

# CLASSIFICAÇÃO E ANÁLISE DE DADOS

*Métodos e Aplicações III - CLADMap III*



CLAD

## Editores

Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Marcelo

Ana Sousa Ferreira

Paulo Infante

Adelaide Figueiredo

# **CLASSIFICAÇÃO E ANÁLISE DE DADOS MÉTODOS E APLICAÇÕES III - CLADMAp III**

**Editores**

Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Marcelo

Ana Sousa Ferreira

Paulo Infante

Adelaide Figueiredo

**Título**

Classificação e Análise de Dados – Métodos e Aplicações III

**Editores**

Helena Bacelar-Nicolau (Universidade de Lisboa)

Fernanda Sousa (Universidade do Porto)

Carlos Marcelo (Instituto Nacional de Estatística)

Ana Sousa Ferreira (Universidade de Lisboa)

Paulo Infante (Universidade de Évora)

Adelaide Figueiredo (Universidade do Porto)

**Impressão**

Instituto Nacional de Estatística

Av. António José de Almeida

1000-043 LISBOA

**1.ª Edição**

Lisboa, Abril de 2019

ISSN 2183-8801

Depósito legal 454535/19

Tiragem: 200 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 III – CLADMap III

Prefácio .....	vii
CLAD's 25 years .....	ix
História das Jornadas .....	xi
Agradecimento aos revisores .....	xiii
Um estudo estatístico para a avaliação do cultivo da vinha (Minho e Galiza) .....	1
<i>Daniela Catalão, A. Manuela Gonçalves, Susana Faria e Jorge Oliveira (JOCLAD2014)</i>	
Classificação hierárquica de dados intervalares da indústria automóvel com o coeficiente de afinidade .....	13
<i>Áurea Sousa, Helena Bacelar-Nicolau, Osvaldo Silva e Fernando C. Nicolau (JOCLAD2014)</i>	
Avaliações internacionais e desempenho dos alunos portugueses .....	23
<i>Ana Sousa Ferreira (JOCLAD2014)</i>	
Alguns motivos para indivíduos entre os 40 e os 49 anos quererem ter filhos .....	37
<i>Anabela Afonso, Paulo Infante e Maria Filomena Mendes (JOCLAD2015)</i>	
Alguns benefícios do envelhecimento ativo: um caso de estudo .....	45
<i>Gonçalo Jacinto, Paulo Infante, Anabela Afonso, José Conde, Teresa Engana, Jorge Vicente, Pedro Silva, Ana Carla Coelho e Paula Costa (JOCLAD2015)</i>	

Percepção de barreiras na carreira e diversidade de género: uma abordagem multivariada numa amostra de trabalhadores do sector energético .....	55
<i>Ana Sousa Ferreira e Rosário Lima (JOCLAD2015)</i>	
O Sono das crianças do 1º ciclo: caso de estudo numa escola do concelho de Évora .....	67
<i>Anabela Afonso, Gonçalo Jacinto, Paulo Infante, Teresa Engana, Filipe Glória e Rosa Espanca (JOCLAD2016)</i>	
Avaliação do efeito do desenho de amostragem em modelos de regressão logística .....	77
<i>Ana Laura Carreiras, Paulo Infante, Anabela Afonso e Maria Filomena Mendes (JOCLAD2016)</i>	
A importância da adequabilidade do modelo no desempenho de cartas de controlo com risco ajustado .....	87
<i>Maria João Inácio, Paulo Infante e Fernanda Otília Figueiredo (JOCLAD2016)</i>	
A análise classificatória na caracterização da produção e consumo de produtos de origem animal a nível mundial .....	101
<i>Manuel Minhoto e Luís Fernandes (JOCLAD2016)</i>	
A operacionalização de uma tipologia socioeconómica para as áreas metropolitanas de Lisboa e Porto .....	111
<i>Cátia Nunes e Francisco Vala (JOCLAD2016)</i>	
Dificuldades na ANOVA com dois fatores não paramétrica com células omissas .....	125
<i>Dulce G. Pereira e Anabela Afonso (JOCLAD2016)</i>	
Perfis de estudantes no contexto do Empreendedorismo: Análise de Correspondências Múltiplas e Análise de Clusters .....	135
<i>Áurea Sousa, Gualter Couto, Nélia Branco, Osvaldo Silva e Helena Bacelar-Nicolau (JOCLAD2016)</i>	
Comparação entre métodos não paramétricos para a análise de variância com dois fatores: um estudo de simulação .....	147
<i>Anabela Afonso e Dulce G. Pereira (JOCLAD2017)</i>	

A WebGIS-based platform for data management in water resources systems ...	159
<i>Marta Castilho Gomes, Reydleon Paulo, Bernardo Varela, Daniel Conde, Ricardo Canelas, Alexandre Gonçalves, Francisco Regateiro e Ana M. Ricardo (JOCLAD2017)</i>	
O uso do web scraping nas Estatísticas Oficiais .....	169
<i>Maria José Fernandes (JOCLAD2017)</i>	
Avaliação estatística da eficácia do projeto AdolesSer .....	181
<i>Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Ana Carla Coelho, Ana Gabriela Pontes, Isabel Fernandes e Edgar Palminhas (JOCLAD2017)</i>	
Labour market attractiveness in the EU at country-level .....	191
<i>João Sollari Lopes, Marco Folgado Moura e Sónia Quaresma Gonçalves (JOCLAD2017)</i>	
Avaliação de métodos de estimação da variância em amostras complexas .....	201
<i>Eládio Muianga e Anabela Afonso (JOCLAD2017)</i>	
O papel mediador da avaliação cognitiva na relação entre a mútua interferência trabalho-família e o <i>Burnout</i> .....	211
<i>Jéssica Rodrigues, A. Manuela Gonçalves, Susana Faria, Clara Simões e A. Rui Gomes (JOCLAD2017)</i>	



# Prefácio

Neste ano em que a CLAD completa os seus 25 anos de existência, publica-se o 3º Volume da série CLASSIFICAÇÃO E ANÁLISE DE DADOS - Métodos e Aplicações, CLADMap, que tem, como principal objetivo, estimular e promover a difusão de informação técnica e científica no campo da Classificação e Análise de Dados.

Os trabalhos que aqui se publicam, apresentados nas JOCLAD – Jornadas de Classificação e Análise de Dados, cobrem e combinam um vasto número de métodos e aplicações da análise de dados multivariados, em áreas tão diversas como a Sociologia, o Ambiente, a Economia, a Demografia, a Saúde ou a Educação. Não esquecendo a inovação metodológica, estes trabalhos ilustram bem a vivacidade da área do *Data Science* entre os investigadores que participam nestas jornadas.

O volume CLADMap III, vem continuar a dar resposta à vontade expressa pelos associados e participantes nas sucessivas JOCLAD, de que a CLAD divulgue os trabalhos nelas apresentados, dando-lhes visibilidade e valorizando-os. Os 20 artigos incluídos neste CLADMap III, após processo de revisão interpares, são desenvolvimentos de trabalhos apresentados nas JOCLAD 2014-2017 e mostram, mais uma vez, a interdisciplinaridade e a diversidade das áreas que integram estas Jornadas. Os textos são maioritariamente em português, refletindo o interesse /pertinência de publicar também na língua materna.

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

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

Aos investigadores teóricos e aplicados nesta área científica, fica o convite para que submetam os seus trabalhos para publicação no próximo CLADMAp IV.

Lisboa, Abril de 2019

Helena Bacelar-Nicolau (Universidade de Lisboa)  
Fernanda Sousa (Universidade do Porto)  
Carlos Marcelo (Instituto Nacional de Estatística)  
Ana Sousa Ferreira (Universidade de Lisboa)  
Paulo Infante (Universidade de Évora)  
Adelaide Figueiredo (Universidade do Porto)

## CLAD's 25 years

In 2019 the Associação Portuguesa de Classificação e Análise de Dados (CLAD), the Portuguese Association for Classification and Data Analysis, completes 25 years. In fact, CLAD was born in 1994 after a historical first Jornadas de Classificação e Análise de Dados (JOCLAD) meeting grouping more than 250 people the previous year at LEAD/FPCE at the University of Lisbon. All of them were interested about the new methods and applications related to what we now call data science. The following year, by the hand of the IFCS President Allan Gordon, CLAD became the 10th IFCS member.

The 25th CLAD anniversary will then be celebrated during the next 26th JOCLAD Annual Meeting. A special short session is planned, where the participation of the current, the last and the first CLAD President is expected.

Some of the most important milestones in CLAD's 25 years are:

- 1993, December: JOCLAD'93 - first Meeting on Classification and Data Analysis (LEAD – FPCE, University of Lisbon)
- 1994, June: CLAD foundation (with SFC, BCS, and CSNA support)
- 1994, December: JOCLAD'94 - first Meeting of the CLAD (LEAD – FPCE, University of Lisbon)
- 1995: CLAD becomes Member of IFCS (the 10th “branch” in the IFCS dendrogram)
- 1995: CLAD and SFC sign a co-operation protocol of associate membership (Paris)
- 1995: CLAD and ABE sign a co-operation protocol of associate membership (Universidade de S. Paulo)
- 1996: CLAD and GfKI sign a co-operation protocol of associate membership (Lisbon, at JOCLAD'96)
- 1997: IPM on Classification at ISI'97 Meeting, on behalf of CLAD, ISI's “Sister Society”, by invitation of the ISI Director, August 97
- 1998: CLAD becomes Member of ECAS (European Courses in Advanced Statistics)
- 1999: CLAD organizes ASMDA-99 International Conference on Applied Stochastic Models and Data Analysis. ASMDA-IS foundation (Lisbon), June 14th – 17th

- 2001: CLAD organizes EMPG2001, 32nd European Mathematical Psychology Group Meeting (Lisbon) and previous short meetings, Introductory Course on Mathematical Psychology and Data Analysis, and Workshop on Teaching and Training Mathematical Psychology in an Interdisciplinary and International Context, September 26th-29th
- 2002: CLAD sponsors the I Workshop in “Estatística e Análise de Dados”, EstAD 2002 (University of Algarve), April 12th
- 2003: CLAD organizes JISS-2003, the IASC-IFCS Joint International Summer School on Classification and Data Mining in Business, Industry and Applied Research - Methodological and Computational Issues (University of Lisbon), July 23th-30th
- 2007: CLAD and SPE sign a co-operation protocol of associate membership
- 2015: CLAD and APD sign a co-operation protocol of associate membership
- 2018: CLAD sponsors the Symbolic Data Analysis Workshop 2018 (Instituto Politécnico de Viana do Castelo), October 18th-20th

Meanwhile many other short courses, round tables, and seminars have been organized, mainly on the scope of JOCLAD and other CLAD organized/sponsored meetings.

CLAD has developed partnerships with several institutions and enterprises along its 25 years, especially with INE (CLAD’s general partner since the beginning) and Banco de Portugal (since 2012). Both institutions regularly support and/or participate at JOCLAD and other CLAD initiatives.

Helena Bacelar-Nicolau (Universidade de Lisboa, Faculdade de Psicologia e ISAMB/FMUL)

Fernanda Sousa (Universidade do Porto, Faculdade de Engenharia)

José Gonçalves Dias (Instituto Universitário de Lisboa (ISCTE-IUL))

## História das Jornadas

<b>DATA</b>	<b>SECRETÁRIO DAS JORNADAS</b>	<b>LOCAL</b>
<b>JOCLAD1993</b> 13 a 15 dez.	Helena Bacelar-Nicolau (Universidade de Lisboa)	Lisboa
<b>JOCLAD1994</b> 12 a 14 dez.	Helena Bacelar-Nicolau (Universidade de Lisboa)	Lisboa
<b>JOCLAD1995</b> 13 a 15 dez.	Helena Bacelar-Nicolau (Universidade de Lisboa)	Lisboa
<b>JOCLAD1996</b> 12 a 14 dez.	Helena Bacelar-Nicolau (Universidade de Lisboa)	Lisboa
<b>JOCLAD1997</b> 4 a 6 dez.	Paula Brito (Universidade do Porto)	Porto
<b>JOCLAD1998</b> 10 a 12 dez.	Jorge de Sá (Universidade Técnica de Lisboa)	Lisboa
<b>JOCLAD2000</b> 23 a 25 mar.	Helena Bacelar-Nicolau (Universidade de Lisboa)	Lisboa
<b>JOCLAD2001</b> 8 a 10 fev.	Pedro Duarte Silva (Universidade Católica do Porto)	Porto
<b>JOCLAD2002</b> 21 a 23 fev.	António Belo (Instituto Politécnico de Lisboa)	Lisboa
<b>JOCLAD2003</b> 10 a 12 abr.	Carlos Ferreira (Universidade de Aveiro)	Aveiro
<b>JOCLAD2004</b> 1 a 3 abr.	Pedro Coelho (Universidade Nova de Lisboa)	Lisboa
<b>JOCLAD2005</b> 20 a 23 abr.	Oswaldo Silva (Universidade dos Açores)	Ponta Delgada
<b>JOCLAD2006</b> 6 a 8 abr.	Leonor Nicolau (Universidade Lusíada de Lisboa)	Lisboa
<b>JOCLAD2007</b> 8 a 10 fev.	Helena Silva (Instituto Politécnico do Porto)	Porto

<b>JOCLAD2008</b> 27 a 29 mar.	Anibal Vieira (Instituto Politécnico de Setúbal)	Setúbal
<b>JOCLAD2009</b> 2 a 4 abr.	Filipe Barros (Universidade do Algarve)	Faro
<b>JOCLAD2010</b> 25 a 27 mar.	José Gonçalves Dias (Instituto Universitário de Lisboa)	Lisboa
<b>JOCLAD2011</b> 7 a 9 abr.	Fátima Ferreira (Universidade de Trás-os-Montes e Alto Douro)	Vila Real
<b>JOCLAD2012</b> 28 a 31 mar.	Luís Grilo (Instituto Politécnico de Tomar)	Tomar
<b>JOCLAD2013</b> 11 a 13 abr.	Arminda M. Gonçalves (Universidade do Minho)	Guimarães
<b>JOCLAD2014</b> 10 a 12 abr.	Fernanda Sousa (Universidade do Porto)	Lisboa
<b>JOCLAD2015</b> 10 a 11 abr.	Otília Dias (Instituto Politécnico de Setúbal)	Barreiro
<b>JOCLAD2016</b> 31 mar. a 2 abr.	Paulo Infante (Universidade de Évora)	Évora
<b>JOCLAD2017</b> 20 a 22 abr.	Isabel Silva Magalhães (Universidade do Porto)	Porto
<b>JOCLAD2018</b> 5 a 7 de abr.	M. Filomena Teodoro (Escola Naval)	Almada
<b>JOCLAD2019</b> 11 a 13 abr.	Ana Cristina Matos (Instituto Politécnico de Viseu)	Viseu

## Agradecimento aos revisores

Os revisores dos artigos submetidos a esta publicação CLADMAp III 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
Pavel Brazdil	Universidade do Porto
Paula Brito	Universidade do Porto
Jorge Cadima	Universidade de Lisboa
Teresa Calapez	Instituto Universitário de Lisboa
Pedro Campos	Universidade do Porto
Margarida Cardoso	Instituto Universitário de Lisboa
José Dias	Instituto Universitário de Lisboa
Pedro Duarte Silva	Universidade Católica Portuguesa
Susana Faria	Universidade do Minho
Carlos Ferreira	Universidade de Aveiro
Fernanda Figueiredo	Universidade do Porto
Rita Gaio	Universidade do Porto
Luis M. Grilo	Instituto Politécnico de Tomar
Carla Henriques	Instituto Politécnico de Viseu
Victor Lobo	Escola Naval e Universidade Nova de Lisboa
Manuela Maia	Universidade Católica Portuguesa
Manuela Neves	Universidade de Lisboa
Rosário Oliveira	Universidade de Lisboa
Irene Oliveira	Universidade de Trás-os-Montes e Alto Douro
Conceição Rocha	INESC -TEC e Universidade do Porto
Fátima Salgueiro	Instituto Universitário de Lisboa
Áurea Sousa	Universidade dos Açores
Fernanda Sousa	Universidade do Porto
Paula Vicente	Instituto Universitário de Lisboa



## Um estudo estatístico para a avaliação do cultivo da vinha (Minho e Galiza)

Daniela Catalão<sup>1</sup> · A. Manuela Gonçalves<sup>2</sup> · Susana Faria<sup>2</sup> · Jorge Oliveira<sup>3</sup>

**Resumo** Neste trabalho foi realizado um estudo estatístico, nas áreas da Estatística Multivariada e da Inferência Estatística, com o objetivo de avaliar e interpretar a variabilidade de um alargado conjunto de informação sobre as características físico-químicas do solo que influenciam o desenvolvimento das videiras, a qualidade do mosto e das uvas e, por consequência, a qualidade dos vinhos. Este estudo incidiu sobre duas parcelas de cultivo de vinha, uma situada no Minho (Portugal) e outra localizada na Galiza (Espanha). A aplicação da Análise Fatorial permitiu obter um pequeno número de fatores que são responsáveis por explicar grande parte da variabilidade total associada aos dados originais relativos ao solo, reduzindo assim a dimensionalidade do problema. A Análise de *Clusters* permitiu identificar regiões com características físico-químicas do solo semelhantes, nas parcelas em estudo.

**Palavras-chave:** Análise Fatorial, Análise de *Clusters*, Cultivo da Vinha, Inferência Estatística, Solo.

### 1 Introdução

Este trabalho resultou da necessidade de se identificar quais as variáveis físico-químicas do solo que influenciam a produtividade das videiras e a qualidade das uvas, do mosto e, conseqüentemente, dos vinhos numa parcela da Estação Vitivinícola Amândio Galhano, no concelho de Arcos de Valdevez (Minho-Portugal) e numa parcela localizada na Bodega Santiago Ruiz, em Tomiño (Região da Galiza-Espanha). Pretende-se com este conhecimento e num mercado cada vez

---

<sup>1</sup> DMAT – Departamento de Matemática e Aplicações, Universidade do Minho, Portugal,  
*danielacatalao@hotmail.com*

<sup>2</sup> CMAT – Centro de Matemática, DMAT - Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, *mneves@math.uminho.pt*; *sfaria@math.uminho.pt*

<sup>3</sup> Sinergeo, Braga, Portugal, *jorgeoliveira@sinergeo.pt*

mais competitivo, otimizar qualitativamente e quantitativamente a produção de vinho. O conhecimento adequado das características do solo e das exigências da vinha é fundamental para se conseguir atingir níveis de produção economicamente rentáveis e de qualidade do vinho (Jordão, 2007).

As vinhas em estudo do Minho eram constituídas apenas pela casta tinta Vinhão, enquanto a vinha da Galiza pela casta branca Alvarinho (Figura 1). Na parcela da Estação Vitivinícola Amândio Galhano, o solo é um solo residual granítico, enquanto na Bodega Santiago Ruiz em Tomiño (Galiza) o solo é um aluvião, constituído na sua maioria por acumulações de seixos, frequentemente cobertos por camadas arenoargilosas e matéria orgânica.



**Figura 1** – À esquerda, exemplo do cacho e da folha da casta tinta Vinhão (fonte: <http://www.winesofportugal.info/>); à direita, exemplo do cacho e da folha da casta branca Alvarinho (fonte: <http://www.winesofportugal.info/>).

Os dados analisados, recolhidos em campo e em laboratório, reportaram-se ao ano de 2011. Na quinta do Minho, a recolha das uvas para o estudo da produtividade, mostos e prova foi efetuada a 16/09/2011, e na quinta da Galiza foi realizada no dia 23/08/2011. Quanto à recolha das amostras do solo para análise, esta foi realizada em Abril de 2010, em ambas as quintas, tendo sido posteriormente analisadas no Laboratório de Análise de Solos da Escola Superior Agrária do Instituto Politécnico de Viana do Castelo.

Os principais objetivos deste trabalho foram descrever o comportamento das diferentes variáveis em estudo e identificar regiões com características físicas e químicas do solo semelhantes, nas parcelas em estudo. Também se pretendeu verificar se existia um pequeno conjunto de variáveis que fosse responsável por explicar uma proporção elevada da variação total associada ao conjunto original das variáveis do solo que influenciam a qualidade do vinho e quais as regiões que apresentam grandes/pequenas concentrações dessas variáveis e, deste modo, reduzir a dimensionalidade do problema (Catalão, 2014).

Para a concretização destes objetivos foram usadas metodologias das áreas da Estatística Multivariada (Análises de *Clusters* e Análise Fatorial) e de Inferência Estatística (em particular, testes de hipóteses não paramétricos, de modo a lidar com o problema da não normalidade dos dados e do número reduzido de observações de algumas variáveis). A Análise de *Clusters* foi aplicada ao conjunto das variáveis do solo, a partir das variáveis originais do solo e a partir das variáveis

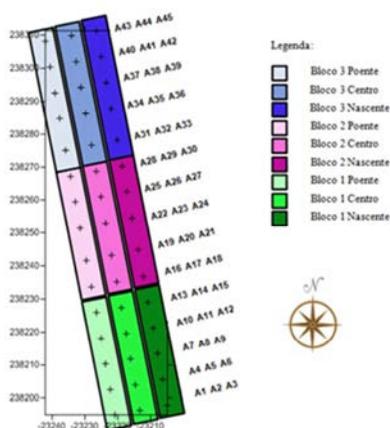
fatoriais retidas pela aplicação de uma Análise Fatorial a essas mesmas variáveis. A qualidade do vinho e o equilíbrio vegetativo da videira foram avaliados através de análise de correlações com as variáveis do solo.

Esta análise foi efetuada nas duas parcelas em estudo (Minho e Galiza) e, no final, os resultados de ambas foram analisados e comparados.

As várias metodologias de análise estatística foram realizadas utilizando o *software* R, com o auxílio de várias *packages*, e o *software* SPSS (*Statistical Package for Social Sciences*).

## 2 Metodologia

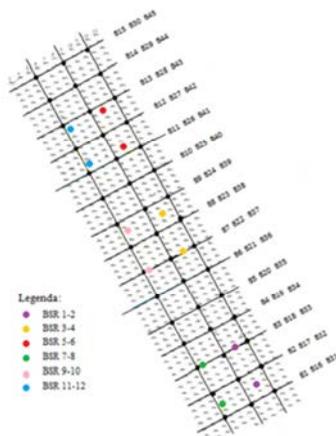
Na parcela do Minho foram georreferenciados 45 pontos com espaçamento regular, em 9 regiões, cada uma contendo 5 destes pontos onde foram recolhidas as amostras do solo, com o objetivo de se analisarem algumas variáveis referentes a características físico-químicas do solo (Figura 2). Nestas localizações também se avaliaram as variáveis referentes à produtividade das videiras. Foram recolhidas 9 amostras de mosto a partir das quais foram avaliadas algumas variáveis relativas às características físico-químicas do mosto; também foi feita uma análise do rendimento em sumo e da composição cromática e aromática do mosto (Araújo, 2004). Por último, foram analisados os vinhos correspondentes às 9 regiões referidas anteriormente.



**Figura 2** – Parcela do Minho com a localização dos 45 pontos georreferenciados bem como a sua divisão em 9 parcelas.

Na parcela da Galiza também foram georreferenciados 45 pontos, de onde foram recolhidas as amostras do solo e analisadas algumas variáveis relativas às

características físico-químicas do solo (Figura 3). Relativamente às variáveis alusivas à produção das videiras (número de cachos por videira, peso médio do cacho por videira, peso bruto por videira e peso de cachos por videira), estas foram avaliadas de acordo com 12 regiões. As 21 variáveis analisadas referentes às uvas (como por exemplo massa, percentagem do rendimento em sumo) e ao mosto (23 variáveis, como por exemplo teor de álcool provável, percentagem do rendimento em sumo) foram avaliadas, apenas a partir de 6 amostras de uvas provenientes de 6 regiões. Por fim, foram analisados os 6 vinhos correspondentes a estas 6 regiões (avaliação final atribuída ao vinho).



**Figura 3** – Parcela da Galiza com a localização dos 45 pontos georreferenciados bem como a sua divisão em 6 parcelas.

A título de exemplo na Tabela 1 apresentam-se as variáveis caracterizadoras das características do solo (16 variáveis) que foram avaliadas em ambas as parcelas.

**Tabela 1** – Variáveis do solo em estudo nas parcelas do Minho e da Galiza.

	Variáveis		Descrição
	Designação	Unidades	
SOLO	DA	(g/vol)	Densidade Aparente
	MO	(%)	Matéria Orgânica
	pH		pH em Estrato Aquoso
	FF	(%)	Fração Fina (%FF+%FG=100%)
	FG	(%)	Fração grosseira (%FF+%FG=100%)
	P2O5	(μg/g)	Fósforo Assimilável
	K2O	(%)	Potássio Assimilável
	Ca	(μg/g)	Cálcio Assimilável

Mg	( $\mu\text{g/g}$ )	Magnésio Assimilável
Azt	( $\mu\text{g/g}$ )	Azoto Total
Ni	( $\mu\text{g/g}$ )	Níquel
Cr	(%)	Crómio
Cd	( $\mu\text{g/g}$ )	Cádmio
N	( $\mu\text{g/g}$ )	Nitratos
B	( $\mu\text{g/g}$ )	Boro
CTC	( <i>m.e./100g</i> )	Capacidade de Troca Catiótica

Numa primeira fase efetuou-se uma análise exploratória dos dados, e verificou-se a existência de valores em falta (*Not Available* (NA)) nas bases de dados referentes ao Minho e à Galiza. De acordo com a opinião do perito na área em estudo, optou-se por substituir os valores em falta pelo método do vizinho mais próximo: a partir do cálculo da distância Euclidiana encontrou-se o ponto georreferenciado, que estava mais próximo do que continha o NA e partilharam o respetivo valor da variável. Dada a grande percentagem de valores em falta na variável Densidade Aparente (DA) do solo da Galiza, cerca de 46,67%, foi decidido em concordância com o perito na área em estudo, remover a variável (na Galiza o solo é constituído por seixos de grande dimensão que impossibilitou a recolha do material para a análise desta variável).

Numa segunda fase desta análise, após se proceder à substituição dos NA, fez-se uma análise preliminar dos dados com o objetivo de maximizar a obtenção de informações sobre os dados e, assim, descobrir tendências, detetar alguns padrões, *outliers*, que pudessem guiar os procedimentos a realizar posteriormente. No entanto, quando se realizou uma Análise Fatorial, esta foi efetuada com e sem os *outliers* severos, e os resultados foram similares, pelo que se optou por trabalhar com estes valores, pois a sua eliminação implicaria perda de informação uma vez que não foram erros de registo, mas dados observados.

Posteriormente foram realizados dois estudos pela aplicação da Análise de *Clusters* (Branco, 2004; Chatefield & Collins, 1980) ao conjunto dos dados do solo: um estudo a partir das variáveis originais do solo e o outro a partir dos fatores retidos pela aplicação de uma Análise Fatorial a essas mesmas variáveis (Everitt & Hothorn, 2011). Esta análise foi efetuada nas duas parcelas em estudo, no Minho e na Galiza e, no final, os resultados de ambas foram analisados e comparados, retirando-se as principais conclusões.

### 3 Resultados

#### 3.1 Estudo 1

A Análise de *Clusters*, realizada pela aplicação de métodos hierárquicos aglomerativos (aplicação da distância euclidiana e do método de Ward) a partir das variáveis originais do solo, permitiu obter um agrupamento dos pontos de amostragem em três grupos homogêneos, relativamente às características do solo em ambas as parcelas (Minho e Galiza).

Assim, relativamente à parcela do Minho, obteve-se um dendrograma que permitiu identificar 3 *clusters* (associado a uma correlação cofenética de 0,74). Observou-se que o 1º *cluster* é constituído pelas parcelas B1P, B1C, B1N e B2N, o 2º *cluster* engloba a zona B2P e B2C e ao 3º *cluster* pertencem os terrenos B3P, B3C e B3N. A parcela escolhida na quinta do Minho é caracterizada por um solo residual granítico, com um ligeiro aumento da cota de Sul para Norte. A drenagem das águas superficiais ocorre de Norte para Sul e de Oeste para Este. Empiricamente, esta parcela é dividida em 3 blocos (Bloco 1, Bloco 2 e Bloco 3). O Bloco 1 corresponde aos lotes B1P, B1C e B1N; o bloco 2 corresponde a B2P, a B2C e a B2N e o Bloco 3 corresponde a B3P, a B3C e a B3N. Face ao exposto, considera-se que os *clusters* 1, 2 e 3 diferenciam-se pela cota que se verifica no terreno. O *cluster* 3 está a uma cota superior ao *cluster* 2 e este está a uma cota superior ao *cluster* 1.

Relativamente à parcela da Galiza, obteve-se um dendrograma que permitiu identificar 4 *clusters* (associado a uma correlação cofenética de 0,62). Apesar dos critérios sugerirem a retenção de 4 *clusters*, apenas foram considerados 3 grupos homogêneos, pois o 4º *cluster* a ser considerado continha apenas uma observação, BSR 5; realizado um estudo a este elemento, verificou-se que este representa um *outlier* moderado nas variáveis P2O5, Ca, Mg e CTC. De acordo com os peritos nas áreas em estudo, não haveria necessidade de se criar um quarto *cluster* contendo esta única videira. Assim, o 1º *cluster* é constituído pelas parcelas BSR 1, BSR 2, BSR 3, BSR 7 e BSR 8, o 2º *cluster* é formado pelas parcelas BSR 4, BSR 5 e BSR 6 e o 3º *cluster* formado por BSR 9, BSR 10, BSR 11 e BSR 12. Sendo caracterizada por terraços fluviais do Rio Minho, a parcela total em estudo na Galiza evoluiu de granulometria baixa (silte e argila) para uma granulometria mais grosseira com seixos que podem ir até aos 10 cm de comprimento. O *cluster* 1 corresponde à zona da parcela constituída exclusivamente por argila e silte. A transição para granulometrias mais extensas (argila, areia, seixos) ocorre nos *clusters* 2 e 3. Sabe-se também que existe uma ligeira subida na cota de Sul para Norte, ou seja, a cota do *cluster* 1 é inferior à dos *clusters* 2 e 3.

Com a aplicação do teste não paramétrico de Kruskal-Wallis (Higgins, 2004), verificou-se que as variáveis do solo da parcela do Minho MO, FF, FG, P2O5, Mg,

AzT, Ni, Cr, B e CTC apresentaram diferenças significativas entre os diferentes *clusters*. Estas diferenças foram ainda detetadas para as variáveis ácido málico do mosto e para a família de compostos em C6 do aroma do mosto na fração livre (FL1M). Relativamente à parcela da Galiza, as variáveis do solo MO, P2O5, K2O, Ca, Mg, AzT, Cd, N, B e CTC foram as que apresentaram diferenças significativas entre os três *clusters*. Foram, ainda, encontradas estas diferenças para a variável acidez total do mosto, para as famílias dos compostos em C6 do aroma das uvas na fração livre (FL1U), para as famílias compostos carbonilados do aroma das uvas na fração livre (FL5U), para as famílias de compostos em C6 do aroma das uvas na fração glicosilada (FG1U), para as famílias de compostos carbonilados do aroma das uvas na fração glicosilada (FG7U) e para a variável representativa da nota final atribuída ao vinho. O teste de comparações múltiplas não paramétrico LSD permitiu, ainda, avaliar em que *clusters* é que as variáveis do solo apresentavam diferenças significativas.

### 3.2 Estudo 2

Com o objetivo de transformar um problema que envolve um elevado número de variáveis originais e correlacionadas num problema com um número reduzido de variáveis, efetuou-se uma Análise Fatorial (por aplicação do método de verosimilhança máxima) a cada conjunto de dados referentes às duas parcelas (Minho e Galiza). Com esta análise pretendeu-se verificar se existia um pequeno número de variáveis que fossem responsáveis por explicar grande parte da variabilidade total associada aos dados originais relativos ao solo e, portanto, reduzir assim a dimensionalidade do problema.

Para dar início à realização de uma Análise Fatorial, analisaram-se as relações existentes entre os diferentes constituintes químicos do solo em estudo e, para tal, calcularam-se os valores das correlações de Spearman (bem como a aplicação do teste de correlação de Spearman) entre cada par de constituintes. Em particular, a produtividade da videira e o seu equilíbrio vegetativo, a qualidade das uvas, do mosto e do vinho foram avaliados através de análise de correlações com as variáveis do solo e identificaram-se as zonas das parcelas que apresentaram défice/excesso nas concentrações dessas variáveis.

Da aplicação das Análises Fatoriais efetuadas aos dados de ambas as parcelas com aplicação da rotação *Varimax*, para facilitar a interpretabilidade dos fatores, retiveram-se quatro fatores, que explicaram aproximadamente 72% da variabilidade total dos dados originais da parcela do Minho e 85,5% da variabilidade total dos dados originais da quinta da Galiza.

**Tabela 2** – *Loadings* dos 4 primeiros factores (rotação *Varimax*) e comunalidades das variáveis em estudo; proporção da variância explicada por cada fator (Minho).

Variável	Fator				Comunalidade
	1	2	3	4	
DA	-0,268	0,121	0,307	<b>-0,770</b>	0,774
MO	<b>0,961</b>	-0,051	0,206	0,029	0,969
pH	-0,169	<b>0,686</b>	-0,297	0,109	0,599
FF	0,288	-0,101	<b>0,756</b>	-0,122	0,679
P2O5	-0,457	<b>0,574</b>	-0,003	0,321	0,641
K2O	0,036	0,196	0,352	<b>0,566</b>	0,484
Mg	-0,191	<b>0,844</b>	0,090	-0,003	0,757
AzT	<b>0,968</b>	0,034	0,165	0,042	0,967
Ni	0,462	-0,246	<b>0,563</b>	0,065	0,594
Cr	-0,093	0,020	<b>0,801</b>	-0,109	0,662
Cd	-0,183	0,327	0,074	<b>0,631</b>	0,544
N	0,332	<b>0,645</b>	-0,110	0,114	0,552
B	<b>0,946</b>	-0,012	0,233	0,016	0,950
CTC	0,113	<b>0,920</b>	-0,028	0,116	0,874
Valor próprio	3,563	3,011	1,986	1,487	
Proporção de variância (%)	25,451	21,509	14,183	10,619	
Proporção de variância acumulada (%)	25,451	46,960	61,143	71,763	

A Tabela 2 representa os resultados da Análise Fatorial aplicada aos dados da parcela do Minho. Foram retidos 4 fatores (aplicação do critério de Kaiser). Analisando os valores dos *loadings* das variáveis latentes, ou seja, avaliando o peso que cada variável original tem em cada um dos fatores extraídos, verifica-se que: o 1º fator está fortemente correlacionado positivamente com os constituintes MO, AzT, e B; o 2º fator apresenta uma correlação positiva elevada com os constituintes pH, P2O5, Mg, N e CTC; o 3º fator apresenta valores positivos de correlações mais elevados para as variáveis FF e Ni e um valor negativo mais elevado para Cr; o 4º fator é influenciado positivamente por K2O e Cd e negativamente por DA.

Observando a comunalidade associada a cada variável original, ou seja, a proporção da variância de cada variável explicada pelos fatores retidos, verifica-se que K2O, N e Cd são as que mais informação perdem; estes quatro fatores apenas conseguem explicar 48,4%, 55,2% e 54,4% da variância destas variáveis, respetivamente.

Para a parcela da Galiza da Análise Fatorial consideraram-se apenas 4 fatores explicando aproximadamente 85,5% da variabilidade total dos dados (aplicação do critério de Kaiser), como se pode observar pela Tabela 3.

**Tabela 3** – *Loadings* dos 4 primeiros fatores (rotação *Varimax*) e comunalidades das variáveis em estudo; proporção da variância explicada por cada fator (Galiza).

Variável	Fator				Comunalidade
	1	2	3	4	
MO	0,521	<b>0,771</b>	0,093	0,255	0,939
pH	0,021	<b>-0,893</b>	0,002	-0,062	0,802
FF	-0,022	-0,011	<b>0,954</b>	0,102	0,921
P2O5	<b>0,890</b>	0,296	-0,038	0,112	0,894
K2O	<b>0,655</b>	0,519	0,070	0,254	0,767
Mg	<b>0,939</b>	0,088	-0,079	-0,156	0,920
AzT	0,577	<b>0,698</b>	0,101	0,278	0,907
Ni	-0,109	0,176	0,083	<b>0,911</b>	0,879
Cr	-0,059	0,606	<b>0,617</b>	-0,032	0,752
Cd	<b>0,643</b>	0,524	0,191	-0,239	0,782
N	0,396	<b>0,724</b>	0,073	0,020	0,687
B	0,528	<b>0,744</b>	0,071	0,284	0,918
CTC	<b>0,953</b>	0,098	-0,065	-0,129	0,939
Valor próprio	4,482	4,003	1,380	1,243	
Proporção de variância (%)	34,474	30,790	10,616	9,561	
Proporção de variância acumulada (%)	34,474	65,264	75,880	85,441	

Observando-se as correlações entre os vários constituintes do solo e os fatores apresentados na Tabela 3, verifica-se que as variáveis que apresentam maior correlação positiva com o 1º fator são as variáveis CTC, Mg, P2O5, K2O e Cd; relativamente ao 2º fator são as variáveis MO, AzT, N e B que apresentam maior correlação positiva. Este fator ainda apresenta uma forte correlação negativa com a variável pH; os constituintes FF e Cr são os que mais influenciam o 3º fator, com correlação positiva; relativamente ao 4º fator é a variável Ni que apresenta uma forte correlação positiva com este fator. Observando a comunalidade associada a cada variável original, no geral, os fatores comuns retidos conseguem explicar uma boa proporção de variância de cada variável original, sendo o constituinte N o que mais informação perde, uma vez que os 4 fatores apenas conseguem explicar aproximadamente 69% da variância desta variável.

