenho é avaliado realizando estudos de simulação.

## 2 Estratégia robusta de selecção de modelos

As abordagens de selecção robusta de modelos mencionadas, que se baseiam em critérios de selecção robustos, podem ser resumidas em dois passos. No primeiro passo, as variáveis explicativas sem importância são eliminadas obtendo-se um modelo de regressão com um conjunto reduzido de variáveis (há uma redução prévia da dimensão). Estas variáveis entram sequencialmente no modelo de regressão, obtendo-se um modelo de estimação com as primeiras  $d$  variáveis consideradas mais importantes para o modelo. Os métodos de redução de dimensionalidade procuram portanto sintetizar em poucas variáveis a informação contida nas variáveis originais. No segundo passo, como a dimensão dos preditores foi reduzida a um tamanho exequível/viável, as técnicas robustas podem ser aplicadas para testarem a significância da inserção dos preditores (do conjunto reduzido de variáveis obtido) no modelo de estimação.

## 3 *Least Angle Regression* (LARS)

O algoritmo da regressão LARS foi proposto por Efron *et al.* (2004). Neste algoritmo as variáveis entram no modelo pela ordem da sua importância. LARS está relacionado com a selecção *forward stagewise*, que adiciona sequencialmente as variáveis no modelo em cada passo. A selecção *forward stagewise* começa com a constante  $\beta_0$  e sequencialmente adiciona ao modelo o preditor mais correlacionado com a variável resposta que melhora o ajustamento. Usualmente, a selecção *forward stagewise* envolve um elevado número de passos usando a mesma variável antes da introdução de uma outra variável, que produz uma maior correlação com o restante resíduo. Assim, a selecção *forward stagewise* não é eficiente em termos computacionais. Pela reformulação de um problema matemático simples de optimização, LARS reduz muito o número de passos do processo e acelera o tempo de computação (LARS é um algoritmo eficiente).

Considere o conjunto de dados

$$(y_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{ip}) = (y_i, \mathbf{x}_i), i = 1, \dots, n$$

onde  $y_i \in IR$  e  $\mathbf{x}_i \in IR^p$  são a variável resposta e as variáveis predictoras, respectivamente. Pretende-se ajustar aos dados o seguinte modelo linear

$$y_i = \beta_0 + \boldsymbol{\beta}^t \mathbf{x}_i + e_i$$

com  $\boldsymbol{\beta} \in IR^p$  e assume-se que os erros  $e_i$  têm  $E(e_i) = 0$  e  $Var(e_i) = \sigma^2$ .

Considere-se  $\beta_0 = 0$ . Esta condição é imposta por forma a centrar as variáveis predictoras e a variável resposta.



De uma forma sucinta, os passos a seguir para a aplicação do algoritmo LARS são os seguintes (Efron *et al.* (2004)):

1. Começar com todos os coeficientes iguais a zero e considerar o vector inicial dos resíduos igual ao vector da variável resposta  $\mathbf{y}$ ;
2. Encontrar o preditor mais correlacionado com a variável resposta, seja  $\mathbf{x}_1$ ;
3. Considerar o maior passo possível na direcção deste preditor até que algum outro preditor, seja  $\mathbf{x}_2$ , esteja fortemente correlacionado com o actual resíduo;
4. Prosseguir na direcção equiangular dos dois preditores até uma terceira variável  $\mathbf{x}_3$  ganhar o seu caminho para o "mais correlacionado" conjunto. A direcção equiangular é a direcção em que a correlação dos preditores diminui ao mesmo ritmo de modo a que essas correlações permanecem iguais em todos os momentos;
5. Repetir os passos até que todos os preditores sejam introduzidos.

### 3.1 LARS robusta

Como o método LARS se baseia em medidas como a média, a variância e a correlação das observações, é um método sensível à presença de contaminação nos dados. Assim, Khan *et al.* (2007) propuseram um LARS robusto (RLARS) substituindo as médias, as variâncias e as correlações dos dados por medidas robustas equivalentes. Como medidas robustas de localização e de escala usaram a mediana e o desvio absoluto médio, (MAD), respectivamente. Para a medida de correlação introduziram um estimador de correlação bivariado robusto baseado na winsorização bivariada (uma generalização da winsorização univariada como a introduzida por Huber (2009)).

A winsorização bivariada de dados bivariados estandardizados de forma robusta é baseada na matriz inicial de correlações bivariadas robustas  $R_0$  e na elipse de tolerância correspondente. Por exemplo, para  $\mathbf{x} = (x_1, x_2)^t \in \mathbb{R}^2$ , considere-se a distância de Mahalanobis  $D(\mathbf{x})$  baseada na matriz inicial de correlações  $R_0$  e considere-se o valor da constante “tuning”  $c = 5,99$ , que é o quantil de 95 % da distribuição qui-quadrado. Aplicando a transformação bivariada  $\mathbf{u} = \min(\sqrt{(c/D(\mathbf{x}))}, 1)\mathbf{x}$  com  $\mathbf{x} = (x_1, x_2)^t$ , os *outliers* são reduzidos ao limite da elipse de tolerância de 95% e, portanto, irão influenciar menos o valor do estimador de correlação obtendo-se um estimador de correlação mais robusto.

## 4 Estudo de simulação

Nesta secção apresenta-se o estudo de simulação implementado por forma a avaliar o desempenho dos métodos LARS e RLARS. Os algoritmos LARS e RLARS foram implementados em R recorrendo aos *packages* “lars” (Hastie *et al.* (2013)) e “robustHD” (Alfons *et al.* (2013)), respectivamente.

Considere-se o estudo de simulação de Khan *et al.* (2007), que se baseia no cenário de Frank e Friedman (1993). O modelo linear é estabelecido da seguinte forma:

$$\mathbf{y} = \mathbf{L}_1 + \mathbf{L}_2 + \dots + \mathbf{L}_k + \sigma \mathbf{e}$$

com  $k$  variáveis latentes (considera-se  $k = 6$ ), onde  $\mathbf{L}_1, \dots, \mathbf{L}_k$  e  $\mathbf{e}$  são variáveis com distribuição Normal e independentes. O valor de  $\sigma$  é escolhido de modo que

$$\sqrt{\text{Var}(\mathbf{L}_1 + \dots + \mathbf{L}_k) / \text{Var}(\sigma \mathbf{e})} = \sqrt{k / \sigma}$$

seja igual a 3. Sejam  $e_1, \dots, e_p$  variáveis independentes com distribuição Normal estandardizada e sejam

$$\begin{aligned} x_i &= L_i + \tau e_i, i = 1, \dots, k \\ x_{k+1} &= L_1 + \delta e_{k+1} \\ x_{k+2} &= L_1 + \delta e_{k+2} \\ &\vdots \\ x_{3k-1} &= L_k + \delta e_{3k-1} \\ x_{3k} &= L_k + \delta e_{3k} \\ x_i &= e_i, i = 3k+1, \dots, p \end{aligned}$$

com  $\delta = 5$  e  $\tau = 0,3$ .

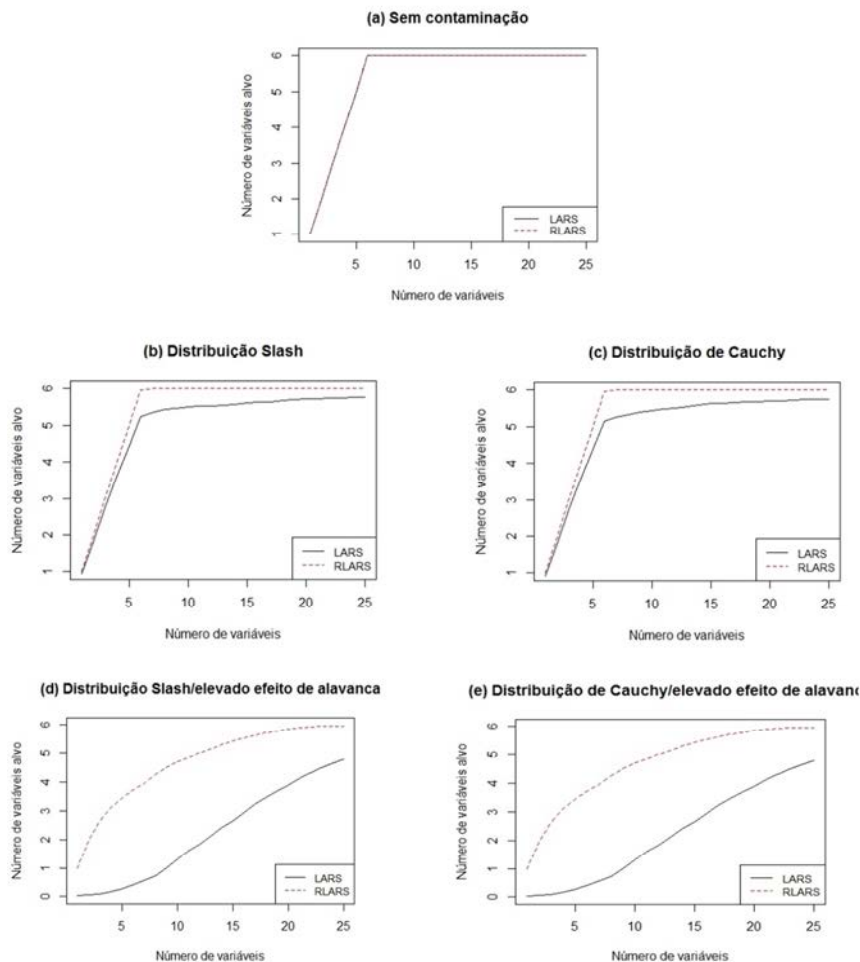
Considera-se a fracção de contaminação  $a = 0,1$ . São consideradas cinco distribuições para a geração da contaminação dos erros:

- (a)  $e \sim N(0,1)$ , sem contaminação;
- (b)  $e \sim (1-a)N(0,1) + aN(0,1)/U(0,1)$ , distribuição *Slash* simétrica;
- (c)  $e \sim \text{Cauchy}(0,1)$ , distribuição de *Cauchy* com caudas pesadas;
- (d) como em (b) mas com  $X \sim N(50,1)$ ;
- (e) como em (c), mas com  $X \sim N(50,1)$ .

Geram-se 200 réplicas de conjuntos de dados independentes com dimensão  $n = 150$  a partir dos 5 cenários de simulação descritos com  $p = 50$  preditores e, em cada cenário, executam-se todos os métodos mencionados no mesmo conjunto de dados.

## 5 Resultados e discussão

Com o objectivo de comparar o desempenho dos métodos descritos, na Figura 1 estão representados o número médio de variáveis alvo *versus* o número de variáveis em cada um dos métodos e em cada situação de contaminação considerada.



**Figura 1** - Número médio de variáveis alvo *versus* o número de variáveis em cada um dos métodos e em cada situação de contaminação.

O desempenho do método proposto de selecção robusta de variáveis em diferentes conjuntos de dados simulados contendo vários tipos de contaminações

(com pontos de elevado efeito de alavanca e com uma fracção de *outliers*) é examinado e comparado.

Os resultados mostram que a técnica de selecção robusta de variáveis considerada apresenta um melhor desempenho na presença nos dados de *outliers* e de pontos com elevado efeito de alavanca. O método robusto para a selecção de variáveis também encontra e sequencia melhor os preditores para integrarem o modelo de regressão, entre o conjunto dos preditores iniciais, em comparação com o método não-robusto.

## Agradecimentos

Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - “Fundação para a Ciência e a Tecnologia”, no âmbito do projecto PEstOE/MAT/UI0013/2014.

## Referências

- AKAIKE, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, 22, 203-217.
- ALFONS, A. (2013). robustHD: Robust methods for high-dimensional data. R package version 0.4.0.
- EFRON, B. E., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32, 407-451.
- FRANK, L.E. & FRIEDMAN, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109-135.
- HASTIE, T. & EFRON, B. (2013). lars: Least Angle Regression, Lasso and Forward Stagewise. R package version 1.2.
- HUBER, P.J. & RONCHETTI, E.M. (2009). *Robust Statistics*. Wiley, New York
- KHAN, J.A., VAN AELST, S. & ZAMAR, R.H. (2007). Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association*, 102, 1289-1299.
- MALLOWS, C.L. (1973). Some comments on  $C_p$ . *Technometrics*, 15, 661-675.
- SCHWARTZ, G. (1978). Estimating the Dimensions of a Model. *The Annals of Statistics*, 6, 461-64.
- MILLER, A.J. (2002). *Subset Selection in Regression*. New York, Chapman-Hall.
- MORGENTHALER, S., WELSCH, R.E. & ZENIDE, A. (2003). Algorithms for Robust Model Selection in Linear Regression in M. Hubert, G. Pison, A. Struyf, and S. Van (Eds). *Theory and Applications of Recent Robust Methods*, Birkhauser-Verlag, 195-206.

- MULLER, S. & WELSH, A.H. (2005). Outlier Robust Model Selection in Linear Regression. *Journal of the American Statistical Association*, 100, 1297-1310.
- RONCHETTI, E. (1985). Robust Model Selection in Regression. *Statistics and Probability Letters*, 3, 21-23.
- RONCHETTI, E. & STAUDTE, R.G. (1994). A Robust Version of Mallow's Cp. *Journal of the American Statistical Association*, 89, 550-559.
- RONCHETTI, E., FIELD, C. & BLANCHARD, W. (1997). Robust Linear Model Selection by Cross Validation. *Journal of the American Statistical Association*, 92, 1017- 1023.
- SALIBIAN-BARRERA, M. & VAN AELST, S. (2008). Robust Model Selection using fast and Robust Bootstrap. *Computational Statistics and Data Analysis*, 52, 5121-5135.
- WEISBERG, S. (1985). *Applied Linear Regression*. (2nd ed.), NewYork: Wiley-Interscience.

# A deeper glance on the percentage warping path distortion measure

Joana Hora<sup>1</sup> · Pedro Campos<sup>2</sup>

© The Author(s) 2017

**Abstract** This work provides a comprehensive description of a new measure – the Percentage Warping Path Distortion (PYPD). PYPD assesses the shape similarity of multivariate time series by quantifying the distortion of the Warping Path obtained from the Dynamic Time Warping (DTW) algorithm. This work includes the analysis of a real world case study on forest fire activity, whose multivariate time series were assessed by comparing the outputs from PYPD, DTW and Diagonal Path.

**Keywords:** Multivariate time series, Shape similarity assessment, DTW, PYPD.

## 1 Introduction

The similarity of time series can be assessed by different features such as the magnitude discrepancy between the time series with error measures (Prestwich et al., 2014) or distance measures (Wang et al., 2013); the information contained in the time series using Information Theory measures such as mutual information or entropy (Niu and Wang, 2015); or using statistical tests to conclude if the time series arrive from similar probability distributions (Teyssèdre et al., 2012).

Another relevant feature to assess similarity of time series is the similarity of the shape they draw. Literature includes several approaches such as the Dynamic Time Warping (DTW) algorithm (Müller, 2007), the Longest Common Subsequence (LCSS) (Vlachos et al., 2003), the Normalized Integral Square Error (NISE) (Sarin

---

<sup>1</sup>Faculdade de Engenharia da Universidade do Porto (FEUP), INESC TEC, [joana.hora@gmail.com](mailto:joana.hora@gmail.com)

<sup>2</sup>Faculdade de Economia da Universidade do Porto (FEP), INESC TEC, [pcampos@fep.up.pt](mailto:pcampos@fep.up.pt)

et al., 2010), the SEA method using the comparison and alignment of quasi-periodic time series (Boucheham, 2008), the Angular Metric for Shape Similarity (AMSS) which applies a variant of cosine similarity (Nakamura et al., 2013). From these methods, the DTW is the most popular (Tavenard and Amsaleg, 2015).

DTW was proposed in the field of speech recognition (Rabiner and Juang, 1993). It has been applied to other areas such as data mining and information retrieval (Keogh and Pazzani, 2000), the indexing of time series (Keogh and Ratanamahatana, 2005), and combined with approaches for distance assessment in multivariate datasets (Góreckia and Łuczakb, 2015), digit recognition (Qu et al., 2015) or fuzzy clustering of time series (Izakian et al., 2015). Some variations of DTW have been proposed, such as adding penalties (Clifford et al., 2009), adding boundaries to reduce the processing time (Lemire, 2009), or applying projection techniques to reach linear time and space complexity (Salvador and Chan, 2007). DTW is adequate to assess the shape similarity of univariate time series (Berndt and Clifford, 1994, Sarin et al., 2010) and multivariate time series (Rath and Manmatha, 2002). For samples of equal length, the time complexity of DTW is  $O(N^2)$ , which can be reduced with window constraints (Parizeau and Plamondon, 1990).

DTW searches for the optimal alignment between two time series even when they are out of phase, aligning peaks and valleys with the compression and expansion of the time axis (Keogh and Ratanamahatana, 2005, Keogh et al., 2009). It returns the so-called DTW distance and the Warping Path (a set of bivariate elements storing the coordinates of the shortest cumulative path between the two time series).

Two time series following similar shapes would return a Warping Path coincident with the Diagonal Path. The Percentage Warping Path Distortion (PYPD) quantifies the shape likeness of two time series by measuring the average distance amid the Warping Path and the corresponding Diagonal Path. The information provided by PYPD is distinct from that of DTW distance (which relates to the cumulative sum of the shortest path between the two series).