Foram, para ambas as vinhas em estudo, calculados os *scores* de cada fator extraídos através do método dos mínimos quadrados ponderados. De seguida,

procedeu-se a uma Análise de *Clusters* com o objetivo de identificar, em cada parcela em estudo, zonas com características semelhantes em relação às variáveis do solo.

Os resultados obtidos a partir desta Análise de *Clusters* não expuseram uma clara estrutura dos grupos subjacentes aos dados da parcela da Galiza. Relativamente à parcela do Minho, a Análise de *Clusters* realizada a partir dos fatores, bem como os testes de comparações múltiplas, evidenciou resultados semelhantes aos obtidos a partir das variáveis originais do solo.

## **4 Conclusões e trabalho futuro**

A análise dos resultados obtidos a partir das duas bases de dados, que incorporavam um conjunto de dados recolhidos em campo e em laboratório relativos ao ano 2011, pôs em evidência a influência do solo sobre o rendimento e qualidade das uvas e do vinho. Foi possível identificar as características mais relevantes do solo que estão correlacionadas com o desenvolvimento das videiras e com a qualidade das uvas e respetivos vinhos, nestas duas vinhas em estudo.

Este trabalho evidenciou um conjunto de indicadores que poderão ajudar no processo de tomada de decisão para a racionalização de utilização dos fatores de produção, como os nutrientes e os produtos fitofármacos, contribuindo, assim, para uma maior produtividade da videira e uma maior qualidade da uva e do vinho.

A produtividade da videira, a qualidade das uvas, do mosto e do vinho foram avaliadas através das suas correlações com as variáveis do solo, permitindo avaliar e identificar as regiões das parcelas que apresentaram um défice/excesso nas concentrações destas variáveis e, assim, sempre que se justifique, efetuar correções diferenciadas para cada local da parcela de cultivo, efetuando-se deste modo, uma agricultura de uma forma mais sustentada.

No entanto, no decorrer deste estudo foram encontradas algumas limitações que impossibilitaram a aplicação de diversas metodologias estatísticas, que seriam do maior interesse para o desenvolvimento e valorização deste trabalho. Estas limitações devem-se, essencialmente, à necessidade de revisão do processo de amostragem que deve permitir que haja uma melhor recolha das amostras, com um número mais elevado de observações e idêntico para cada variável em estudo; a realização do estudo em vários anos para assim se poder tirar conclusões fidedignas e até mesmo perceber-se melhor a evolução da qualidade do fruto e da videira e, conseqüentemente, do vinho, caso se altere as concentrações de alguns constituintes do solo; a introdução no estudo da variável água disponível no solo, uma vez que se trata de um elemento que influencia de forma significativa a absorção dos nutrientes por parte das plantas, podendo, assim, influenciar a produtividade da videira e a qualidade das uvas, do mosto e do vinho; a recolha e

estudo de variáveis referentes à climatologia (precipitação, exposição solar, etc.), visto que a sua interação com o fator solo resulta na produção de vinhos de qualidade e, portanto, uma boa avaliação destes fatores em conjunto ajudaria a aumentar e a melhorar a rentabilidade dos sistemas produtivos e a minimizar a degradação do recurso do solo, para garantir a sustentabilidade dos sistemas de produção vitivinícola.

## Agradecimentos

Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - “Fundação para a Ciência e a Tecnologia”, no âmbito do projeto PEstOE/MAT/UI0013/2017.

## Referências

- ARAÚJO, J. (2004). *Características aromáticas e cromáticas das castas Amarel e Vinhão*. Tese de Mestrado em Viticultura e Enologia. Universidade do Porto e Universidade Técnica de Lisboa, Porto.
- BRANCO, J.A. (2004). *Uma Introdução à Análise de Clusters*, Sociedade Portuguesa de Estatística, Évora.
- CATALÃO, D.B. (2014). *Metodologia Estatística para a Avaliação de um Recurso Natural (Minho e Galiza)*. Tese de Mestrado em Estatística. Universidade do Minho, Guimarães.
- CHATEFIELD, C. & COLLINS, A.J. (1980). *Introduction to Multivariate Analysis*, Chapman and Hall, Cambridge.
- EVERITT, B. & HOTHORN, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer, USA.
- HIGGINS, J.J. (2004). *Introduction to Modern Nonparametric Statistics*, Thomson, Duxbury Advanced Series, Toronto.
- JORDÃO, A.J. (2007). *Gestão do Solo na Vinha*. Texto elaborado no âmbito do Plano de Acção para a Vitivinicultura da Alta Estremadura.



# Classificação hierárquica de dados intervalares da indústria automóvel com o coeficiente de afinidade

Áurea Sousa<sup>1</sup> · Helena Bacelar-Nicolau<sup>2</sup> · Osvaldo Silva<sup>3</sup> · Fernando C. Nicolau<sup>4</sup>

**Resumo** São mostradas as estruturas classificatórias mais relevantes, obtidas com a Análise Classificatória Hierárquica de um conjunto de dados retirado da literatura da análise de dados complexos, referentes a modelos de carros (*Car data set*), com base no coeficiente de afinidade generalizado, ponderado, centrado e reduzido pelo método de Wald e Wolfowitz, e em critérios de agregação probabilísticos no âmbito da Metodologia VL. Comparam-se os resultados obtidos com os de outros autores e diferentes métodos, bem como com uma partição definida *a priori*. Os métodos utilizados foram capazes de detetar classes consistentes e consonantes de macro-dados, quer com a partição *a priori* quer com os resultados de outros autores, e enquadram-se nas metodologias de Análise de Dados Simbólicos (ADS), sendo de salientar que as agregações implícitas nessas abordagens permitem o tratamento de bases de micro-dados (dados não agregados) de grande dimensão.

**Palavras-chave:** Análise Classificatória Hierárquica, Coeficiente de Afinidade Generalizado Ponderado, Dados Simbólicos ou Complexos, Metodologia VL.

## 1 Introdução

Na sociedade atual, devido aos contínuos avanços computacionais, é cada vez mais frequente a utilização de bases de dados de elevada dimensão, sendo fundamental efetuar a síntese de conjuntos de dados em termos dos seus conceitos subjacentes, os quais têm de ser descritos por dados complexos, também designados por dados simbólicos. Estes dados podem ser representados em tabelas, cujas células incluem

---

<sup>1</sup> Universidade dos Açores, CEEApIA, [aurea.st.sousa@uac.pt](mailto:aurea.st.sousa@uac.pt)

<sup>2</sup> Universidade de Lisboa, Faculdade de Psicologia e ISAMB/FMUL, [hbacelar@psicologia.ulisboa.pt](mailto:hbacelar@psicologia.ulisboa.pt)

<sup>3</sup> Universidade dos Açores, Centro Interdisciplinar de Ciências Sociais – CICS.UAc/CICS.NOVA.UAc, [osvaldo.dl.silva@uac.pt](mailto:osvaldo.dl.silva@uac.pt)

<sup>4</sup> Universidade Nova de Lisboa, FCT, Departamento de Matemática e DataScience

um ou mais valores, tais como subconjuntos de categorias, intervalos da reta real, ou distribuições de frequências (e.g., Bock e Diday, 2000; Diday e Noirhomme-Fraiture, 2008; Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau *et al.*, 2009, 2010; Sousa *et al.*, 2010, 2013a). As linhas da tabela de dados correspondem geralmente às unidades de dados (frequentemente grupos de indivíduos) e as colunas às variáveis, referidas então, respetivamente, como objetos simbólicos e variáveis simbólicas.

Uma solução viável para a análise de conjuntos de dados de dimensão elevada é a identificação prévia de classes nesses dados. Trata-se, geralmente, de classes de indivíduos, cuja descrição é depois feita através de variáveis simbólicas, que sintetizam as suas propriedades em matrizes de macro-dados (e.g., Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau *et al.*, 2009, 2014, 2015; Bock e Diday, 2000; Sousa, 2005). Os métodos ilustrados no presente trabalho, com base num *case-study*, são assim apropriados para a Análise Classificatória (*Cluster Analysis*) de bases de micro-dados de dimensão elevada: a matriz de dados resultou da agregação dos dados originais, relativos a um número elevado de carros dos modelos selecionados, pelo que cada modelo representa uma classe de indivíduos (carros) e as variáveis que descrevem os modelos são variáveis intervalares.

A Análise Classificatória tem como objetivo identificar grupos (classes) de entidades (indivíduos, objetos, etc.), relativamente homogêneos e, de preferência, bem separados, com base nas semelhanças ou dissemelhanças entre essas entidades. Existem dois grupos principais de métodos de análise classificatória: os hierárquicos e os não hierárquicos. Os primeiros geram uma sequência de partições encaixadas (estrutura hierárquica), enquanto os segundos fornecem uma única partição dos elementos a classificar num número  $k$  de classes. Neste artigo, são utilizados métodos de Análise Classificatória Hierárquica Ascendente (ACHA), e selecionamos depois uma ou mais partições relevantes da hierarquia, através do cálculo de estatísticas apropriadas e da visualização dos dendrogramas obtidos.

Têm sido referidas na literatura várias medidas de dissemelhança entre objetos simbólicos (e.g., Bock e Diday, 2000; Chavent e Lechevallier, 2002; Chavent *et al.*, 2003; Souza e De Carvalho, 2004; De Carvalho *et al.*, 2006a, 2006b), assim como algumas medidas de semelhança (e.g., Bacelar-Nicolau, 2000; Bacelar Nicolau *et al.*, 2009, 2010, 2014, 2015). Uma vez obtida a matriz de proximidades entre os elementos do conjunto a classificar, podem ser aplicados critérios de agregação clássicos ou probabilísticos. Destes últimos, foram usados critérios incluídos na Metodologia VL – onde a sigla VL se refere a “*Vraisemblance du Lien*” e “*Validade da Ligação*” - que recorrem a noções probabilísticas, para a definição das funções de comparação (e.g., Lerman, 1981, 2000, 2016; Nicolau, 1983; Bacelar-Nicolau, 1987; Nicolau e Bacelar-Nicolau, 1998; Bacelar-Nicolau *et al.*, 2009, 2010; Sousa *et al.*, 2013a).

No presente artigo, são apresentados os principais resultados obtidos com a ACHA de trinta e três modelos de carros (*Car data set*, conjunto de dados retirado da literatura da análise de dados complexos), com base no coeficiente de afinidade

generalizado ponderado (Bacelar-Nicolau, 2000; Bacelar-Nicolau *et al.*, 2009, 2010, 2015), centrado e reduzido pelo método de Wald e Wolfowitz, e em três critérios de agregação probabilísticos, *AVL* (*Algoritmo da Validade da Ligação*), *AVI* e *AVB*, incluídos numa família paramétrica de métodos no contexto da Metodologia *VL* (Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1970, 1972, 1981, 2016). As partições resultantes foram comparadas com as obtidas por outros autores (e.g., De Carvalho *et al.*, 2006a, 2006b; Souza *et al.*, 2007) e com uma partição definida *a priori*, a qual será descrita na Secção 3. A Secção 2 refere-se ao coeficiente de afinidade generalizado ponderado, centrado e reduzido sob uma hipótese de referência baseada no teorema limite de Wald e Wolfowitz. Na Secção 3, são resumidos os resultados mais preponderantes relativos à ACHA do conjunto de dados em análise, o qual tem uma estrutura conhecida. Enfim, da Secção 4 constam algumas considerações finais sobre os métodos e os resultados obtidos.

## 2 O coeficiente de afinidade generalizado ponderado centrado e reduzido pelo método de Wald e Wolfowitz

A partir do coeficiente de afinidade entre duas distribuições de probabilidade discretas, proposto por Matusita (1951), Bacelar-Nicolau (e.g., 1980, 1988) introduziu o coeficiente de afinidade no domínio da Análise Classificatória, para avaliar a semelhança entre pares de colunas ou pares de linhas de uma matriz de dados, ou seja, entre variáveis ou entre indivíduos, conforme o conjunto que se pretende classificar, desenvolvendo um estudo teórico aprofundado sobre as propriedades deste coeficiente. O coeficiente de afinidade foi depois estendido a diferentes tipos de dados, incluindo os heterogéneos e de natureza complexa ou simbólicos (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau *et al.*, 2009, 2010, 2015; Sousa *et al.*, 2013a), frequentemente resultantes da agregação de micro-dados, por vezes de elevada dimensão. É neste contexto que aqui utilizamos o coeficiente de afinidade generalizado ponderado,  $a(k, k')$ , entre duas unidades de dados  $k$  e  $k'$ .

Seja  $E = \{1, \dots, N\}$  um conjunto de unidades de dados descritas por  $p$  variáveis,  $Y_1, \dots, Y_p$ , cujos valores são intervalos da reta real, isto é cada entrada da matriz de dados contém um intervalo,  $I_{kj} = [a_{kj}, \dots, b_{kj}]$ , com  $k=1, \dots, N$  e  $j=1, \dots, p$ , tendo-se que a  $k$ -ésima unidade de dados é descrita pela linha:  $[a_{k1}, \dots, b_{k1}] [a_{k2}, \dots, b_{k2}] \dots [a_{kp}, \dots, b_{kp}]$ . Neste caso, é possível decompor o domínio,  $I_j = \cup I_{kj}$  ( $k = 1, \dots, N$ ), de cada variável  $V_j$  (abreviadamente,  $j$ ), num número  $m_j$  de intervalos elementares e disjuntos,  $\{I_{j\ell}: \ell = 1, \dots, m_j\}$ , com  $\ell = 1, \dots, m_j$ , tais que para  $\ell, \ell' = 1, \dots, m_j$ ,  $\ell \neq \ell'$ ;  $k = 1, \dots, N$ , são verificadas as seguintes propriedades: i)  $I_j = \cup I_{j\ell}$ ; ii)  $I_{j\ell} \cap I_{j\ell'} = \emptyset$ , isto é  $|I_{j\ell} \cap I_{j\ell'}| = 0$ ; iii)  $|I_{kj} \cap I_{j\ell}| = |I_{j\ell}|$ , se  $|I_{kj} \cap I_{j\ell}| \neq 0$ ;  $|I_{kj} \cap I_{j\ell}| = 0$ , caso contrário, onde  $| \cdot |$  denota

a amplitude do intervalo (e.g., Bacelar-Nicolau *et al.*, 2009, 2014, 2015; Sousa *et al.*, 2013a). Consequentemente, os intervalos originais,  $I_{kj}$  ( $k = 1, \dots, N; j = 1, \dots, p$ ) também podem ser decompostos em intervalos elementares disjuntos. Nessas condições, o coeficiente de afinidade generalizado ponderado, assintoticamente centrado e reduzido pelo método de Wald e Wolfowitz,  $a_{WW}(k, k')$ , entre duas unidades de dados  $k$  e  $k'$  ( $k, k' = 1, \dots, N$ ), é dado por:

$$a_{WW}(k, k') = \sum_{j=1}^p \pi_j \sqrt{m_j - 1} \frac{\sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell} \cdot x_{k'j\ell}} - \frac{1}{m_j} \sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell}} \sum_{\ell=1}^{m_j} \sqrt{x_{k'j\ell}}}{\sqrt{\left[ x_{kj\bullet} \cdot \frac{1}{m_j} \left( \sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell}} \right)^2 \right] \left[ x_{k'j\bullet} \cdot \frac{1}{m_j} \left( \sum_{\ell=1}^{m_j} \sqrt{x_{k'j\ell}} \right)^2 \right]}}$$

onde  $\pi_j$  são pesos, tais que  $0 \leq \pi_j \leq 1$  e  $\sum \pi_j = 1$ ;  $x_{kj\ell} = |I_{kj} \cap I_{j\ell}|$  e  $x_{k'j\ell} = |I_{k'j} \cap I_{j\ell}|$ , tendo-se, por exemplo,  $x_{kj\ell} = |I_{j\ell}|$  se  $I_{kj} \cap I_{j\ell} = I_{j\ell}$ , e  $x_{kj\ell} = 0$ , caso contrário, representando  $|I_{j\ell}|$  a amplitude do intervalo  $I_{j\ell}$ ;  $x_{kj\bullet} = \sum_{\ell=1}^{m_j} x_{kj\ell}$  e  $x_{k'j\bullet} = \sum_{\ell=1}^{m_j} x_{k'j\ell}$ .

O coeficiente assintoticamente centrado e reduzido, sob uma hipótese de referência permutacional baseada no teorema limite de Wald e Wolfowitz permite, por sua vez, definir um coeficiente probabilístico no contexto da metodologia VL, na linha iniciada por Lerman (1970, 1972, 1981) e desenvolvida por Bacelar-Nicolau (e.g., 1980, 1987, 1988) e Nicolau (e.g., 1983, 1998). A utilização do coeficiente probabilístico VL associado a  $a_{WW}(k, k')$ , em vez do coeficiente  $a(k, k')$ , faculta a utilização de valores da medida de semelhança comparáveis numa escala probabilística, os quais são realizações de variáveis aleatórias com a mesma distribuição (assintoticamente, normal reduzida). Estudos de simulação (Sousa, 2005; Sousa *et al.*, 2013b) mostraram que a convergência dos valores do coeficiente de afinidade para a distribuição normal é rápida.

Aplicações da extensão do coeficiente de afinidade para o caso de dados simbólicos podem ser encontradas, por exemplo, em Bacelar-Nicolau *et al.* (2009, 2010, 2014, 2015) e em Sousa *et al.* (2010, 2012, 2013a, 2015a, 2015b).

### 3 Exemplo da indústria automóvel: “Car data set”

A matriz de dados “Car data set” é referida na literatura da análise de dados simbólicos (e.g., De Carvalho *et al.*, 2006a, 2006b; Souza *et al.*, 2007) e contém trinta e três modelos, onde cada modelo representa um conjunto de carros, descritos por oito variáveis cujos valores são intervalos da reta real (“Preço”, “Cilindrada”, “Velocidade Máxima”, “Aceleração”, “Distância entre os eixos”, “Comprimento”, “Largura” e “Altura”), duas variáveis categóricas (“Alimentação” e “Tração”) com categorias não ordenadas que podem assumir múltiplos valores (subconjuntos de

categorias) e uma variável nominal (“*Categoria do Carro*”). Esta última variável, com as modalidades “*Utilitário*” (U), “*Berlina*” (B), “*Desportivo*” (S) e “*Luxo*” (L), é considerada como sendo a classificação *a priori* dos modelos de carros (De Carvalho *et al.*, 2006a, 2006b). A Tabela 1 mostra uma parte da matriz de dados simbólicos, sendo de referir que a matriz de dados completa está disponível no software SODAS (*Symbolic Official Data Analysis System*). A primeira linha da tabela, por exemplo, descreve as características dos carros que, na matriz de dados originais, pertencem ao modelo “*Alfa 145*”.

**Tabela 1** – Parte da matriz de dados simbólicos -“*Car data set*”.

<i>Modelo</i>	<i>Preço</i>	<i>Cilindrada</i>	...	<i>Categoria</i>
<i>Alfa 145</i>	[27806, 33596]	[1370, 1910]	...	<i>Utilitário</i>
<i>Alfa 156</i>	[41593, 62291]	[1598, 2492]	...	<i>Berlina</i>
...	...	...	...	...
<i>Passat</i>	[39676, 63455]	[1595, 2496]	...	<i>Luxo</i>

Para efeito do cálculo dos valores do coeficiente  $\alpha_{ww}(k, k')$ , os domínios das oito variáveis de intervalo (“*Preço*”, “*Cilindrada*”, “*Velocidade Máxima*”, “*Aceleração*”, “*Distância entre os eixos*”, “*Comprimento*”, “*Largura*” e “*Altura*”) foram decompostos, conforme explicado na secção anterior, em intervalos elementares e disjuntos, satisfazendo as condições enunciadas. Os resultados, para a presente base de dados, foram descritos em Sousa *et al.* (2015b).

A aplicação da ACHA sobre a sub-matriz resultante, utilizando o método AVL, cujo dendrograma se pode encontrar em Sousa *et al.* (2015b), permitiu realçar quatro classes bem definidas, cuja estrutura pode ser visualizada na Figura 1:

Classe 1: {1/U, 14/B, 26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L};

Classe 2: {6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S};

Classe 3: {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U};

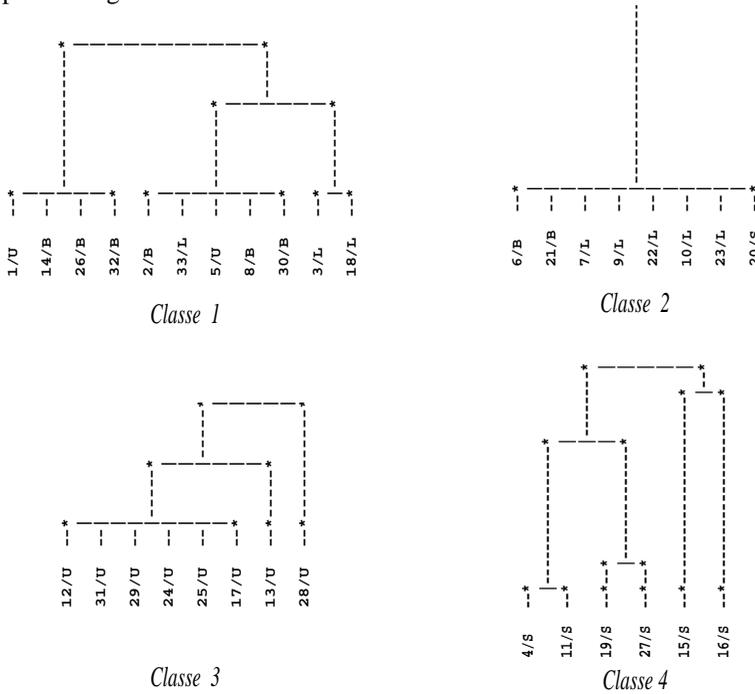
Classe 4: {4/S, 11/S, 19/S, 27/S, 15/S, 16/S}.

Conforme mostra a Figura 1, a Classe 1 resulta da reunião das subclasses {1/U, 14/B, 26/B, 32/B} e {2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L} e a Classe 4 resultou da ligação das subclasses {4/S, 11/S, 19/S, 27/S} e {15/S, 16/S}, observando-se um efeito de cadeia na Classe 3.

As classes acima indicadas estão em consonância com as obtidas por outros autores (e.g., De Carvalho *et al.*, 2006a, 2006b; Souza *et al.*, 2007), com base em diferentes métodos, diferindo unicamente no que respeita às classes em que são inseridas as unidades de dados 20/S e 21/B.

No que se refere às partições obtidas pelos três métodos utilizados (AVI, AVB e AVL), é de realçar que as classes {6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S} e {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U} foram encontradas em todos os dendrogramas, tratando-se assim de duas classes consistentes. Por outro lado, a classe {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U} foi também identificada

pelos autores citados. Importa, enfim, referir que a partição em quatro classes, fornecida pelo método de ACHA descrito, é compatível com a partição definida *a priori* pelas categorias referentes aos modelos de carro.

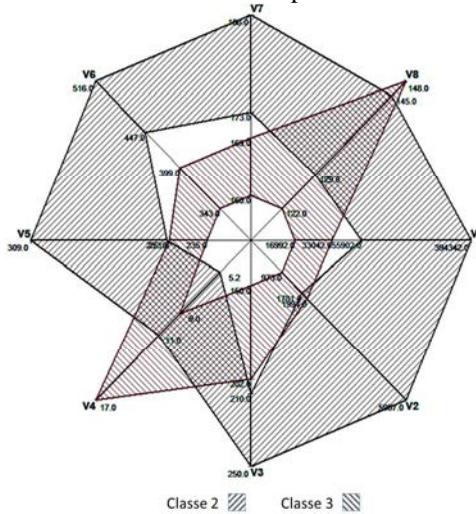


**Figura 1** – Quatro classes evidenciadas no dendrograma obtido pelo AVL (níveis 21 a 32).

Os gráficos Zoom-star a duas dimensões (2 D) correspondentes às duas classes consistentes, referidas acima, são representados na Figura 2 e mostram, de uma forma visual, a informação descrita na Tabela 2, a qual corresponde à matriz de dados simbólicos que contém as descrições dos modelos de carros incluídos em cada uma das classes consistentes supracitadas. As notações V1, V2, V3, V4, V5, V6, V7 e V8 designam, respetivamente, as variáveis “Preço”, “Cilindrada”, “Velocidade Máxima”, “Aceleração”, “Distância entre os eixos”, “Comprimento”, “Largura” e “Altura”.

A Figura 2 e a Tabela 2 mostram que os modelos de carros incluídos na Classe 2 apresentam características bastante diferenciadas das dos incluídos na Classe 3. Note-se que os modelos incluídos na Classe 2 apresentam valores mais elevados nas variáveis V1, V2, V3, V5, V6 e V7, comparativamente aos da Classe 3. Os valores das variáveis V4 e V8 apresentam uma maior variabilidade no caso dos

modelos de carros englobados na Classe 3, notando-se no caso destas variáveis uma sobreposição parcial entre os intervalos correspondentes das duas classes.



**Figura 2** – Sobreposição dos gráficos Zoom star (2D) relativos às classes consistentes.

**Tabela 2** – Descrição das classes consistentes com base nas variáveis V1 a V8.

Classe 2: {6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S}			
V1	V2	V3	V4
[55902, 394342]	[1781, 5987]	[210, 250]	[5.2, 11]
V5	V6	V7	V8
[252, 309]	[447, 516]	[173, 188]	[129, 145]
Classe 3: {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U}			
V1	V2	V3	V4
[16992, 33042]	[973, 1994]	[150, 202]	[9, 17]
V5	V6	V7	V8
[235, 251]	[343, 399]	[160, 169]	[122, 148]

Os resultados obtidos evidenciam que o coeficiente de afinidade centrado e reduzido pelo método de Wald e Wolfowitz captou, juntamente com os critérios de agregação probabilísticos, a estrutura classificatória inerente a este conjunto de dados.

#### 4 Considerações finais

A decomposição dos intervalos originais em intervalos elementares e disjuntos é particularmente interessante no caso do coeficiente  $a_{WW}(k, k')$ , devido à

correspondente hipótese de referência associada a este coeficiente, a qual é baseada no teorema limite de Wald e Wolfowitz.

Neste artigo, ilustrámos a aplicação da ACHA de unidades de dados descritas por variáveis de diferentes magnitudes cujos valores são intervalos da reta real. Os métodos utilizados conduziram à obtenção de estruturas classificatórias que estão próximas das obtidas (uma delas idêntica) por outros autores, com base em diferentes métodos. É ainda de referir que a partição em quatro classes, obtida pelo método AVL, é compatível com a partição definida *a priori* pelas categorias referentes aos modelos de carros.

A aplicação destes métodos a diversos conjuntos de dados, com estrutura conhecida, tem vindo a confirmar a sua capacidade de reproduzirem as diversas estruturas classificatórias, mesmo no caso de conjuntos de dimensão elevada.

## Referências

- BACELAR-NICOLAU, H. (1980). *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória, Tese de Doutoramento, FCL, Universidade de Lisboa.*
- BACELAR-NICOLAU, H. (1987). *On the distribution equivalence in cluster analysis.* In Devijver, P.A. & Kittler, J. (Eds.) Pattern Recognition Theory and Applications, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, Springer - Verlag, New York, 73-79.
- BACELAR-NICOLAU, H. (1988). *Two probabilistic models for classification of variables in frequency tables.* In: BOCK, H.-H. (Eds.), Classification and related methods of data analysis, Elsevier Sciences Publishers B.V., North Holland, pp. 181-186.
- BACELAR-NICOLAU, H. (2000). *The Affinity Coefficient.* In: BOCK, H.-H. & DIDAY, E. (Eds.), Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, 160-165.
- BACELAR-NICOLAU, H. (2002). On the Generalised Affinity Coefficient for Complex Data. *Biocybernetics and Biomedical Engineering*, 22(1), 31-42.
- BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á. & BACELAR-NICOLAU, L. (2009). Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets. *Biocybernetics and Biomedical Engineering*, 29 (2), 9-18.
- BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á. & BACELAR-NICOLAU, L. (2010). Clustering Complex Heterogeneous Data Using a Probabilistic Approach, *Proceedings of the Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)*, 85-93 (electronic publication).
- BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á. & BACELAR-NICOLAU, L. (2014). Clustering of Variables with a Three-Way Approach for Health Sciences.

- Testing, Psychometrics, Methodology in Applied Psychology (TPM)*, 21(4 Special Issue), 435-447. doi: 10.4473/TPM21.4.5.
- BACELAR-NICOLAU, H., NICOLAU, F.C., SOUSA, Á. & BACELAR-NICOLAU, L. (2015). *On Cluster Analysis of Complex and Heterogeneous Data*. In: MANCA R., MCCLEAN, S., SKIADAS, C.H. (Eds), *New Trends in Stochastic Modeling and Data Analysis*, ISAST, Athens, 123-134.
- BOCK, H.-H. & DIDAY, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Series: Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Berlin.
- CHAVENT, M. & LECHEVALLIER, Y. (2002). *Dynamical clustering algorithm of interval data: Optimization of an adequacy criterion based on Hausdorff distance*. In: JAJUGA, K., SOKOLOWSKI, A. & BOCK H.-H. (Eds.), *Classification, clustering, and data analysis*, Springer-Verlag, Berlin, 53-60.
- CHAVENT, M., DE CARVALHO, F.A.T., LECHEVALLIER, Y. & VERDE, R. (2003). *Trois Nouvelles Méthodes de Classification Automatique de Données Symboliques de Type Intervalle. Revue de Statistique Appliquée*, vol. LI, no. 4, 5-29.
- DE CARVALHO, F.A.T., BRITO, P. & BOCK, H.-H. (2006a). *Dynamic Clustering For Interval Data Based on L2 Distance. Computational Statistics*, 21(2), 1-19.
- DE CARVALHO, F.A.T., SOUZA, R.M.C.R. DE, CHAVENT, M. & LECHEVALLIER, Y. (2006b). *Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters*, 27(3), 167-179.
- DIDAY, E. & NOIRHOMME-FRAITURE, M. (Eds.) (2008). *Symbolic Data Analysis and the SODAS Software*, Chichester, John Wiley & Sons.
- LERMAN, I.C. (1970). *Sur l'analyse des données préalable à une classification automatique (Proposition d'une nouvelle mesure de similarité). Rev. Mathématiques et Sciences Humaines*, 32, 5-15.
- LERMAN, I.C. (1972). *Étude Distributionnelle de Statistiques de Proximité entre Structures Algébriques Finies du Même Type: Application à la Classification Automatique, Cahiers du B.U.R.O.*, 19, PARIS.
- LERMAN, I.C. (1981). *Classification et analyse ordinaire des données*, Dunod, Paris.
- LERMAN, I.C. (2000). *Comparing Taxonomic Data. Math. & Sci. Hum.*, 38<sup>e</sup> année, n° 150, 37-51.
- LERMAN, I.C. (2016). *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series: Advanced Information and Knowledge Processing*, Springer-Verlag, London. doi: 10.1007/978-1-4471-6793-8.
- MATUSITA, K. (1951). *On the Theory of Statistical Decision Functions. ANN. Instit. Stat. Math*, III, 1-30.
- NICOLAU, F.C. (1983). *Cluster Analysis and Distribution Function. Methods of Operations Research*, 45, 431-433.

- NICOLAU, F.C. & BACELAR-NICOLAU, H. (1998). *Some trends in the classification of variables*. In: HAYASHI, C., OHSUMI, N., YAJIMA, K., TANACA, Y., BOCK, H.-H & BABA, Y. (Eds) *Data Science, Classification, and Related Methods*, Springer-Verlag, 89-98.
- SOUSA, Á. (2005). *Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa, Tese de Doutoramento*, Universidade dos Açores.
- SOUSA, Á., NICOLAU, F.C., BACELAR-NICOLAU, H. & SILVA, O. (2010). Weighted Generalised Affinity Coefficient in Cluster Analysis of Complex Data of the Interval Type. *Biometrical Letters*, 47(1), 45-56.
- SOUSA, Á., SILVA, O., BACELAR-NICOLAU, H. & NICOLAU, F.C. (2012). Aplicação do Coeficiente de Afinidade a Dados de Natureza Complexa. *Boletim da Sociedade Portuguesa de Matemática - Número especial - Encontro Nacional da SPM*, 171-174.
- SOUSA, Á., NICOLAU, F.C., BACELAR-NICOLAU, H. & SILVA, O. (2013a.) Clustering of Symbolic Data based on Affinity Coefficient: Application to a Real Data Set. *Biometrical Letters*, 50 (1), 27-38. doi: 10.2478/bile-2013-0015
- SOUSA, Á., SILVA, O., BACELAR-NICOLAU, H. & NICOLAU, F.C. (2013b). Distribution of the affinity coefficient between variables based on the Monte Carlo simulation method. *Asian Journal of Applied Sciences*, 1(5), 236-245.
- SOUSA, Á., BACELAR-NICOLAU, H., & SILVA, O. (2015a). *Hierarchical Cluster Analysis of Groups of Individuals: Application to Business Data*. In: FILUS, L., OLIVEIRA, T. & SKIADAS, C. H. (Eds), *Stochastic Modeling, Data Analysis and Statistical Applications*, ISAST, Athens, 195 - 203.
- SOUSA, Á., BACELAR-NICOLAU, H., NICOLAU, F.C. & SILVA, O. (2015b). On Clustering Interval Data with Different Scales of Measures: Experimental Results. *Asian Journal of Applied Science and Engineering*, 4(1), 17-25.
- SOUZA, R.M.C.R., & DE CARVALHO, F.A.T. (2004). Clustering of interval data based on City-Block distances. *Pattern Recognition Letters*, 25(3), 353-365. doi: 10.1016/j.patrec.2003.10.016
- SOUZA, R.M.C.R., DE CARVALHO, F.A.T. & PIZZATO, D.F. (2007). A Partitioning Method for Mixed Feature-Type Symbolic Data Using a Squared Euclidean Distance. In: FREKSA, C., KOHLHASE, M. & SCHILL, K. (Eds.) *KI 2006, Advances in Artificial Intelligence. 29TH ANNUAL GERMAN CONFERENCE ON AI, KI 2006, BREMEN, GERMANY, JUNE 14-17, 2006, Proceedings. Series: Lecture Notes in Computer Science*, VOL. 4314, Berlin Heidelberg, Springer-Verlag, 260-273.

## Avaliações internacionais e desempenho dos alunos portugueses

Ana Sousa Ferreira<sup>1</sup>

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

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

### 1 Introdução

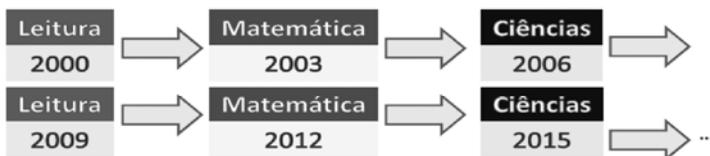
A OCDE (Organização para a Cooperação e Desenvolvimento Económico) e a IEA (*International Association for the Evaluation of Educational Achievement*) promovem programas internacionais de avaliação do desempenho de crianças e jovens do mundo inteiro em várias áreas como a matemática, ciências e leitura. Portugal tem vindo a participar em diversos destes estudos internacionais como o *PISA* (*Programme for International Students Assessment*), o *TIMSS* (*Trends in International Mathematics and Science Study*) e o *PIRLS* (*Progress in International Reading Literacy Study*). Todos estes projetos avaliam a proficiência de alunos de vários países, em diversos domínios, contextualizada pela informação recolhida em diversos questionários, ao aluno, à escola, aos pais,...

O *PISA*, desenvolvido pela OCDE, visa avaliar se os alunos de 15 anos, aqueles que na maior parte dos países participantes se aproximam do final da escolaridade obrigatória, estão bem preparados para enfrentarem os desafios da vida quotidiana. A prova *PISA* foi concebida para avaliar, em ciclos trienais, a literacia de Matemática, de Leitura e de Ciências, de jovens de 15 anos que frequentem pelo

---

<sup>1</sup> Faculdade de Psicologia, Universidade de Lisboa e Business Research Unit (BRU-IUL), Lisboa, Portugal, [asferreira@psicologia.ulisboa.pt](mailto:asferreira@psicologia.ulisboa.pt)

menos o 7.º ano de escolaridade,<sup>2</sup> elegendo em cada ciclo, uma das três áreas referidas como domínio principal a avaliar (Figura 1).



**Figura 1** - Sequência de ciclos *PISA*.

Mais recentemente, foram considerados alguns domínios secundários, como a capacidade de resolução de problemas em 2012, e de resolução de problemas em contexto colaborativo em 2015. Opcionalmente, foi ainda considerado o domínio secundário de literacia financeira, não tendo os alunos portugueses participado nesta prova. Portugal tem vindo a participar no projeto *PISA* desde o primeiro ciclo, em 2000, até ao último ciclo aplicado, em 2015.

O *TIMSS* é uma avaliação internacional do desempenho dos alunos do 4.º e do 8.º ano de escolaridade em Matemática e Ciências, desenvolvida pela *IEA*. O *TIMSS Advanced* é uma versão do estudo que tem o objetivo de avaliar as tendências do desempenho dos alunos no final do ensino secundário, em Matemática avançada e Física. Desde 1995, os testes *TIMSS* são aplicados de quatro em quatro anos com a finalidade de gerar informação de qualidade sobre os resultados do desempenho dos alunos e sobre os contextos em que estes aprendem. Portugal participou no primeiro ciclo do *TIMSS* em 1995, e voltou a participar na prova do 4º ano em 2011, e em 2015. Em 2015, os alunos portugueses que frequentavam Matemática A e Física do 12º ano participaram, pela primeira vez, no *TIMSS Advanced*.

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

Em 2016, Portugal participou novamente no projeto *PIRLS* e, estreou-se na nova prova, o *ePIRLS*, o novo programa de literacia de Leitura em formato digital desenvolvido pela *IEA*. Em 2011, coincidiu a aplicação dos dois estudos da *IEA*, o *TIMSS* e o *PIRLS*, o que permitiu um novo olhar sobre o desempenho dos alunos

<sup>2</sup> Em 2000, na primeira ronda *PISA*, foram selecionados alunos desde o 5.º ao 11.º ano de escolaridade.

do 4º ano de escolaridade, relacionando os seus resultados a literacia de Leitura com as pontuações alcançadas em literacia de Matemática e Ciências (ProjAVI, 2013 a).

## **2 As provas *PISA*, *TIMSS* e *TIMSS Advanced*, *PIRLS* e *ePIRLS***

A OCDE e a IEA planeiam cuidadosamente todo o processo de desenvolvimento das provas internacionais, para garantir a comparabilidade dos resultados, definindo procedimentos padronizados e implementando um exigente controlo de qualidade. Por exemplo, é realizada uma extensa série de verificações para garantir a comparabilidade das traduções dos itens de avaliação e dos questionários; é fornecida documentação detalhada sobre os padrões de amostragem; são minuciosamente verificadas as taxas de participação das escolas e dos alunos, e aplica-se um rigoroso programa de garantia da qualidade para monitorizar a recolha dos dados.

As provas em análise têm diversas características comuns:

- Todas consideram um grande número de itens nas provas, de forma a poder avaliar os diversos domínios e conteúdos;
- Os itens não são públicos, permitindo, assim, a comparação de resultados ao longo das diversas edições e a identificação de tendências. Habitualmente, em cada ciclo, são libertos alguns itens que exemplificam o tipo de perguntas apresentadas aos alunos, substituindo-se, no ciclo seguinte, esses itens libertos por novos itens;
- A par dos testes, são também aplicados questionários visando recolher informação de contexto que permite descrever as situações e os fatores que influenciam a aprendizagem;
- O desempenho global de cada aluno, em cada domínio, é estimado, como se tivesse respondido ao conjunto de todos os itens, usando métodos complexos denominados de metodologia dos valores plausíveis, tendo em conta a dificuldade dos itens e a dispersão dos resultados;
- Em cada país participante, a amostragem é realizada em duas fases: num primeiro momento são seleccionadas aleatoriamente as escolas e, num segundo momento, são seleccionados aleatoriamente os alunos (*PISA*) ou as turmas de alunos (*TIMSS* e *PIRLS*) de cada escola participante.

Existem, porém, algumas diferenças entre a prova implementada pela OCDE e as provas da responsabilidade da IEA. No caso do *PISA*, o estudo é ancorado na idade dos alunos e visa comparar os diversos sistemas educativos através da avaliação das competências adquiridas pelos alunos, de forma a estarem preparados

para o ingresso na vida ativa. No caso dos projetos *TIMSS* e *PIRLS*, são ancorados no ano de escolaridade (4º, 8º ou ano terminal do ensino secundário) e pretendem avaliar os conhecimentos adquiridos pelos alunos em relação aos conteúdos de cada curriculum.

Após a apresentação dos relatórios internacionais sobre os resultados de um ciclo, as bases de dados dos vários estudos são disponibilizadas e podem ser descarregadas gratuitamente nos *sites* do *PISA*, *TIMSS* e *PIRLS*<sup>3</sup>.

## 2.1 A prova *PISA*

A prova do *PISA* é constituída por vários cadernos de teste que combinam itens dos diferentes domínios avaliados, respondendo cada aluno a um único caderno de teste, que lhe é atribuído aleatoriamente. Entre 2000 e 2012, o teste *PISA* foi aplicado de forma tradicional, com papel e lápis, com uma duração total de duas horas, ocupando os itens do domínio principal cerca de 2/3 do tempo total da prova. Em 2015, 57 dos 72 países participantes realizaram a prova em formato digital, após um estudo piloto, em 2014, ter analisado a invariância de modo da prova, isto é, tendo sido verificado que não existe um efeito do modo de aplicação da prova sobre o desempenho dos alunos, pelo que é possível manter o estudo da série de resultados *PISA* entre 2000 e 2015 (Marôco *et al.*, 2016a).

Em 2012, a prova foi organizada em 13 cadernos diferentes e em 2015, já aplicada em formato digital, consideraram-se 66 versões distintas (Sousa Ferreira *et al.*, 2017). O teste *PISA* considera itens relativos quer aos processos cognitivos necessários para a sua resolução quer aos conteúdos. Por exemplo, em 2012, os 65 países e economias participantes realizaram a avaliação em literacia matemática, tendo sido aplicados itens relativos a três processos cognitivos - *Formular; Aplicar e Interpretar* – e quatro conteúdos – *Quantidade, Incerteza, Mudança e relações, Espaço e forma*.

O *PISA* não se limita a avaliar se um aluno reproduz eficazmente os conhecimentos adquiridos, procura antes aferir se os alunos conseguem aplicar, em contextos diferenciados, o que aprenderam. Uma abordagem desta natureza procura perceber se as sociedades contemporâneas reconhecem e valorizam os indivíduos não por aquilo que eles sabem mas por aquilo que eles conseguem fazer com o que sabem (OCDE, 2014a).

Visando estimar a literacia de uma população, o desempenho global de cada aluno é estimado em *PISA*, como nos outros estudos de larga escala, usando métodos complexos, considerando uma escala de 0 a 1000 com um valor médio de 500 e um desvio-padrão de 100 (OECD, 2014b). Neste projeto, são também

---

<sup>3</sup> Ver <http://www.oecd.org/pisa/data/>; <https://timssandpirs.bc.edu/timss2015/international-database/> e <https://timssandpirs.bc.edu/pirls2016/international-database/>.

recolhidas informações que permitem contextualizar os resultados, através da aplicação de questionários aos alunos, aos pais e aos diretores das escolas e, em 2015, também aos professores. O *PISA* permite, pois, identificar os fatores que influenciam os níveis de desempenho nos vários domínios de literacia, nomeadamente, a informação sociodemográfica dos alunos e a sua relação com a aprendizagem e, ainda, diversas características das escolas participantes, como a sua organização e recursos.

## 2.2 As provas *TIMSS* e *TIMSS Advanced*

A prova do *TIMSS* é constituída por um conjunto de cadernos, numa composição de itens de matemática e de ciências que abrangem os três domínios cognitivos – *Conhecer*, *Aplicar* e *Raciocinar*. Cada aluno responde a um único caderno de prova. A avaliação em matemática considera duas dimensões: uma relativa ao conteúdo – *Números*, *Formas Geométricas e Medida*, *Apresentação de Dados* – e a outra ao domínio cognitivo. A avaliação em ciências segue o mesmo desenho, sendo contempladas as seguintes áreas: *Ciências da Vida*, *Ciências Físicas e Ciências da Terra*. Neste estudo, a par dos testes, são também aplicados questionários visando recolher informação de contexto que permite descrever os fatores que influenciam a aprendizagem da matemática e das ciências (ProjAVI, 2012a, 2012b).

O *TIMSS* 2015 contou com a participação de cerca de 60 países e economias, tendo Portugal participado no *TIMSS* 4º ano e, pela primeira vez, no *TIMSS Advanced*. No *TIMSS* 2015 consideraram-se 14 cadernos de teste diferentes, respondendo cada aluno apenas a um deles. Os catorze cadernos foram distribuídos pelos alunos de cada turma do 4º ano, de forma a obter-se um número de respostas equivalentes a cada um. No *TIMSS* 2011 e 2015 existia um total de 28 blocos de itens (com aproximadamente 10-14 itens cada um, na prova do 4º ano), catorze de matemática e os restantes catorze de ciências, sendo cada caderno de teste constituído por diversas combinações destes blocos de itens. Espera-se que os alunos do 4º ano gastem cerca de 18 minutos em cada bloco de itens, durando a prova um máximo de 72 minutos, seguidos de um período de 30 minutos para a resposta ao Questionário ao Aluno (Marôco *et al.*, 2016b)

O *TIMSS Advanced* é a única avaliação internacional que fornece informação sobre os conhecimentos dos alunos em Matemática avançada e em Física. Os alunos avaliados frequentam o ano terminal do ensino secundário em cursos que dão acesso a carreiras *STEM* (*Science, Technology, Engineering and Mathematics*). Em 2015, a prova de Matemática avançada considerou as áreas de *Álgebra*, *Cálculo* e *Geometria* e a de Física incidiu sobre *Mecânica* e *Termodinâmica*, *Eletricidade* e *Magnetismo*, *Fenómenos Ondulatórios* e *Física Atómica*,

consistindo a prova delineada em aproximadamente 100 itens para cada um dos dois domínios. Neste projeto participaram nove países: Eslovénia, Estados Unidos da América, França, Itália, Noruega, Portugal, Rússia e Suíça (Marôco *et al.*, 2016c).

### 2.3 As provas *PIRLS* e *ePIRLS*

O *PIRLS* é uma prova de avaliação da literacia de Leitura. No relatório produzido pelo Instituto de Avaliação Educativa (IAVE) a propósito dos resultados do ciclo de 2016, pode ler-se:

“...A literacia de leitura nos sistemas educativos modernos é particularmente relevante para a faixa etária dos alunos que participam no *PIRLS* (9-10 anos): até essa idade, a criança aprende a ler; daí em diante, lê para aprender. É, pois, fundamental que o domínio da leitura seja alcançado nessa idade. Ao avaliar as competências da leitura, o *PIRLS* afere áreas onde é necessário intervir de modo a que a leitura se transforme num meio eficaz para adquirir conhecimento.”

(IAVE, 2017, p. 3).

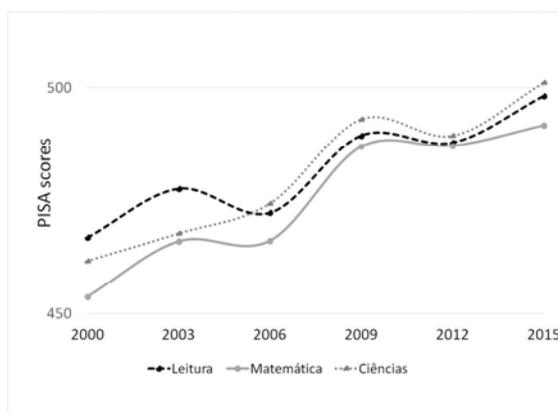
A prova do *PIRLS* foi desenhada de modo a contemplar duas finalidades de leitura, a *literária* - ler como experiência literária - e a *informativa* - ler para adquirir e utilizar informação -, assim como os processos de compreensão da leitura - *Refer, Fazer inferências diretas, Interpretar e integrar e Avaliar* -. A prova é constituída por um conjunto de cadernos, numa composição de itens que envolvem as diferentes finalidades e processos de compreensão da leitura. Cada aluno responde apenas a um caderno de prova. Em 2011, a prova *PIRLS* foi constituída por 13 cadernos. No último ciclo desta prova, em 2016, foram concebidos 16 cadernos de teste, dos quais quinze apresentavam duas partes: a primeira incluía um texto literário seguido de um conjunto de questões; a segunda incluía um texto informativo e as respetivas questões. O 16º caderno incluía um texto literário, um texto informativo e um conjunto de questões relativas aos dois textos (ProjAVI, 2012c; IAVE, 2017).

No *ePIRLS*, aplicado pela primeira vez em 2016, avalia-se a literacia de Leitura *online*. As tarefas propostas aos alunos baseiam-se, deste modo, em simulações de páginas da *internet*, incidindo sobre *Ciências sociais e humanas* e *Ciências da terra, da vida e do espaço*. Estas páginas de *internet* simuladas, são muito semelhantes às verdadeiras páginas da *internet*, utilizando recursos interativos, de multimédia, janelas de navegação e *links* para recursos digitais. Na prova *ePIRLS* avalia-se apenas a compreensão da leitura e não as competências de navegação pelo que surgem também tutoriais e até “professores *Avatar*” que pretendem orientar a navegação dos alunos. Nesta prova, as questões incidem sobre as cinco temáticas de ciências já referidas, combinadas duas a duas. Também nesta prova, cada aluno responde apenas a um caderno de prova, composto por duas temáticas e respetivas questões. Neste estudo, também são aplicados questionários visando recolher

informação de contexto que permite perceber que fatores influenciam a literacia de Leitura *online* (IAVE, 2017).

### 3 O desempenho dos alunos portugueses

A análise do desempenho dos alunos portugueses pode ser efetuada por valores de referência internacionais (*benchmarks*), ou por conteúdos e domínios cognitivos, ou ainda usando a informação de contexto recolhida por qualquer das avaliações internacionais em que Portugal participa. Como referido, o projeto *PISA* teve início em 2000 e Portugal participou em todos os ciclos *PISA* até à última aplicação em 2015. Em 2000, participaram na prova 4585 alunos portugueses de quinze anos e 149 escolas participantes, e na última aplicação, em 2015, cerca de 7325 alunos e 246 agrupamentos de escolas/escolas não agrupadas distribuídos por todo o território nacional. Desde o início, em 2000, até 2015, os resultados médios de Portugal têm vindo consistentemente a melhorar nos três domínios considerados, aproximando-se dos *scores* médios da OCDE. O Gráfico 1 mostra a tendência dos resultados médios dos alunos portugueses:

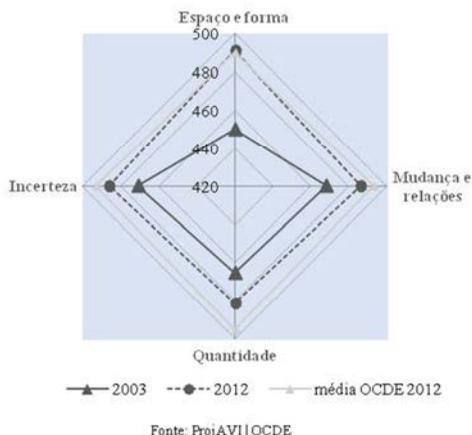


**Gráfico 1** – Resultados médios dos alunos portugueses ao longo dos ciclos *PISA* 2000 - 2015 nos três domínios avaliados.

Neste gráfico, pode observar-se que se registou uma melhoria do desempenho médio dos alunos portugueses entre 2000 e 2003, uma certa estagnação de resultados entre 2003 e 2006, seguida de nova melhoria acentuada, em 2009. Posteriormente, entre 2009 e 2012, verificou-se também uma certa estabilidade dos resultados, voltando a presenciar-se uma subida dos

scores médios em 2015, aproximando-se da linha de referência (500 pontos) a Matemática e Leitura, e ultrapassando mesmo essa linha a Ciências (ProjAVI, 2013b; Marôco *et al.*, 2016a).

Em 2012, apesar da pouca evolução registada, os alunos portugueses alcançaram, pela primeira vez, pontuações médias a uma distância não significativa da média da OCDE a Leitura e a Matemática. Em 2015, os resultados médios aumentaram e revelaram-se ligeiramente superiores à média da OCDE quer em literacia de Leitura, quer de Ciências, mantendo-se a literacia de Matemática ao nível da média da OCDE. Como referido inicialmente, o desempenho médio dos alunos portugueses também pode ser avaliado por conteúdos ou domínios cognitivos, e, em cada ciclo, a OCDE avalia com maior detalhe as pontuações obtidas no domínio principal. Assim, por exemplo, analisou-se o desempenho médio dos alunos portugueses por conteúdos em 2003 e 2012 (ciclos cujo domínio principal foi Matemática), como pode observar-se no gráfico seguinte:

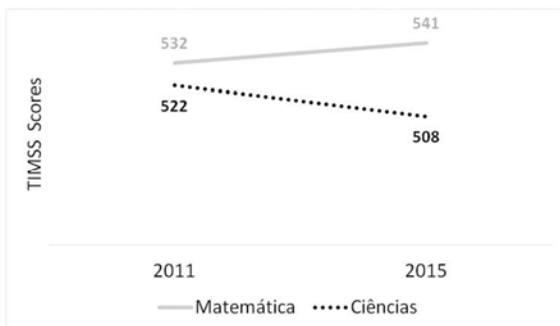


**Gráfico 2** – Comparação dos resultados médios dos alunos portugueses no PISA 2003 e 2012, por áreas de conteúdo de Matemática.

Neste gráfico, pode-se constatar que, entre 2003 e 2012, os alunos portugueses melhoraram consideravelmente os seus resultados médios por conteúdos, destacando-se particularmente a subida da pontuação média a “*Espaço e forma*”, ultrapassando mesmo a respetiva média da OCDE em 2012.

No TIMSS 2011, participaram 63 países e 14 participantes em *Benchmarking* e, em 2015, 50 países e 4 regiões em *Benchmarking*. Em Portugal, em 2011, participaram cerca de 4000 alunos portugueses do 4º ano de escolaridade e 150 escolas e, no ciclo de 2015, participaram 4693 alunos e 217 escolas. O desempenho

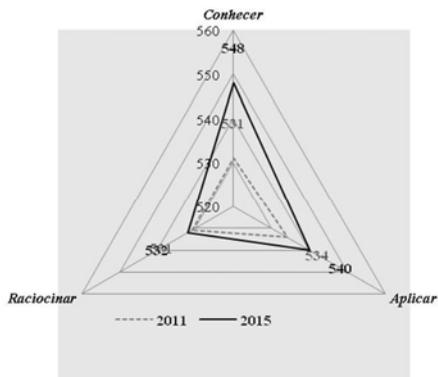
médio dos alunos portugueses neste estudo é apresentado no Gráfico 3 (ProjAVI 2012a, 2012b; Marôco *et al.*, 2016b).



**Gráfico 3** – Resultados médios dos alunos portugueses nos ciclos *TIMSS* 2011 e 2015 nos dois domínios avaliados.

Em 2015, Portugal registou uma subida significativa em Matemática (9 pontos) revelando um excelente resultado (13<sup>a</sup> posição) mas também uma descida significativa a Ciências (14 pontos).

Tal como no estudo *PISA*, os resultados médios dos alunos portugueses no *TIMSS* também podem ser analisados por domínios cognitivos ou por áreas de conteúdo. O Gráfico 4 ilustra este estudo de tendências, comparando os resultados dos ciclos *TIMSS* 2011 e 2015 a Matemática, por domínios cognitivos.



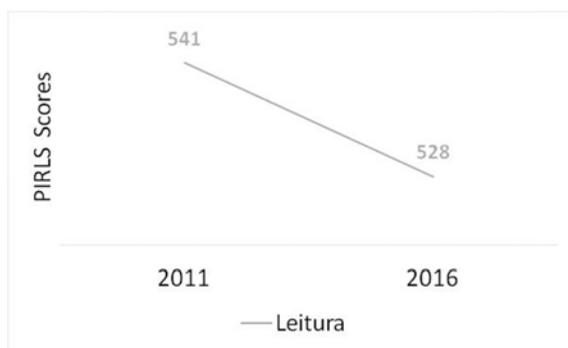
**Gráfico 4** – Comparação dos resultados médios dos alunos portugueses nos ciclos *TIMSS* 2011 e 2015, por domínios cognitivos de Matemática.

Os alunos portugueses mostraram no domínio *Conhecer* uma subida significativa (de 531 para 548 pontos) mas quer no domínio *Raciocinar* quer em *Aplicar* não registaram melhorias relevantes, registando-se apenas mais 1 e 6 pontos, respetivamente.

No *TIMSS Advanced*, como já referido, estiveram envolvidos 9 países e, em Portugal, participaram 5748 alunos do 12º ano (4009 realizaram o teste de Matemática avançada e 1739 realizaram o teste de Física) e 229 escolas/agrupamentos de escolas. Portugal obteve um resultado médio a Matemática avançada de 482 pontos (5ª posição) e a Física 467 pontos (4ª posição). Na avaliação de Matemática, os alunos portugueses revelaram melhor desempenho a *Álgebra* e pior a *Geometria*, seguindo a tendência internacional. Em Física, os alunos portugueses revelaram melhor desempenho a *Mecânica e Termodinâmica*, seguindo a tendência dos restantes países, e pior a *Eletricidade e Magnetismo* enquanto a tendência internacional revelou desempenhos mais fracos nos *Fenómenos Ondulatórios e Física Atómica* (Marôco *et al.*, 2016c).

No *PIRLS 2011*, participaram 48 países e 9 participantes em *Benchmarking* e, no novo ciclo, em 2016, participaram 55 países/sistemas educativos e 16 países/sistemas educativos no *ePIRLS*.

Em Portugal, na primeira participação do país no *PIRLS*, estiveram envolvidas cerca de 150 escolas e, aproximadamente, 4000 alunos do 4.º ano de escolaridade e, em 2016, Portugal participou no *PIRLS* e no *ePIRLS*, contando com a participação de 220 escolas e 5324 alunos do 4º ano. O desempenho médio dos alunos portugueses no *PIRLS* é apresentado no Gráfico 5 (ProjAVI, 2012c; IAVE, 2017).



**Gráfico 5** – Resultados médios dos alunos portugueses nos ciclos *PIRLS* 2011 e 2016.

Tendo Portugal alcançado 528 pontos em 2016, alcança uma pontuação média significativamente superior ao ponto central da escala *PIRLS* (500 pontos) mas, relativamente a 2011, regista-se uma descida significativa de 13 pontos. Na escala ordenada dos resultados, entre 2011 e 2016, Portugal passou do 19º para o 30º lugar. Entre 2011 e 2016, 21 países participantes melhoraram significativamente os

seus desempenhos, destacando-se entre o conjunto de países com pontuações médias acima dos 500 pontos, como subidas mais acentuadas, as protagonizadas pela Lituânia (22 pontos) e pela Bulgária (20 pontos). No mesmo período, onze participantes apresentaram descidas consideráveis, das quais a mais acentuada foi do Irão (29 pontos), integrando Portugal este grupo de países. Apesar do desempenho médio ter diminuído em 2016, verifica-se que Portugal obteve nos dois ciclos em que participou resultados médios acima da linha de referência (500 pontos). Assim como nos estudos *TIMSS Ciências* e *TIMSS Matemática*, os resultados médios dos vários países participantes no *PIRLS* também são, em cada ciclo, analisados por domínios cognitivos ou por áreas de conteúdo<sup>4</sup>.

No *ePIRLS*, Portugal registou um resultado médio de 522 pontos, ocupando a 12ª posição na lista ordenada de resultados e Singapura foi o país que obteve o melhor resultado com 588 pontos. Contrariamente à tendência internacional, Portugal obteve um resultado médio a literacia de Leitura em papel superior ao resultado médio na avaliação de Leitura *online*.

## 4 Conclusões

Os projetos de avaliação internacional de alunos por amostragem de larga escala, como os anteriormente apresentados, visam estimar a proficiência de alunos, em vários domínios e graus de ensino, dos sistemas de educação de diversos países. Tendo esse objetivo em mente, estes estudos definem processos complexos para selecionar os alunos participantes, para delinear a distribuição das diferentes versões das provas pelos vários alunos, para estimar o desempenho médio dos alunos em cada país, por domínios avaliados, áreas de conteúdo e processos cognitivos e, finalmente, para estimar o erro padrão de qualquer estimativa populacional.

A metodologia de amostragem usada para a seleção dos alunos participantes, em cada país, é multietápica: numa primeira etapa são selecionadas aleatoriamente as escolas, e, na segunda, são selecionados aleatoriamente os alunos ou as turmas de alunos de cada escola participante. Este tipo de metodologia de amostragem tende a aumentar o erro padrão de qualquer estimativa populacional, sendo necessário recorrer a métodos de reamostragem como o *Bootstrap* ou o *Jackknife* para estimar a variabilidade. Desta forma, analisar os dados recolhidos nestas provas internacionais, acarreta um particular cuidado na sua análise, uma vez que

---

<sup>4</sup> Ver *Resultados Globais PIRLS 2016 & ePIRLS 2016 – Portugal. Literacia de Leitura & Literacia de Leitura online*, disponível no site do IAVE, I.P.

as técnicas estatísticas disponíveis nos *softwares* estatísticos mais comuns não são adequadas.

Por outro lado, a avaliação dos alunos num determinado domínio como a matemática ou as ciências, exigiria um grande conjunto de questões, o que tornaria inexecutável realizar uma prova com tão grande dimensão (no *TIMSS* seriam necessárias cerca de 8h30m minutos de prova para o aluno responder aos vários itens). Assim, é necessário um delineamento de vários blocos de itens, de forma a garantir que cada aluno responde apenas a um conjunto de itens num tempo executável (no *TIMSS* 2015 cerca de 72 minutos) mas cobrindo todos os domínios avaliados e áreas de conteúdo e ainda, a ligação com os resultados dos ciclos anteriores através dos itens *trend*<sup>5</sup>.

Para estimar o desempenho médio dos alunos de um país participante é, pois, necessário estimar a pontuação global de cada aluno no conjunto de todos os itens, recorrendo ao conceito de *valor plausível*. Isto é, em vez de estimar o *score* global de um aluno num certo domínio, são estimados um conjunto de valores possíveis (*plausible values*) e a respetiva probabilidade, recorrendo à Teoria da Resposta ao Item (TRI).

Os dados recolhidos e os resultados obtidos pelos estudos internacionais apresentados são tornados públicos pelas instituições responsáveis após a divulgação dos relatórios internacionais e, permitem aos investigadores o acesso a um alargado leque de informação sobre educação, contribuindo para que estes instrumentos de avaliação internacional possam ser reguladores do sistema de ensino-aprendizagem. Estes estudos lançam, pois, um verdadeiro desafio aos investigadores não só pela colossal informação educacional gerada mas também pelo desafio metodológico no âmbito da Estatística e da Análise de Dados.

## Referências

- IAVE (2017). *Resultados Globais PIRLS 2016 & ePIRLS 2016 – Portugal. Literacia de Leitura & Literacia de Leitura online*. ed.1, ISBN: 978-989-99971-1-0 Lisboa: IAVE, I.P.
- Marôco, J. P.; Gonçalves, C.; Lourenço, V.; Mendes, R. (2016a). *PISA 2015 - Portugal. Volume I: Literacia Científica, Literacia de Leitura, Literacia Matemática*. ed. 1, ISBN: 978-989-99741-2-8. Lisboa: IAVE, I. P.
- Marôco, J. P.; Lourenço, V.; Mendes, R.; Gonçalves, C. (2016b). *TIMSS 2015 - Portugal. Volume I: Desempenhos em Matemática e em Ciências*. ed. 1, 1 vol., ISBN: 978-972-8866-96-9. Lisboa: IAVE, I. P.

---

<sup>5</sup> Os itens *trend* são itens que são sucessivamente aplicados em todas as edições das provas para permitir estudar as tendências.

- Marôco, J. P.; Lourenço, V.; Mendes, R.; Gonçalves, C. (2016c). *TIMSS Advanced 2015 – PORTUGAL. Volume 1: Desempenhos em Matemática e em Física*. ed. 1, 1 vol., ISBN: 978-989-99741. Lisboa: IAVE, I. P.
- OCDE (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (vol. 1). OECD Publishing (<http://doi.org/10.1787/9789264201118-en>).
- OCDE (2014b). *PISA 2012 Technical Report*. Paris: OECD Publishing.
- ProjAVI (2012a). *TIMSS 2011: Principais Resultados – Ciências*. Disponível em: [http://www.dgeec.mec.pt/np4/%7B\\$clientServletPath%7D/?newsId=371&fileName=TIMSS2011\\_PrincipaisResultados\\_Ciencias.pdf](http://www.dgeec.mec.pt/np4/%7B$clientServletPath%7D/?newsId=371&fileName=TIMSS2011_PrincipaisResultados_Ciencias.pdf)
- ProjAVI (2012b). *TIMSS 2011: Principais Resultados – Matemática*. Disponível em: [http://www.dgeec.mec.pt/np4/%7B\\$clientServletPath%7D/?newsId=371&fileName=TIMSS2011\\_PrincipaisResultados\\_Matematic.pdf](http://www.dgeec.mec.pt/np4/%7B$clientServletPath%7D/?newsId=371&fileName=TIMSS2011_PrincipaisResultados_Matematic.pdf)
- ProjAVI (2012c). *PIRLS 2011: Principais Resultado – Leitura*. Disponível em: [http://www.dgeec.mec.pt/np4/%7B\\$clientServletPath%7D/?newsId=371&fileName=PIRLS2011\\_PrincipaisResultados\\_Leitura.pdf](http://www.dgeec.mec.pt/np4/%7B$clientServletPath%7D/?newsId=371&fileName=PIRLS2011_PrincipaisResultados_Leitura.pdf)
- ProjAVI (2013 a). *TIMSS & PIRLS 2011 – Relações entre desempenhos*. Disponível em: [http://www.dgeec.mec.pt/np4/%7B\\$clientServletPath%7D/?newsId=371&fileName=TIMSS\\_PIRLS\\_2011\\_RelacoesEntreDesempenho.pdf](http://www.dgeec.mec.pt/np4/%7B$clientServletPath%7D/?newsId=371&fileName=TIMSS_PIRLS_2011_RelacoesEntreDesempenho.pdf)
- ProjAVI (2013b). *PISA 2012 – Primeiros Resultados – Portugal*. Disponível em: [http://www.dgeec.mec.pt/np4/%7B\\$clientServletPath%7D/?newsId=371&fileName=PISA\\_Primeiros\\_Resultados\\_PORTUGAL.pdf](http://www.dgeec.mec.pt/np4/%7B$clientServletPath%7D/?newsId=371&fileName=PISA_Primeiros_Resultados_PORTUGAL.pdf)
- Sousa Ferreira, A.; Flores, I.; Casas-Novas, T. (2017). *Porque melhoraram os resultados PISA em Portugal. Estudo longitudinal e comparado (2000-2015)*. Disponível em: <https://www.ffms.pt/FileDownload/9857244f-4dfb-48ad-b196-0448dc444865/porque-melhoraram-os-resultados-pisa-em-portugal>.



## Alguns motivos para indivíduos entre os 40 e os 49 anos quererem ter filhos

Anabela Afonso<sup>1</sup> · Paulo Infante<sup>2</sup> · Maria Filomena Mendes<sup>3</sup>

**Resumo** A fecundidade portuguesa tem-se tornado tema de debate alargado não só entre a comunidade académica, como também na sociedade civil, na comunicação social e mesmo no seio dos governo e oposição, pois atravessamos um período de poucos nascimentos. Com base nos dados do Inquérito à Fecundidade realizado pelo INE em 2013, neste trabalho caracterizamos a população do grupo etário dos 40 aos 49 anos que ainda pensa vir a ter filhos e apresentamos alguns motivos que estão na base dessa decisão.

**Palavras-chave:** Amostras Complexas, Análise Classificatória, Análise de Correspondências Múltiplas, Demografia, Fecundidade.

### 1 Introdução

Nunca se registaram tão poucos nascimentos em Portugal como nos últimos anos. A fecundidade portuguesa tem-se tornado tema de debate alargado não só entre a comunidade académica, como também na sociedade civil, na comunicação social e mesmo no seio do governo e da oposição.

Com o aumento da escolaridade e da participação no mercado de trabalho, muitas pessoas adiaram as suas intenções de fecundidade. Por outro lado, as restrições económico-financeiras são muitas vezes apontadas como um constrangimento para a concretização da fecundidade ao longo do curso de vida.

Apesar do grupo etário dos 40 aos 49 anos ter um impacto residual em termos de fecundidade intencional (o número de filhos que pretende ainda vir a ter), a sua caracterização e a identificação de motivos que levam pessoas nesta classe etária a

---

<sup>1</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>2</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>3</sup> CIDEHUS/IIFA e Departamento de Sociologia/ECS, da Universidade de Évora, [mmendese@uevora.pt](mailto:mmendese@uevora.pt)

quererem ainda ter filhos ou vir a ter mais filhos parece-nos muito importante para a compreensão dos comportamentos de fecundidade dos portugueses.

Em 2013, no âmbito de um protocolo entre a Fundação Francisco Manuel dos Santos (FFMS) e o Instituto Nacional de Estatística (INE), foi realizado o Inquérito à Fecundidade (IFEC). Com base nos dados recolhidos, neste trabalho caracterizamos a população do grupo etário dos 40 aos 49 anos que ainda pensa vir a ter filhos, apresentamos alguns motivos que estão na base dessa decisão, identificam-se perfis de motivos e grupos de pessoas que correspondem aos perfis identificados.

## 2 Metodologia

Neste trabalho analisam-se os dados obtidos através do Inquérito à Fecundidade (IFEC), realizado pelo Instituto Nacional de Estatística (INE) entre 16 de janeiro e 15 de abril de 2013. A amostra foi selecionada utilizando um esquema de amostragem complexo, i.e., multi-etápico e com ajustamentos para compensar as não respostas e outras pós-estratificações, o que permitiu obter uma amostra representativa ao nível regional (NUTS II) para as mulheres, dos 18 aos 49 anos, e ao nível nacional para os homens, dos 18 aos 54 anos (INE, 2013). Para cada indivíduo  $i$ ,  $i = 1, \dots, n$ , da amostra foi fornecido o peso final de amostragem ( $w_i$ ) que pode ser intuitivamente interpretado como o número de indivíduos na população que são representados por esse indivíduo observado.

O ponderador  $w_i$  foi utilizado para se obterem estimativas para os totais, médias e proporções populacionais (Lorh, 2010). Para o cálculo do erro padrão estimado associado às estimativas, utilizou-se um estimador de variância do tipo Jackknife (Lohr, 2010, p. 382).

Para se identificarem associações entre a importância atribuída aos motivos para a intenção de terem ou terem mais filhos, foi utilizada a análise de correspondências múltiplas (Greenacre & Blasius, 2006), associando a cada indivíduo  $i$  da amostra o peso  $w_i$ ,  $i = 1, \dots, n$ . Para a escolha do número de dimensões a reter foram considerados os seguintes critérios: i) análise da representação gráfica dos valores próprios; ii) número de valores próprios superiores a  $1/Q$ , sendo  $Q$  o número de variáveis (Greenacre, 2007: 140); iii) Percentagem total da variabilidade explicada. Tendo em conta as categorias mais discriminantes das variáveis que mais contribuía para cada uma das dimensões consideradas, foram identificados alguns perfis. Posteriormente, para identificar grupos de residentes que possam corresponder aos perfis, aplicou-se uma análise classificatória, considerando como variáveis as coordenadas dos indivíduos nas dimensões retidas (Carvalho, 2008). Inicialmente foi efetuada uma análise

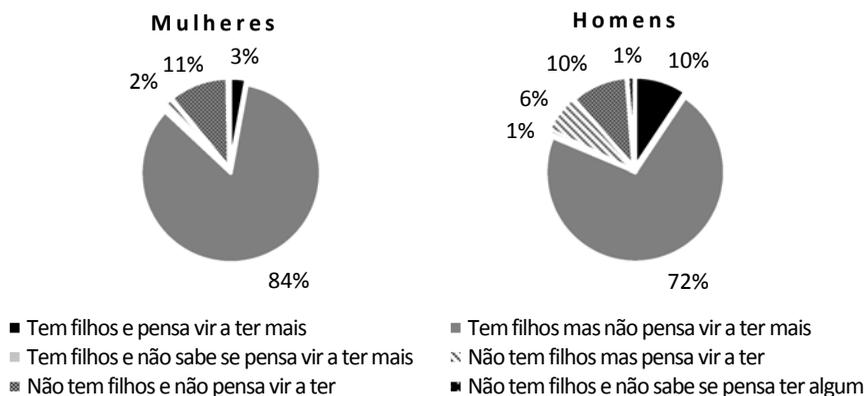
classificatória hierárquica ascendente usando os critérios de agregação (ligação simples, ligação completa e método de Ward), para identificar o número de grupos a considerar, através da análise gráfica do coeficiente de fusão contra o número de grupos, seguida de uma análise classificatória não hierárquica recorrendo ao método *K-means* para formar os grupos.

A análise estatística foi realizada com auxílio dos *software* R e SPSS.