This work starts with a review of the DTW algorithm in its univariate and multivariate approaches (Section 2). It proceeds with the description of the PYPD measure for univariate and multivariate applications (Section 3). Section 4 provides an application to a real world case study, considering the assessment of multivariate time series from the management of forest fires, using PYPD, DTW and Diagonal Path distances. Conclusions are drawn in Section 5.

## 2 The DTW algorithm

The univariate approach of the Dynamic Time Warping (DTW) algorithm assesses the shape similarity between two time series of one variable and  $N$  observations each:  $\mathbf{x} = [x_1, \dots, x_N]$  and  $\mathbf{y} = [y_1, \dots, y_N]$ , here treated as vectors. Without loss of generality we consider time series of similar length (i.e., they have the same number of observations).

A distance function such as the Euclidean Distance ( $ED$ ) is applied to each combination of elements in  $\mathbf{x}$  and  $\mathbf{y}$ , as specified in expression (1). The Cost Matrix  $\mathbf{A}$  stores all distances at the corresponding coordinates, as specified in expressions (2). The elements composing  $\mathbf{A}$  relate to valid steps (i.e. meaningful combinations), and not valid steps should be left in blank.

The DTW algorithm can also be applied to multivariate time series with  $P$  variables and  $N$  observations each:  $\mathbf{X} = [X_{ij}]_{N \times P}$  and  $\mathbf{Y} = [Y_{ij}]_{N \times P}$ , here treated as matrixes, see (3).

In the multivariate case, the Cost Matrix  $\mathbf{A}$  is calculated with a distance function (e.g.,  $ED$ ) for each combination of lines composing  $\mathbf{X}$  and  $\mathbf{Y}$  (Rath and Manmatha, 2002), see expressions (4-5).

$$d(x_i, y_j) = \sqrt{(x_i - y_j)^2} \quad (1)$$

$$A_{ij} = d(x_i, y_j); \quad \mathbf{A} = \begin{bmatrix} A_{11} & \dots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{N1} & \dots & A_{NN} \end{bmatrix} \quad (2)$$

$$\mathbf{X} = \begin{bmatrix} X_{11} & \dots & X_{1P} \\ \vdots & \ddots & \vdots \\ X_{N1} & \dots & X_{NP} \end{bmatrix}; \quad \mathbf{Y} = \begin{bmatrix} Y_{11} & \dots & Y_{1P} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \dots & Y_{NP} \end{bmatrix} \quad (3)$$

$$d(X_i, Y_j) = \sum_{p=1}^P (X_{ip} - Y_{jp})^2 \quad (4)$$

$$A'_{ij} = d(X_i, Y_j); \quad \mathbf{A}' = \begin{bmatrix} A'_{11} & \dots & A'_{1N} \\ \vdots & \ddots & \vdots \\ A'_{N1} & \dots & A'_{NN} \end{bmatrix} \quad (5)$$

The univariate and multivariate approaches only differ from calculation of  $\mathbf{A}$ . In a multivariate approach, if the variables considered are measured with different units, the data must be normalized prior to the calculus of  $\mathbf{A}$ . If the case study includes not meaningful combinations, their positions in  $\mathbf{A}$  should be left in blank.



The classic Shortest Path Problem is applied to  $\mathbf{A}$  to calculate the minimum cumulative cost to reach each position from the starting position  $i = 1 \wedge j = 1$ . This procedure can be implemented using the Dijkstra algorithm (Johnson, 1973). All cumulative costs are stored in Matrix  $\mathbf{B} = [B_{ij}]_{N \times N}$ . The definition of each element in  $\mathbf{B}$  is provided in expressions (6).  $DTW$  distance between  $\mathbf{x}$  and  $\mathbf{y}$  is finally calculated as the square root of the element  $B_{NN}$ , see expression (7).

Lower values of  $DTW$  indicate a better fit between the two vectors.  $DTW$  distance is sensitive to scale, returning values in the same units of observations, and is sensitive to the length of time series. This distance does not satisfy the triangle inequality since  $DTW(\mathbf{a}, \mathbf{b}) + DTW(\mathbf{b}, \mathbf{c})$  is not always higher than  $DTW(\mathbf{a}, \mathbf{c})$  (Müller, 2007).

$$\begin{cases} B_{11} = A_{11} \\ B_{ij} = A_{ij} + \min(B_{i-1,j-1}, B_{i-1,j}, B_{i,j-1}), \text{ if } i \wedge j \neq 1 \end{cases} \quad (6)$$

$$DTW(\mathbf{x}, \mathbf{y}) = \sqrt{B_{NN}} \quad (7)$$

In addition to the  $DTW$  distance, the algorithm also returns the Warping Path storing a set of  $K$  bivariate elements:  $\mathbf{W} = \langle w_1, w_2, \dots, w_K \rangle$ . The number of bivariate elements integrating  $\mathbf{W}$  can vary between  $K \in [N, 2N - 1]$ . Each element  $w_k = [i_k^w, j_k^w]$ , with  $i_k^w, j_k^w \in [1, N]$ , stores the coordinates of  $\mathbf{B}$  corresponding to the shortest cumulative path between  $B_{11}$  and  $B_{NN}$ , respecting three conditions:

- boundary:  $w_1 = [1, 1]$  and  $w_K = [N, N]$
- continuity:  $i_k^w - i_{k-1}^w \leq 1$  and  $j_k^w - j_{k-1}^w \leq 1$
- monotonicity:  $(i_k^w - i_{k-1}^w > 0 \wedge j_k^w - j_{k-1}^w \geq 0) \vee (i_k^w - i_{k-1}^w \geq 0 \wedge j_k^w - j_{k-1}^w > 0)$

For the particular case of a perfect shape similarity,  $\mathbf{W}$  would be coincident with the path included in  $diag(\mathbf{B})$ . In that case, the corresponding  $DTW$  distance would be equal to square root of the trace of  $\mathbf{A}$ , as specified in equation (8).

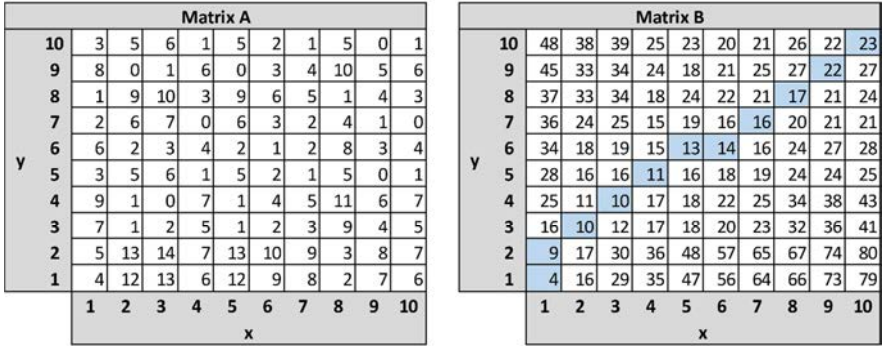
$$\sqrt{tr(\mathbf{A})} = \sqrt{\sum_{i=1}^N A_{ii}} \quad (8)$$

Note that the  $\sqrt{tr(\mathbf{A})}$  always relates to an equal or more expansive distance than the  $DTW$  distance (i.e., the  $DTW$  is calculated from the Shortest Path). A demonstrative example is provided in Figure 1. Figure 1 – top provides an example of two univariate time series. Expression (1) was used to calculate  $\mathbf{A}$ , see Figure 1 – bottom left. For example  $d(x_1, y_1) = \sqrt{(10 - 14)^2} = 4$ , and therefore the element  $A_{11} = 4$ .