### 3 Resultados

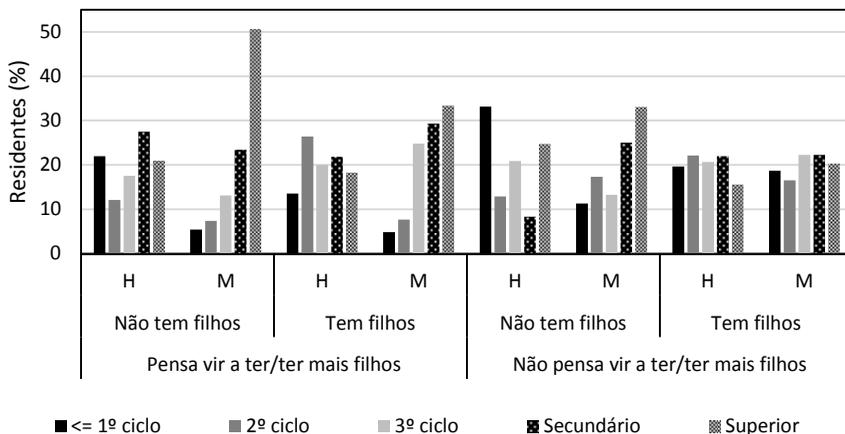
Na base de dados do IFEC 2013 existem 2730 pessoas no grupo etário dos 40 aos 49 anos, sendo mais de metade delas mulheres (69,3%).

Estima-se que a maioria dos residentes em Portugal com idade entre os 40 aos 49 anos já transitou na parentalidade (Figura 1), sendo essa proporção estimada maior entre as mulheres (87,4%, IC<sub>95%</sub>: [85,6%; 89,1%]) do que entre os homens (82,2%, IC<sub>95%</sub>: [79,5%; 84,9%]). As mulheres revelam uma menor intenção de vir a ter filhos (4,6%, IC<sub>95%</sub>: [3,5%; 5,6%]) relativamente aos homens (15,7%, IC<sub>95%</sub>: [13,3%; 18,1%]).



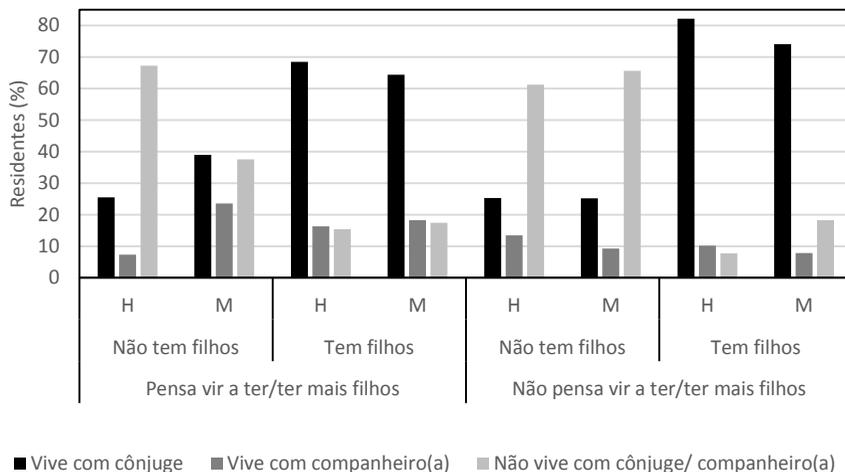
**Figura 1** – Fecundidade realizada e expectativa futura, por sexo.

As mulheres que não têm filhos têm níveis de instrução mais elevados do que as que têm filhos (21% vs. 37% têm formação superior). Metade das mulheres que não têm filhos e ainda pensam vir a ter filhos têm nível de instrução superior (Figura 2). Entre os homens que não têm filhos e não pretendem tê-los, a maioria tem formação ao nível do 1º ciclo.



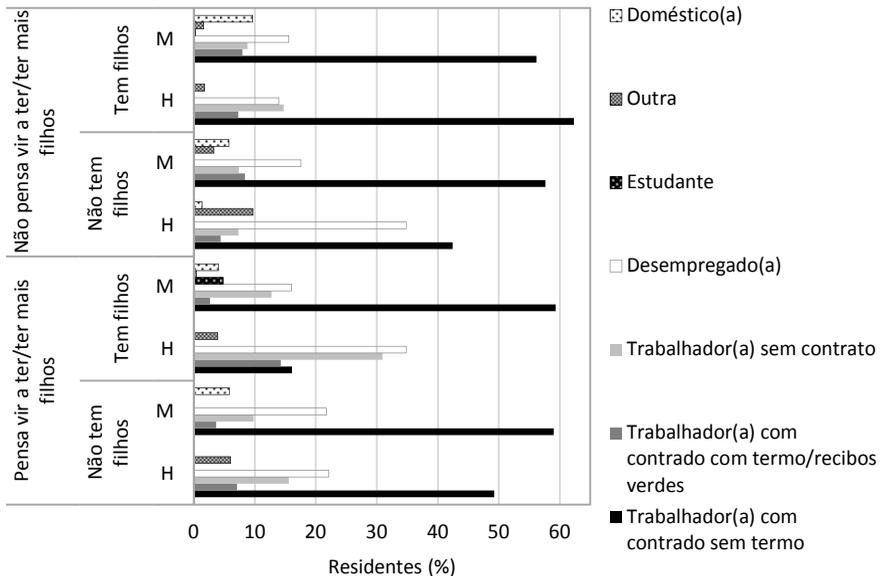
**Figura 2** – Distribuição do nível de escolaridade dos residentes por existência de filhos biológicos, intenção futura e sexo.

O padrão de conjugalidade é semelhante entre homens e mulheres que têm filhos (a maioria vive com cônjuge), quer ainda tencionem ter ou não mais filhos (Figura 3). Cerca de 70% dos homens que não têm filhos e não vivem em conjugalidade ainda pensam vir a ter filhos. A maior parte dos homens e das mulheres que não têm filhos nem tencionam tê-los não vivem com cônjuge/companheiro(a).



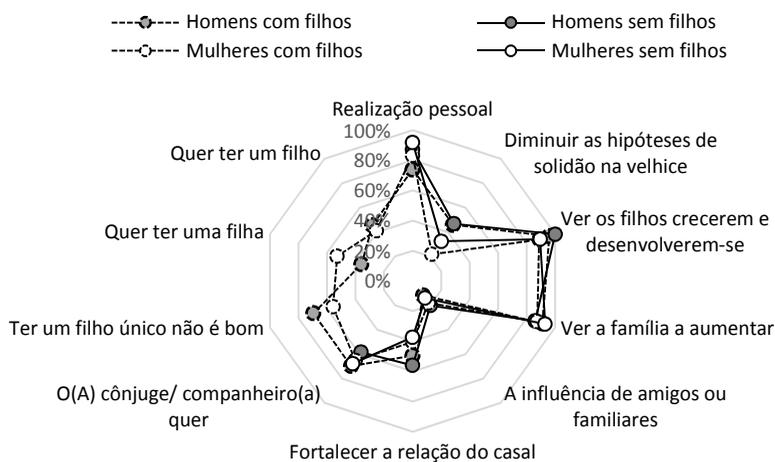
**Figura 3** – Distribuição da situação conjugal atual dos residentes por existência de filhos biológicos, intenção futura e sexo.

Mais de metade das mulheres têm uma situação profissional estável (contrato sem termo, Figura 4). O padrão da condição perante o trabalho é semelhante entre as mulheres, quer tenham ou não filhos, e na sua intensão futura de fecundidade. Os homens que não têm filhos têm uma situação profissional mais precária do que os que têm filhos (mais desempregados ou noutra situação). É curioso observar que, por um lado a maioria dos homens que já têm filhos e ainda pensam vir a ter estão desempregados, e por outro lado é também muito elevada a proporção de homens que não têm filhos nem tencionam tê-los e que estão desempregados. Parecem existir aqui duas situações distintas: os homens que consideram que a sua condição profissional precária é uma situação temporária (possivelmente devido ao momento de crise que o país atravessava na altura do inquérito) e os que não têm esperança de melhorar essa condição.



**Figura 4** – Condição perante o trabalho dos residentes que querem ter ou ter mais filhos.

Os três motivos mais indicados pelos homens e pelas mulheres como importantes para quererem ter ou ter mais filhos foram a realização pessoal, ver os filhos crescer e desenvolverem-se e ver a família aumentar (Figura 5). Os motivos ter um filho único não é bom, fortalecer a relação do casal e diminuir as hipóteses de solidão na velhice, foram mais vezes mencionados como importantes pelos homens do que pelas mulheres. Pelo contrário foram as mulheres que mais mencionaram o facto de quererem ter uma filha.



**Figura 5** – Motivos considerados importantes para quererem ter ou ter mais filhos.

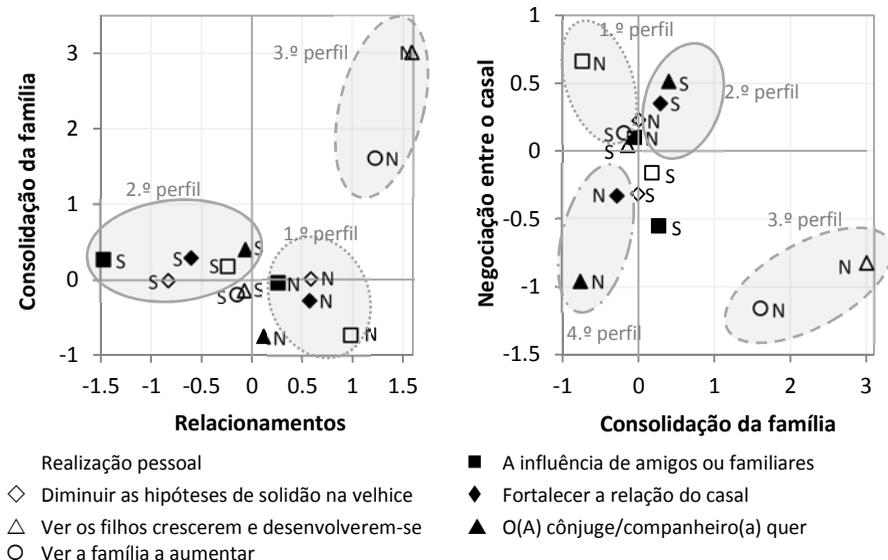
Foi possível resumir a informação relacional entre os motivos em três dimensões que explicam 58% da variância (Tabela 1). As variáveis Sexo e Tem filhos biológicos foram consideradas como variáveis passivas (i.e., não são usadas na construção das dimensões), mas não mostraram capacidade discriminante nestas dimensões. Com base nos resultados obtidos, os eixos podem ser nomeados como:

1. Relacionamentos (pessoais e conjugais);
2. Consolidação da família (através dos filhos);
3. Negociação entre o casal (na constituição e crescimento da família).

**Tabela 1** – Medidas de discriminação dos motivos para quererem ter ou ter mais filhos.

Motivos	Dimensão		
	1	2	3
Realização pessoal	,236	,132	,107
Diminuir as hipóteses de solidão na velhice	,490	,000	,072
Ver os filhos crescerem e desenvolverem-se	,118	,421	,032
Ver a família a aumentar	,180	,311	,160
A influência de amigos ou familiares	,388	,013	,055
Fortalecer a relação do casal	,348	,081	,116
O(A) cônjuge/companheiro(a) quer	,008	,303	,489
<b>Inércia</b>	<b>,252</b>	<b>,180</b>	<b>,147</b>

Com base nas 3 dimensões retidas foram identificados quatro perfis de pessoas, com idade entre os 40 e os 49 anos (Figura 6).



**Figura 6** – Representação das dimensões a) primeira vs. segunda b) segunda vs. terceira, da análise multivariada da importância atribuída aos motivos para quererem ter ou ter mais filhos (N = não importante, S = importante).

Seguidamente, procedeu-se à classificação desses residentes identificando-se 4 grupos homogêneos que correspondem aos perfis identificados, e cujas principais características diferenciadoras são:

- Grupo 1 (43% dos residentes): desvalorizam os relacionamentos e suportam a sua decisão na importância da consolidação da família e da negociação entre o casal;
- Grupo 2 (32% dos residentes): preservam todos os valores tradicionalmente enraizados na sociedade. São indivíduos que valorizam os relacionamentos pessoais. Este grupo é o que apresenta a maior percentagem de homens e os menores níveis de escolaridade;
- Grupo 3 (17% dos residentes): valorizam a consolidação da família, mas desvalorizam os relacionamentos. Mais de metade dos indivíduos deste grupo não têm cônjuge ou companheiro(a);
- Grupo 4 (8% dos residentes): valorizam a negociação entre o casal e a realização pessoal e não dão importância a todos os outros motivos. Este grupo é o que tem maior percentagem de indivíduos com filhos e têm uma elevada escolaridade.

## 4 Conclusão

Porque, por um lado, aprendemos com as experiências do passado e, por outro, ainda temos alguma expectativa de que se venham a recuperar nascimentos anteriormente adiados, caracterizámos, em termos de fecundidade, os indivíduos com 40 ou mais anos, os quais se encontram muito próximo do final do seu ciclo reprodutivo.

De acordo com os resultados obtidos, foi possível, em particular, identificar a realização pessoal, ver os filhos crescer e desenvolverem-se e ver a família aumentar, como os 3 motivos principais para indivíduos nesta idade quererem ainda ter filhos ou ter mais filhos. Foi, ainda, possível resumir a informação relacional entre os motivos em três dimensões: relacionamentos, consolidação da família e Negociação entre o casal.

## Agradecimentos

Este trabalho inseriu-se no projeto “Determinantes da Fecundidade em Portugal”, financiado pela Fundação Francisco Manuel dos Santos.

## Referências

- CARVALHO, H. (2008). *Análise multivariada de dados qualitativos: utilização da análise de correspondências múltiplas com o SPSS*, 2ª Edição, Edições Sílabo, Lisboa.
- GREENACRE, M. (2007). *Correspondence analysis in practice*, 2nd edition, Chapman & Hall/ CRC Press, Boca Raton.
- GREENACRE, M. J & BLASIUS, J. (2006). *Multiple correspondence analysis and related methods*, Chapman & Hall/CRC, Boca Raton.
- LOHR, S. L. (2010). *Sampling: design and analysis*, 2nd Edition, Brooks/Cole, Cengage learning, Boston.
- INE (2014). *Inquérito à fecundidade 2013*, Instituto Nacional de Estatística e Fundação Francisco Manuel dos Santos, Lisboa.

## Alguns benefícios do envelhecimento ativo: um caso de estudo

Gonçalo Jacinto<sup>1</sup> · Paulo Infante<sup>2</sup> · Anabela Afonso<sup>3</sup> · José Conde<sup>4</sup> · Teresa Engana<sup>4</sup> · Jorge Vicente<sup>4</sup> · Pedro Silva<sup>4</sup> · Ana Carla Coelho<sup>5</sup> · Paula Costa<sup>5</sup>

**Resumo** Cerca de 18% da população do Concelho de Évora tem 65 ou mais anos (Censos 2011) e por cada 100 jovens com menos de 15 anos existem 140 pessoas idosas com pelo menos 65 anos. A Câmara Municipal de Évora oferece o programa *Seniores Ativos* que visa promover a atividade física e estilos de vida saudáveis entre os seniores (indivíduos com pelo menos 55 anos). Este estudo teve como objetivo avaliar alguns dos benefícios médicos e físicos proporcionados por este programa após 6 meses de participação, observando-se várias melhorias, tanto nos parâmetros médicos como nos testes de aptidão física.

**Palavras-chave:** Análise de Covariância, Atividade física, Envelhecimento ativo.

### 1 Introdução

A Organização Mundial da Saúde (OMS) reconhece a importância da atividade física (AF) na saúde física, mental e social e afirma que as políticas de intervenção são necessárias para promover a prática da AF em todas as idades e grupos sociais (WHO, 2010). A prática regular da AF é essencial na prevenção de várias doenças coronárias, como diabetes tipo 2, cancro, hipertensão, obesidade, depressão, osteoporose e morte prematura (Warburton, Nicol e Bredin, 2006). Evidências recentes sugerem que as pessoas fisicamente ativas têm um risco reduzido de demência (Hamer & Chida, 2009) e declínio cognitivo (Sofi et al., 2011). A AF promove o bem-estar, aumenta a interação social, a qualidade de vida dos indivíduos que a praticam e proporciona benefícios à sociedade (Bull et al., 2010).

---

<sup>1</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [gjcj@uevora.pt](mailto:gjcj@uevora.pt)

<sup>2</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>3</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>4</sup> Divisão da Juventude e Desporto, Câmara Municipal de Évora, [teresaeangana@cm-evora.pt](mailto:teresaeangana@cm-evora.pt)

<sup>5</sup> Administração Regional de Saúde do Alentejo, [AnaCarla.Coelho@alentejocentral.min-saude.pt](mailto:AnaCarla.Coelho@alentejocentral.min-saude.pt)

A OMS declarou que a AF é um dos principais fatores a considerar para um envelhecimento ativo e um aumento da qualidade de vida.

Em Portugal, existem vários programas de AF desenvolvidos pelo governo local para pessoas idosas. Contudo, que tenhamos conhecimento, nenhuma autarquia avalia a eficiência desses programas na saúde e qualidade de vida dos participantes, não existindo estudos que afirmam a qualidade destes programas (Marques *et al.*, 2011). O programa Seniores Ativos (SA) é desenvolvido pela Câmara Municipal de Évora (CME), sendo um programa gratuito cujos principais objetivos são melhorar, desenvolver, adquirir e manter estilos de vida ativos em idosos. De acordo com os dados do Censos de Portugal de 2011, cerca de 18% da população do município de Évora tem 65 anos ou mais e, para cada 100 jovens menores de 15 anos, existem 140 pessoas idosas com 65 anos ou mais.

Durante o Ano Europeu do Envelhecimento Ativo e da Solidariedade entre Gerações, a CME, em parceria com a Universidade de Évora (através do Centro de Investigação em Matemática e Aplicações - CIMA), realizou um estudo sobre os aspectos físicos, psicológicos e sociais do programa SA. Neste artigo, apresentamos os resultados deste estudo, referentes aos parâmetros médicos e físicos, no início do estudo e após 6 meses de seguimento, em adultos com 55 ou mais anos.

## **2 Métodos**

### **2.1 Programa seniores ativos**

O programa SA, criado em 2005 pela CME em parceria com algumas instituições sociais locais, é um programa anual com uma sessão de AF de uma hora por semana para a população com mais de 55 anos. As atividades são desenvolvidas por professores de educação física e ocorrem nos espaços verdes da cidade (circuitos de manutenção, trilhos naturais, jardins públicos), nas instalações dos parceiros e na piscina municipal. Durante 2013 e 2014 foram matriculados no programa 21 instituições e um total de 442 participantes.

### **2.2 Participantes**

Para este estudo, a população-alvo é o conjunto de pessoas com idade igual ou superior a 55 anos residentes no município de Évora no momento do estudo. Foram divulgados dois tipos de pedidos de participação: um pedido foi dirigido a indivíduos inativos, não participantes no programa SA, que quisessem aderir ao programa ou que pretendiam permanecer inativos, apesar de terem sido motivados para participar do programa. O segundo pedido foi dirigido ao grupo de

participantes atuais do programa SA que desejavam manter ou aumentar a frequência semanal atual de AF. Uma vez que houve algum conhecimento prévio sobre o estado de saúde deste grupo de participantes e sua disponibilidade, o pedido para aumentar a frequência da prática foi dirigido a pessoas que se sabia que podiam fazê-lo. Para integrar este estudo, todos os voluntários concordaram em testar os seus parâmetros físicos e médicos no início do estudo e 6 meses depois. Todos eles assinaram um termo de consentimento livre e comprometeram-se a frequentar pelo menos 75% das aulas.

A amostra final de 91 participantes consiste numa amostra de conveniência não probabilística e foi dividida em 4 grupos:

1. **Inativos:** 11 indivíduos inativos que, apesar de terem sido motivados por nós para entrar no grupo de iniciantes, declararam que permaneceriam inativos durante o período de acompanhamento;
2. **Iniciantes:** 15 adultos inativos que se tornaram fisicamente ativos com a participação no programa SA no início do estudo, com uma frequência semanal igual à do grupo padrão;
3. **1x Ativos** (grupo padrão): 37 membros do programa SA que mantiveram a frequência atual de uma vez por semana;
4. **3x Ativos:** 28 membros do programa SA que se ofereceram para aumentar a frequência de AF para três vezes por semana.

## 2.3 Medições

No início do estudo, os participantes responderam a um questionário sobre as suas características sociodemográficas como sexo, idade, ocupação, presença de algumas doenças e consumo de medicamentos bem como sobre a AF realizada fora do programa SA nos últimos 6 meses. Os participantes foram submetidos a avaliação antropométrica e exames de sangue para medir a pressão arterial sistólica e diastólica, colesterol (bom: HDL, mau: LHL, todo o mau: Não HDL) e triglicéridos. Para todos os participantes, as amostras de sangue foram obtidas antes dos testes de exercícios. Os exames de sangue e a avaliação antropométrica foram realizados pela equipa de enfermagem da Administração Regional de Saúde do Alentejo. Para avaliar o desempenho da aptidão funcional, os participantes foram submetidos à bateria de testes Senior Fitness Tests (SFT) (Rikli & Jones, 2013) no início do estudo e 6 meses depois. O SFT inclui medidas de força dos membros superiores (flexão do antebraço) e inferiores (levantar e sentar na cadeira), resistência aeróbia (correr 6 minutos), agilidade/equilíbrio dinâmico (sentado, levanta e caminha 2,44m e voltar a sentar), e flexibilidade dos membros superiores (alcançar atrás das costas) e inferiores (sentar e alcançar o calcanhar).

## 2.4 Análise Estatística

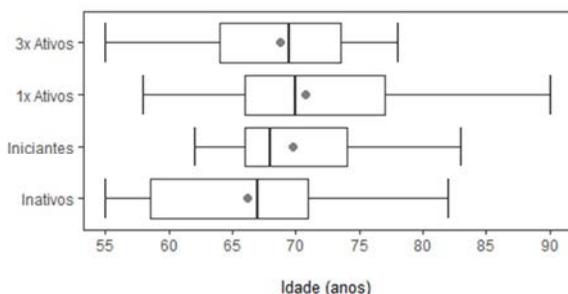
Realizou-se um estudo longitudinal para compreender os benefícios em participar no programa SA durante 6 meses. Para cada grupo, foram utilizados testes t-student para amostras emparelhadas para examinar as diferenças nos testes de aptidão médica e física entre o início do estudo e 6 meses após a prática de AF. No caso de se verificar a violação do pressuposto das diferenças não se afastarem da distribuição normal, foram utilizados o teste não paramétrico de Wilcoxon, no caso de ser possível assumir a simetria da distribuição em estudo, ou o teste não paramétrico dos Sinais no caso da distribuição não ser simétrica. Para comparar os grupos que realizaram AF no período de estudo (Iniciantes, 1x Ativos e 3x Ativos) foi aplicada uma análise da covariância (ANCOVA) com todas as variáveis médicas e físicas, ajustando os valores obtidos para o valor inicial. Com esta abordagem, é possível concluir quais são os grupos que diferem de acordo com a evolução que os participantes tiveram em 6 meses. A ANCOVA é o melhor modelo estatístico para analisar os dados e comparar os grupos entre si, quando temos dois momentos de medição com grupos que não podem ser considerados completamente randomizados (Dimitrov & Rumrill Jr, 2003). Os pressupostos da ANCOVA foram verificados, isto é, verificou-se que o efeito da medição inicial se pode considerar independente do grupo, verificou-se também a normalidade dos erros, a homogeneidade das variâncias nos diferentes grupos e a linearidade entre a variável resposta e as covariáveis. Quando o pressuposto de normalidade não foi validado e não foi encontrada uma transformação adequada, realizou-se uma ANCOVA não paramétrica (Logan, 2010). Para os testes em que o fator grupo foi considerado significativo, foram realizadas comparações múltiplas para comparar cada nível de AF com o anterior e posterior, ou seja, o grupo 1x Ativos foi comparado com os grupos Iniciantes e 3x Ativos, mas o grupo 3x Ativo só foi comparado com o grupo 1x Ativos.

## 3 Resultados

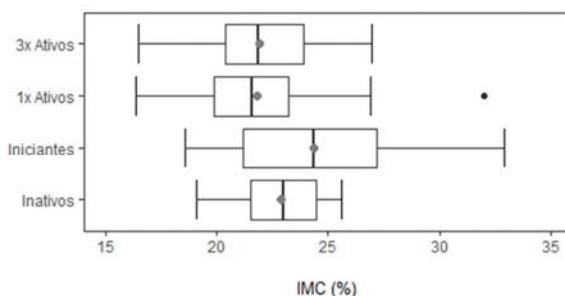
### 3.1 Caracterização dos grupos

A percentagem de participantes do sexo feminino foi maior do que a do sexo masculino nos grupos Iniciantes (67%), 1x Ativos (86%) e 3x Ativos (68%), enquanto no grupo Inativo predominaram os indivíduos do sexo masculino (73%). A idade dos participantes varia entre os 55 e os 90 anos de idade (Figura 1). No grupo 1x Ativos, um em cada quatro participantes tinha mais de 75 anos. Um em cada três participantes dos grupos Iniciantes e 3x Ativos indicaram que praticaram AF nos últimos 6 meses fora do programa AS. Nos outros dois grupos, a

percentagem foi menor. Mais de 90% dos participantes relataram tomar medicamentos.



**Figura 1** – Idade dos participantes, por grupo, no início do estudo (os pontos dentro das caixas representam a média).



**Figura 2** – IMC dos participantes, por grupo, no início do estudo (os pontos dentro das caixas representam a média).

Verificou-se que os participantes geralmente tinham um Índice de Massa Corporal (IMC) normal (valores no intervalo  $18,5 \leq \text{IMC}\% < 24,9$ ), porém os participantes no grupo Inativo apresentaram valores mais parecidos entre si, contrariamente ao observado no grupo Iniciantes (Figura 2). Somente nos grupos Iniciantes e 1x Ativos houve participantes com obesidade moderada ( $30 \leq \text{IMC}\% < 35$ ), mas em todos os grupos cerca de 90% dos participantes apresentaram riscos metabólicos (perímetro abdominal  $>94$  para homens e  $>80$  para mulheres). Cerca de 5% dos participantes no grupo 1x Ativos e 11% no grupo 3x Ativos estavam abaixo do peso ideal ( $\text{IMC}\% < 18,5$ ). Os valores de referência para IMC e perímetro abdominal podem ser encontrados em WHO (2011).

Nos resultados apresentados na Tabela 1 pode-se observar que, no início do estudo, os participantes de todos os grupos apresentavam valores médios normais de HDL (40-59 mg/dL). O grupo 3x Ativos apresentava ótimos níveis médios de mau colesterol (LDL  $< 100$  mg/dL) e os restantes grupos tinham níveis médios quase ótimos (101-130 mg/dL). Todos os grupos apresentavam resultados médios

normais da pressão arterial diastólica (<85 mmHg). Quase todos os grupos tinham um valor médio da pressão arterial sistólica no primeiro estágio da hipertensão (140-159 mmHg). Apenas o grupo Iniciantes apresentava resultados desejáveis de triglicédeos (<150 mg/dL), enquanto os demais grupos apresentavam valores médios elevados (>150 mg/dL).

**Tabela 1** – Média ± desvio padrão (DP) de alguns parâmetros médicos, por grupo, medido no início do estudo e após 6 meses, e valor p (p) do teste de t-Student para a diferença entre o início do estudo e após 6 meses (se (a) teste de Wilcoxon, (b) teste de sinais).

Parâmetro	Instante	Inativos	Iniciantes	p	1x Ativos	p	3x Ativos	p
		Média±DP	Média±DP		Média±DP		Média±DP	
Peso (kg)	Início	74,6±10,0	76,9±14,5	<0,001 <sup>b</sup>	67,0±10,4	0,006 <sup>b</sup>	69,0±10,2	0,092 <sup>b</sup>
	6 meses	73,7±10,3	73,9±15,4		65,6±11,1		69,0±10,7	
IMC (%)	Início	22,9±2,0	24,4±4,2	<0,001 <sup>b</sup>	21,8±3,1	0,006 <sup>b</sup>	21,9±2,8	0,092 <sup>b</sup>
	6 meses	22,6±2,1	23,6±4,3		21,3±3,3		21,9±3,0	
Perímetro abdominal (cm)	Início	99,5±10,3	103,4±10,9	0,015 <sup>a</sup>	94,0±11,5	0,184 <sup>a</sup>	96,2±11,1	0,228
	6 meses	95,7±10,2	98,5±12,1		93,2±10,5		94,9±10,2	
HDL (mg/dL)	Início	48,5±10,2	54,7±9,0	0,006	52,5±12,3	0,011 <sup>a</sup>	50,8±11,8	0,460
	6 meses	42,9±14,3	51,1±8,4		48,4±12,3		50,6±12,5	
LDL (mg/dL)	Início	119,2±30,5	101,5±23,0	0,151 <sup>b</sup>	109,3±30,4	0,237	99,2±23,8	0,086
	6 meses	119,5±20,4	102,9±30,6		105,1±30,6		91,7±27,0	
Não HDL (mg/dL)	Início	160,0±36,9	124,0±33,0	0,001 <sup>b</sup>	138,0±27,5	0,341	130,0±26,7	0,147
	6 meses	152,0±33,5	123,0±31,1		135,0±29,6		125,0±25,3	
Pressão diastólica (mmHg)	Início	81,4±10,3	71,2±13,5	0,377	75,2±10,6	0,250	72,9±8,3	0,727
	6 meses	81,1±12,5	72,1±8,2		73,9±8,9		74,4±12,0	
Pressão sistólica (mmHg)	Início	147,5±16,6	137,7±21,6	0,005	141,9±19,4	<0,001	142,1±14,7	<0,001
	6 meses	135,4±14,1	128,3±12,3		129,7±17,2		129,8±18,1	
Triglicéridos (mg/dL)	Início	223,0±132,8	127,0±33,3	0,018 <sup>b</sup>	160,0±101,9	0,250 <sup>b</sup>	150,0±75,5	0,679
	6 meses	174,0±127,1	132,0±41,4		172,0±118,2		157,0±70,3	

Relativamente à aptidão funcional, em geral, os participantes do grupo Inativo tiveram pontuações mais parecidas entre si nos testes de força dos membros inferiores, resistência aeróbia e flexibilidade dos membros superiores do que os participantes dos outros grupos (Tabela 2). O grupo Inativo apresentou o pior desempenho na prova de flexibilidade dos membros inferiores e o melhor desempenho na prova de flexibilidade dos membros superiores e na prova de agilidade. O grupo Iniciantes apresentou um pior desempenho no teste de resistência aeróbia e flexibilidade dos membros superiores. Alguns dos resultados foram similares no grupo Inativo e nos outros grupos, mas tal deveu-se ao facto de os participantes desse grupo serem, em geral, mais jovens.

**Tabela 2** – Média  $\pm$  desvio padrão (DP) de alguns parâmetros de aptidão física, por grupo, medido no início do estudo e após 6 meses, e valor p (p) do teste de t-Student para a diferença entre o início do estudo e após 6 meses (se (a) teste de Wilcoxon, (b) teste de sinais).

Parâmetro	Instante	Inativos	Iniciantes	p	1x Ativos	p	3x Ativos	p
		Média $\pm$ DP	Média $\pm$ DP		Média $\pm$ DP		Média $\pm$ DP	
Força membros superiores (n.º rep.)	Início	17,7 $\pm$ 4,3	17,4 $\pm$ 3,9	<0,001	19,6 $\pm$ 4,7	<0,001	19,5 $\pm$ 3,8	<0,001
	6 meses	18,8 $\pm$ 4,9	19,9 $\pm$ 3,5		22,9 $\pm$ 4,4		25,3 $\pm$ 4,2	
Força membros inferiores (n.º levant.)	Início	14,8 $\pm$ 2,4	15,1 $\pm$ 3,2	0,001	15,5 $\pm$ 3,3	<0,001	17,4 $\pm$ 3,2	0,011
	6 meses	17,6 $\pm$ 4,7	16,7 $\pm$ 2,7		17,7 $\pm$ 3,3		18,7 $\pm$ 3,7	
Resistência aeróbia (min)	Início	552 $\pm$ 56	535 $\pm$ 70	<0,001	546 $\pm$ 66	0,024 <sup>b</sup>	589 $\pm$ 58	<0,001
	6 meses	646 $\pm$ 82	611 $\pm$ 78		584 $\pm$ 100		678 $\pm$ 84	
Agilidade (s)	Início	5,0 $\pm$ 1,3	6,4 $\pm$ 1,3	0,001 <sup>b</sup>	5,9 $\pm$ 1,0	0,045 <sup>a</sup>	5,1 $\pm$ 0,9	0,092 <sup>b</sup>
	6 meses	4,8 $\pm$ 1,2	6,0 $\pm$ 1,4		5,8 $\pm$ 0,9		5,0 $\pm$ 0,9	
Flex. Membros superiores (cm +/-)	Início	-6,8 $\pm$ 7,4	-17,8 $\pm$ 12,3	<0,001	-12,8 $\pm$ 10,7	<0,001	-9,7 $\pm$ 10,0	0,004 <sup>b</sup>
	6 meses	-5,7 $\pm$ 6,1	-10,0 $\pm$ 10,0		-8,2 $\pm$ 9,2		-8,5 $\pm$ 9,2	
Flex. Membros inferiores (cm +/-)	Início	-9,9 $\pm$ 11,4	3,1 $\pm$ 6,1	<0,001	5,4 $\pm$ 6,7	<0,001	1,5 $\pm$ 11,8	<0,001
	6 meses	-3,5 $\pm$ 11,7	9,0 $\pm$ 8,1		10,5 $\pm$ 8,8		8,4 $\pm$ 11,9	

## 3.2 Benefícios médicos e físicos após 6 meses de AF

### Diferenças dentro de grupos

Para entender o benefício da AF após 6 meses, realizou-se um estudo longitudinal, comparando os resultados dos exames médicos e físicos antes e após 6 meses de prática de AF, para os grupos que praticam AF (Iniciantes, 1x Ativos e 3x Ativos). Observou-se que ocorreram algumas alterações significativas tanto nos testes de aptidão médica como nos de aptidão física (Tabela 1).

Comparando as duas medições feitas, registam-se diminuições significativas no peso, IMC, HDL e pressão sistólica dos membros dos grupos Iniciantes e 1x Ativos, no Não HDL e Triglicéridos no grupo Iniciantes. No grupo 3 x Ativos apenas se registou uma diminuição significativa na pressão sistólica.

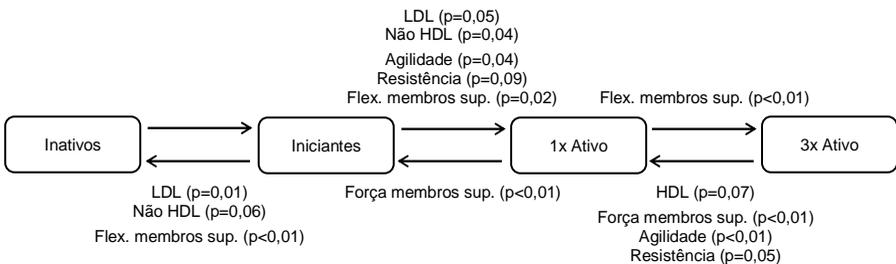
Ao nível dos testes de aptidão funcional, registaram-se diferenças significativas (no sentido de melhorias) em todos os SFT e todos os grupos, com exceção de uma diferença apenas marginalmente significativa na prova de agilidade do grupo 3x Ativos.

## Diferenças entre grupos

Através da ANCOVA foram encontradas diferenças significativas entre alguns parâmetros médicos e físicos após 6 meses, considerando os valores no início do estudo como covariáveis.

Observou-se um valor médio final de Não HDL e LDL significativamente menor para o grupo Iniciantes do que para o grupo Inativos ( $p=0,06$  e  $p=0,04$ , respectivamente) e para o grupo 1x Ativos ( $p=0,04$  e  $p=0,05$ , respectivamente) (Figura 3). Os valores do colesterol bom (HDL), aos 6 meses, parecem ser inferiores, em média, no grupo 3x Ativos em comparação com o grupo 1x Ativos, quando comparamos participantes com os mesmos valores iniciais de HDL ( $p=0,07$ ).

Em relação à aptidão física, comparando participantes em iguais circunstâncias no início do estudo, o grupo que aumentou a intensidade de AF (3x Ativos) obteve melhor desempenho médio do que o grupo que manteve a intensidade da AF (1x Ativos) na resistência aeróbia ( $p<0,01$ ), força dos membros superiores ( $p<0,01$ ) e agilidade ( $p=0,05$ ) (Figura 3). O grupo Iniciantes apresentou um desempenho significativamente melhor na resistência aeróbia ( $p=0,04$ ) e na flexibilidade dos membros superiores ( $p=0,02$ ) do que o grupo 1x Ativos, e também um desempenho significativamente melhor na prova de agilidade ( $p=0,09$ ). O grupo Iniciantes apresentou um desempenho significativamente superior ao grupo Inativo na flexibilidade dos membros superiores ( $p<0,01$ ). O grupo 1x Ativos teve uma melhor flexibilidade dos membros superiores após 6 meses do que o grupo 3x Ativos ( $p<0,01$ ) e maior força dos membros superiores do que o grupo Iniciantes ( $p<0,01$ ). Nos testes de resistência aeróbia e flexibilidade dos membros inferiores, não foram detectadas diferenças entre os grupos, mas houve diferença entre os três grupos quanto à flexibilidade dos membros superiores.



**Figura 3** – Resultados das comparações múltiplas que foram significativas nos testes médicos e físicos com base na ANCOVA (o grupo com o melhor resultado está associado à origem da seta). As diferenças de significância marginal ( $p<0,1$ ) também estão apresentadas.

## 4 Conclusão

Este estudo teve como objetivo avaliar, pela primeira vez, os efeitos do programa SA sobre saúde e aptidão física dos seus participantes. No nosso estudo, observamos que em apenas 6 meses após o início da AF regular no programa SA, foram notadas várias melhorias, tanto nos parâmetros médicos como nos testes de aptidão física.

Os participantes que introduziram AF regular melhoraram todos os resultados nos testes físicos após 6 meses. Nos testes médicos apenas não apresentaram melhorias significativas nos valores de HDL e da pressão diastólica. Este grupo conseguiu uma redução significativa do mau colesterol e um aumento da flexibilidade dos membros superiores, relativamente ao grupo dos participantes que permaneceram inativos.

Os participantes que mantiveram a mesma intensidade de AF melhoraram em todos os parâmetros físicos e reduziram significativamente a sua gordura corporal, o HDL e a pressão arterial sistólica, não se registando alterações significativas nos restantes parâmetros médicos. Este último resultado revela que com a prática de AF o estado de saúde estabilizou evitando a deterioração da sua qualidade de vida (Envelhecimento Ativo).

Aumentar a intensidade da AF levou ao controlo do colesterol bom e melhorou a flexibilidade e força dos membros superiores, bem como a resistência aeróbia e a agilidade. Finalmente, refira-se que a AF pode ser vista como um complemento terapêutico para diversas doenças como a hipertensão, glicemia e doenças psicológicas (depressão, ansiedade).

Estes resultados não devem ser vistos apenas numa perspectiva de existência, ou não, de diferenças nos desempenhos dos grupos, quer entre si quer longitudinalmente. Devido à idade avançada dos indivíduos, muitas vezes a inexistência de diferenças traduz-se num abrandamento da perda de funções (Envelhecimento Ativo).

Este trabalho é um caso de estudo de uma iniciativa municipal, não houve um treino individualizado nem testes laboratoriais, onde os resultados poderiam ser diferentes, dado poder haver um maior controlo sobre as unidades de medida e as condições experimentais. Os resultados não podem ser generalizados para toda a população uma vez que as unidades amostrais foram selecionadas por conveniência.

## Agradecimentos

Anabela Afonso, Gonalo Jacinto e Paulo Infante sao membros do CIMA, centro de investigaao financiado pela Fundaao Nacional para a Ciencia e Tecnologia (FCT), Portugal, pelo projeto «UID/MAT/04674/2019 (CIMA)».

## Referencias

- BULL, F. C., GAUVIN, L., BAUMAN, A., SHILTON, T., KOHL III, H. W. & SALMON, A. (2010). The Toronto Charter for Physical Activity: a global call for action. *Journal of Physical Activity and Health*, 7:4, 421–422.
- DIMITROV, D. M. & RUMRILL JR, P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, 20:2, 159–165.
- HAMER, M., & CHIDA, Y. (2009). Physical activity and risk of neurodegenerative disease: a systematic review of prospective evidence. *Psychological medicine*, 39:1, 3–11.
- LOGAN, M. (2010). *Biostatistical design and analysis using R: a practical guide*. Hoboken, Wiley-Blackwell.
- MARQUES, A. I., SOARES, P., SOARES-MIRANDA, L., MOREIRA, C., OLIVEIRA-TAVARES, A., CLARA-SANTOS, P., VALE, S., SANTOS, S. & CARVALHO, J. (2011). Evaluation of physical activity programmes for the elderly-exploring the lessons from other sectors and examining the general characteristics of the programmes. *BMC Research Notes*, 4:1, 368.
- RIKLI, R. E. & JONES, C. J. (2013). *Senior Fitness Test Manual, Second edition*. Champaign, Human Kinetics.
- SOFI, F., VALECCHI, D., BACCI, D., ABBATE, R., GENSINI, G. F., CASINI, A. & MACCHI, C. (2011). Physical activity and risk of cognitive decline: a meta-analysis of prospective studies. *Journal of Internal Medicine*, 269:1, 107–117.
- WARBURTON, D. E. R., NICOL, C. W. & BREDIN, S. S. D. (2006). Health benefits of physical activity: the evidence. *Canadian Medical Association Journal*, 174:6, 801–809.
- WHO (2011). *Waist circumference and waist–hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008*. Geneva, World Health Organization. [http://apps.who.int/iris/bitstream/handle/10665/44583/9789241501491\\_eng.pdf;jsessionid=67DBD38F51DF89BB752E2CF41A63FC92?sequence=1](http://apps.who.int/iris/bitstream/handle/10665/44583/9789241501491_eng.pdf;jsessionid=67DBD38F51DF89BB752E2CF41A63FC92?sequence=1) (Acedido a 8 de julho de 2013).
- WHO (2010). *Global recommendations on physical activity for health*. Geneva, World Health Organization. <http://www.who.int/dietphysicalactivity/publications/9789241599979/en> (Acedido a 8 de julho de 2013).

# Percepção de barreiras na carreira e diversidade de género: uma abordagem multivariada numa amostra de trabalhadores do sector energético

Ana Sousa Ferreira<sup>1</sup> · Rosário Lima<sup>2</sup>

**Resumo** A presente investigação sobre percepção de barreiras na carreira e a sua relação com o género, visou o estudo das barreiras que interferem no processo de desenvolvimento da carreira das mulheres. Foi utilizado o Inventário de Percepção de Barreiras da Carreira - Revisto (IPBC-R) numa amostra de 325 colaboradores de uma empresa do sector energético, onde a diversidade de género constitui uma prioridade estratégica. Os resultados obtidos revelam que as mulheres percecionam mais significativamente as barreiras na carreira do que os homens.

**Palavras-chave:** Análise Fatorial, Desenvolvimento de Carreira, Diversidade de Género, Percepção de Barreiras na Carreira.

## 1 Introdução

O conceito de trabalho como oportunidade de desenvolvimento pessoal e profissional, tem sido cada vez mais relevante não só para os colaboradores, como também para as organizações, que, ao promoverem o desenvolvimento e gestão das carreiras dos seus trabalhadores, estão igualmente a contribuir para a sua satisfação, motivação e empenhamento. Desta forma, compreender as barreiras que interferem com o processo de desenvolvimento de carreira, é importante para determinar os factores que podem influenciar o sucesso dos indivíduos e das organizações das quais fazem parte (Swanson & Woitke, 1997). Este objectivo, associado ao aumento do número de mulheres no mercado de trabalho, e ao facto de estas percecionarem mais barreiras na carreira do que os homens, torna cada vez mais pertinente o estudo da percepção das barreiras na carreira desta população, de

---

<sup>1</sup> Faculdade de Psicologia, Universidade de Lisboa e Business Research Unit (BRU-IUL), Lisboa, Portugal, [asferreira@psicologia.ulisboa.pt](mailto:asferreira@psicologia.ulisboa.pt)

<sup>2</sup> Faculdade de Psicologia, Universidade de Lisboa, [rosariolimapt@gmail.com](mailto:rosariolimapt@gmail.com)

forma a compreender melhor não só o papel da mulher como também a influência da sua participação nas organizações (Subramaniam, Arumugam & Akeel, 2014).

No presente estudo é adoptada a definição de Swanson e Woitke (1997), que definem barreiras na carreira como vivências ou condições da pessoa ou do meio, que dificultam o seu progresso na carreira. Tem-se demonstrado que a percepção de barreiras na carreira é influenciada por diversos factores como estratégias de *coping*, maturidade de carreira, locus de controlo, traços de personalidade, indecisão face à carreira, identidade vocacional, e o género, sendo este último o foco do presente estudo (Reis, 2014).

Entre outros aspectos, a aposta na diversidade de género nas organizações, ajuda a atrair e reter nas empresas mulheres competentes, apesar de se verificarem apenas ligeiras mudanças na representação deste género nas organizações (Washington, 2010). O desenvolvimento de ambientes de trabalho que promovam a igualdade de género, não são conseguidos simplesmente pelo aumento do número de mulheres na organização, mas sim, através da adopção de políticas e procedimentos sensíveis às necessidades destas, ou mesmo através de acções de formação e de sensibilização (Patel & Buiting, 2013; Bukhari & Saharma, 2014).

Apesar de se verificarem algumas mudanças no mercado de trabalho relativamente à igualdade de género, as mulheres ainda não são representadas e tratadas da mesma forma do que os homens. Subsistem, assim, barreiras responsáveis pela não ascensão das mulheres a cargos de topo nas organizações (Erford & Crockett, 2012), que devem ser trabalhadas e exploradas tanto pelo indivíduo como pela organização, numa relação que vise o mútuo benefício.

Deste modo, colocam-se naturalmente as seguintes questões: As mulheres percebem mais as barreiras na carreira do que os homens? Qual a barreira, ou tipo de barreira na carreira mais percebida pelas mulheres? E pelos homens?

## **2 Metodologia**

### **2.1 Introdução**

Tal como foi já referido, a presente investigação foi desenvolvida numa empresa portuguesa do sector energético, tendo o estudo da percepção de barreiras na carreira, representado o primeiro passo para o desenvolvimento de medidas estratégicas, direccionadas e fundamentadas no âmbito da gestão dos Recursos Humanos. Deste modo, foi seleccionada uma amostra estratificada e representativa da população da empresa, considerando todos os segmentos da organização: Macro Estrutura ou Superior (11%), Quadro Superior ou Gestor de Equipa (37%) e Técnico Operacional (52%).

## 2.2 Amostra

O presente estudo abrangeu uma amostra de 325 colaboradores (46% do total da população em causa), sendo 31% do sexo feminino e 69% do sexo masculino, com idades compreendidas entre os 27 e os 65 anos (Média= 55.33 anos;  $s = 7.26$  anos). Foram também considerados dados relevantes, como a antiguidade na empresa (Média= 32.06 anos;  $s = 8.76$  anos) e as habilitações literárias, verificando-se que 2% dos participantes concluiu o ensino básico, 47% o ensino secundário/técnico profissional, e 51% o ensino superior.

## 2.3 Instrumento

Foi utilizado o Inventário de Percepção de Barreiras da Carreira - Revisto (IPBC-R), que consiste numa tradução e adaptação para a população portuguesa do *Career Barriers Inventory - Revised (CBI-R)*, desenvolvido por Swanson, Daniels e Tokar (1996). O IPBC-R é um instrumento multidimensional que avalia um vasto universo de barreiras da carreira, dirigido a estudantes universitários e adultos, sendo composto por 70 itens organizados em 13 escalas: discriminação sexual (itens 7, 26, 43, 46, 48, 49 e 63); falta de confiança (itens 18, 21, 41 e 45); conflito de papéis múltiplos (itens 4, 11, 17, 28, 40, 44, 60 e 69); conflito entre as exigências da carreira e os filhos (itens 2, 15, 19, 23, 33, 37 e 53); discriminação racial (itens 3, 12, 24, 36, 51 e 54); preparação inadequada (itens 29, 35, 58, 62 e 66); desaprovação por pessoas significativas (itens 20, 47 e 70); dificuldades na tomada de decisão (itens 1, 5, 14, 27, 38, 39, 55 e 59); insatisfação com a carreira (itens 9, 22, 30, 32 e 57); desencorajamento em escolher carreiras não tradicionais (itens 10, 31, 52, 64 e 67); deficiências/problemas de saúde (itens 6, 13 e 25); restrições do mercado de trabalho (itens 16, 34, 56 e 68), e dificuldades de socialização/*networking* (itens 8, 42, 50, 61 e 65). Para responderem ao Inventário, os participantes assinalam numa escala de *Likert* de sete pontos o grau em que a barreira apresentada dificultaria o seu progresso na carreira (1- não dificultaria de todo; 4 - dificultaria um pouco; 7 - dificultaria totalmente). O CBI-R, no seu estudo original, apresentou índices de precisão elevados, com base no cálculo do coeficiente *alpha de Cronbach*, que variam entre .64 e .86 (*alpha* médio = 0.77), e verificaram-se relações significativas ao nível dos 13 fatores que se diferenciam entre .27 e .80 ( $r_p$  médio = 0.60) (Swanson *et al.*, 1996).

Realizada a tradução do instrumento original *CBI-R* de acordo com os padrões internacionais, e autorizada a sua adaptação, a versão portuguesa do Inventário foi aplicada presencialmente a uma amostra piloto. A ausência de sugestões de alteração levou à conservação da tradução do instrumento na aplicação a ser efectuada no âmbito do estudo principal.

Para avaliar a consistência interna do IPBC-R, procedeu-se à análise global do seu coeficiente de precisão *alpha* de Cronbach, que foi de .97, expressando o grau de fiabilidade do instrumento utilizado, e a análise por exclusão de itens (*alpha if item deleted*) não revelou o aumento da consistência interna do instrumento, caso algum dos itens fosse eliminado.

## 2.4 Análise de dados

Aplicado o IPBC – R a 325 trabalhadores, estudou-se a sua estrutura fatorial visando concluir se se trata de um instrumento multidimensional.

A análise da estrutura fatorial deste tipo de escalas é habitualmente realizada, em Psicologia, por recurso à Análise Fatorial Exploratória, baseando-se no estudo da relação entre os itens usando o coeficiente de correlação linear de Pearson, não obstante a natureza ordinal dos itens que habitualmente as compõem (Holgado – Tello *et al.*, 2010).

Neste trabalho, comparam-se esses resultados com a aplicação de outros métodos de análise fatorial adequados à natureza dos dados, como a Análise Ordinal (Lebart *et al.*, 1995) ou a Análise Fatorial baseada no coeficiente de correlação policórica (Babakus *et al.*, 1987).

Como é sabido, o coeficiente de correlação de Pearson avalia a relação entre duas variáveis numéricas sendo essa relação linear. Por outro lado, o coeficiente de correlação de Spearman, uma estatística baseada nas ordens proposta pelo psicometrista Charles Spearman em 1904, e comumente usada em estudos descritivos e inferenciais sobre variáveis expressas em escala ordinal, é pouco utilizado em análise multivariada.

O coeficiente de Spearman, por seu lado, avalia a relação linear entre duas variáveis expressas em escala ordinal.

Recentemente, vários autores propuseram que as análises fatoriais com variáveis ordinais não sejam baseadas em nenhum destes coeficientes de correlação, de Pearson ou de Spearman, mas sim no coeficiente de correlação policórica (Marôco, 2010).

Genericamente, quando se considera uma escala ou instrumento psicológico, pretende-se analisar uma variável latente, variável essa que é contínua, sendo os itens que compõem essa escala, a operacionalização da variável latente em estudo. Deste modo, a questão não é estimar a relação entre os itens ordinais, mas sim entre as variáveis latentes que lhes estão subjacentes.

O coeficiente de correlação policórica estima essa associação. Pearson, definiu este coeficiente no início do séc. XX, supondo que qualquer variável ordinal, por exemplo com 3 categorias pode ser entendida como uma variável quantitativa representada em 3 classes ou categorias. A correlação policórica estima a associação entre duas variáveis latentes, que se assumem ter distribuição normal bivariada, subjacentes a duas variáveis ordinais manifestas.

Este coeficiente tornou-se, contudo, pouco popular por envolver cálculos computacionais pesados quer na estimação pelo método da máxima verosimilhança, quer usando um método iterativo, computacionalmente mais rápido, designado por algoritmo *two-step* (Martinson & Hamdan, 1975).

Assim, neste trabalho, confrontamos os resultados do estudo da dimensionalidade do instrumento IPBC-R, recorrendo a métodos de análise fatorial sobre a matriz de correlações de Pearson (AFE) ou sobre a matriz de correlações de Spearman (AO) ou sobre a matriz de correlações policóricas (AP).

Apresentar-se-ão, ainda, as barreiras mais percecionadas pelas mulheres e pelos homens, e a análise da diversidade de género, onde serão utilizadas técnicas de análise inferencial.

### 3 Resultados

A análise da estrutura fatorial do IPBC – R, considerando os 70 itens que o constituem, e o estudo da diversidade de género são apresentados seguidamente.

#### 3.1 Resultados da análise da estrutura fatorial do IPBC-R

A aplicação do método de análise fatorial sobre as várias matrizes de correlação – de Pearson, de Spearman e policóricas – conduziu a dois resultados diferentes relativamente à estrutura fatorial do Inventário, embora semelhantes.

Isto é, apresentamos seguidamente os resultados da análise fatorial exploratória (AFE), com estimação dos factores comuns pelo método da análise em componentes principais, e rotação *varimax*, de acordo com a metodologia usada no instrumento original. Os primeiros resultados deste modelo de AFE, seleccionando o número de factores pelo critério de Kaiser, revelou uma estrutura com catorze factores, no entanto, o peso (isto é, a correlação entre o item e o fator) dos itens nos últimos factores era extremamente fraco e, seguindo a indicação do *scree plot* e a percentagem de variabilidade explicada (54.36%), optou-se por uma solução com cinco factores, tendo-se excluído os itens com pesos de 0.4 ou inferior (itens 10, 34, 60). Na Tabela 1, apresentamos os cinco factores obtidos no modelo AFE, indicando para cada fator: o número de itens e os itens que o constituem, o valor do *alpha de Cronbach* e a variabilidade explicada por esse fator.

**Tabela 1** – Resultados da análise fatorial exploratória (AFE).

Fatores	Nº de itens	Itens	<i>alpha de Cronbach</i>	Var. explicada
Mentalidade individual	20	1, 5, 8, 14, 18, 21, 22, 27, 29, 32, 35, 38, 39, 41, 45, 58, 59, 62, 66, 68	.95	36.23%
Discriminação sexual e racial	19	3, 6, 7, 12, 13, 24, 25, 26, 33, 36, 37, 43, 46, 48, 49, 51, 54, 63, 67	.96	6.85%
Desaprovação por pessoas signif.	12	20, 28, 31, 40, 42, 44, 47, 52, 53, 64, 69, 70	.92	4.42%
Dificuldades de social./net.	7	30, 50, 55, 56, 57, 61, 65	.88	3.61%
Conflito trab.- família	9	2, 4, 9, 11, 15, 16, 17, 19, 23	.87	3.25%
Perceção global de barreiras na carreira	67	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 67, 68, 69 e 70	.93	54.36%

Como podemos observar, embora não se espelhe a estrutura em treze tipos de barreiras definidas originalmente, a interpretação dos factores permite concluir que estes representam barreiras que fazem parte do fator geral Perceção global de barreiras na carreira.

Por outro lado, apresentamos seguidamente, na Tabela 2, os resultados obtidos com a aplicação da Análise Ordinal (AO), tendo-se excluído os itens com pesos de .4 ou inferior (itens 33 e 60). Os resultados da análise baseada nas correlações policóricas (AP) são iguais aos resultados obtidos com a AO, uma vez que na nossa amostra a matriz de correlações de Spearman e policóricas são extremamente semelhantes. Na Tabela 2, apresentamos os cinco factores obtidos nos modelos AO e AP, indicando para cada fator: o número de itens e os itens que o constituem, o valor do *alpha de Cronbach* e a variabilidade explicada por esse fator.

**Tabela 2** – Resultados da análise ordinal (AO) e da análise policórica (AP).

Fatores	Nº de itens	Itens	<i>alpha de Cronbach</i>	Var. explicada
Insegurança/Falta de confiança	22	1, 5, 8, 9, 14, 18, 21, 22, 27, 29, 32, 35, 38, 39, 41, 45, 58, 59, 61, 62, 66, 68	.94	37.83%
Discriminações múltiplas	17	3, 6, 7, 10, 11, 12, 13, 16, 24, 25, 36, 37, 43, 46, 48, 54, 63	.94	8.55%
Desaprovação por pessoas significativas	13	20, 28, 31, 40, 42, 44, 47, 52, 53, 64, 67, 69, 70	.91	5.60%
Insatisfação e dificuldades na carreira	10	26, 30, 34, 49, 50, 51, 55, 56, 57, 65	.89	4.79%
Conflitos múltiplos	6	2, 4, 15, 17, 19, 23	.82	2.93%
Percepção global de barreiras na carreira	68	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 67, 68, 69 e 70	.97	59.70%

Comparando os resultados apresentados na Tabela 1 e Tabela 2 podemos notar que os fatores obtidos não são exatamente os mesmos, mas os itens que compõem cada um dos fatores são muito semelhantes. Contudo, o segundo modelo fatorial (obtido com a AO e AP) tem uma variabilidade explicada superior, como poderia ser esperado, dado que em geral, as estimativas das relações entre variáveis ordinais obtidas pelo coeficiente de Pearson tendem a ser inferiores às verdadeiras relações e, deste modo, os fatores obtidos tendem a ter menor capacidade explicativa da variabilidade dos dados. Este modelo apresenta também uma fiabilidade um pouco mais elevada (.97) apesar de terem sido retirados dois itens que certamente contribuem pouco para a precisão do instrumento.

### 3.2 Comparação das barreiras percebidas por género

De modo a poder responder às questões inicialmente colocadas:

- As mulheres percebem mais barreiras na carreira do que os homens?
- Qual a barreira, ou tipo de barreira na carreira mais percebida pelas mulheres? E pelos homens?

recorreu-se a técnicas de inferência para, por um lado, comparar a perceção de barreiras nos dois géneros e, por outro, perceber quais as barreiras mais percecionadas pelas mulheres e pelos homens.

Verificadas as condições de aplicação, aplicou-se um teste de igualdade de valores médios para populações normalmente distribuídas e amostras independentes para comparar as perceções de barreiras na carreira por género. As hipóteses testadas são as seguintes:

$H_0$ : A perceção média na barreira  $X$  é igual nos dois géneros vs.

$H_a$ : A perceção média na barreira  $X$  é superior no género feminino à registada no género masculino.

Na Tabela 3, apresentam-se os resultados da comparação das perceções de barreiras na carreira por género, indicando para cada tipo de barreira, a perceção média e o desvio padrão por género bem como o valor da estatística de teste  $t$  com distribuição  $t$  de *Student* ( $n_1+n_2-2=323$  g. l.), o valor- $p$  e o respetivo valor da medida  $d$  de *Cohen*.

**Tabela 3** – Resultados da comparação da perceção de barreiras por género, segundo os fatores obtidos no modelo AFE.

Barreiras	Género	n	Média	s	t	valor-p <sup>1</sup>	d de Cohen																																																								
Mentalidade individual	Feminino	101	5,11	.94	2.11	.02*	.24																																																								
	Masculino	224	4.84	1.24				Discriminação sexual e racial	Feminino	101	5.18	1.12	4.59	.00**	.52	Masculino	224	4.53	1.36	Desaprovação por pessoas significativas	Feminino	101	4.04	1.12	2.28	.01*	.28	Masculino	224	3.71	1.23	Dificuldades de social./net.	Feminino	101	4.87	.97	1.71	.04*	.19	Masculino	224	4.66	1.23	Conflito trabalho-família	Feminino	101	4.92	1.00	4.04	.00**	.51	Masculino	224	4.36	1.20	Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41
Discriminação sexual e racial	Feminino	101	5.18	1.12	4.59	.00**	.52																																																								
	Masculino	224	4.53	1.36				Desaprovação por pessoas significativas	Feminino	101	4.04	1.12	2.28	.01*	.28	Masculino	224	3.71	1.23	Dificuldades de social./net.	Feminino	101	4.87	.97	1.71	.04*	.19	Masculino	224	4.66	1.23	Conflito trabalho-família	Feminino	101	4.92	1.00	4.04	.00**	.51	Masculino	224	4.36	1.20	Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41	Masculino	224	4.44	1.15								
Desaprovação por pessoas significativas	Feminino	101	4.04	1.12	2.28	.01*	.28																																																								
	Masculino	224	3.71	1.23				Dificuldades de social./net.	Feminino	101	4.87	.97	1.71	.04*	.19	Masculino	224	4.66	1.23	Conflito trabalho-família	Feminino	101	4.92	1.00	4.04	.00**	.51	Masculino	224	4.36	1.20	Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41	Masculino	224	4.44	1.15																				
Dificuldades de social./net.	Feminino	101	4.87	.97	1.71	.04*	.19																																																								
	Masculino	224	4.66	1.23				Conflito trabalho-família	Feminino	101	4.92	1.00	4.04	.00**	.51	Masculino	224	4.36	1.20	Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41	Masculino	224	4.44	1.15																																
Conflito trabalho-família	Feminino	101	4.92	1.00	4.04	.00**	.51																																																								
	Masculino	224	4.36	1.20				Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41	Masculino	224	4.44	1.15																																												
Perceção global de barreiras na carreira	Feminino	101	4.86	.88	3.60	.00**	.41																																																								
	Masculino	224	4.44	1.15																																																											

<sup>1</sup> O valor- $p$  apresentado é para um teste unilateral.

\* valor- $p < .05$  \*\* valor- $p < .01$

Como se pode observar na Tabela 3, quer nos cinco tipos de barreiras quer na Percepção global de barreiras na carreira, o género feminino percebeu em média significativamente mais a barreira do que o género masculino. Note-se, contudo, que se determinássemos uma medida de dimensão do efeito como o *d de Cohen*, poderíamos dizer que esses efeitos eram todos pequenos ou médios, destacando-se as diferenças observadas em Discriminação sexual e racial e Conflito trabalho-família (Cohen, 1988).

E, no modelo AO (ou AP)? Na Tabela 4, apresentam-se os resultados da comparação das percepções de barreiras na carreira por género, indicando para cada tipo de barreira, a percepção média e o desvio padrão por género bem como o valor da estatística de teste t com distribuição t de *Student* ( $n_1+n_2-2 = 323$  g. l.), valor-p e o respetivo valor da medida *d de Cohen*.

**Tabela 4** – Resultados da comparação da percepção de barreiras por género, segundo os fatores obtidos no modelo AO e AP.

Barreiras	Género	n	Média	s	t	valor-p <sup>1</sup>	<i>d de Cohen</i>																																																								
Insegurança/Falta de confiança	Feminino	101	5.13	.86	1.66	.04*	.21																																																								
	Masculino	224	4.92	1.15				Discriminações múltiplas	Feminino	101	4.91	1.06	3.58	.00**	.49	Masculino	224	4.34	1.25	Desaprovação por pessoas significativas	Feminino	101	3.94	.99	3.13	.00**	.41	Masculino	224	3.51	1.12	Insatisfação e dific. na carreira	Feminino	101	4.86	.96	2.32	.01*	.27	Masculino	224	4.56	1.21	Conflitos múltiplos	Feminino	101	4.48	1.11	4.57	.00**	.58	Masculino	224	3.79	1.26	Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54
Discriminações múltiplas	Feminino	101	4.91	1.06	3.58	.00**	.49																																																								
	Masculino	224	4.34	1.25				Desaprovação por pessoas significativas	Feminino	101	3.94	.99	3.13	.00**	.41	Masculino	224	3.51	1.12	Insatisfação e dific. na carreira	Feminino	101	4.86	.96	2.32	.01*	.27	Masculino	224	4.56	1.21	Conflitos múltiplos	Feminino	101	4.48	1.11	4.57	.00**	.58	Masculino	224	3.79	1.26	Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54	Masculino	224	4.36	.98								
Desaprovação por pessoas significativas	Feminino	101	3.94	.99	3.13	.00**	.41																																																								
	Masculino	224	3.51	1.12				Insatisfação e dific. na carreira	Feminino	101	4.86	.96	2.32	.01*	.27	Masculino	224	4.56	1.21	Conflitos múltiplos	Feminino	101	4.48	1.11	4.57	.00**	.58	Masculino	224	3.79	1.26	Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54	Masculino	224	4.36	.98																				
Insatisfação e dific. na carreira	Feminino	101	4.86	.96	2.32	.01*	.27																																																								
	Masculino	224	4.56	1.21				Conflitos múltiplos	Feminino	101	4.48	1.11	4.57	.00**	.58	Masculino	224	3.79	1.26	Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54	Masculino	224	4.36	.98																																
Conflitos múltiplos	Feminino	101	4.48	1.11	4.57	.00**	.58																																																								
	Masculino	224	3.79	1.26				Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54	Masculino	224	4.36	.98																																												
Percepção global de barreiras na carreira	Feminino	101	4.81	.65	3.75	.00**	.54																																																								
	Masculino	224	4.36	.98																																																											

<sup>1</sup> O valor-p apresentado é para um teste unilateral.

\* valor-p < .05. \*\* valor-p < .01

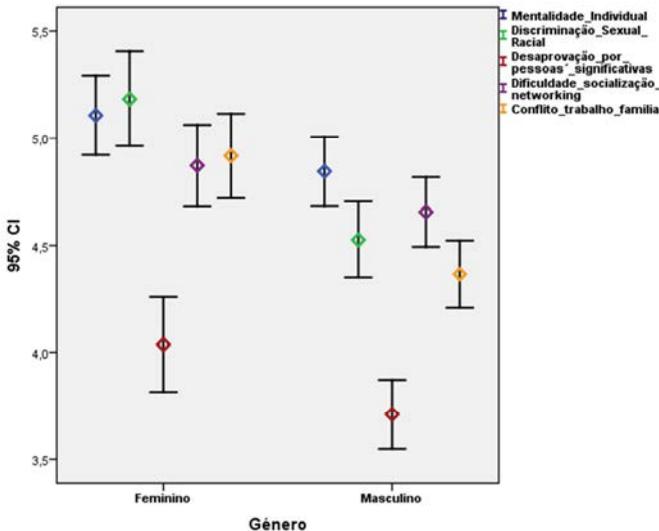
Como se pode observar, também no modelo AO e AP, quer nos cinco tipos de barreiras quer na Percepção global de barreiras na carreira, o género feminino percebeu em média significativamente mais a barreira do que o género masculino. Note-se, contudo, que se determinássemos uma medida de dimensão do efeito como o *d de Cohen*, poderíamos dizer que esses efeitos eram todos pequenos

ou médios, destacando-se as diferenças observadas em Conflitos múltiplos e Perceção global de barreiras na carreira (Cohen, 1988).

Finalmente, vamos tentar compreender quais as barreiras que as mulheres mais percecionam, bem como quais as barreiras que os homens mais percecionam, tendo-se optado por apresentar estes resultados graficamente.

No Gráfico 1, apresentamos uma representação num *error bar* onde cada barra representa o intervalo de confiança a 95% para a perceção média em cada tipo de barreira.

É, então, possível verificar que as mulheres percecionam em média mais a Barreira de Discriminação Sexual e Racial, enquanto os homens percecionam em média mais a Barreira Mentalidade Individual. Note-se, contudo, que nas mulheres não existem diferenças significativas entre a perceção em média na Barreira de Discriminação Sexual e Racial e na Barreira Mentalidade Individual. Pode-se ainda concluir que, quer para as mulheres quer para os homens a barreira menos percecionada em média é a Barreira de Desaprovação por pessoas significativas.



**Gráfico 1** – Comparação das barreiras de carreira por género.

Isto é, as barreiras, agora identificadas, poderão certamente guiar as intervenções ao nível da gestão dos recursos humanos.

## 4 Conclusões

O estudo comparativo sobre modelos de análises fatoriais para estudar relações entre variáveis ordinais apontam no sentido das conclusões de alguns autores,

segundo os quais, quando o número de categorias da variável ordinal aumenta (7, 10,...), como é o caso dos itens do Inventário IPBC, as diferenças entre as correlações obtidas pelo coeficiente de Pearson e outros coeficientes mais adequados atenuam-se (Bollen, 1989; Finney & Distefano, 2006).

Por outro lado, constata-se a importância de completar os resultados de comparações de grupos (no caso a comparação de barreiras na carreira entre gêneros) com medidas de dimensão do efeito. Estas medidas geram informação respeitante à intensidade da divergência entre a hipótese nula e os resultados observados, completando assim a decisão, relativa à veracidade da hipótese nula, que os testes de significância estatística fornecem a um determinado nível de significância.

Relativamente às intervenções ao nível da gestão dos recursos humanos, com base nos resultados obtidos, torna-se evidente a necessidade das empresas apostarem em medidas que atenuem as barreiras percebidas pelas mulheres, providenciando a este grupo minoritário as ferramentas necessárias ao seu desenvolvimento de carreira. Ou seja, as barreiras identificadas poderão traduzir a importância de intervir ao nível da gestão dos recursos humanos em aspetos como a preparação inadequada das mulheres, os seus níveis de satisfação com a carreira, falta de confiança e dificuldades na tomada de decisão (Reis, 2014).

Uma vez que a diversidade de gênero, no âmbito do planeamento e desenvolvimento estratégico dos recursos humanos, pode ser considerada um fator crítico para a implementação eficaz de estratégias que contribuem para o sucesso das organizações, pode-se afirmar que da investigação desenvolvida resultou a definição de linhas estratégicas a assumir pela empresa alvo do estudo, no âmbito da diversidade de gênero. Nesse sentido, o foco deve incidir na identificação dos problemas que sistematicamente afetam negativamente a diversidade de gênero, e prejudicam o desempenho organizacional (Bukhari & Sharma, 2014).

## Referências

- BABAKUS, E., FERGUSON, C. E., & JÖRESKOG, K. G. (1987). The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumption. *Journal of Marketing Research*, 24(2), 222-228.
- BOLLEN, K. A. (1989). *Structural equations with latent variables*, Wiley, New York.
- BUKHARI, S. S. & SHARMA, B. C. (2014). Workplace Gender Diversity & Inclusive Growth in Public & Private Organizations. *Indian Journal of Industrial Relations*, 49(3), 551-563.
- COHEN, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum Associates, USA.

- ERFORD, B. T., & CROCKETT, S. A. (2012). Practice and research in career counseling and development - 2011. *The Career Development Quarterly*, 60(4), 290-332.
- FINNEY, S. J., & DISTEFANO, C. (2006). Non-normal and Categorical Data in Structural Equation Modelling, in G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: a second course (pp. 269-314)*. Greenwich, Connecticut: Information Age Publishing.
- HOLGADO-TELLO, F.P., CHACÓN-MOSCOSO, S., BARBERO-GARCÍA, I. & VILA-ABAD, E.(2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44(1), 153-166.
- LEBART, L.; MORINEAU, A. & PIRON, M. (1995): *Statistique Exploratoire Multidimensionnelle*, Dunod Ed., Paris.
- MARÔCO, J. (2010). Integração do R nos menus do PASW Statistics: Um exemplo de aplicação com o package 'polycor' do R, in *Boletim da Sociedade Portuguesa de Estatística, Primavera de 2010*, 71-80. Portugal: Sociedade Portuguesa de Estatística.
- MARTINSON, E. O., & HAMDAN, M. A. (1975). Algorithm AS 87: Calculation of the Polychoric Estimate of Correlation in Contingency Tables. *Applied Statistics*, 24(2), 272-278.
- PATEL, G. & BUITING, S., (2013, Junho). Gender differences in leadership styles and the impact within corporate boards, in *The commonwealth secretariat, social transformation program division*.
- REIS, A. C. (2014). *Percepção de Barreiras na Carreira e Diversidade de Género: Estudo com uma amostra de trabalhadores do sector energético*. Tese de Mestrado, Universidade de Lisboa.
- SUBRAMANIAM, I. D., ARUMUGAM, T., & AKEEL, A. B. A. A. B. (2014). Demographic and Family Related Barriers on Women Managers' Career Development. *Asian Social Science*, 10(1), 86-94.
- SWANSON, J. L., DANIELS, K. K., & TOKAR, D. M. (1996). Assessing perceptions of career-related barriers: The Career Barriers Inventory. *Journal of Career Assessment*, 4(2), 219-244.
- SWANSON, J. L., & WOITKE, M. B. (1997). Theory into practice in career assessment for women: Assessment and interventions regarding perceived career barriers. *Journal of Career Assessment*, 5(4), 443-462.
- WASHINGTON, C. E. (2010, June). Mentoring, Organizational Rank, and Women's Perceptions of Advancement Opportunities in the Workplace, in *Forum on Public Policy: A Journal of the Oxford Round Table*.

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

Anabela Afonso<sup>1</sup> · Gonçalo Jacinto<sup>2</sup> · Paulo Infante<sup>3</sup> · Teresa Engana<sup>4</sup> ·  
Filipe Glória<sup>5</sup> · Rosa Espanca<sup>6</sup>

**Resumo** A privação do sono contribui para vários problemas de saúde, emocionais comportamentais, bem como para o insucesso escolar. Estas consequências têm sido alertadas por diversas entidades de saúde devido à crescente prevalência da privação de sono em crianças. Neste trabalho avaliamos os hábitos e problemas do sono de um grupo de crianças do 1º ciclo e comparam-se estes resultados com os obtidos com outros estudos realizados em Portugal, em idades similares.