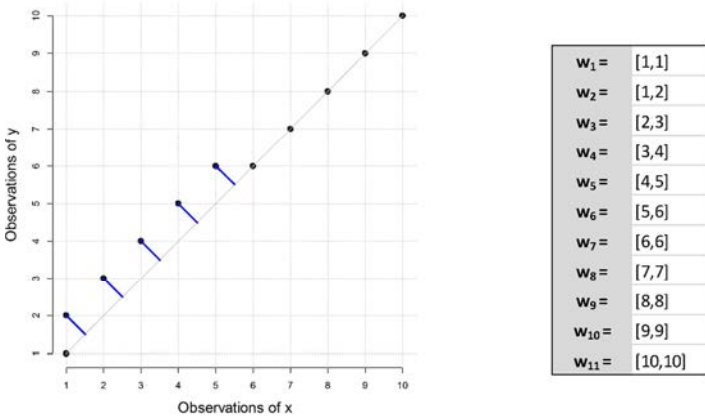
The matrix  $\mathbf{B}$  is then calculated from  $\mathbf{A}$ , following expressions (6). Accordingly, the first element is calculated as  $B_{11} = A_{11} = 4$ . For the remaining elements, we take as an example  $B_{21} = A_{21} + \min(B_{10}, B_{11}, B_{20}) = 12 + 4 = 16$ . Note that the elements with index 0 do not exist, and therefore are discarded. For this example, the distance  $DTW = \sqrt{23} = 4.796$ , the  $\sqrt{\text{tr}(\mathbf{A})} \approx 6.40$  and the ratio  $DTW / \sqrt{\text{tr}(\mathbf{A})} \approx 0.75$ . Figure 1 (bottom right) highlights  $\mathbf{W}$  with shaded cells in  $\mathbf{B}$ , including 11 bivariate elements:

$$\mathbf{W} = \langle [1,1], [1,2], [2,3], [3,4], [4,5], [5,6], [6,6], [7,7], [8,8], [9,9], [10,10] \rangle.$$

Observation no.	1	2	3	4	5	6	7	8	9	10
x	10	2	1	8	2	5	6	12	7	8
y	14	15	3	1	7	4	8	11	2	7



**Figure 1** – Demonstrative example: values of the two time series (top); cost matrix A (bottom left); matrix B (bottom right).



**Figure 2** – Visualization of the distance of each  $\mathbf{W}$  element to the diagonal (left) and the bivariate elements composing  $\mathbf{W}$  (right).

### 3 Percentage Warping Path Distortion (PWPD)

The Percentage Warping Path Distortion (*PWPD*) measure quantifies the distortion of the Warping Path obtained from the *DTW* algorithm, taking the Diagonal Path as a reference. The calculus of the *PWPD* measure takes the Warping Path  $\mathbf{W}$  as an input, and therefore this process is similar for univariate and multivariate approaches (as they only differ from calculation of  $\mathbf{A}$ ).

The calculus of *PWPD* starts with the estimation of the average distance between  $\mathbf{W}$  and Diagonal Path (*DP*), which we refer as Warping Path Distortion (*WPD*). The distance between each element of  $\mathbf{W}$  (i.e.,  $w_k = [i_k^w, j_k^w]$ ) to the corresponding nearest position in *DP* (i.e.,  $p_k = [i_k^p, j_k^p]$ ) is determined, and can be visualized by tracing a line amid these two positions (see Figure 2).

Both coordinates of each  $p_k$  are defined from the corresponding element in  $\mathbf{W}$  as specified in equation (9). The set of nearest diagonal positions relating to an arbitrary warping path  $\mathbf{W}$  is defined as  $\mathbf{P} = \langle p_1, p_2, \dots, p_K \rangle$ . The distance between an arbitrary  $w_k$  to the corresponding nearest position in *DP*,  $p_k$ , is then calculated using the Manhattan Distance  $d_{r=1}(w_k, p_k)$ . *WPD* is then estimated as defined in equation (10).

$$i_k^p = \left\lfloor \frac{(i_k^w - j_k^w)}{2} \right\rfloor + \min(i_k^w, j_k^w) \quad (9)$$

$$WPD = \frac{1}{K-2} \sum_{k=2}^{K-1} d_{r=1}(w_k, p_k) \quad (10)$$

*WPD* is scale independent, however it is dependent on the sample size. The values returned by *WPD* are included within the range  $WPD \in [0, (N-1)^2/(K-2)[$ . A lower value of *WPD* is associated with a better fit of the estimated model. The perfect fit would return a *WPD* value of zero (when the positions stored within  $\mathbf{W}$  are coincident with *DP*). *PWPD* is calculated with the ratio of *WPD* by the maximum possible value that *WPD* may achieve, as defined in equation (11).

$$PWPD = \frac{WPD}{(N-1)^2/K-2} \quad (11)$$

*PWPD* indicates the percentage distortion of the  $\mathbf{W}$  to *DP* and is independent on size and scale of samples. *PWPD* values are within the range  $[0, 1[$ . A value of 0 is obtained when  $\mathbf{W}$  is coincident with *DP*, representing a good similarity between the two series. The asymptotic value of 1 occurs for the maximum distortion of  $\mathbf{W}$  from *DP*. Both *DTW* and *PWPD* respect the commutative property (i.e.,  $PWPD(\mathbf{x}, \mathbf{y}) = PWPD(\mathbf{y}, \mathbf{x})$  and  $DTW(\mathbf{x}, \mathbf{y}) = DTW(\mathbf{y}, \mathbf{x})$ ).

## 4 Shape likeness of forest fires multivariate time series

We aim to assess which Portuguese municipalities follow a similar pattern concerning the forest fires activity over time. The results obtained can support forest fire management actions and fire prevention policies. All Portuguese municipalities within the mainland territory (i.e., 278) were considered between 1980 and 2010. The data were collected by the Portuguese governmental agency “*Instituto da Conservação da Natureza e das Florestas, ICNF, I.P.*” (ICNF, 2012). Four variables are available for each municipality: annual number of fires ( $V_1$ ), annual overall area burned ( $V_2$ ), annual area burned within forests ( $V_3$ ), and annual area burned within residential zones ( $V_4$ ). The results were calculated with the R software, with all functions calculated by the authors.

Table 1 includes the results for 4 pairs of municipalities randomly selected from the total set of comparable pairs. Each pair was analyzed with multivariate *PWPD* (MPWPD) alongside with the values obtained for the univariate *PWPD* for each variable and the corresponding average ( $\sum_{i=1}^4 V_i/4$ ). Moreover, the homologous values obtained with multivariate *DTW* (MDTW) were calculated. All calculations were made with the previous normalization of data. The visualization of the multivariate  $\mathbf{W}$  of each pair is provided in Figure 3. Considering the measure MPWPD, the pair of municipalities returning the best shape similarity is Pair 1, followed by Pair 4, Pair 3 and finally Pair 2. This order of similarity is in accordance with the visual analysis retrieved from Figure 3. Moreover, the same conclusions would be drawn using the average of univariate *PWPD* values. Note that Pair 1 returned a value of  $MPWPD = 0$ , indicating that its multivariate  $\mathbf{W}$  is coincident with the corresponding diagonal, which is shown in Figure 3 – top left. However, we can see that the conclusions retrieved from approaches MPWPD and MDTW are not coincident. When considering the measure MDTW, the best shape similarity would be attributed to Pair 4, followed by Pair 1, Pair 2 and finally Pair 3.

Both measures MPWPD and MDTW assess shape similarity of multivariate time series, however they do not assess the same information. The MPWPD assesses the average distance from the Warping Path to the diagonal, and MDTW assesses the cumulative cost of the Shortest Path obtained.

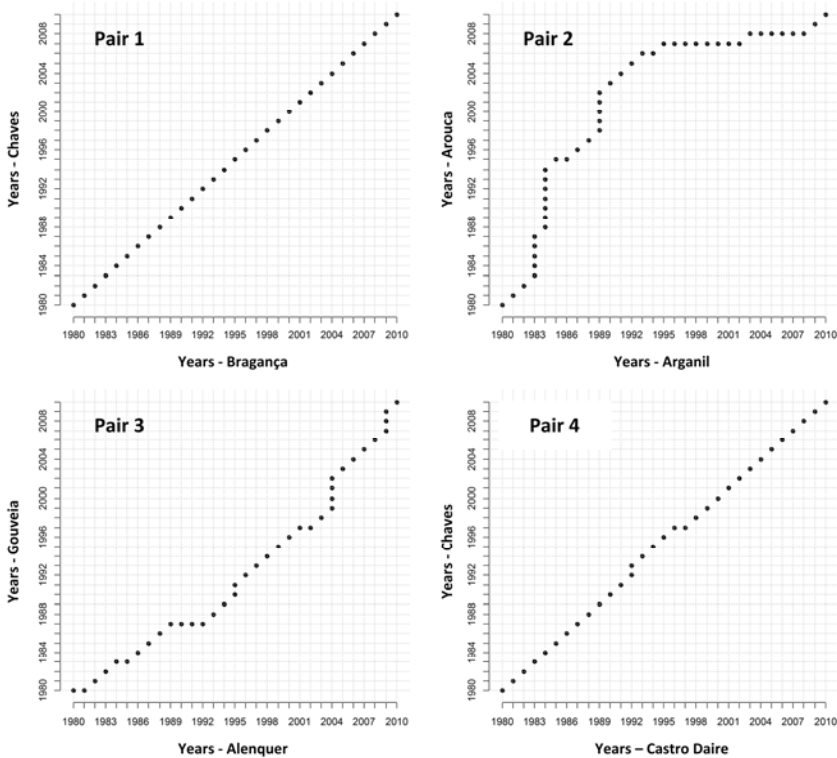
## 5 Conclusions

This work investigates the quantification of shape similarity between two multivariate datasets. The *DTW* algorithm has been widely used to address this topic. Normally, the *DTW* distance is used as a quantitative indicator on shape

similarity and a visual analysis of the corresponding Warping Path is used to support the conclusions achieved.

**Table 1** - Results for 4 pairs randomly selected.

Pair		Univariate <i>PWPD</i>					<i>MPWPD</i>	<i>MDTW</i>
No.	Municipalities	$V_1$	$V_2$	$V_3$	$V_4$	$\frac{\sum_{i=1}^4 V_i}{4}$		
1	<i>Bragança - Chaves</i>	0.040	0.029	0.057	0.014	0.035	0	4.552
2	<i>Arganil - Arouca</i>	0.271	0.294	0.241	0.326	0.283	0.331	5.128
3	<i>Alenquer - Gouveia</i>	0.087	0.241	0.302	0.053	0.171	0.116	5.578
4	<i>Castro Daire – Chaves</i>	0.071	0.018	0.061	0.104	0.064	0.006	4.335



**Figure 3** – Visualization of the Multivariate Warping Path of pairs 1-4.

The visual information contained within the  $\mathbf{W}$  returned by the *DTW* algorithm is normally analysed by eye, but its quantification was not yet proposed in literature. This study proposes the *PWPD* measure to quantify the distortion of the

Warping Path having the diagonal as a reference. PWPDP is adequate to assess the shape similarity of univariate and multivariate time series.

The PWPDP returns the percentage average distance amid  $W$  and the Diagonal Path, with base on the reasoning that a larger distance between  $W$  and DP relates to a larger dissimilarity of shape. The results obtained with the application of PWPDP are aligned with the visual comparison of Warping Paths – which is not true when using DTW.

## Acknowledgements

This research was partially supported by FCT doctoral scholarship with reference PD/BD/113761/2015.

## References

- BERNDT, D. J. & CLIFFORD, J. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. Workshop on Knowledge Discovery in Databases (KDD-94), AAAI Technical Report WS-94-03, pp. 359-370, .
- BOUCHEHAM, B. 2008. Matching of quasi-periodic time series patterns by exchange of block-sorting signatures. *Pattern Recognition Letters*, 29, 501–514.
- CLIFFORD, D., STONE, G., MONTOLIU, I., REZZI, S., MARTIN, F. P., GUY, P., BRUCE, S. & KOCHHAR, S. 2009. Alignment Using Variable Penalty Dynamic Time Warping. *Analytical Chemistry*, 81, 1000-1007.
- GÓRCKIA, T. & ŁUCZAKB, M. 2015. Multivariate time series classification with parametric derivative dynamic time warping. *Expert Systems with Applications*, 42, 2305–2312.
- ICNF. 2012. *Instituto da Conservação da Natureza e das Florestas, I. P.* [Online]. Portugal: Instituto da Conservação da Natureza e das Florestas, I. P. Available: <http://www.icnf.pt/portal/icnf> [Accessed September 2012].
- IZAKIAN, H., PEDRYCZ, W. & JAMAL, I. 2015. Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39, 235-244.
- JOHNSON, D. B. 1973. A note on Dijkstra's shortest path algorithm. *Journal of the ACM (JACM)*, 20, 385-388.
- KEOGH, E. & RATANAMAHATANA, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7, 358-386.
- KEOGH, E., WEI, L., XI, X. P., VLACHOS, M., LEE, S. H. & PROTOPAPAS, P. 2009. Supporting exact indexing of arbitrarily rotated shapes and periodic time

- series under Euclidean and warping distance measures. *The VLDB Journal—The International Journal on Very Large Data Bases*, 18, 611-630.
- KEOGH, E. J. & PAZZANI, M. J. Scaling up Dynamic Time Warping for Datamining Applications. Sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000 Boston, MA, USA. 285-289.
- LEMIRE, D. 2009. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern Recognition*, 42, 2169-2180.
- MÜLLER, M. 2007. Dynamic Time Warping. In: MÜLLER, M. (ed.) *Information Retrieval for Music and Motion*. Berlin: Springer.
- NAKAMURA, T., TAKI, K., NOMIYA, H., SEKI, K. & UEHARA, K. 2013. A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, 16, 535-548.
- NIU, H. & WANG, J. 2015. Quantifying complexity of financial short-term time series by composite multiscale entropy measure. *Communications in Nonlinear Science and Numerical Simulation*, 22, 375-382.
- PARIZEAU, M. & PLAMONDON, R. 1990. A Comparative-Analysis of Regional Correlation, Dynamic Time Warping, and Skeletal Tree Matching for Signature Verification. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 12, 710-717.
- PRESTWICH, S., ROSSI, R., ARMAGAN TARIM, S. & HNICH, B. 2014. Mean-based error measures for intermittent demand forecasting. *International Journal of Production Research*, 52, 6782-6791.
- QU, C., ZHANG, D. & TIAN, J. 2015. Online Kinect handwritten digit recognition based on dynamic time warping and support vector machine. *Journal of Information and Computational Science*, 12, 413-422.
- RABINER, L. & JUANG, B. H. 1993. *Fundamentals of Speech Recognition*, Prentice Hall.
- RATH, T. M. & MANMATHA, R. 2002. Lower-bounding of dynamic time warping distances for multivariate time series. *University of Massachusetts Amherst Technical Report MM*, 40.
- SALVADOR, S. & CHAN, P. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11, 561-580.
- SARIN, H., KOKKOLARAS, M., HULBERT, G., PAPALAMBROS, P., BARBAT, S. & YANG, R. J. 2010. Comparing time histories for validation of simulation models: Error measures and metrics. *Journal of Dynamic Systems, Measurement, and Control*, 132, 1-10.
- TAVENARD, R. & AMSALEG, L. 2015. Improving the efficiency of traditional DTW accelerators. *Knowledge and Information Systems*, 42, 215-243.
- TEYSSÈDRE, S., ELSEN, J.-M. & RICARD, A. 2012. Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. *Genetics Selection Evolution*, 44, 1-17.

- VLACHOS, M., HADJIELEFThERIOU, M., GUNOPULOS, D. & KEOGH, E. 2003. Indexing multi-dimensional time-series with support for multiple distance measures. *9th ACM SIGKDD international conference on Knowledge discovery and data mining*. Washington DC, USA: ACM.
- WANG, X. Y., MUEEN, A., DING, H., TRAJCEVSKI, G., SCHEUERMANN, P. & KEOGH, E. 2013. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26, 275-309.





## Elevados níveis de ferro nos doentes alcoólicos: um contributo para esclarecer esta correspondência

Ana Matos<sup>1</sup> · Carla Henriques<sup>2</sup> · Luís Costa Matos<sup>3</sup> · Nuno Monteiro<sup>4</sup> · Paulo Batista<sup>5</sup>

© The Author(s) 2017

**Resumo** O mecanismo que estimula o aumento dos níveis de ferro nos pacientes com doença hepática alcoólica (DHA) não está ainda completamente esclarecido, sendo este o cerne deste estudo.

Desde a recente descoberta da hepcidina como proteína com papel crucial na regulação do mecanismo do ferro, várias investigações em animais foram efetuadas, sendo ainda escassos estudos em humanos. Este trabalho envolve um estudo observacional, sobre a expressão da hepcidina em tecido hepático humano em indivíduos com DHA e em indivíduos saudáveis.

**Palavras-chave:** Doença Hepática Alcoólica, Hepsidina, Regressão Múltipla.

### 1 Introdução

A doença hepática alcoólica (DHA) define-se como a lesão do fígado causada pela ingestão excessiva de álcool. Em geral, a quantidade de álcool consumido determina a probabilidade e a importância da lesão hepática. Contudo, nem todas as pessoas que bebem excessivamente desenvolvem DHA, ocorrendo também casos de DHA em pessoas que bebem pouco. Apesar do alto consumo de álcool estar

---

<sup>1</sup>Centro de Estudos em Educação Tecnologias e Saúde, Escola Superior de Tecnologia e Gestão de Viseu, [amatos@estv.ipv.pt](mailto:amatos@estv.ipv.pt)

<sup>2</sup>Centro de Matemática da Universidade de Coimbra (CMUC) e Escola Superior de Tecnologia e Gestão de Viseu, [carlahenriq@estv.ipv.pt](mailto:carlahenriq@estv.ipv.pt)

<sup>3</sup>Centro Hospitalar Tondela-Viseu, [costamatos.luis@gmail.com](mailto:costamatos.luis@gmail.com)

<sup>4</sup>Centro Hospitalar Tondela-Viseu, [nunoricardomonteiro@gmail.com](mailto:nunoricardomonteiro@gmail.com)

<sup>5</sup>Centro Hospitalar Tondela-Viseu, [paulobat@gmail.com](mailto:paulobat@gmail.com)

associado à doença hepática alcoólica, não se sabe exatamente como a doença se desenvolve.