**Palavras-chave:** Análise em Componentes Principais, Regressão Logística, Sono.

### 1 Introdução

O sono e os problemas associados ao reduzido número de horas que as crianças dormem, está cada vez mais na ordem do dia e é alvo de diversas alertas pelas entidades de saúde. Para chamar a atenção para a sua importância, foi definido que na sexta-feira da segunda semana completa de março se comemoraria o dia Mundial do Sono.

Vários estudos alertam para os efeitos negativos associados à privação do sono nas crianças (Clemente, 1997, Silva *et al.*, 2013, Silva, 2014, Silva *et al.*, 2014,

---

<sup>1</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>2</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [gjcj@uevora.pt](mailto:gjcj@uevora.pt)

<sup>3</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>4</sup> Associação de Pais e Encarregados de Educação da Escola Básica do Chafariz D'El Rei, [teresaengana@cm-evora.pt](mailto:teresaengana@cm-evora.pt)

<sup>5</sup> Centro da Criança e do Adolescente, Hospital Cuf Descobertas, Lisboa, [fs.sono@gmail.com](mailto:fs.sono@gmail.com)

<sup>6</sup> Agrupamentos de Centros de Saúde do Alentejo Central, [Rosa.Espanca@alentejocentral.min-saude.pt](mailto:Rosa.Espanca@alentejocentral.min-saude.pt)

Owens, Spirito & McGuinn, 2000) que, para além das consequências na regulação emocional e do comportamento, nas funções cognitivas, no sucesso académico, na obesidade e no risco de quedas acidentais, também se repercutem na vida dos pais (Silva *et al.*, 2013). Os padrões e ciclos de sono variam à medida que as crianças se desenvolvem, sendo influenciados por fatores de ordem cultural e ambiental. Para dormir bem é fundamental que a criança tenha bons hábitos de sono. Segundo a World Association of Sleep Medicine, uma criança entre os 6 e os 10 anos deve dormir entre 10 a 11 horas por noite.

O objetivo deste trabalho foi avaliar os hábitos e problemas do sono de um grupo de crianças do 1º ciclo da cidade de Évora e comparar alguns resultados com os obtidos com outros estudos realizados em Portugal, em idades similares.

## 2 Métodos

### 2.1 População e recolha de informação

Para este estudo foi definida como população alvo as crianças do 1º ciclo de uma escola no concelho de Évora, no ano letivo 2014/15.

Durante o mês de maio de 2015 foi solicitado aos pais das crianças que respondessem a dois questionários. Um primeiro questionário que visava a caracterização das crianças (idade, sexo, ano de escolaridade), e a caracterização sociodemográfica do agregado familiar e do encarregado de educação. Um segundo questionário que consistia na “versão-curta” portuguesa do Children’s Sleep Habits Questionnaire (Silva *et al.*, 2013) com vista a recolher informação sobre o padrão do sono do aluno durante a última semana, ou numa semana recente considerada mais normal. O questionário é composto por questões sobre as horas de deitar e de acordar (nos dias de semana e no fim de semana), a duração habitual do sono e a frequência de diversos comportamentos relacionados com o sono, identificados com base em sintomas clínicos comuns dos diagnósticos pediátricos mais usuais (Owens, Spirito & McGuinn, 2000). As respostas aos 33 itens do comportamento são do tipo ordinal com 3 opções pontuadas de 1 a 3, de forma a que a pontuação mais elevada corresponda aos comportamentos-problema mais frequentes (Silva *et al.*, 2013).

Os comportamentos estão agrupados em 8 subescalas de forma desigual: Resistência em ir para a cama (6 itens), Início do sono (1 item), Duração do sono (3 itens), Ansiedade associada ao sono (4 itens), Despertares noturnos (3 itens), Parassónias (7 itens), Perturbações respiratórias do sono (3 itens), e Sonolência diurna (8 itens). Deste modo, é possível avaliar diferentes tipos de problemas de sono e obter o Índice de Perturbação do Sono (IPS), que corresponde à soma dos 33 itens cujo valor varia entre 33 e 99, onde um valor mais elevado corresponde a

maiores problemas nos hábitos do sono. Para Portugal considera-se que uma criança que apresente um IPS de 48 ou mais pontos tem “problemas do sono” (Silva *et al.*, 2013).

## 2.2 Análise estatística

Inicialmente a informação recolhida foi analisada usando métodos de estatística descritiva univariada e multivariada.

Para testar a normalidade das variáveis quantitativas foi usado o teste de Shapiro-Wilk. Para comparar as distribuições do número de horas de sono e do Índice de Perturbação do Sono por sexo e por idades foram usados o teste de Mann-Whitney-Wilcoxon e Kruskal-Wallis. A associação entre as subescalas do hábito do sono e o IPS foi avaliada com o coeficiente de correlação de Spearman.

Para averiguar associações entre os comportamentos realizou-se uma análise em componentes principais (Johnson & Wichern, 2007), com base na matriz de correlações, considerando a idade, o ano de escolaridade e o sexo como variáveis passivas (i.e., não são usadas na construção das componentes principais). Os valores omissos nas subescalas foram substituídos pela média da subescala. Para a escolha do número de dimensões a reter foram considerados seguintes critérios: i) análise da representação gráfica dos valores próprios; ii) valores próprios superiores à média dos valores próprios; iii) percentagem total da variabilidade explicada. Tendo em conta as variáveis que mais contribuíam para cada uma das componentes consideradas, foram interpretadas as componentes principais, bem como alguns perfis de crianças. Posteriormente foi usada uma análise classificatória não hierárquica (*K-means*) com as coordenadas das crianças nas componentes principais.

Para identificar os fatores de risco que contribuem para um reduzido número de horas de sono nas crianças ajustaram-se modelos de regressão logística (Hosmer *et al.*, 2013) à variável horas de sono diárias da criança reportadas pelos pais, dicotomizada pelas crianças que dormem menos de 10 horas e pelas que dormem pelo menos 10 horas (o valor mínimo que uma criança deve dormir segundo a World Association of Sleep Medicine). Para as variáveis explicativas foram consideradas variáveis caraterizadoras das crianças, do seu meio socioeconómico e os valores dos índices das subescalas do questionário. O ajustamento do modelo foi feito usando a estratégia referida em Hosmer *et al.* (2013). Além de uma avaliação da forma funcional das variáveis contínuas e de uma análise de resíduos para pesquisa de observações influentes e *outliers*, foi feita uma validação do modelo por *Bootstrap*. A bondade do ajustamento foi testada através do teste de Cessie-Von-Houvelingen, e a sua capacidade discriminativa avaliada pela área AUC da curva ROC.

Toda a análise estatística foi realizada com auxílio do *software* R versão 3.2.3 (R Core Team, 2015).

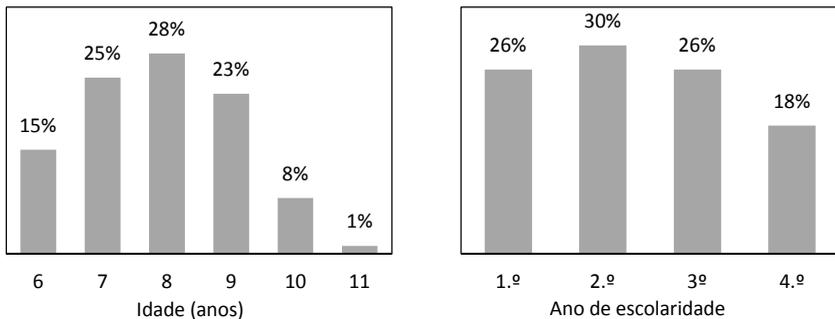
### 3 Resultados

#### 3.1 Caracterização geral

Foram distribuídos 98 inquéritos tendo-se obtido 88 respostas, ou seja, uma taxa de resposta de 89,8%.

A maioria das respostas obtidas refere-se a alunos do sexo masculino (54%), com média de idades igual a 7,88 anos (desvio padrão - DP - igual a 1,22), estando os alunos do 2º ano mais representados, com 30% dos alunos, seguidos do 1º e 3º anos, com 26% dos alunos em cada um dos anos (Figura 1). Três em cada quatro crianças tem atividades fora da escola durante a semana, usualmente entre 1 a 3 vezes por semana (64% do total das crianças). Nos dias de semana, fora da escola e das atividades, as crianças brincam, em média, 1h 40m (DP = 1h).

Menos de 7% destes alunos têm necessidades educativas especiais. O encarregado de educação é na maior parte das vezes a mãe (84%), tem habilitação superior (63%), trabalha por conta de outrem (75%), não tem horário de trabalho flexível (56%) e a entidade patronal não possibilita horários adaptados (53%).

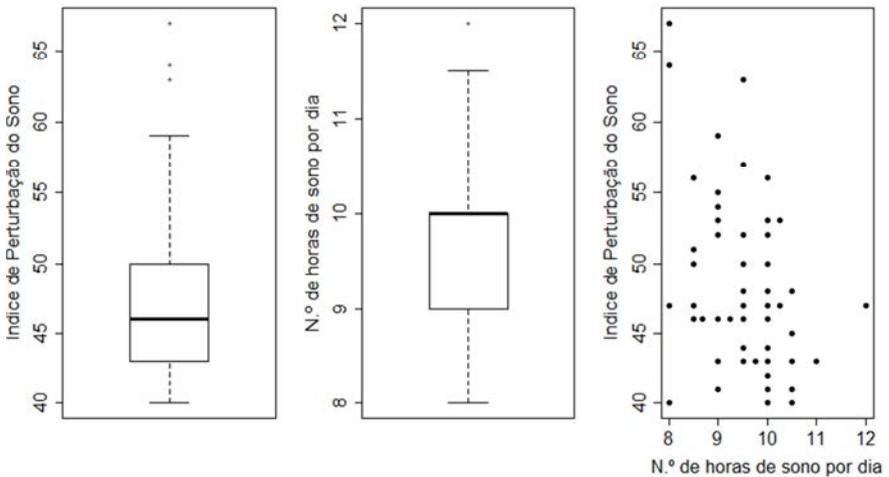


**Figura 1** – Idade e ano de escolaridade dos alunos da amostra.

#### 3.2 Análise das horas de sono e do IPS

O número de horas de sono das crianças reportadas pelos pais foi, em média, 9h 40m (DP = 48m), não existindo diferenças significativas entre sexos ( $p$ -value = 0,348) nem entre idades ( $p$ -value = 0,342). Durante a semana as crianças dormem, em média, 9h 53m (DP = 31m), verificando-se que 57% das crianças dormem entre 10 e 11 horas por noite enquanto as restantes dormem menos de 10

horas. O tempo de sono durante a semana difere entre os sexos ( $p$ -value = 0,025), em que as raparigas dormem em média mais 16m do que os rapazes. Ao fim de semana as crianças dormem mais horas (10h 24m, DP = 55m), no entanto 23% delas dormem menos de 10 h e apenas 12% dormem mais de 11h. Durante a semana as crianças tendem, em média, a acordar às 7h 40m (DP = 26m) e a deitar mais cedo 21h 45m (DP = 25m) do que ao fim de semana (acordar: 9h, DP = 56m; deitar: 22h 35m, DP = 38m). Nos dias de semana, há diferença significativa na hora de deitar por sexo ( $p < 0,001$ ), em que os rapazes se deitam, em média, 23m mais tarde do que as raparigas. O diagrama de caixa e bigodes do número de horas de sono encontra-se na Figura 2b.



**Figura 2** – a) Índice de Perturbação do Sono (esquerda); b) Número de horas de sono por dia reportadas pelos pais (centro) e c) gráfico dispersão entre ambos (direita).

Cerca de 6% dos pais referiram que consideram que o seu “filho/filha tem algum problema com o sono”. Esta baixa prevalência contrasta com o valor elevado do Índice de Perturbação do Sono (IPS), uma vez que 36% apresentam um IPS igual ou superior a 48 (Figura 2a), contraste que parece traduzir uma tolerância cultural em relação aos problemas do sono.

Não foram identificadas diferenças significativas no Índice de Perturbação do Sono (IPS) por sexo ( $p$ -value = 0,109), nem por idades ( $p$ -value = 0,676). O IPS mostrou estar correlacionado negativamente com o número de horas de sono diárias da criança ( $r = -0,39$ ), como se pode observar na Figura 2c.

A cotação de todas as subescalas está correlacionada de forma positiva com a cotação total (IPS), sendo que as que apresentam uma correlação mais forte são as *Parassónias*, a *Duração do sono*, a *Sonolência diurna*, a *Resistência em ir para a*

*cama* e a *Ansiedade associada ao sono*, e as que apresentam uma correlação mais fraca o *Início do sono* e as *Perturbações respiratórias do sono* (Tabela 1).

O máximo observado nos índices *Início do Sono* e *Despertares noturnos* coincide com o valor máximo que pode ser obtido destes índices, o que indica que existem crianças com todos os comportamentos-problemas descritos nestas subescalas (Tabela 1). Tendo em conta o número de itens que compõem a subescala *Parassónias* (8), estas crianças não parecem apresentar comportamentos de parassónias frequentes.

**Tabela 1** – Principais medidas resumo dos índices (cotação) das subescalas de problemas do sono e correlação linear com o Índice de Perturbação do Sono (IPS).

Índices das subescalas	N	Min.	Q <sub>1</sub>	Q <sub>2</sub>	$\bar{x}$	Q <sub>3</sub>	Máx.	DP	r <sub>IPS</sub>
Resistência em ir para a cama	84	6	6	7	7,4	8	17	2,0	0,55*
Início do sono	87	1	1	2	2,1	3	3	0,9	0,27*
Duração do sono	85	3	3	3	3,8	4	8	1,3	0,59*
Ansiedade associada ao sono	81	4	4	5	5,3	6	11	1,6	0,48*
Despertares noturnos	80	3	3	3	3,4	3	9	0,9	0,36*
Parassónias	83	7	7	8	8,4	9	13	1,5	0,67*
Perturbações respiratórias do sono	83	3	3	3	3,4	3	7	0,9	0,28*
Sonolência Diurna	79	12	15	16	16,2	18	22	2,5	0,58*
Índice de Perturbação do Sono	72	40	43	46	47,3	50	67	6,0	-

\* Valor significativo ao nível de significância de 5%.

### 3.3 Análise de Correspondências Múltiplas

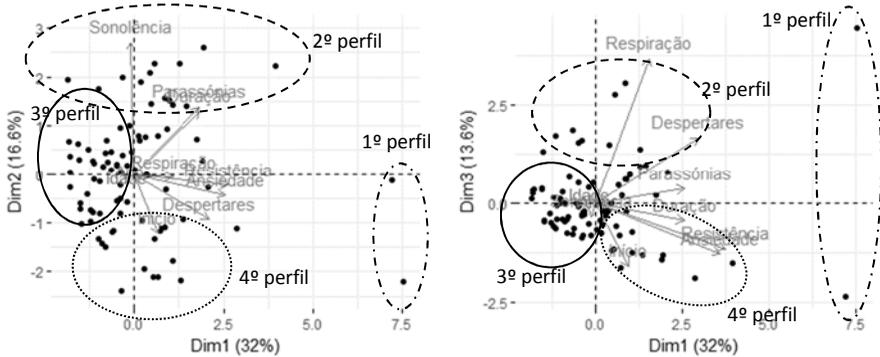
Foi possível resumir a informação relacional entre os índices em três componentes ortogonais que explicam 62% da variância total das variáveis originais. As variáveis passivas sexo, idade e ano de escolaridade não mostraram estar relacionadas com as subescalas dos comportamentos-problema do sono.

Tendo em conta os índices com maior peso (correlação com a componente) em cada componente principal foi possível interpretá-las como:

1. Dificuldade em adormecer: esta componente relaciona a *Resistência em ir para a cama* e a *Ansiedade associada ao sono* com os *Despertares noturnos*;

2. Sonolência diurna e sono agitado: há uma forte correlação positiva desta componente com a *Sonolência diurna*. Além disso, há uma oposição das *Parassónias* e a *Duração do sono* com o *Início do sono*;
3. Distúrbios respiratórios do sono: a subescala *Perturbações respiratórias do sono* é a que apresenta maior correlação positiva com esta componente. Os *Despertares noturnos* são contrastados com o *Início do sono*.

Com base nestas componentes principais identificaram-se 4 perfis de crianças (Figura 3).



**Figura 3** – Posicionamento de cada criança no mapa bidimensional definido pelas três componentes retidas, etiquetados pela idade, e alguns perfis identificados.

Aplicou-se a análise classificatória (*K-means*) às coordenadas das crianças em cada uma das componentes. Note-se que, atendendo à baixa variância explicada - 62% da variância total - a análise classificatória utilizará todas as componentes principais, no desenvolvimento deste trabalho.

As crianças foram agrupadas em 4 grupos homogêneos, que correspondem, aproximadamente, aos perfis identificados, e cujas principais características são:

Grupo 1 (2 crianças): destacam-se das restantes crianças por terem apresentado maiores problemas do sono devido à ansiedade (medo do escuro, medo de dormir sozinha, precisa de companhia para adormecer, dificilmente consegue dormir fora de casa) e, talvez como consequência, ao facto de apresentarem maior resistência em ir para a cama (“luta” contra o sono e não se deita sempre à mesma hora, tem dificuldade em adormecer sozinha na sua própria cama pelo que precisa dos pais no quarto ou de adormecer na cama dos pais ou dos irmãos) e acordarem durante a noite. Estas crianças dormem apenas 8h por dia;

Grupo 2 (22 crianças): sofrem de sonolência diurna (de manhã têm que ser acordadas pelos pais ou irmãos porque não acordam por si próprias, têm dificuldade em sair da cama, acordam mal-humoradas, demoram a ficar

bem acordadas, parecem cansadas, adormecem a ver televisão ou a andar de carro), que pode ser devido ao facto de demorarem a adormecer, o que origina a que não durmam o suficiente. Estas crianças dormem, em média, 9h 25m por dia e mais de 60% delas dormem menos de 10 h por dia;

Grupo 3 (30 crianças): apenas apresentam problemas de sonolência diurna. Metade destas crianças dorme entre 10 e 11h por dia;

Grupo 4 (34 crianças): apenas apresentam alguma dificuldade em adormecer. As crianças deste grupo são as que têm menores valores de IPS e são as que mais horas dormem por dia.

O Grupo 1, correspondente ao 1º perfil da ACM, é constituído apenas por duas crianças, que surgem, através dos resultados da análise multivariada dos dados, como *outliers*. Posteriormente, foi feita a análise multivariada sem estes dois indivíduos (resultados não apresentados) que resultou na retenção de mais componentes principais, com menor percentagem de variância total explicada, e um maior número de perfis e grupos.

### 3.4 Alguns fatores potenciadores de défice de sono

Para identificar fatores potenciadores de uma duração de sono adequada, ajustou-se um modelo de Regressão Logística (Hosmer *et al.*, 2013), considerando como variável resposta o número de horas de sono da criança reportadas pelos pais, dicotomizada nas seguintes categorias: 0 – para as crianças “com duração do sono reduzida” (menos de 10 horas), e 1 – para as crianças “com duração do sono adequada” (com pelo menos 10 horas). Foram consideradas as seguintes variáveis explicativas:

- Caraterizadoras da criança: idade, sexo, tem necessidades educativas especiais, já reprovou algum ano, tem problemas em adormecer;
- Meio socioeconómico da criança: tem irmãos, o encarregado de educação (EE) tem horário flexível, a entidade patronal do EE possibilita horários adaptados;
- Problemas do sono: índices das subescalas do questionário (exceto “duração do sono”).

O modelo obtido apresentou uma boa capacidade discriminativa e revelou estar bem ajustado, tendo-se obtido um valor de AUC igual a 0,74 e um valor p para o teste de Cessie-Von-Houvelingn igual a 0,93.

Nenhuma das variáveis do meio socioeconómico da criança se mostrou significativa. Das variáveis caraterizadoras da criança apenas o sexo se mostrou significativo e dos oito índices das subescalas de problemas do sono apenas a *Resistência em ir para a cama* e *Sonolência diurna* se mostraram significativas. (Tabela 2).

Podemos concluir que os fatores que potenciam um sono adequado:

- ser do sexo feminino: as crianças do sexo Feminino têm cerca de 3 vezes mais possibilidades de terem um sono adequado do que as crianças do sexo masculino;
- ter um menor índice de resistência em ir para a cama: por cada unidade de aumento do índice de resistência em ir para a cama, as possibilidades da criança ter um sono adequado diminuem 25%;
- Ter um índice de sonolência diurna: por cada unidade de aumento do índice de sonolência diurna, as possibilidades da criança ter um sono adequado diminuem 20%.

Note-se que estes factores, complementam bem alguns dos resultados da ACM (ver as componentes principais).

**Tabela 2** – Modelo de regressão logística para o reduzido número de horas de sono.

Variáveis	Coefficiente	Erro Padrão	p	OR	OR <sub>IC95%</sub>
Sexo					
Feminino (ref. Masculino)	1,14	0,55	0,040	3,12	(1,10; 9,89)
Resistência em ir para a cama	-0,29	0,15	0,047	0,75	(0,54; 0,97)
Sonolência diurna	-0,23	0,11	0,044	0,80	(0,62; 0,98)

## 4 Discussão de resultados

A análise realizada neste trabalho permitiu identificar padrões homogêneos nos comportamentos-problema do sono, bem como fatores mais potenciadores do défice de sono entre as crianças da escola em estudo.

Apesar de pouco reconhecidos pelos pais, os problemas comportamentais do sono revelados com os dados desta amostra são frequentes em algumas crianças e têm consequências em termos de sonolência diurna, com índices superiores aos descritos por Silva *et al.* (2013), pelo que não podem ser negligenciados.

A média do número de horas de sono reportada neste estudo é semelhante à descrita por Clemente (1997) e Silva *et al.* (2013), no grupo etário em estudo. A percentagem de pais que referiram que consideram que o seu “filho/filha tem algum problema com o sono” é inferior à mencionada por Silva *et al.* (2014): 9,1% na classe etária dos 6-8 anos e 11,7% na classe etária dos 9-10 anos. Neste estudo obteve-se um Índice de Perturbação do Sono (IPS) um pouco superior (47,3) ao observado no grupo de crianças com a mesma idade escolar (45,5) do estudo de Lopes *et al.* (2016). O IPS assume valores similares em ambos os sexos, ao contrário do mencionado no estudo de Silva *et al.* (2013) onde se verificou maior IPS nas raparigas. O contraste entre a baixa prevalência de problemas do sono reconhecidos pelos pais e os valores elevados do IPS é indicador de uma

desvalorização desta questão. Por este motivo, torna-se necessário promover maior sensibilização sobre o tema do sono e as características do sono adequado nesta faixa etária.

As conclusões deste trabalho referem-se às crianças da escola em estudo. No futuro, pretendemos replicar o desenvolvimento deste trabalho a outras escolas do concelho de Évora.

## Agradecimentos

Anabela Afonso, Gonçalo Jacinto e Paulo Infante são membros do CIMA, centro de investigação financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal, pelo projeto «UID/MAT/04674/2019 (CIMA)».

## Referências

- CLEMENTE, V. (1997). *Sono e Vigília em Crianças de Idade Escolar: Hábitos, Comportamentos e Problemas*, Dissertação de Mestrado não publicada, Faculdade de Psicologia e Ciências da Educação, Universidade de Coimbra, Coimbra.
- HOSMER, D., LEMESHOW, S. & STURDIVANT, R. (2013). *Applied Logistic Regression*, 3rd Edition, Wiley, New York.
- JOHNSON, R. A. & WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.
- LOPES, S., ALMEIDA, F., JACOB, S., FIGUEIREDO, M., VIEIRA, C. & CARVALHO, F. (2016). Diz-me como dormes: hábitos e problemas de sono em crianças portuguesas em idade pré-escolar e escolar, *Nascer e Crescer*, 25(4), 211-216.
- OWENS, J., SPIRITO, A. & MCGUINN, M. (2000). The Children's Sleep Habits Questionnaire (CSHQ): Psychometric Properties of a Survey Instrument for School-Aged Children, *Sleep*, 23:8.
- R CORE TEAM (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, <http://www.R-project.org/>
- SILVA, F. G. (2014). *Hábitos e Problemas do Sono das Crianças dos 2 aos 10 anos*, Dissertação de Doutoramento não publicada, Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Lisboa.
- SILVA, F. G., SILVA, C. R., BRAGA, L. B. & NETO, A. S. (2013). Hábitos e Problemas do Sono dos Seis aos Dez Anos: Estudo Populacional, *Acta Pediátrica Portuguesa*, 44(5), 196-202.
- SILVA, F. G., SILVA, C. R., BRAGA, L. B. & NETO, A. S. (2014). Portuguese Children's Sleep Habits Questionnaire – validation and cross-cultural comparison, *Jornal de Pediatria (Rio J.)*, 90(1), 78-84.

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

Ana Laura Carreiras<sup>1</sup> · Paulo Infante<sup>2</sup> · Anabela Afonso<sup>3</sup> · Maria Filomena Mendes<sup>4</sup>

**Resumo** As amostras complexas resultam da combinação de vários métodos de amostragem para a seleção de uma amostra representativa da população. Uma das estratégias mais usadas para corrigir as estimativas obtidas com base no pressuposto da amostra complexa é considerar que a amostra é proveniente de um esquema de amostragem aleatória simples usando os pesos normalizados corrigidos pelo efeito do desenho (*deff*). Neste trabalho, analisa-se o impacto de uma estimação incorreta do *deff* na significância das variáveis e no seu efeito em modelos de regressão logística.

**Palavras-chave:** Amostras Complexas, Efeito do Desenho, *Jackknife*, Pesos, Regressão Logística.

## 1 Introdução

Em muitos estudos é necessário usar delineamentos de amostragem complexos para selecionar uma amostra representativa da população. Uma amostra complexa possui pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais e ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008). Alguns autores ainda tratam este tipo de amostras sob a suposição de amostragem aleatória simples, ignorando o desenho de amostragem. Esta abordagem pode produzir incorreções, tanto para as estimativas, como para as respetivas variâncias, comprometendo os resultados e as conclusões da pesquisa (Osborne, 2011).

Contudo, por motivos de confidencialidade, muitas vezes não é facultada a

---

<sup>1</sup> Mestrado em Modelação Estatística e Análise de Dados, [analaura.carreiras@gmail.com](mailto:analaura.carreiras@gmail.com)

<sup>2</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>3</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>4</sup> CIDEHUS e DSOC/ECS, Universidade de Évora, [mmendes@uevora.pt](mailto:mmendes@uevora.pt)

informação relativa às variáveis associadas ao desenho de amostragem. Algumas vezes são fornecidos os pesos de replicação que contêm informação sobre o desenho e salvaguardam a confidencialidade (Sturgis, 2004). Mas nem todas as técnicas estatísticas que se pretendem utilizar e que estão implementadas em *software*, permitem a utilização destes pesos.

Existem várias estratégias para corrigir as estimativas obtidas, com base no pressuposto da amostra ser proveniente de um esquema de amostragem aleatória simples (Osborne, 2011). A mais usual é ponderar as observações pelos pesos normalizados corrigidos pelo efeito do desenho (*deff*). Deste modo, quando o valor do *deff* é superior a 1 então a amostra aleatória correspondente deverá ter uma dimensão menor, para se obter a mesma precisão nas estimativas da obtida com a amostra complexa.

O *deff* será descrito no ponto seguinte, mas é de referir que depende do número de observações e dos pesos de amostragem. Contudo, nem sempre é fornecido o valor do *deff* e este, usualmente, varia consoante a variável em estudo (Sturgis 2004). Baseando-nos no facto de não nos ser fornecido nenhum tipo de informação sobre o desenho da amostragem temos de estimar o *deff* para corrigir os pesos.

Assim, o objetivo deste trabalho prende-se com a avaliação das consequências de uma estimação incorreta do efeito do desenho na significância das variáveis e no seu efeito sobre a resposta em modelos de regressão logística. Para tal, foram utilizadas duas abordagens: 1) comparar os vários modelos ajustados quando se consideram as observações ponderadas pelos pesos normalizados corrigidos por vários valores para o *deff* e pelos pesos de replicação; 2) considerar o modelo ajustado com base nos pesos de replicação e comparar as estimativas obtidas para os parâmetros, e respetivos erros padrão, das variáveis significativas quando se consideram os pesos normalizados corrigidos com diferentes valores para o *deff*. Esta segunda abordagem permite verificar qual o impacto de uma incorreta estimação do *deff* na significância dos parâmetros associados às variáveis que sabemos serem significativas.

## **2 Efeito do desenho e pesos**

Ao contrário do que acontece numa amostra aleatória simples, numa amostra complexa nem todos os indivíduos da amostra representam o mesmo número de indivíduos da população, sendo essa informação dada por pesos. Existem vários tipos de pesos, como o peso do delineamento, o peso da pós estratificação, o peso da população, entre outros.

Em modelação estatística e análise dos dados é preciso incluir os pesos, sendo o investigador alertado para isso aquando da cedência dos dados. Mas nem sempre o investigador está familiarizado com os pesos disponibilizados, o que dificulta a identificação dos pesos a utilizar na análise que pretende realizar, bem como se estes devem ou não ser alvo de alguma transformação. Por outro lado, alguns

investigadores optam por não considerar os pesos, abordagem que pode produzir incorreções, tanto para as estimativas, como para as respectivas variâncias, enviesando os resultados e as conclusões (Osborne, 2011).

Para se tentar evitar estes erros, como já foi referido, aplica-se o efeito do desenho. Este efeito é uma medida que quantifica a perda ou o ganho de precisão na estimação devido ao uso de uma amostra complexa em vez de uma amostra aleatória simples e é definido pelo quociente entre a estimativa da variância determinada pelo plano amostral complexo e a estimativa da variância obtida por uma amostra aleatória simples do mesmo tamanho (Kish, 1965).

Seja  $\hat{\theta}$  um estimador para  $\theta$ , o *deff* é definido por:

$$deff(\hat{\theta}) = \frac{Var_{amostra\ complexa}(\hat{\theta})}{Var_{amostra\ aleatória\ simples}(\hat{\theta})}$$

Sem termos as variáveis associadas ao desenho do desenho nem os pesos de replicação, a solução é utilizar o efeito do desenho aplicando-o aos pesos normalizados. Seja  $w_i$  o peso de amostragem do indivíduo  $i$  na amostra, o peso normalizado  $w'_i$  e o peso corrigido  $w_i^*$  são definidos, respetivamente, por:

$$w'_i = \frac{w_i}{\sum_{j=1}^n w_j} n \quad e \quad w_i^* = \frac{w'_i}{\sqrt{deff}}$$

Esta correção leva à estimação de erros padrão muito próximos dos que se obteriam considerando as variáveis do desenho.

### 3 Material e métodos

#### 3.1 Dados amostrais

Os dados utilizados para esta análise de sensibilidade aos valores do *deff* pertencem ao Inquérito à Fecundidade 2013, realizado no âmbito de um protocolo entre a Fundação Francisco Manuel dos Santos e o Instituto Nacional de Estatística.

Selecionou-se um conjunto de 27 variáveis com características sociodemográficas dos indivíduos e, a partir destas, ajustaram-se modelos de regressão logística à variável dicotómica tem filhos vs. não tem filhos (Hosmer et al., 2013).

Por motivos de confidencialidade, as variáveis associadas ao desenho não foram facultadas, mas, em contrapartida, foram fornecidos os pesos de replicação que contêm toda a informação necessária para se obter uma estimativa do erro padrão do estimador do parâmetro pelo estimador de variância do tipo *Jackknife* (Lohr, 2010).

## 3.2 Regressão logística

O ajustamento de modelos de regressão logística foi efetuado com recurso aos pacotes *survey*, *rms*, *mfp*, *EPI* e *epiR* do programa *R Project* (R Core Team, 2012). Nesta etapa consideraram-se os pesos de replicação (pacote *survey*) e os pesos normalizados corrigidos para o efeito do desenho (restantes pacotes).

Para ajustar os modelos seguimos a seguinte estratégia (Hosmer & Lemeshow, 2013): (1) para o modelo inicial foram selecionadas todas as covariáveis que se revelaram significativas na fase univariada (valor  $p < 0,20$ ); (2) a partir deste modelo foram eliminadas sucessivamente, e por ordem decrescente dos valores  $p$ , todas as covariáveis não significativas (valor  $p > 0,05$ ); (3) verificámos se alguma(s) das covariáveis que não foram incluídas no modelo inicial se mostra(m) agora significativa(s) na presença das que estão no modelo, caso em que foram adicionadas ao modelo; (4) avaliámos a forma funcional das covariáveis contínuas, através do alisamento em diagrama de dispersão e ajustamento de um modelo aditivo generalizado (GAM), sendo aplicado o método dos polinómios fracionários em casos de não linearidade; (5) foram testadas as interações que faziam sentido no contexto do estudo (valor  $p < 0,05$ ); (6) foi feita uma análise de resíduos por padrões para pesquisa de observações influentes ou *outliers*, através dos resíduos da desviância, distância de Cook e estatísticas DFBETAS.

A significância das covariáveis e das interações foi testada recorrendo ao teste de Wald modificado (Hosmer e Lemeshow, 2013). A adequabilidade do ajustamento foi feita recorrendo aos testes de bondade de ajustamento de Hosmer e Lemeshow e de Cessie-van Houwelingen (ou teste de Wald no caso de se usarem os pesos de replicação) e a capacidade discriminativa do modelo avaliada pelo valor da AUC da curva ROC.

## 4 Resultados

### 4.1 Abordagem 1: comparação dos modelos

A variável resposta dicotómica considerada, tem filhos *vs.* não tem filhos, foi avaliada para todos os indivíduos da amostra e a categoria de referência (não tem filhos) representa 33,2% da dimensão da amostra ( $n = 7624$ ).

Os modelos de regressão logística ajustados apresentam um valor  $p$  do teste de Hosmer e Lemeshow (H&L) semelhante, um valor da AUC idêntico, mas o valor de  $R^2$  (Nagelkerque) diminui com o aumento do valor do *deff*, assim como o valor  $p$  do teste Cessie Van Howelingen (Tabela 1).

**Tabela 1** – Medidas da bondade do ajustamento e capacidade discriminativa dos modelos ajustados com base nos diferentes valores de *deff* e nos pesos de replicação.

Modelos	H&L (p)	$R^2$	AUC	Valor p
<i>deff</i> = 0,5	0,05	0,88	0,98	0,75*
<i>deff</i> = 1	0,06	0,84	0,98	0,55*
<i>deff</i> = 1,5	0,14	0,81	0,98	0,18*
<i>deff</i> = 2	0,06	0,81	0,98	0,13*
<i>deff</i> = 2,5	0,08	0,79	0,98	0,05*
<i>deff</i> = 3	0,08	0,79	0,98	0,05*
Pesos de replicação	0,11	0,72	0,98	0,93**

\* Teste de Cessie van Howelingen, \*\* Teste de Wald

Com a diminuição do valor do *deff* aumenta o número de variáveis significativas e também o número de interações (Tabela 2). Os modelos ajustados considerando valores do *deff* = 2, 2,5 e 3 têm o mesmo número de variáveis significativas que o modelo ajustado com base nos pesos de replicação, mas diferente número de interações significativas. O nível de significância dos coeficientes varia com o valor considerado para o *deff*. Para alguns coeficientes o nível de significância mantém-se igual em todos os *deff* (por ex., variável A e C) e há também situações em que o nível de significância diminui com o valor do *deff* (por ex., variável B). Este resultado deve-se ao facto de quanto menor o valor do *deff* maior a dimensão da amostra aleatória e, por conseguinte, menor o erro-padrão das estimativas.

**Tabela 2** – Significância das variáveis comuns, número de variáveis e de interações significativas nos modelos ajustados para cada *deff* e pesos de replicação.

Variável	A	B	C	N.º de variáveis significativas a 5%	N.º de interações significativas a 1%
<i>deff</i> = 0,5	***	*	***	13	14
<i>deff</i> = 1	***	*	***	11	10
<i>deff</i> = 1,5	***	†	***	9	8
<i>deff</i> = 2	***	***	***	8	7
<i>deff</i> = 2,5	***	**	***	8	6
<i>deff</i> = 3	***	**	***	8	6
Pesos de replicação	***	**	***	8	5

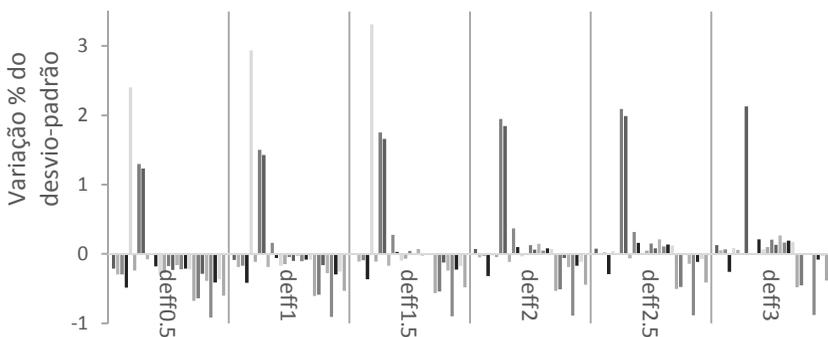
Significativa a: † 0,10; \* 0,05; \*\* 0,01; \*\*\* <0,001.