Matos (2006), num estudo sobre a epidemiologia da doença hepática alcoólica, refere que a ingestão de bebidas alcoólicas é um hábito enraizado na cultura portuguesa, sendo o álcool a droga de abuso mais frequente em Portugal. Em doentes internados nos EUA, 20 a 40% têm problemas relacionados com o álcool, e, em idosos, o alcoolismo motiva tantos internamentos como o enfarte agudo do miocárdio. No que respeita a Portugal, numa análise retrospectiva efetuada num serviço de Medicina Interna durante 12 anos, verificou-se que, em 7,8% dos internamentos, os diagnósticos principais e/ou secundários são relacionados com o álcool (Matos, 2006).

De facto, o álcool é das drogas mais conhecidas e aceite socialmente e, apesar do elevado número de trabalhos e pesquisas feitos em torno deste problema, muitas questões continuam por esclarecer.

É do conhecimento clínico que os pacientes com doença hepática alcoólica apresentam níveis elevados de ferro no fígado, o que se julga contribuir para o agravamento da doença.

O ferro é um sal mineral essencial para o bom funcionamento do organismo. Desempenha um papel crucial na formação da hemoglobina contida nos glóbulos vermelhos do sangue (responsável pelo transporte de oxigénio para as células e pela eliminação do dióxido de carbono), na formação da mioglobina dos músculos e em numerosas enzimas indispensáveis para o correto funcionamento do organismo.

Um adulto saudável tem de 40 a 160 microgramas de ferro no sangue, que é o nível recomendado. Índices abaixo ou acima dos valores de referência são um sinal de problemas de saúde. A deficiência deste mineral no organismo leva a um quadro conhecido como anemia. O seu excesso pode ser bastante prejudicial à saúde: no fígado, altos níveis do mineral podem causar cirrose; no pâncreas, diabetes; no coração, insuficiência cardíaca; nas glândulas, mau funcionamento e problemas na produção hormonal.

A acumulação de ferro em quantidades superiores às necessárias pode ter causas genéticas - por exemplo o caso dos portadores de hemocromatose, ou podem ser adquiridas - caso dos doentes alcoólicos.

O organismo humano não tem nenhum mecanismo de rejeição do ferro (à exceção das mulheres em idade fértil através da menstruação). Recentemente foi descoberta uma proteína produzida no fígado, denominada hepcidina (Krause *et al.*, 2000; Park *et al.*, 2001), que impede a absorção intestinal do ferro, regulando a quantidade de ferro que entra no organismo.

Existem vários fatores que podem influenciar a expressão da hepcidina. A anemia, a carência de ferro e a perda da função hepática na cirrose, entre outros, diminuem a expressão da hepcidina, enquanto que a inflamação e a sobrecarga de ferro aumentam-na. Em 2007 (Takaaki *et al.*, 2007) foi provado, em experimentação animal com ratos, que o álcool exerce influência nos níveis de hepcidina, diminuindo a expressão desta proteína.

O agravamento da DHA julga-se estar relacionado com os elevados níveis de ferro no fígado destes pacientes. Investigar sobre o que origina tal cenário é de crucial importância para tentar controlar a evolução da DHA. De facto, é conhecida a relação entre os níveis elevados de ferro e a doença alcoólica do fígado, no entanto o que provoca este aumento dos níveis de ferro é um mecanismo desconhecido. É nesta questão que se centra a investigação apresentada neste trabalho, pretendendo-se contribuir para clarificar este mecanismo nos pacientes com DHA.

O estudo observacional envolveu pacientes selecionados na consulta ambulatória de doenças Hepáticas do Centro Hospitalar Tondela-Viseu.

Recorreu-se a análises estatísticas univariadas e multivariadas de modo a identificar os fatores associados com os níveis de hepcidina. A comparação dos controlos com os pacientes de DHA foi feita através de testes paramétricos, sempre que as condições de aplicabilidade dos mesmos eram satisfeitas, e testes não paramétricos, caso contrário. Foram utilizados os testes de Mann-Whitney, teste t, teste de Kruskal-Wallis e testes de comparações múltiplas. A análise conjunta dos fatores relacionados com a hepcidina foi levada a cabo através da estimação de modelos de regressão linear múltipla. Toda a análise estatística foi efetuada com suporte no *software* estatístico SPSS – versão 22.

## 2 Métodos e resultados

O estudo envolveu 61 pacientes com DHA e 20 controlos saudáveis. Os pacientes foram selecionados da consulta ambulatória de doenças Hepáticas do Centro Hospitalar Tondela-Viseu. As amostras de fígado normal (controlos) foram recolhidas em pacientes admitidos para colecistectomia eletiva. Durante o procedimento cirúrgico, uma pequena amostra de fígado foi recolhido. A avaliação da biópsia do fígado foi realizada no Departamento de Anatomopatologia do Hospital da Universidade de Coimbra.

Na Tabela 1 comparam-se os pacientes com DHA e os controlos relativamente às variáveis em análise neste estudo. A ferritina é usada para quantificar os níveis de ferro. Trata-se de uma proteína globular que se localiza essencialmente no fígado.

A fibrose é o resultado da manutenção por longo período da inflamação do fígado, que vai formando fibras (cicatrizes), as quais, por sua vez, passam a dificultar a passagem do sangue pelas veias do órgão. A manutenção da fibrose por um longo período leva ao estado de cirrose. A fibrose classifica-se em 5 níveis: 0- ausência de fibrose; 1 – Fibrose periportal ou perissinusoidal; 2 Fibrose periportal e perissinusoidal; 3 – Fibrose em ponte; 4 – Cirrose.

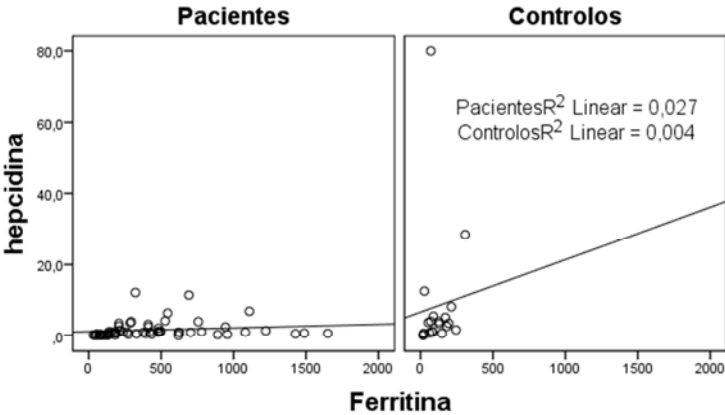
**Tabela 1** - Apresentação sumária da amostra.

	<b>Pacientes com DHA (n=61)</b>				<b>Controlos (n=20)</b>	<b>p<sup>(*)</sup></b>
<b>Ferritina</b>	434,7±387,9				115,7±81,8	<0,0005
<b>Hepcidina</b>	1,56±2,4				8,32±18,0	=0,002
<b>Grau fibrose</b>	0	1	2	3	0	
<b>%</b>	18	8,2	11,5	21,3	100%	

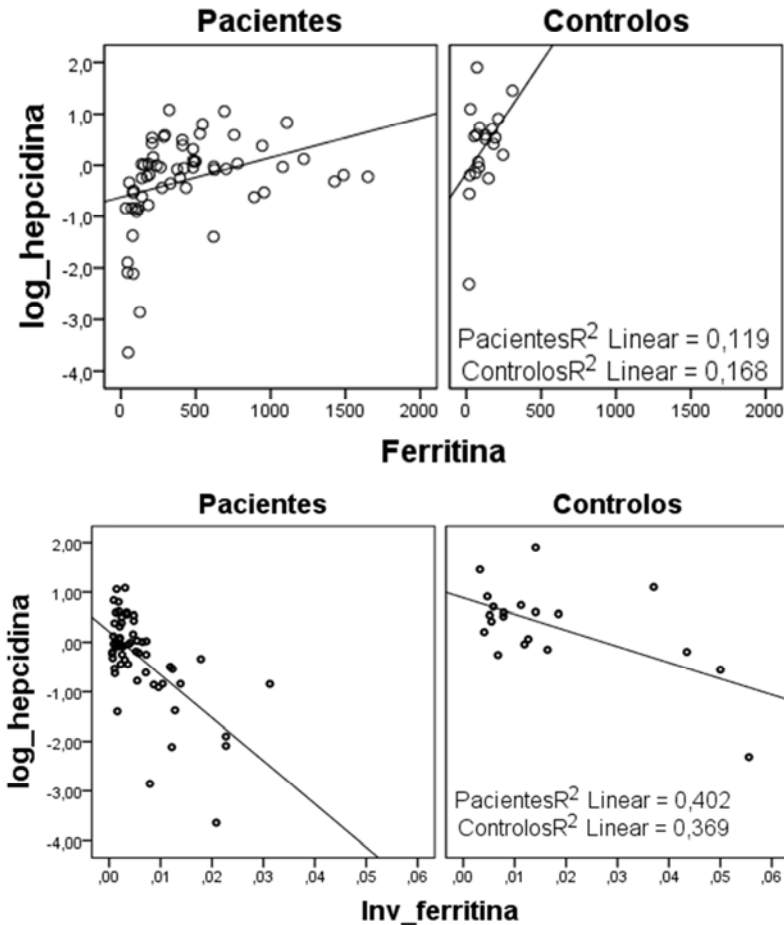
Nas variáveis contínuas são apresentados valores médios ± desvio padrão  
(\*) Teste de Mann-Whitney

A comparação dos dois grupos em estudo permite constatar que os pacientes com doença hepática alcoólica apresentam níveis de ferritina significativamente superiores aos controlos ( $p<0,0005$ ) e níveis de hepcidina inferiores ( $p=0,002$ ). Quer na hepcidina quer na ferritina, observam-se elevados desvios em relação à média, causados pela presença de observações atipicamente elevadas. Sendo estes registos reais e clinicamente admissíveis, a sua remoção do conjunto de dados em análise estava fora de questão. Para contornar esta dificuldade, as análises estatísticas foram efetuadas com as variáveis transformadas: logaritmo da hepcidina ( $\log\_hepcidina$ ) e inverso da ferritina ( $Inv\_ferritina$ ).

Começou-se por avaliar a relação entre os níveis de hepcidina e ferritina recorrendo a modelos de regressão linear simples. As Figuras 1 e 2 ilustram o efeito da transformação das variáveis na identificação das associações:



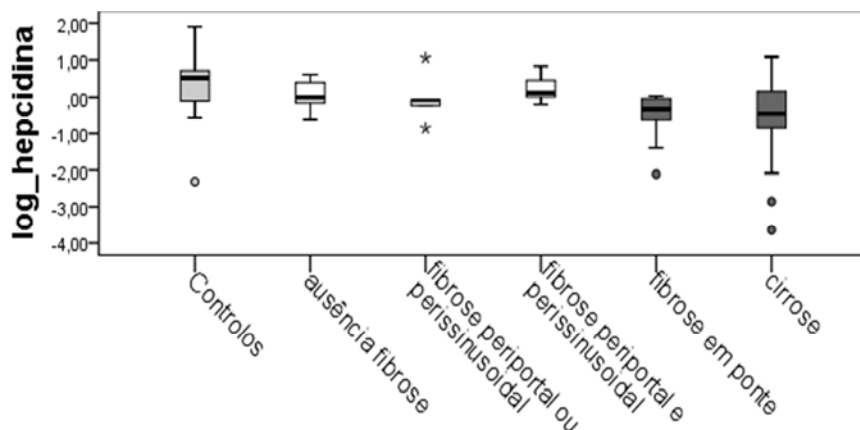
**Figura 1** - Relação linear entre hepcidina e ferritina.



**Figura 2** - Relação linear entre transformações da hepcidina e ferritina.

Os modelos confirmam a relação conhecida entre a hepcidina e os níveis de ferro (ferritina), sustentando a hipótese de que um dos fatores que incentiva a produção desta proteína é o excesso de ferro.

Também a perda da função hepática, resultante do aumento dos níveis de fibrose se revelou associada a uma menor expressão da hepcidina, conforme Figura 3 e teste de Kruskal-Wallis ( $p=0,003$ ). Observaram-se diferenças estatisticamente significativas entre o grupo de controlo e os dois níveis mais graves de fibrose.



**Figura 3** - Relação entre hepcidina e os níveis de fibrose.

Num esforço para analisar conjuntamente as variáveis em estudo e o seu efeito sobre os níveis de hepcidina, recorreu-se à regressão linear múltipla.

**Tabela 2** - Modelo de regressão linear múltipla.

Modelo	Coeficientes		T	(p)	Intervalo de confiança		Colineariedade	
	B	Erro padrão					Tolerância	VIF
Constante	1,125	0,213	5,291	(<0,00005)	0,701	1,548		
Fibr.graus 3e4	-0,409	0,191	-2,138	(0,036)	-0,790	-0,028	0,550	1,818
Tipo	-0,890	0,247	-3,609	(0,001)	-1,381	-0,399	0,768	1,302
Inv_ferritina	-46,72	8,425	-5,545	(<0,0005)	-63,5	-29,945	0,681	1,468

O modelo estimado da Tabela 2 traduz-se na forma:

$$\widehat{\log y} = 1,125 - 0,409 x_1 - 0,89 x_2 - 46,72 x_3,$$

onde, y representa a hepcidina e  $x_1$ ,  $x_2$  e  $x_3$  representam as variáveis independentes, respetivamente, Fibr\_graus 3 e 4, Tipo e Inv\_ferritina. O valor estimado da hepcidina pode, então, ser obtido através de:

$$\hat{y} = 13,33 \times 10^{-0,409x_1} \times 10^{-0,89x_2} \times 10^{-46,72x_3}.$$

O modelo estimado reafirma a relação entre a hepcidina e a ferritina (aumento significativo dos níveis de hepcidina com aumento da ferritina) e a diminuição da hepcidina nos dois níveis graves de fibrose hepática (variável Fibr. graus 3 e 4: 1 - paciente com fibrose em ponte ou cirrose; 0-paciente não apresenta fibrose em ponte nem cirrose). Estima-se que, em indivíduos com características semelhantes nas variáveis Tipo e Inv\_ferritina, pelo facto de apresentarem grau de fibrose 3 ou

4, têm significativamente menos 61%  $((1-10^{-0,409}) \times 100\%)$  de hepcidina.

O aumento dos níveis de hepcidina com aumento da ferritina não é de interpretação tão imediata. No entanto, observe-se que a passagem dos níveis de ferritina de 500 para 600 conduz a uma subida na hepcidina de 0,8064 para 0,8359, correspondendo a um aumento da hepcidina de 3,6%; já a passagem dos níveis de ferritina de 600 para 700, provoca uma subida na hepcidina de 0,8359 para 0,8575, correspondendo a um aumento de 2,6%. De facto, o aumento percentual na hepcidina devido a um acréscimo de 100 unidades de ferritina, mantendo as outras variáveis constantes, é estimado em:

$$c_1 \frac{100}{f^2 + 100f} - 1,$$

onde  $c_1 = 10^{-46,72}$  e  $f$  é o valor base de ferritina. Em resumo, a subida nos valores de ferritina conduz, em média, a um aumento no valor de hepcidina, embora esse aumento dependa do valor base de ferritina.

O modelo permite, pois, não só confirmar os fatores conhecidos de influência na expressão da hepcidina (níveis de ferro e grau de fibrose) como também acrescentar que a DHA exerce um papel crucial nos níveis de hepcidina (variável tipo: 1-paciente com DHA; 0-controlo saudável). Indivíduos com características semelhantes, pelo facto de desenvolverem DHA, têm significativamente menos hepcidina e este decréscimo estima-se em 87,2%  $((1-0,128) \times 100\%)$ . Fica, assim, facultada uma evidência de que os doentes de DHA têm, para os mesmos níveis de ferro, menores níveis de hepcidina, podendo esta relação ajudar a esclarecer os elevados níveis de ferro nos pacientes de DHA.

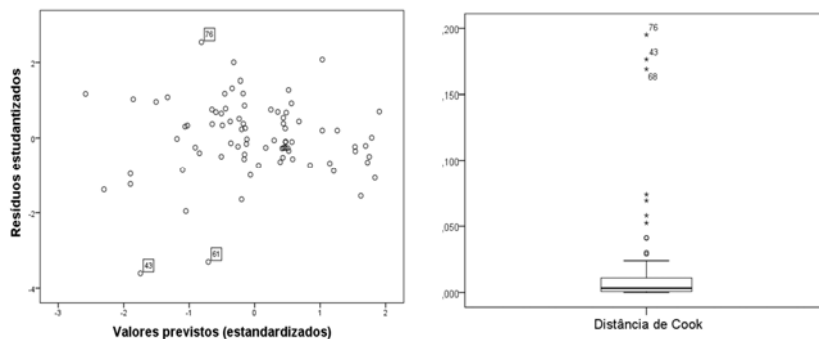
Cerca de 41% da variabilidade da hepcidina ( $R^2=0,411$ ) é explicada pela variabilidade da ferritina, pela presença/ausência de níveis graves de fibrose hepática e pela presença/ausência de DHA. O valor do coeficiente de determinação, inferior a 0,5, sugere que existem outras variáveis explicativas, não consideradas neste estudo, que influenciam a expressão da hepcidina. Em desenvolvimentos futuros deste trabalho deverá ter-se em conta esta questão.