A Tabela 2 foi elaborada de forma a permitir ter uma noção da complexidade dos modelos sem torná-la muito densa, dado o número de variáveis e interações envolvidas em alguns dos modelos. Com o objetivo de obter modelos mais parcimoniosos optou-se por considerar um nível de significância inferior para inclusão das interações no modelo (1%). As comparações foram apenas realizadas

para as variáveis que não estão incluídas nas interações, pois os coeficientes e significância das restantes variáveis são influenciados pelas interações em que as mesmas estão envolvidas. As 3 variáveis apresentadas na Tabela 2 são as que não intervêm em interações no modelo ajustado com os pesos de replicação, mas que com o aumento do valor do *deff* e consequente aumento de interações passam também a estar incluídas em interações. As variáveis significativas para o modelo ajustado com os pesos de replicação são 8. Todos os outros modelos incluem essas 8 variáveis. Os modelos ajustados com os pesos normalizados corrigidos pelo *deff* = 0,5, *deff* = 1 e *deff* = 1,5 são os que incluem mais variáveis significativas. As interações significativas para o modelo ajustado com os pesos de replicação foram 5. Todos os outros modelos ajustados com os pesos normalizados corrigidos pela estimação do *deff* incluem essas 5 interações e mais as reportadas na Tabela 2.

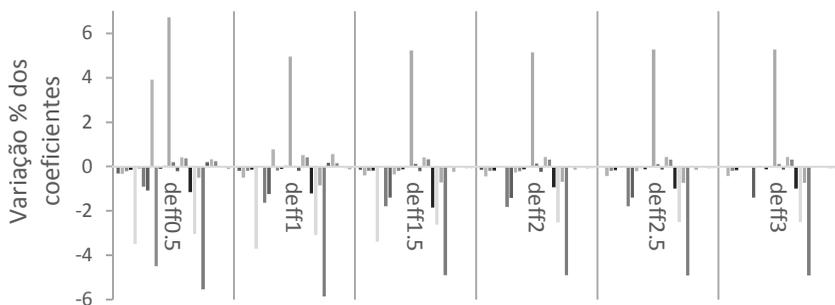
O efeito nas estimativas dos parâmetros dos modelos de regressão logística foi medido através de uma comparação da variação percentual entre desvio-padrão, coeficientes e *odd ratios* (OR) estimados com base nos modelos com os pesos normalizados corrigidos pelo *deff*, e estimado com base nos pesos de replicação (Figura 1 e Figura 2). Esta análise apenas é possível realizar entre as variáveis comuns às do modelo ajustado com os pesos de replicação. O número de barras é idêntico em todos os valores de *deff* e representa as categorias das variáveis, contemplando também as interações que são idênticas entre os modelos.

Comparando a variação percentual entre o desvio-padrão estimado com base nos modelos com os pesos normalizados corrigidos pelo *deff* e o estimado com base nos pesos de replicação, a sub ou sobrestimação do desvio-padrão é maior nos valores de *deff*s mais reduzidos (Figura 1). Na maior parte dos casos ocorre uma subestimação para *deff*s mais baixos e quando aumenta o valor do *deff* algumas estimativas do desvio-padrão reduziram, o que em alguns casos originou uma sobrestimação dos mesmos.



**Figura 1** – Variação percentual do desvio-padrão das estimativas dos parâmetros dos modelos ajustados considerando os pesos normalizados corrigidos pelo *deff* e a do modelo ajustado com os pesos de replicação.

Para todos os valores do  $deff$  considerados, os coeficientes dos modelos ajustados parecem mostrar uma subestimação relativamente aos coeficientes do modelo ajustado com os pesos de replicação (Figura 2). Consequentemente, ocorre maioritariamente uma subestimação dos ORs, pois estes obtêm-se recorrendo à exponencial das estimativas dos coeficientes.



**Figura 2** – Variação percentual das estimativas dos coeficientes dos modelos ajustados considerando os pesos normalizados corrigidos pelo  $deff$  e as do modelo ajustado com os pesos de replicação.

## 4.2 Abordagem 2: modelo base ajustado com os pesos de replicação

Nesta abordagem consideram-se apenas as variáveis no modelo ajustado a partir dos pesos de replicação e estimam-se os coeficientes deste modelo considerando os pesos normalizados corrigidos pelo  $deff$ .

Todos os modelos parecem estar ajustados corretamente aos dados, todos apresentam valores de AUC semelhantes e, tal como na abordagem anterior, o valor do coeficiente  $R^2$  diminui com o aumento do valor do  $deff$  (Tabela 3).

**Tabela 3** – Medidas da bondade do ajustamento e capacidade discriminativa dos modelos ajustados com base nos diferentes valores de  $deff$  e nos pesos de replicação.

Modelos	Hosmer	$R^2$	AUC	Valor p
$deff = 0,5$	0,45	0,64	0,93	0,05*
$deff = 1$	0,45	0,59	0,92	0,05*
$deff = 1,5$	0,45	0,57	0,91	0,05*
$deff = 2$	0,45	0,55	0,91	0,05*
$deff = 2,5$	0,45	0,54	0,91	0,05*
$deff = 3$	0,45	0,53	0,91	0,05*
Pesos de replicação	0,45	0,46	0,91	0,18**

\* Teste de Cessie van Howelingen, \*\* Teste de Wald

O nível de significância para algumas variáveis pode ser bastante diferente do obtido no modelo dos pesos de replicação sendo que diferem mais quanto menor o *deff* (Tabela 4). O modelo que mostrou menos diferenças em relação ao modelo ajustado com os pesos de replicação foi o obtido quando *deff* = 2 (Tabela 4). A significância das variáveis é comparável apenas para as variáveis que não estão incluídas em interações.

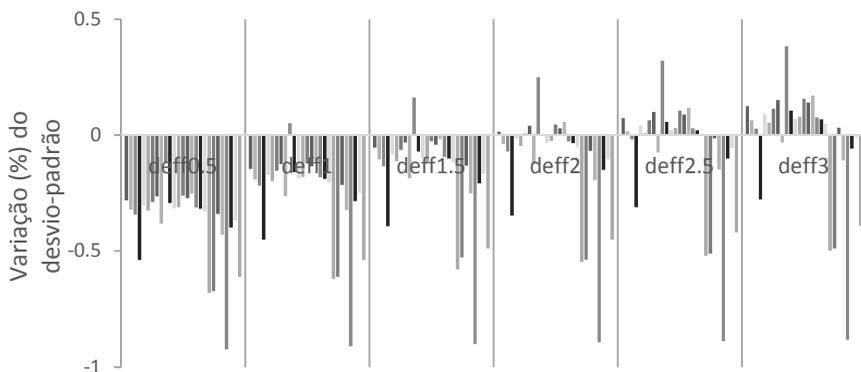
**Tabela 4** – Significância das variáveis, das suas interações e categorias nos modelos ajustados.

Variável ou interação	Categoria	<i>deff</i> =0,5	<i>deff</i> =1	<i>deff</i> =1,5	<i>deff</i> =2	<i>deff</i> =2,5	<i>deff</i> =3	Pesos de replicação
1	1	***	***	***	***	***	***	***
2	1	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	2	***	***	***	***	***	***	***
	3	*	.	.	n.s.	n.s.	n.s.	n.s.
3	1	***	***	***	***	**	**	**
4	1	***	***	***	***	***	***	***
5	1	***	***	***	**	**	**	**
	2	***	***	***	***	***	***	***
6	1	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
7	1	***	***	***	***	***	***	***
8		***	***	***	***	***	***	**
2*8	1	***	***	***	***	***	***	***
	2	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	3	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
7*8		***	***	***	***	***	***	***
2*5	1	**	*	*	*	*	.	*
	2	***	***	***	***	***	***	***
	3	.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	4	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	5	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	6	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
2*6	1	***	***	***	***	***	***	***
	2	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	3	**	*	*	*	.	.	n.s.
2*7	1	***	***	***	***	***	***	***
	2	***	***	***	***	***	***	***
	3	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Significativa : . - 0,10; \* - 0,05; \*\* - 0,01; \*\*\* - <0,001.

Verificamos que para valores de *deff*s mais baixos ocorre uma subestimação dos desvios-padrão e para valores de *deff*s mais elevados ocorre maioritariamente uma sobrestimação, embora em algumas categorias das interações ocorra uma

subestimação do desvio-padrão (Figura 3). De referir que a variável 8 é contínua e por isso a célula da categoria se encontra em branco.



**Figura 3** – Variação percentual do desvio-padrão das estimativas dos parâmetros dos modelos ajustados considerando os pesos normalizados corrigidos pelo *deff* e do modelo ajustado com os pesos de replicação.

As estimativas pontuais dos parâmetros, bem como dos ORs, não são afetadas pelo valor dos pesos considerados, uma vez que os modelos ajustados são idênticos.

Refira-se, por fim, que para todos os modelos o agrupamento das categorias das variáveis foi sempre idêntico, concluindo-se que uma incorreta estimação do valor do *deff* não trará nenhuma influência ao agrupamento das categorias. Tal também se concluiu na primeira abordagem.

## 5 Considerações Finais

Neste trabalho estudou-se o efeito de vários valores de *deff* no ajustamento dos modelos de regressão logística e na significância das variáveis e suas interações, com vista a alertar para as consequências de uma incorreta estimação do valor do *deff*, bem como para a sua omissão no processo de inferência.

Os modelos resultantes podem diferir bastante consoante o valor do *deff* considerado, quer nas estimativas dos coeficientes, quer nos seus desvios-padrão. Contudo, o número de variáveis a incluir no modelo aumenta com a diminuição do valor do *deff*, havendo uma tendência para a subestimação do desvio-padrão com valores de *deff* mais reduzidos. As estimativas dos coeficientes e dos OR's são enviesadas sempre que são incluídas no modelo novas variáveis por consequência da diminuição do *deff*. Os intervalos de confiança dos ORs parecem não ser

sensíveis a uma incorreta estimação do valor do *deff*, principalmente se a estimação deste for abaixo do valor de *deff* correto. Quando se ajusta o modelo obtido com os pesos de replicação, as estimativas obtidas para os coeficientes nos modelos ajustados com pesos normalizados corrigidos pelo *deff* são idênticas.

Conclui-se, assim, que uma estimação incorreta do valor do *deff* afeta a significância das variáveis e das interações entre as variáveis, mas não parece afetar a capacidade discriminativa nem a bondade do ajustamento.

A terminar, refira-se que o investigador poderá estar perante um modelo que se parece ajustar aos dados e com uma boa capacidade discriminativa, mas que, no entanto, não explica convenientemente o evento em estudo. A discussão de resultados e a sua teorização sobre o fenómeno em estudo pode ser condicionada por uma não inclusão do efeito do delineamento (*deff*) ou por uma incorreta estimação desse efeito.

## **Agradecimentos**

Anabela Afonso e Paulo Infante são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2019), financiado pela Fundação para a Ciência e Tecnologia (FCT).

## **Referências**

- KISH, L. (1965). *Survey sampling*, John Wiley & Sons, New York.
- LAVRAKAS, P. J. (2008). *Encyclopedia of survey research methods*, SAGE Publications, Thousand Oaks.
- LORH, S. L. (2010). *Sampling: Design and Analysis*, Second Edition, Michelle Julet, Boston.
- OSBORNE, J. W. (2011). Best practices in using large, complex samples: the importance of using appropriate weights and design effect compensation, *Practical Assessment, Research & Evaluation*, 16, 1-7.
- STURGIS, P. (2004). Analysing complex survey data: clustering, stratification and weights, *Social Research Update*, 43, Autumn Issue.
- HOSMER, D., LEMESHOW, S. & STURDIVANT, R. (2013). *Applied logistic regression*, 3rd Edition, Wiley, New York.

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

Maria João Inácio<sup>1</sup> · Paulo Infante<sup>2</sup> · Fernanda Otilia Figueiredo<sup>3</sup>

**Resumo** Várias cartas de controlo têm sido adaptadas para uma utilização mais eficiente na área da Saúde, incorporando na análise, um modelo que estima, para cada paciente, o risco de ocorrência do acontecimento que se pretende monitorizar. A adequabilidade do modelo é muito importante no delineamento destas cartas na medida em que afeta o desempenho das mesmas. Neste trabalho, estudámos o efeito do erro de estimação e predição no desempenho das cartas de controlo CUSUM adaptadas para risco ajustado.

**Palavras-chave:** ARL, Erro de Estimação dos Parâmetros do Modelo, Monitorização de Processos de Saúde, RACUSUM.

### 1 Introdução

As cartas de controlo foram desenvolvidas por Walter Shewhart em 1920 e posteriormente popularizadas por Deming, e ao longo dos anos têm sido feitas várias adaptações às cartas existentes e desenvolvidas novas metodologias, por forma a tornar esta ferramenta mais potente e/ou mais adequada em diversas situações (Montgomery, 2013).

Apesar de criadas inicialmente com o intuito de serem utilizadas em processos industriais, as técnicas de controlo estatístico do processo têm-se revelado muito úteis em todas as áreas.

No início dos anos 90, surge a ideia de aplicar cartas de controlo na área da saúde, para monitorizar o desempenho de instituições de saúde e dos seus profissionais. No entanto, ao contrário dos processos industriais, onde as matérias primas são relativamente homogéneas, em aplicações médicas os indivíduos (pacientes) frequentemente variam muito em termos do risco de ocorrência do acontecimento que se pretende monitorizar, e, portanto, o uso das cartas de controlo

---

<sup>1</sup> Instituto Politécnico de Tomar e CIMA/IIFA, [mjantunes@ipt.pt](mailto:mjantunes@ipt.pt)

<sup>2</sup> Universidade de Évora, DMAT/ECT e CIMA/IIFA, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>3</sup> Faculdade de Economia da Universidade do Porto e CEAUL, [otilia@fep.up.pt](mailto:otilia@fep.up.pt)

tradicionais, regra geral, não é adequado para monitorizar este tipo de processos. Neste contexto, e com intuito de controlar esta fonte de variação, evitando assim erros indesejados, foram desenvolvidas as cartas de controlo com risco ajustado.

Nas cartas de controlo com risco ajustado são incorporados modelos de regressão, usualmente designados por modelos de risco, que estimam, para cada paciente, o risco de ocorrência do evento que se está a monitorizar. Tipicamente, são ajustados modelos de regressão logística baseados em scores conhecidos, tais como o *Parsonnet Score*, o *APACHE* e o *Euroscore* (Cook *et al.*, 2008; Woodall *et al.*, 2015). Os resultados observados são comparados com o que se poderia esperar com base no modelo de risco. O objetivo é detetar períodos de tempo nos quais os resultados dos cuidados de saúde não correspondem às expectativas que se tinham no momento da admissão. Pontos fora dos limites de controlo significam que o processo não se ajusta às expectativas, dando assim indicação de alterações na qualidade dos serviços prestados e consequentemente alertam para a necessidade de se proceder a alguma intervenção. Se por um lado, a deteção antecipada de um deterioramento na qualidade dos serviços prestados permite a eliminação de problemas, por outro, a deteção de uma melhoria permite a implementação e o uso de melhores práticas clínicas. Assim, a monitorização destes processos permite uma melhoria na qualidade dos serviços prestados, quer por parte das instituições quer pelos próprios profissionais de saúde.

Ao longo dos últimos anos têm sido propostas várias cartas de controlo adaptadas para risco ajustado. Atendendo a que, em contexto médico, os eventos binários são os mais frequentes, são as cartas por atributos que mais se utilizam, e se têm desenvolvido. Entre as mais populares, salientamos as cartas *Risk-Adjusted P-chart* (Alemi *et al.*, 1996), *Variable Life-Adjust Display* ou *Cumulative Risk-Adjusted Mortality* (Lovegrove *et al.*, 1997 e Poloniecki *et al.*, 1998), *Risk-Adjusted Exponentially Weighted Moving Average* (Cook *et al.*, 2011), *Risk-adjusted sequential probability ratio tests* (Spiegelhalter *et al.*, 2003) e *Risk-Adjusted CUMulative SUM control chart* (Steiner *et al.*, 2000). Para uma visão geral sobre estas cartas podem ser consultados, por exemplo, os artigos de Grigg & Farewell (2004), Cook *et al.* (2008) e Woodall *et al.* (2015).

Com exceção da carta *Risk-Adjusted P-chart*, que designaremos por carta RAP, todas as restantes são atualizadas sempre que dispomos de informação sobre um paciente. Dada a natureza dos processos a monitorizar é extremamente importante que, caso ocorram alterações no processo, estas sejam detetadas rapidamente, e, a menos que disponhamos de muitos casos, o que não é usual nestes processos, as cartas indivíduo a indivíduo são preferíveis à carta RAP, na medida em que se aumenta a hipótese de detetar mais rapidamente alterações nos cuidados de saúde.

De entre as várias cartas adaptadas para risco ajustado, a mais popular (a que tem recebido mais atenção na literatura, sendo também a mais aplicada) é a carta *Risk-Adjusted CUMulative SUM*, RACUSUM. Novick *et al.* (2006) usaram cartas CUSUM tradicionais e RACUSUM para monitorizar os resultados de cirurgia de

revascularização miocárdia, e compararam os seus resultados, tendo concluído que a RACUSUM apresenta vantagens relativamente à carta CUSUM tradicional, na medida em que esta não emite sinal de deterioração do processo quando o risco pré-operatório é elevado. Chen *et al.* (2011) aplicaram a carta RACUSUM para monitorizar a taxa de mortalidade de pacientes com paragem cardíaca assistidos pelas equipas de emergência médica. Para outras aplicações podem ver-se, por exemplo, os artigos de Harris *et al.* (2005), Axelrod *et al.* (2006), Moore *et al.* (2007) e Bottle & Aylin (2008).

O ajuste do modelo de risco é uma questão importante no delineamento das cartas de controlo com risco ajustado, na medida em que afeta o desempenho das mesmas.

Cook *et al.* (2008) apresentam algumas considerações relacionadas com o modelo de risco, referindo, por exemplo, que este deve ter uma boa calibração e capacidade de discriminação. Calibração é a capacidade do modelo estimar o risco real do acontecimento que estamos a monitorizar. Um modelo com uma fraca calibração pode alterar os resultados. Por exemplo, admita-se que o acontecimento de interesse é a morte do paciente após uma intervenção cirúrgica. Se o modelo sobrestimar o risco de morte de um paciente, podemos (erradamente) concluir que a qualidade dos serviços prestados melhorou; se, por outro lado, o modelo subestimar o risco de morte do paciente somos levados a concluir (erradamente) que existe uma deterioração do processo. Discriminação é a habilidade do modelo em atribuir uma probabilidade de morte mais elevada a um paciente que morre do que a um que sobrevive. De acordo com estes autores, se a capacidade de discriminação do modelo for baixa então o modelo é pouco melhor que o acaso, independentemente da sua calibração. Assim, se o modelo não for adequado, isto é, se não estimar de forma fidedigna o risco do evento que estamos a monitorizar, toda a análise pode estar comprometida.

De acordo com a nossa pesquisa, muito pouco tem sido feito para avaliar o impacto do uso de um modelo incorreto. Steiner *et al.* (2001) avaliaram o efeito no desempenho das cartas RACUSUM, considerando, na fase de monitorização, duas distribuições extremas de pacientes: uma constituída apenas pelos pacientes de menor risco e outra constituída apenas pelos pacientes de maior risco. Assumindo que o modelo de risco inicialmente obtido era o correto, avaliaram as alterações ocorridas no *Average Run Length* (ARL), para vários valores de *odds ratio* (OR), nas duas populações consideradas, tendo concluído que estes foram claramente afetados. Tian *et al.* (2014) seguindo uma abordagem semelhante, mas utilizando apenas os dados utilizados para definir o modelo de risco, estudam as alterações que ocorrem no ARL sob controlo para 5 subpopulações diferentes. No seu estudo mostram que o ARL sob controlo varia significativamente nas 5 subpopulações, e concluem ainda que o ARL sob controlo diminui à medida que a média dos *Parsonnet scores* aumenta. Estes dados foram também trabalhados por Jones & Steiner (2012), mas utilizando uma abordagem diferente. Estes avaliaram as

alterações decorrentes no ARL sob e fora de controlo, este último assumindo uma duplicação do OR, utilizando processos de reamostragem, considerando quer amostras de dimensão diferentes, quer diferentes valores do ARL sob controlo. Avaliaram ainda o desempenho da carta RACUSUM provocando alterações no termo independente do modelo por forma a obter diferentes taxas de mortalidade. Também estes autores concluíram que o erro de estimação tem um impacto significativo no desempenho destas cartas. Loke & Gan (2012), utilizando também uma base de dados reais, investigaram a sensibilidade das cartas RACUSUM para alterações na distribuição de risco, usando distribuições beta simuladas com diferentes parâmetros. Mostraram que o ARL sob controlo poderia ser até 13% menor ou 31% maior do que o especificado.

Neste trabalho, utilizando uma base de dados real, e uma abordagem diferente das apresentadas anteriormente, avaliamos o efeito do erro de estimação dos parâmetros e do erro de predição no desempenho das cartas de controlo RACUSUM.

## 2 Cartas de Controlo CUSUM com risco ajustado (RACUSUM)

A carta de controlo de somas acumuladas binária adaptada para risco ajustado, vulgarmente designada por carta RACUSUM, foi desenvolvida por Steiner *et al.* (2000), motivado pela necessidade de monitorizar apropriadamente o desempenho cirúrgico na presença de riscos diferentes de paciente para paciente.

Uma vez que cada paciente tem um nível de risco base diferente, na carta de controlo RACUSUM as hipóteses em teste são formuladas em termos do OR, assumindo a forma:

$$H_0: OR = R_0 \quad vs \quad H_1: OR = R_1.$$

Tipicamente define-se  $R_0 = 1$ , que reflete o atual desempenho médico, i.e., o desempenho definido pelo modelo de risco estimado. Na hipótese alternativa, estabelecemos  $R_1 > R_0$  para detetar um deterioramento na qualidade dos serviços prestados e  $R_1 < R_0$  para detetar uma melhoria dos mesmos. Para  $R_0 = 1$ , é comum definir-se  $R_1 = 2$  ou  $R_1 = 0,5$ , consoante se trate, respetivamente, de uma carta RACUSUM unilateral superior ou inferior.

Representando-se por  $p_j$  a probabilidade do evento ocorrer para o paciente  $j$ , estimada pelo modelo de risco, e por  $y_j$  o resultado observado para o paciente  $j$ , definido por:

$$y_j = \begin{cases} 1 & , \text{ se o evento ocorre para o paciente } j \\ 0 & , \text{ caso contrário} \end{cases}.$$

A estatística RACUSUM unilateral superior é definida por

$$S_j^+ = \max(0, S_{j-1}^+ + w_j) \quad \text{com } S_0^+ = 0,$$

e a estatística RACUSUM unilateral inferior por

$$S_j^- = \min(0, S_{j-1}^- - w_j) \quad \text{com } S_0^- = 0,$$

onde  $w_j$  são os pesos atribuídos aos pacientes, determinados com base no logaritmo da razão de verosimilhanças, definidos por

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

$$= \ln \left[ \left( \frac{1 - p_j + R_0 p_j}{1 - p_j + R_1 p_j} \right) \left( \frac{R_1}{R_0} \right)^{y_j} \right]; \quad y_j = 0, 1.$$

Estes pesos terão um valor absoluto maior no caso de termos um resultado não esperado, i.e., se o evento não ocorre para um paciente com elevada probabilidade de ocorrência *à priori* ou se ocorre para um paciente com baixa probabilidade de ocorrência *à priori*. No caso do resultado ser o esperado estes pesos terão, em módulo, um valor baixo. Assim, a estatística RACUSUM unilateral superior sofre um maior aumento quando o acontecimento ocorre para um indivíduo com baixa probabilidade de risco. Por outro lado, a estatística é limitada a valores não negativos, não permitindo acumular muita informação quando ocorre uma sucessão de casos bem sucedidos, sendo assim possível detetar mais rapidamente uma alteração de deterioração na qualidade dos serviços prestados, se esta vier a ocorrer. De forma semelhante, a estatística RACUSUM unilateral inferior sofre um maior decréscimo quando o acontecimento não ocorre para um indivíduo com alta probabilidade de risco. Neste caso, é limitada a valores não positivos, por forma a não acumular muita informação quando ocorre uma sucessão de casos mal sucedidos, detetando mais rapidamente uma alteração de melhoria na qualidade dos serviços prestados, se esta vier a ocorrer.

Moustakides (1986) mostrou que a escolha ótima dos pesos  $w_j$  para a carta CUSUM tradicional é baseada no logaritmo da razão de verosimilhanças. É frequente, na literatura, ver-se justificada a escolha destes pesos na carta RACUSUM com base no mesmo resultado, afirmando-se que desta forma a carta tem um desempenho ótimo, no sentido de que de entre todas as cartas com o mesmo ARL sob controlo, a escolha do logaritmo da razão de verosimilhanças para os pesos dá o menor ARL fora de controlo. Woodall *et al.* (2015) chamam a atenção de que este resultado não está provado, na medida em que os resultados de otimização são baseados na hipótese de uma sequência de observações independentes e identicamente distribuídas, e, no nosso caso, as observações são assumidas como independentes, mas não identicamente distribuídas, já que os pacientes têm diferentes fatores de risco.

A carta RACUSUM emite sinal quando  $S_j^+ > LSC$  ou  $S_j^- < LIC$ , consoante se trate respetivamente da carta unilateral superior ou inferior. Os limites de controlo

$LIC$  e  $LSC$  são definidos de modo a que a carta tenha um determinado ARL sob controlo pré-definido. É comum, neste tipo de aplicações, considerarem-se valores bastantes elevados para o ARL sob controlo, como no caso que apresentamos neste trabalho em que a carta é desenhada para ter um ARL sob controlo na ordem dos 7000 pacientes. Vários métodos podem ser usados para determinar os limites de controlo, incluindo o uso de cadeias de Markov ou métodos de simulação de Monte Carlo.

A técnica mais utilizada para aplicação da carta RACUSUM assenta na construção das duas cartas unilaterais, atrás definidas, permitindo desta forma detetar quer melhorias quer deterioramentos na qualidade dos serviços prestados. Atendendo a que  $S_j^+$  é sempre não-negativo e  $S_j^-$  é sempre não-positivo, podemos representá-las sobre o mesmo eixo horizontal.

Grigg & Farewell (2004) referem a utilidade da fórmula definida por Khan (1984) para obter uma aproximação do ARL da carta RACUSUM bilateral:

$$\frac{1}{ARL} \cong \frac{1}{ARL_S} + \frac{1}{ARL_I}$$

onde  $ARL_S$  e  $ARL_I$  representam, respetivamente o ARL da carta RACUSUM unilateral superior e inferior. Megahed *et al.* (2011) apresentam condições sob as quais esta relação é exata.

### 3 Metodologia e Resultados

Para avaliar o impacto do uso de um modelo incorreto no desempenho da carta RACUSUM vamos utilizar uma base de dados real. Estes dados já foram trabalhados, no âmbito de cartas de controlo com risco ajustados, por Steiner *et al.* (2000, 2001), Grigg *et al.* (2003), Grigg & Farewell (2004), Gombay (2011), Jones & Steiner (2012), Tian *et al.* (2014) e Woodall (2015). Esta base de dados contém informação relativa a todas as cirurgias, realizadas num centro de cirurgia cardíaca do Reino Unido no período de 1992 a 1998, com informação relativa ao paciente e ao tipo de intervenção realizada. De acordo com Steiner *et al.* (2000), durante este período não foi realizado nenhum processo de monitorização. O acontecimento de interesse é a morte do paciente nos 30 dias após a cirurgia. No período em estudo, foram realizadas um total de 6994 cirurgias, tendo-se observado que 461 pacientes faleceram nos primeiros 30 dias após a cirurgia, o que resulta numa taxa bruta de mortalidade de aproximadamente 6,6%.

Nos dois primeiros anos (1992 a 1993) foram realizadas 2218 cirurgias tendo-se verificado no *follow-up* 143 mortes, o que corresponde a uma taxa bruta de mortalidade de aproximadamente 6,4%. Com base nestes dados estimámos a probabilidade de morte nos primeiros 30 dias após a cirurgia, usando um modelo de

regressão logística. O método de seleção usado foi o *Backward* e o modelo obtido é definido por:

$$\text{logit}(p_j) = -3,68 + 0,077X_j$$

onde  $X_j$  representa o *Parsonnet Score* para o paciente  $j$ . O facto de apenas uma variável se ter revelado estatisticamente significativa não nos surpreendeu, na medida em que o *Parsonnet Score* é um índice definido com base em características individuais, tais como idade, género, diabetes e hipertensão, variáveis estas que também estavam disponíveis na nossa base de dados.

Admitimos que o processo de monitorização começou em 1994 e estudámos unicamente a carta RACUSUM unilateral superior. Provocando alterações quer no parâmetro do modelo quer na probabilidade estimada de morte, estudámos a sensibilidade destas cartas, aferindo o seu impacto no desempenho das mesmas. Neste trabalho, à semelhança do que é usual em cartas de controlo com risco ajustado, o desempenho das cartas é medido em termos do ARL, que representa o número médio de pacientes até que um sinal seja emitido. Idealmente, o ARL deve ser grande quando o processo está sob controlo e pequeno quando ocorre uma alteração significativa no processo.

A análise foi realizada no *software* R e o valor do ARL determinado via Cadeias de Markov, utilizando o *script* disponibilizado para o efeito pelo Professor Stefan Steiner na sua página pessoal (Steiner, 2001).

### 3.1 Erro de estimação dos parâmetros do modelo

Numa primeira fase avaliámos o efeito do erro de estimação dos parâmetros do modelo. Considerámos que a carta foi desenhada para detetar uma duplicação do OR de morte. O limite superior de controlo considerado para a carta foi de 4,5, o que se traduz num ARL sob controlo de aproximadamente 7000 pacientes. Atendendo a que por mês se efetuam aproximadamente 80 cirurgias, em média temos um falso alarme a cada 87,5 meses ( $\cong 7,2$  anos). Os ARL's fora de controlo, assumindo que o verdadeiro valor do OR é 1,5, 2, 3 e 10, são respetivamente, 533 ( $\cong 6,7$  meses), 202 ( $\cong 2,5$  meses), 91 ( $\cong 1,1$  meses) e 26 ( $\cong 10$  dias).

Provocámos alterações no parâmetro  $\beta_1$  (coeficiente da variável *Parsonnet Score*) que variaram entre -50% e 100%, considerando incrementos de 5%. Para cada uma das alterações provocadas, calculámos o ARL sob e fora de controlo, este último considerando que o verdadeiro OR era 1,5, 2, 3 e 10. Os resultados são apresentados na forma de um gráfico, onde no eixo das abcissas colocámos a alteração provocada no parâmetro e no eixo das ordenadas a correspondente alteração no ARL, ambas expressas em termos percentuais. Os resultados obtidos são apresentados nas Figuras 1 e 2.

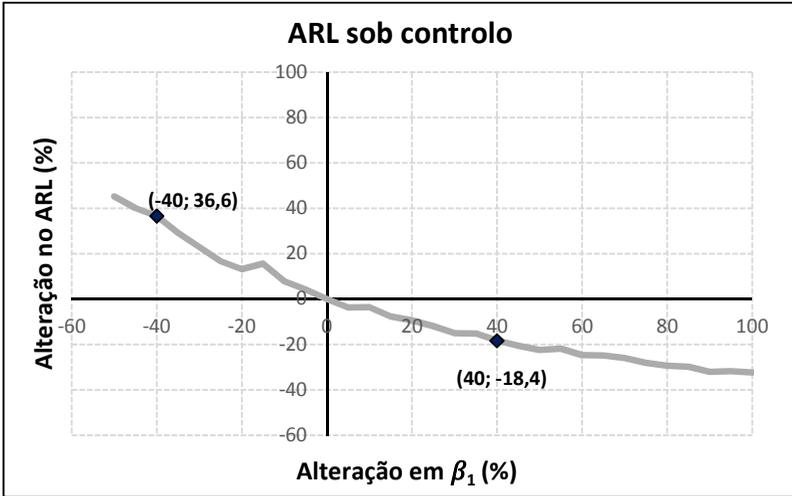


Figura 1 - Variação no ARL sob controle devido a alterações no parâmetro do modelo.

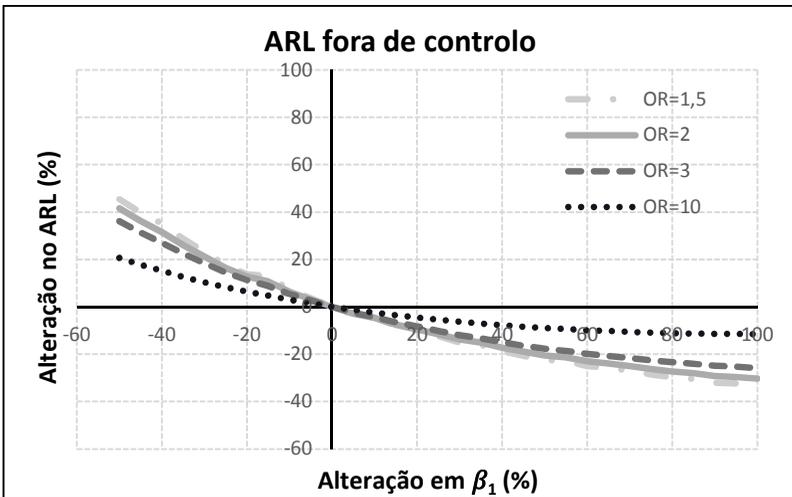


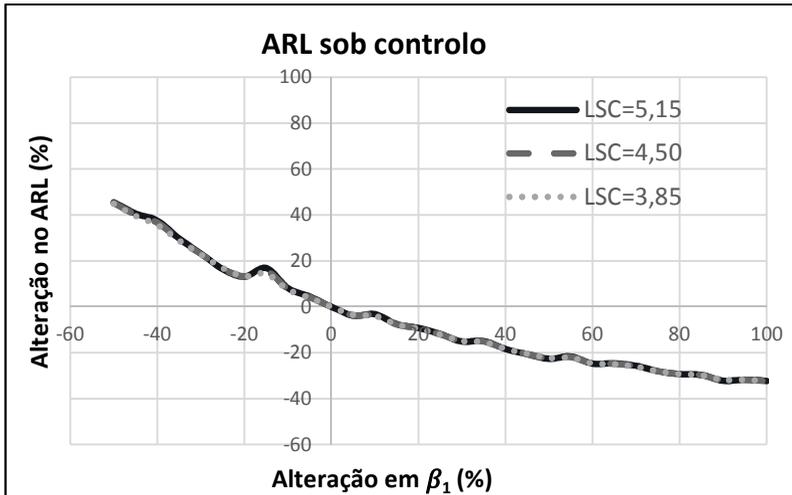
Figura 2 - Variação no ARL fora de controle devido a alterações no parâmetro do modelo.

A análise destes gráficos permite-nos em primeiro lugar concluir que alterações no parâmetro provocam variações no ARL de sentido contrário. Podemos também observar que estimativas por defeito provocam maiores alterações no ARL do que estimativas por excesso. A título de exemplo foram assinalados dois pontos na Figura 1: podemos ver que uma diminuição de 40% em  $\beta_1$  provoca um aumento no

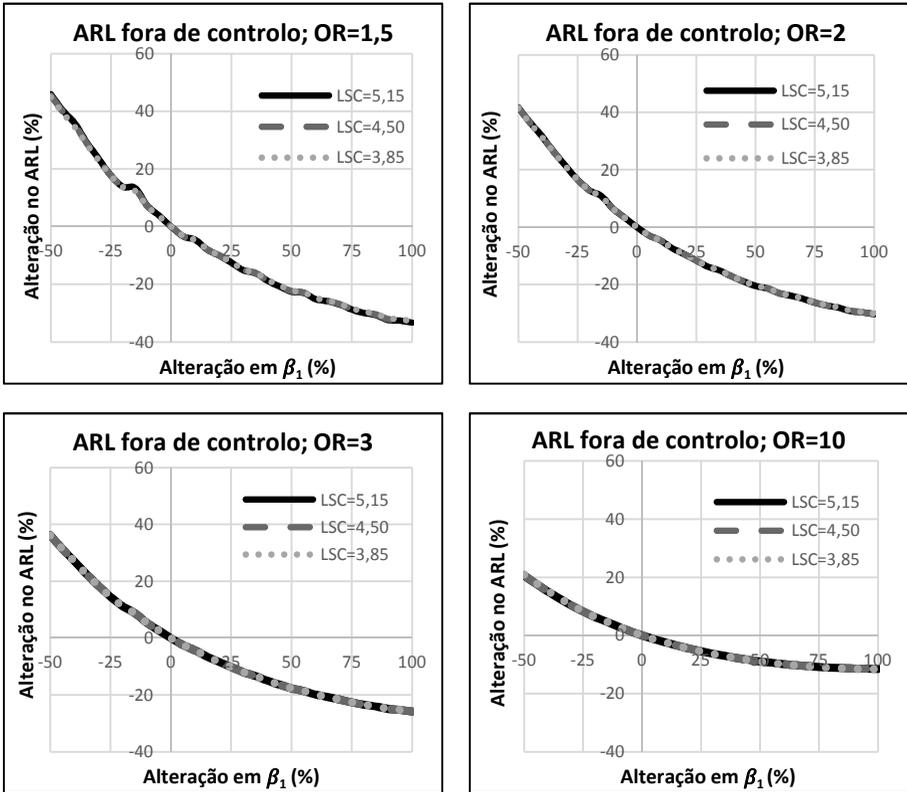
ARL na ordem dos 36,6%, enquanto que um aumento de 40% em  $\beta_1$  provoca uma diminuição no ARL na ordem dos 18,4%. A análise da Figura 2 permite-nos ainda concluir que quanto maior for o OR menores são as alterações ocorridas no ARL fora de controlo.