Os pressupostos do modelo de regressão anterior foram avaliados através da análise dos resíduos. O teste de Durbin-Watson foi utilizado para avaliar o pressuposto de independência, não se tendo rejeitado a hipótese de independência entre resíduos sucessivos (estatística de Durbin-Watson=1,95). Recorrendo ao teste de Kolmogorov-Smirnov, com correção de Lilliefors, não obtivemos evidência de que o pressuposto de normalidade fosse violado ( $p>0,2$ ). Pela análise do gráfico residual (Figura 4) não há indicação de violação do pressuposto de homocedasticidade.

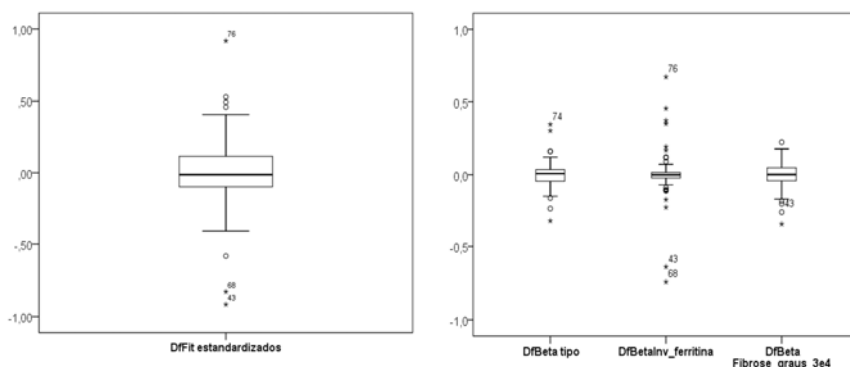
Utilizámos a tolerância e o factor de inflação da variância (do anglo-saxónico *Variance Inflation Factor* - VIF), para quantificar o quanto cada uma das variáveis independentes está associada às outras. A tolerância da variável  $X_i$  mede a proporção da sua variação que não é explicada pelas restantes variáveis independentes. Definem-se tolerância <sub>$i$</sub> = $1-R_i^2$  e VIF <sub>$i$</sub> = $1/(1-R_i^2)$ , onde  $R_i^2$  é o coeficiente de determinação do modelo entre a variável  $X_i$  (variável dependente) e todas as outras variáveis independentes. Observa-se que, para qualquer uma das



variáveis, a tolerância é superior a 0,55 e os valores de VIF são inferiores a 5 (Montgomery & Peck, 1982), refletindo ausência de forte associação entre as variáveis independentes do modelo, sendo este um indicador de não existência de problemas de multicolineariedade (Tabela 2).



**Figura 4** - Gráfico residual (à esquerda); caixa de bigodes da distância de Cook (à direita).



**Figura 5** - Caixa de bigodes dos DFFITS e DFBETAS.

Não parece, pois, haver problemas com a não satisfação dos pressupostos do modelo de regressão linear. Contudo, através da análise dos gráficos da Figura 4 – gráfico residual e a caixa de bigodes (onde se representa a distribuição dos valores da distância de Cook) – detetamos a presença de algumas observações com valores extremos, que poderão exercer influência na estimação do modelo. Os valores dos DFFITS e DFBETAS (Freund *et al.*, 2006) confirmam que as observações com números 43, 68 e 76 são potencialmente influentes na estimação dos coeficientes do modelo (Figura 5). Estimando o modelo de regressão sem cada uma destas observações (Tabela 3: modelo 43, modelo 68 e modelo 76, respetivamente), não se notam grandes alterações nos parâmetros estimados, mantendo-se todas as

conclusões tiradas do modelo original (Tabela 3: Modelo 0).

**Tabela 3** - Modelos de regressão linear múltipla.

Modelo	Modelo 0		Modelo 43		Modelo 68		Modelo 76	
	B	Erro padrão	B	Erro padrão	B	Erro padrão	B	Erro padrão
Constante	1,125	0,213	1,042	0,196	1,080	0,216	1,112	0,205
Fibr.graus 3 e 4	-0,409	0,191	-0,363	0,176	-0,443	0,193	-0,402	0,186
Tipo	-0,890	0,247	-0,821	0,227	-0,862	0,248	-0,863	0,238
Inv_ferritina	-46,72	8,425	-41,77	7,833	-40,66	9,576	-51,60	8,419

A avaliação da estabilidade do modelo foi levada a cabo através do *Bootstrap*, cujos resultados são apresentados na Tabela 4. Os valores dos erros padrão e dos intervalos de confiança estimados por *Bootstrap* são muito semelhantes aos apresentados na Tabela 2.

**Tabela 4** - Análise *Bootstrap*.

Modelo	B	<i>Bootstrap</i> <sup>a</sup>				
		Viés	Erro padrão	p	Intervalo de Confiança 95%	
1 Constante	1,125	0,014	0,204	,001	0,770	1,558
Fibr_graus_3e4	-0,409	0,001	0,158	,014	-0,723	-0,110
Tipo	-0,89	-0,011	0,205	,001	-1,351	-0,540
Inv_ferritina	-46,72	-1,125	12,195	,002	-73,682	-26,097

a. resultados baseados em 1000 amostras *bootstrap*

### 3 Conclusões

Este estudo sugere que a sobrecarga de ferro no organismo dos doentes alcoólicos do fígado está associada a uma expressão da hepcidina muito diminuída, o que permite uma maior absorção de ferro no intestino. De facto os pacientes com DHA apresentam níveis significativamente mais baixos de hepcidina o que vai permitir uma absorção excessiva do ferro. Estes resultados encorajam a continuação de estudos, no sentido de extrapolar para o ser humano a conclusão já verificada em animais pelas investigações de Takaaki *et al.*, 2007 e de Harrison *et al.*, 2007, evidenciando que DHA só por si influencia os níveis de hepcidina, o que leva a um aumento de ferro nestes pacientes.

## Referências

- FREUND, R. J., WILSON, W. J. & SA, P. (2006). *Regression Analysis, Statistical Modeling of a response variable*, 2<sup>nd</sup> edition, Academic Press.
- HARRISON, D., DUYGU, F., CRIST, C., KLEIN, E., EVANS, J., TIMCHENKO. N. & GOLLAN, J. (2007). Iron Mediated Regulation of Liver Hecpidin Expression in Rats and Mice is Abolished by Alcohol, *Hepatology*, 46 (6), 1979-1985.
- MATOS, L.C. (2006). Doença Hepática Alcoólica (DHA), *Medicina Interna*, 13, (3), 207-216.
- MONTGOMERY, D. C. & PECK, E. C. (1982). *Introducion to Linear Regression Analysis*, New York, John Wiley & Sons.
- KRAUSE, A., NEITZ, S., MÄGERT, H., SCHULZ, A., FORSSMANN, W., SCHULZ-KNAPPE, P. & ADERMANN, K. (2000). LEAP-1, A Novel Highly Disulfide-bonded Human Peptide, Exhibits Antimicrobial Activity, *FEBS Letters*, Sep 1; 480 (2-3), 147-150.
- PARK, C., VALORE, E., WARING, A. & GANZ, T. (2001). Hecpidin, a urinary antimicrobial peptide synthesized in the liver, *Journal of Biological Chemistry*, Mar 16, 276 (11), 7806-7810.
- TAKAAKI, O., HIROYUKI, S., YAYOI, H., MITSUTAKA, I., SHIGEKI, M., YASUAKI, S., YOSHINORI, F. & YUTAKA, K. (2007). Hecpidin Is Down-Regulated in Alcohol Loading, *Alcoholism: Clinical and Experimental Research*, Jan (31), No. S1.

**Editores:** Helena Bacelar-Nicolau, Fernanda Sousa, Fátima Ferreira, Luís M. Grilo, A. Manuela Gonçalves, Carlos Marcelo

O segundo volume da série CLASSIFICAÇÃO E ANÁLISE DE DADOS – Métodos e Aplicações, CLADMap II, vem dar continuidade à vontade expressa de associados e participantes nas sucessivas JOCLAD – Jornadas de Classificação e Análise de Dados, de que a CLAD divulgasse os trabalhos nelas apresentados. Os artigos incluídos neste CLADMap II, após processo de revisão inter pares, são desenvolvimentos de trabalhos apresentados nas JOCLAD 2011-2013 e espelham a interdisciplinaridade e a diversidade de áreas que integram estas Jornadas.



Associação Portuguesa de  
Classificação e Análise de Dados

ISSN 2183-8801