Estes primeiros resultados mostram-nos que a carta RACUSUM é sensível a alterações no parâmetro, mesmo quando o verdadeiro valor de OR é demasiado elevado, embora neste caso o impacto seja menor.

Com intuito de verificar se as alterações ocorridas no ARL dependiam do limite superior escolhido para a carta, o que é o mesmo que dizer, se dependiam do ARL sob controlo escolhido no delineamento da carta, e em caso afirmativo perceber como variavam, repetimos a análise anterior, mas considerando que a carta era concebida para ter um ARL sob controlo aproximadamente igual a 3500 ( $LSC=3,85$ ) e 14000 ( $LSC=5,15$ ), ou seja, respetivamente, metade e o dobro do valor considerado inicialmente. Os resultados obtidos apresentam-se nos gráficos seguintes (Figuras 3 e 4).



**Figura 3** - Comparação da variação no ARL sob controlo devido a alterações nos parâmetros do modelo, considerando cartas com diferentes limites superiores de controlo.



**Figura 4** - Comparação da variação no ARL fora de controlo devido a alterações no parâmetro do modelo, considerando cartas com diferentes limites superiores de controlo.

Como podemos observar as linhas praticamente sobrepõem-se, sendo que as reduzidas diferenças que existem se prendem com o facto de os valores serem obtidos por aproximação. Podemos assim concluir que o impacto do erro de estimação dos parâmetros no ARL da carta é independente do ARL sob controlo escolhido para a carta.

### 3.2 Erro de predição

Numa segunda fase, utilizando um procedimento análogo, avaliámos o impacto do erro de predição no desempenho da carta RACUSUM. Provocámos alterações na probabilidade de morte ( $p_j$ ) que variaram entre -50% e 100%, considerando incrementos de 5%.

Os resultados obtidos são similares aos obtidos anteriormente e naturalmente conduzem-nos às mesmas conclusões. A carta RACUSUM é sensível a alterações no modelo de risco sendo que o desempenho da carta sofre maiores alterações quando o modelo estima por defeito a probabilidade de morte do que quando estima por excesso, sendo o efeito tanto menor quanto maior é o verdadeiro OR. Além disso, também neste caso conseguimos mostrar que as alterações são independentes do ARL sob controle definido para a carta. Para melhor visualizar estes resultados, na Tabela 1 apresentamos apenas alguns dos valores obtidos nesta análise.

**Tabela 1** - Variação no ARL sob e fora de controle devido a alterações na probabilidade de morte estimada ( $p_j$ ).

Alteração na probabilidade de morte estimada ( $p_j$ )		ARL sob Controle ( $ARL_0=7040$ )	ARL fora de controle		
			OR=1.5 ( $ARL_1=533$ )	OR=2 ( $ARL_1=201$ )	OR=3 ( $ARL_1=91$ )
-50%	Variação (%)	11914 (69,2%)	919 (72,5%)	344 (70,9%)	152 (67,1%)
-10%	Variação (%)	7395 (5,0%)	568 (6,7%)	216 (7,2%)	97 (7,2%)
10%	Variação (%)	6566 (-6,8%)	497 (-6,8%)	188 (-6,6%)	85 (-6,2%)
50%	Variação (%)	5252 (-25,4%)	399 (-25,1%)	153 (-24,2%)	70 (-22,8%)
100%	Variação (%)	4411 (-37,3%)	336 (-37,0%)	129 (-23,9%)	60 (-33,9%)

## 4 Conclusões

O modelo de risco desempenha um papel importante no delineamento das cartas RACUSUM, na medida em que afeta o desempenho das mesmas. Neste trabalho, provocando alterações quer no parâmetro do modelo quer na probabilidade de morte estimada, avaliamos as alterações ocorridas no desempenho da carta RACUSUM, desempenho esse medido em termos dos ARL's sob e fora de controle. Observamos que a carta RACUSUM é sensível a alterações do modelo. Para ambas as situações analisadas, os resultados obtidos são concordantes e permitem-nos concluir que os ARL's diminuem à medida que aumenta a probabilidade de morte (e vice-versa); além disso, modelos que estimam por defeito a probabilidade de morte têm maior impacto no desempenho da carta do

que modelos que estimam por excesso. Foi ainda possível observar que as alterações são maiores quando ocorrem pequenas variações no processo. Por último, constatámos ainda que, o impacto foi independente do ARL sob controlo escolhido para a carta.

A monitorização da qualidade de processos de saúde é importante e sabemos que nenhum modelo é perfeito: assim, pensamos ser de grande importância o desenvolvimento de novas cartas, ou adaptações às já existentes, de modo a que estas se revelem mais robustas.

## Agradecimentos

Os autores agradecem ao professor Stefan Steiner da Universidade de Waterloo por ter facultado os dados que serviram de base a este estudo.

Este trabalho foi parcialmente financiado por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, projetos UID/MAT/04674/2019(CIMA) e UID/MAT/00006/2019(CEAUL).

## Referências

- ALEMI, F., ROM, W. & EISENSTEIN, E. (1996). Risk-Adjusted Control Charts for Health Care Assessment, *Annals of Operations Research*, 67, 45-60.
- AXELROD, D., GUIDINGER, M., METZGER, R., WIESNER, R., WEBB, R. & MERION, R. (2006). Transplant Center Quality Assessment Using a Continuously Updatable, Risk-Adjusted Technique (CUSUM), *American Journal of Transplantation*, 6, 313-323.
- BOTTLE, A. & AYLIN, P. (2008). Intelligent Information: A National System for Monitoring Clinical Performance, *Health Services Research*, 43, 1-31.
- CHEN, T., CHUNG, K., HU, F., FAN, C. & YANG, M. (2011). The use of statistical process control (risk-adjusted CUSUM, risk-adjusted RSPRT and CRAM with prediction limits) for monitoring the outcomes of out-of-hospital cardiac arrest patients rescued by the EMS system, *Journal of Evaluation in Clinical Practice*, 17, 71-77.
- COOK, D., COORY, M. & WEBSTER, A. (2011). Exponentially Weighted Moving Average Charts to Compare Observed and Expected Values for Monitoring Risk- Adjusted Hospital Indicators, *BMJ Quality and Safety*, 20, 469-474.
- COOK, D., DUKE, G., HART, G., PILCHER, D. & MULLANY, D. (2008). Review of the application of risk-adjusted charts to analyse mortality outcomes in critical care, *Critical Care and Resuscitation*, 10, 239-251.

- GOMBAY, E., HUSSEIN, A. & STEINER, S. (2011). Monitoring Binary Outcomes Using Risk-Adjusted Charts: A Comparative Study, *Statistics in Medicine*, 30, 2815–2826.
- GRIGG, O., FAREWELL, V. & SPIEGELHALTER, D. (2003). The Use of Risk-Adjusted CUSUM and RSPRT Charts for Monitoring in Medical Contexts, *Statistical Methods in Medical Research*, 12, 147–170.
- GRIGG, O. & FAREWELL, V. (2004). An overview of risk adjusted charts, *Journal of the Royal Statistical Society*, 167, 523-539.
- HARRIS, J., FORBES, T., STEINER, S., LAWLOR, K., DEROSE, G. & HARRIS, K. (2005). Risk-Adjusted Analysis of Early Mortality After Ruptured Abdominal Aortic Aneurysm Repair, *Journal of Vascular Surgery*, 42, 387-391.
- JONES, M. & STEINER, S. (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance, *International Journal for Quality in Health Care*, 24(2), 176-178.
- KHAN, R. (1984). On cumulative sum procedures and the SPRT with applications, *Journal of the Royal Statistical Society. Series B*, 46, 79–85.
- LOKE, C. & GAN, F. (2012). Joint Monitoring Scheme for Clinical Failures and Predisposed Risks, *Quality Technology and Quantitative Management*, 9(1), 3–21.
- LOVEGROVE, J., VALENCIA, O., TREASURE, T., SHERLAW-JOHNSON, C. & GALLIVAN, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display, *Lancet*, 18, 1128-1130.
- MEGAHED, F., KENSLER, J., BEDAIR, K. & WOODALL, W. (2011). A Note on the ARL of Two-Sided Bernoulli-based CUSUM Control Charts, *Journal of Quality Technology*, 43(1), 43–49.
- MONTGOMERY, D.C. (2013). *Introduction to Statistical Quality Control*, 7th Edition. John Wiley & Sons, Inc, New York.
- MOORE, R., NUTLEY, M., CINA, C., MOTAMEDI, M., FARIS, P. & ABUZNADAH, W. (2007). Improved Survival After Introduction of an Emergency Endovascular Therapy Protocol for Ruptured Abdominal Aortic Aneurysms, *Journal of Vascular Surgery*, 4, 443-450.
- MOUSTAKIDES, G. (1986). Optimal Stopping Times for Detecting Changes in Distribution, *The Annals of Statistics*, 14, 1379–1387.
- NOVICK, R., FOX, S., STITT, L., FORBES, T. & STEINER, S. (2006). Direct comparison of risk-adjusted and non-risk-adjusted CUSUM analyses of coronary artery bypass surgery outcomes, *The Journal of Thoracic and Cardiovascular Surgery*, 132(2), 386-391.
- POLONIECKI, J., VALENCIA, O. & LITTLEJOHNS, P. (1998). Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery, *British Medical Journal*, 316, 1697-700.

- R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.r-project.org/>.
- SPIEGELHALTER, D., GRIGG, O., KINSMAN, R. & TREASURE, T. (2003). Sequential probability ratio tests: applications to Bristol, Shipmann and adult cardiac surgery, *International Journal for Quality Health Care*, 15, 1-7.
- STEINER, S., COOK, R., FAREWELL, V. & TREASURE, T. (2000). Monitoring Surgical Performance Using Risk Adjusted Cumulative Sum Charts, *Biostatistics*, 1, 441-452.
- STEINER, S., COOK, R. & FAREWELL, V. (2001). Risk-Adjusted Monitoring of Binary Surgical Outcomes, *Medical Decision Making*, 21(3), 163–169.
- STEINER, S. (2001). Software to support risk-adjusted CUSUM charts: Findarl.r. URL [https://uwaterloo.ca/statistics-and-actuarial-science/sites/ca.statistics\\_and-actuarial-science/files/uploads/files/findarl.r](https://uwaterloo.ca/statistics-and-actuarial-science/sites/ca.statistics_and-actuarial-science/files/uploads/files/findarl.r) (acedido em 12 de Janeiro de 2016).
- TIAN, W., SUN, H., ZANG, X. & WOODALL, W. (2014). The impact of varying patient populations on the in-control performance of the risk-adjusted CUSUM chart, *International Journal for Quality in Health Care*, 27(1), 31-36.
- WOODALL, W., FOGEL, S. & STEINER, S. (2015). The Monitoring and Improvement of Surgical-Outcome Quality, *Journal of Quality Technology*, 47(4), 383-399.

# A análise classificatória na caracterização da produção e consumo de produtos de origem animal a nível mundial

Manuel Minhoto<sup>1</sup> · Luís Fernandes<sup>2</sup>

**Resumo:** Este trabalho tem por objetivo contribuir para a análise da evolução da produção e do consumo dos cinco principais produtos de origem animal – leite, carnes de bovinos, de suínos e de aves e ovos –, com base nos dados estatísticos da FAO relativos aos triénios 1961/62/63 e 2011/12/13 de 70 países. Realizou-se Análise Classificatória Hierárquica Ascendente a fim de obter o número de grupos recomendado, seguindo-se o algoritmo das k-médias para refinar e interpretar os grupos constituídos.

**Palavras-chave:** Análise Classificatória, Algoritmo das K-médias, Produtos de Origem Animal, Produção e Consumo Mundial.

## 1 Introdução

A chamada “Revolução Verde”, decorrente da Segunda Guerra Mundial, intensificou fortemente a produção agrícola e, sobretudo a partir dos anos 60, a generalidade dos países foi incrementando a sua produtividade na sequência de inovações estruturais e tecnológicas implementadas nas empresas agropecuárias. Esta situação foi fundamental para que a oferta de alimentos acompanhasse o crescimento demográfico exponencial da segunda metade do século XX, com a população mundial a duplicar no curto período de 40 anos (1960-2000).

No subsector dos alimentos de origem animal regista-se crescimento notoriamente superior ao verificado a nível demográfico, o que permitiu forte aumento no consumo *per capita* daqueles produtos a nível mundial. No entanto,

---

<sup>1</sup> Universidade de Évora, CIMA-UE, [minhoto@uevora.pt](mailto:minhoto@uevora.pt)

<sup>2</sup> Universidade de Évora, ICAAM, [ladsf@uevora.pt](mailto:ladsf@uevora.pt)

este cenário não terá sido idêntico para todos os produtos nem para todos os países, quer ao nível da produção, quer ao nível do consumo.

O presente trabalho tem por objetivo contribuir para a análise da evolução nos últimos 50 anos relativamente à produção e ao consumo dos principais produtos de origem animal destinados à alimentação humana.

Os produtos considerados foram o leite, ovos e as carnes de suíno, de bovino e de aves, com registos referentes a 70 países com população superior a 3 milhões de habitantes no ano 2013 (ver em Anexo I a lista dos países e respetivo código de 3 letras da norma ISO 3166).

## 2 Metodologia

O trabalho teve por base os dados estatísticos de 70 países selecionados pelo critério de população superior a 3 milhões de habitantes em 2013, último ano de registo de dados disponíveis para o trabalho. Esta seleção foi condicionada a que os países participantes não deveriam ter a produção e o consumo restringidos fortemente por razões de natureza religiosa, o que acontece sobretudo relativamente à carne de suíno. O conjunto de 70 países que integrou o estudo representava em 2013 cerca de 80% da população mundial.

Os dados foram obtidos a partir do *website* da *Food and Agriculture Organization of the United Nations* (FAO), e respeitam aos triénios 1961/62/63 e 2011/12/13.

O consumo de produtos de cada país foi obtido diretamente da base de dados estatísticos da FAO em valor anual (kg) *per capita*. No caso da produção optou-se por converter a informação da quantidade anual produzida por cada país em valor anual (kg) por habitante no respetivo ano.

Os valores utilizados foram obtidos a partir da média de cada triénio atrás referido, visto os coeficientes de variação, quer da produção, quer do consumo, entre anos de cada triénio e para os 5 produtos considerados apresentarem valores baixos a nível mundial (cerca de 2% para produção e para consumo no triénio 1961/62/63 e menos de 1% no triénio 2011/12/13).

A Análise Classificatória iniciou-se pela Hierárquica Ascendente para as variáveis observadas previamente normalizadas (centradas e reduzidas), obtendo-se os respetivos dendogramas. Para tal recorreu-se à função *hclust* do Programa Estatístico R e considerou-se como método de agregação o método “*Ward*”, mais concretamente o método “*Ward.D2*”, que utiliza a distância euclidiana quadrática.

Procurou-se que o número de grupos fosse idêntico para as bases “produção” e “consumo” e para ambos os triénios analisados. Apesar de se reconhecer que esta opção por igual número de grupos poder implicar algumas limitações, considerou-

se que seria adequada, quer para a caracterização dos próprios grupos, quer para a análise comparativa de resultados entre produção e consumo e entre triénios.

Estabelecido o número de grupos, passou-se à Análise Classificatória não Hierárquica. Para esta Análise Classificatória recorreu-se ao Algoritmo das *k*-médias. Utilizou-se o algoritmo de Hartigan and Wong (1979) que funciona como algoritmo por *default* da função *kmeans* do programa Estatístico R. Este algoritmo tem em geral um bom desempenho, embora seja recomendado que se considerem várias soluções iniciais (no presente caso consideraram-se 10 soluções iniciais). Cada solução inicial é constituída por quatro centróides iniciais escolhidos aleatoriamente.

Cada país é comparado com cada centróide inicial através da distância euclidiana habitual, sendo alocado ao grupo de menor distância. Atualizam-se então os centróides dos grupos acabados de formar e volta-se a comparar cada país com os novos centróides, tornando-se a alocar o país ao grupo de menor distância. O processo continua até que nenhum país seja realocado ou se atinja o número máximo de iterações. O número máximo de iterações utilizado foi 40. Comparam-se para a Análise Classificatória Hierárquica e Não Hierárquica a soma de quadrados entre os grupos. Nas cinco variáveis adicionam-se então os totais da soma dos quadrados entre os grupos. Calcula-se igualmente (em percentagem) o quociente entre os totais da soma dos quadrados entre os grupos e da soma de quadrados total.

Por fim, e apenas para a Análise Classificatória não Hierárquica, apresenta-se o valor da média de cada uma das variáveis observadas em cada grupo, bem como o quociente entre o valor dos quadrados médios entre os grupos e o valor dos quadrados médios dentro dos grupos (Estatística *F* de Snedecor). A variável cujo valor de *F* seja mais elevado será a que melhor discrimina entre os grupos. Apresenta-se ainda a soma de quadrados dentro de cada grupo (*SQR*) e a sua soma (*Total – SQR*). Termina-se com uma tabela que indica os países que integram cada um dos grupos.

### **3 Resultados**

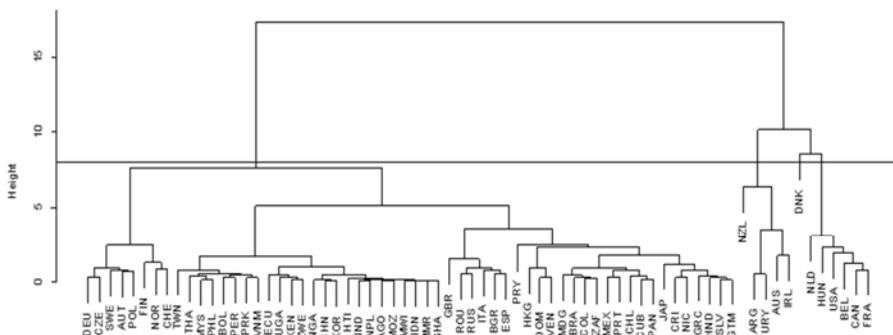
#### **3.1 Análise classificatória hierárquica**

Nesta secção apresentam-se os dendogramas relativos à Análise Classificatória Hierárquica, obtidos considerando como distância entre países a distância euclidiana habitual e como método de agregação dos grupos, o método “*Ward*” (Figuras 1, 2, 3 e 4). Sobre os dendogramas foi traçado um segmento de reta que

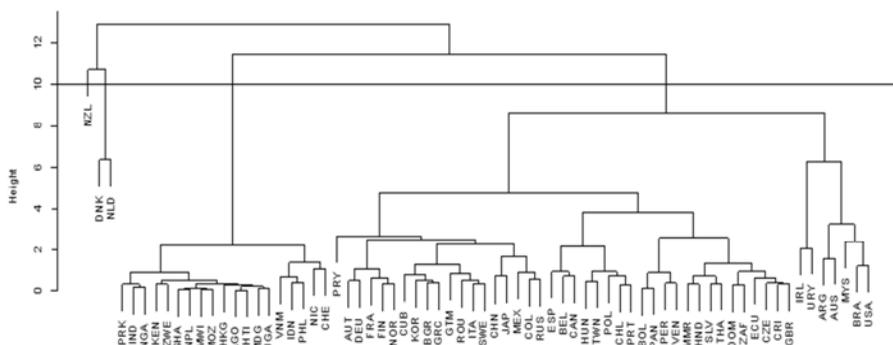
intersecta quatro “ramos da árvore” de modo a facilitar a visualização dos países que irão formar os quatro grupos considerados.

Nas variáveis da produção (Figuras 1 e 2), para o triénio 1961/62/63 constituíram-se grupos formados por 1, 5, 6 e 58 países, enquanto para o triénio 2011/12/13 os grupos apresentavam 1, 2, 19 e 48 países.

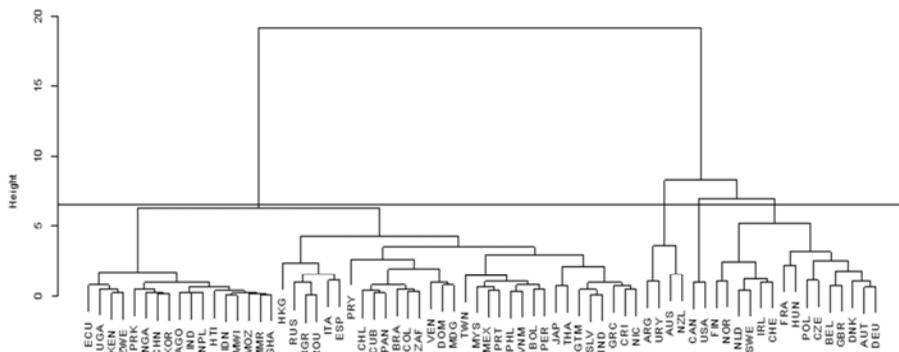
No triénio 1961/62/63 para o consumo (Figura 3) formaram-se grupos de 2, 4, 15 e 49 países. 50 anos mais tarde, os grupos tornam-se mais homogêneos, com 15, 15, 16 e 24 países.



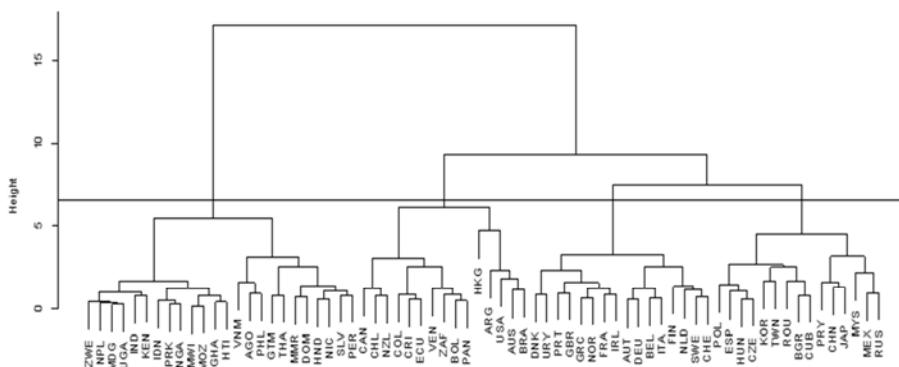
**Figura 1** – Dendrograma para a produção no triénio 1961/62/63.



**Figura 2** – Dendrograma para a produção no triénio 2011/12/13.



**Figura 3** – Dendrograma para o consumo no triénio 1961/62/63.



**Figura 4** – Dendrograma para o consumo no triénio 2011/12/13.

### 3.2 Análise classificatória não hierárquica – o algoritmo das k-médias

A Tabela 1 respeita aos 4 grupos para cada um dos triénios das variáveis da produção e do consumo. Como seria de esperar, a soma de quadrados entre os grupos é sempre superior na classificação não hierárquica que utiliza o algoritmo das k-médias.

**Tabela 1** - Comparação da soma de quadrados entre grupos (e da sua percentagem da soma de quadrados total) para a classificação hierárquica e não hierárquica.

Dados	Classificação Hierárquica	Classificação Não Hierárquica
Produção 1961/62/63	237.032 (68.70%)	242.707 (70.35%)
Produção 2011/12/13	206.964 (59.99 %)	218.266 (63.27%)
Consumo 1961/62/63	242.330 (70.24%)	250.282 (72.65%)
Consumo 2011/12/13	218.093 (63.22 %)	228.900 (66.35%)

Conforme se observa na Tabela 2 relativa à produção, no triénio dos anos 60 as Carnes de Suíno e de Bovino eram claramente as que mais discriminavam, enquanto no triénio da corrente década só a Carne de Bovino mantém valor elevado. O Leite e a Carne de Aves são as que menos discriminam em qualquer dos triénios.

Na Tabela 3, com as mesmas variáveis e triénios mas ao nível do consumo, verifica-se que os valores F estão mais próximos entre si do que na produção, mas mantendo-se a Carne de Bovino, seguida da de Suíno, como as que mais discriminam no triénio 1961/62/63; no triénio mais recente destaca-se o Leite como variável com maior capacidade discriminante.

**Tabela 2** – Centros dos 4 grupos nas variáveis da produção nos dois triénios.

Triénio 1961/62/63	Grupos				Estatística F
	1	2	3	4	
Ovos	3.067	1.004	1.263	-0.481	46.889
Leite	2.846	1.866	0.638	-0.402	28.665
C. Bovino	1.420	3.135	0.148	-0.373	122.96
C. Suíno	6.482	0.171	0.911	-0.376	150.09
C. Aves	4.373	0.012	0.998	-0.341	26.964
Nº países	1	5	13	51	Total - SQR
SQR	0.000	28.149	38.404	35.740	102.293
Triénio 2011/12/13					
Ovos	2.626	0.233	-1.043	0.399	40.245
Leite	1.080	2.595	-0.348	-0.116	19.6
C. Bovino	0.004	3.636	-0.459	-0.106	115.92
C. Suíno	4.658	-0.077	-0.406	-0.002	47.182
C. Aves	1.624	0.834	-1.006	0.370	24.129
Nº países	2	4	22	42	Total - SQR
SQR	20.302	38.015	8.873	59.543	126.697

**Tabela 3** – Centros dos 4 grupos nas variáveis do consumo nos dois triénios.

Triénio 1961/62/63	Grupos				Estatística F
	1	2	3	4	
Ovos	1.448	0.853	0.770	-0.597	44.798
Leite	0.902	1.522	0.627	-0.608	50.293
C. Bovino	0.543	0.146	2.832	-0.468	84.945
C. Suíno	1.729	0.950	-0.045	-0.590	70.399
C. Aves	1.873	0.008	-0.038	-0.414	52.514
Nº de países	4	16	5	45	Total - SQR
SQR	9.642	33.527	12.484	39.065	94.718
Triénio 2011/12/13					
Ovos	-1.275	0.349	0.435	0.573	35.889
Leite	-0.941	0.393	-0.542	1.1811	76.31
C. Bovino	-0.869	1.711	-0.268	0.227	44.402
C. Suíno	-0.926	0.169	-0.265	1.010	29.223
C. Aves	-1.067	1.571	0.189	0.037	47.191
Nº de países	19	10	20	21	Total - SQR
SQR	15.404	27.526	44.142	29.026	116.098

Reportando ao triénio 1961/62/63 das variáveis da produção, o grupo 1 é caracterizado por valores muito acima da média em todas as variáveis e é formado apenas pela Dinamarca, conforme se observa na Tabela 4 (1ª posição na produção por habitante em Carne de Suíno e de Aves, 2ª posição em Leite e Ovos e 7ª em Carne de Bovino). Em oposição está o grupo 4, com valores abaixo da média em todas as variáveis e constituído pela grande maioria dos países (51). O grupo 2, com valores elevados na Carne de Bovino e acima da média nas restantes, pode designar-se por países onde a produção de bovinos tem grande destaque (Argentina, Austrália, Irlanda, Nova Zelândia e Uruguai). O grupo 3 pode considerar-se como os médios produtores (relembra-se que o valor da produção utilizado neste trabalho está indexado ao número de habitantes do país), onde surgem os Estados Unidos da América, Canadá e um conjunto de países europeus, entre os quais a França, Alemanha e Reino Unido.

**Tabela 4** – Países que integram os grupos da produção no triénio 1961/62/63.

Grupo	Países
1	DNK
2	ARG, AUS, IRL, NZL, URY,
3	AUT, CAN, FIN, FRA, DEU, HUN, NLD, POL, SUE, GBR, USA, CZE
4	Todos os restantes 51 países

Cinquenta anos depois (ver Tabela 5), o grupo 1, onde a Holanda se juntou à Dinamarca, mantém-se com grande destaque na produção de Carne de Suíno e Aves e de Ovos, surgindo um grupo 2 associado a grandes produtores de Carne de Bovino e Leite. O grupo 4 deixou de ser o grupo com menor produção, podendo dizer-se que engloba os pequenos/médios produtores e os muito pequenos produtores constituem o grupo 3.

Na análise por países mantem-se relativa semelhança entre os dois primeiros grupos, com a particularidade de a Holanda ter transitado do Grupo 3 para o Grupo 1 e a Argentina ter passado de grande produtor de carne de bovino para médio produtor. O número de países que se consideram pequenos/médios produtores manteve-se inalterado (64), no entanto formou-se um grupo de 22 países maioritariamente africanos e asiáticos (Grupo 3), com níveis de produção muito baixos, que pouco evoluíram nos últimos 50 anos.

**Tabela 5** – Países que integram os grupos da produção no triénio 2011/12/13.

Grupo	Países
1	DNK, NLD
2	AUS, IRL, NZL, URY
3	AGO, CUB, GHA, GRC, HTI, HND, HGK, IND, IDN, KEN, PRK, MDG, MWI, MOZ, NPL, NIC, NGA, PHL, ZWE, CHE, UGA, VNM
4	Todos os restantes 42 países

A Tabela 6 apresenta os grupos de países referente à década de 60 para as variáveis do consumo. O grupo 1 caracteriza-se pelos valores mais elevados em carne de suíno e de aves e em ovos, podendo designar-se por grandes consumidores (USA, Canadá, vários países europeus e Hong Kong). O grupo 3 representa os grandes consumidores de carne de bovino e o grupo 2 (todos países europeus) está sobretudo associado ao consumo de leite. O grupo 4, integrando a grande maioria dos países (45), representa os pequenos e muito pequenos consumidores.

**Tabela 6** – Países que integram os grupos do consumo para o triénio 1961/12/13.

Grupo	Países
1	CAN, FRA, HUN, USA
2	AUT, BEL DNK, DEU, FIN, GBR, HKG, IRL, ITA, NLD, NOR, POL, SWE, CHE, RUS, CZE
3	ARG, AUS, NZL, PRY, URY
4	Todos os restantes 45 países

Cinquenta anos depois alterou-se o panorama no consumo, conforme se constata na Tabela 7. Passaram a existir dois grupos de grandes consumidores num total de 31 países (grupos 2 e 4) e dois grupos de pequenos consumidores num total de 39 países (grupos 1 e 3).

Os grandes consumidores de carne de bovino são também grandes consumidores de carne de aves (grupo 2 com 10 países onde se incluem o Canadá, USA, Brasil, Argentina, Austrália, Nova Zelândia) e os grandes consumidores de carne de suíno são também grandes consumidores de leite (grupo 4, constituído praticamente só por países europeus).

Nos países com menores consumos por habitante, constituindo o grupo 1 (19 países predominantemente africanos e asiáticos), destaca-se pelo muito baixo consumo. Verifica-se que muitos países integram simultaneamente o grupo mais fraco quer da produção, quer do consumo.

**Tabela 7** – Países que integram os grupos do consumo para o triénio 2011/12/13.

Grupo	Países
1	AGO, GHA, HTI, HND, IND, IDN, KEN, PRK, MDG, MWI, MOZ, NPL, NIC, NGA, PER, PHL, ZWE, UGA, VNM
2	ARG, AUS, BRA, CAN, CHL, HKG, NZL, USA, URY, VEN
3	BOL, BGR, MMR, CHL, CHN, COL, CRI, CUB, DOM, ECU, SLV, GTM, JAP, KOR, MYS, MEX, PAN, PRY, ZAF, TWN, THA
4	Todos os restantes 21 países

## 4 Conclusões

Nos últimos 50 anos observa-se que os países com maior quantidade de produção relativamente à sua dimensão demográfica registaram também os maiores consumos anuais *per capita*, sobretudo nos produtos onde são mais especializados.

Quanto aos grupos de países, no triénio 1961/62/63 surge, quer na produção, quer no consumo, um grande grupo com registos muito baixos (51 países na produção e 45 no consumo), que se contrapunha aos restantes 3 grupos com valores praticamente todos positivos.

Passados 50 anos, a composição dos grupos alterou-se, reduzindo-se a dimensão do grupo de muito baixo nível de produção e de consumo (na produção passou para 22 e no consumo para 19), o que mostra que um número considerável de países promoveu inovações estruturais e tecnológicas nos seus sistemas de produção agropecuária e que o poder de compra das famílias também permitiu incrementar o consumo de alimentos de origem animal.

Três notas complementares:

- Os padrões de consumo, referentes ao triénio 2011/12/13, dos grupos de países 2, 4 e mesmo parcialmente do 3, estão muito acima do recomendado pela Organização Mundial de Saúde para este tipo de alimentos;
- Manter o “hábito” já instalado de consumos excessivos obriga a produzir intensivamente e ao menor custo possível, com as consequentes externalidades negativas e seus efeitos no ambiente e nos territórios;
- O grupo de países que nos últimos anos ainda se caracteriza por muito baixos níveis de produção e de consumo, também na maioria deles se registam carências alimentares graves, pelo que seria conveniente que a resolução deste problema não passasse pelo aumento sistemático do consumo de produtos de origem animal.

## Referências

- GOLUB, G. & VAN LOAN, C. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD.
- JOHNSON, R. A. & WICHERN, D. W. (2002). (2002). *Applied Multivariate Statistical Analysis*, 4th Edition, Prentice-Hall, New Jersey.
- CADIMA, J., CERDEIRA, J. O. & MINHOTO, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47, 225-236.
- HARTIGAN, J. A. & WONG, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* **28**, 100–108.
- RAMSEY, J. O., TEN BERGE, J. & STYAN, G. P. H. (1984). Matrix correlation. *Psychometrika*. 49(3), 403-423.

Food and Agriculture Organization of the U.N. (FAO). Databases Production Livestock Primary and Food Balance, <http://www.fao.org/faostat/en/#data/QL> and <http://www.fao.org/faostat/en/#data/FBS> (acedido em Outubro 2017).

R Development Core Team (2012), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-9300051-07-0, URL <http://www.R-project.org>.

## **ANEXO I – LISTA DE PAÍSES, COM CÓDIGO DE 3 LETRAS**

AGO-Angola, ARG-Argentina, AUS-Austrália, AUT-Áustria, BEL-Bélgica, BGR-Bulgária, BOL-Bolívia, BRA-Brasil, CAN-Canadá, CHE-Suíça, CHL-Chile, CHN-China, COL-Colômbia, CRI-Costa Rica, CUB-Cuba, CZE-República Checa (em 1961-2-3 Checoslováquia), DEU-Alemanha, DNK-Dinamarca, DOM-República Dominicana, ECU-Ecuador, ESP-Espanha, FIN-Finlândia, FRA-França, GBR-Reino Unido, GHA-Gana, GRC-Grécia, GTM-Guatemala, HKG-Hong-Kong, HND-Honduras, HTI-Haiti, HUN-Hungria, IDN-Indonésia, IND-Índia, IRL-Irlanda, ITA-Itália, JAP-Japão, KEN-Quênia, KOR-Coreia do Sul, MDG-Madagascar, MEX-México, MMR-Myanmar, MOZ-Moçambique, MWI-Malawi, MYS-Malásia, NGA-Nigéria, NIC-Nicarágua, NLD-Holanda, NOR-Noruega, NPL-Nepal, NZL-Nova Zelândia, PAN-Panamá, PER-Perú, PHL-Filipinas, POL-Polónia, PRK-Coreia do Norte, PRT-Portugal, PRY-Paraguai, ROU-Roménia, RUS-Rússia (em 1961-2-3 URSS), SLV-El Salvador, SWE-Suécia, THA-Tailândia, TWN-Taiwan, UGA-Uganda, URY-Uruguai, USA-Estados Unidos da América, VEM-Venezuela, VNM-Vietnam, ZAF-África do Sul, ZWE-Zimbabué

## A operacionalização de uma tipologia socioeconómica para as áreas metropolitanas de Lisboa e Porto

Cátia Nunes<sup>1</sup> · Francisco Vala<sup>2</sup>

**Resumo** Este estudo tem como principal objetivo a elaboração de um quadro de referência para a análise das características territoriais das áreas metropolitanas de Lisboa e Porto, através da produção de uma tipologia de padrões socioeconómicos com base em análises multivariadas aplicadas a indicadores censitários operacionalizados à escala da subsecção estatística para o ano de 2011. Os resultados obtidos evidenciam as semelhanças e as diferenças entre os dois espaços metropolitanos e salientam as potencialidades da informação censitária no que respeita ao detalhe da escala territorial.

**Palavras-chave:** Análise Multivariada, Áreas Metropolitanas, Censos, Tipologia.

### 1 Introdução

A elaboração deste estudo deu continuidade ao trabalho iniciado pelo Instituto Nacional de Estatística (INE) para as duas áreas metropolitanas, primeiro com base nos resultados dos Censos de 1991 (INE, 2000a e 2000b) e, seguidamente, com base na informação censitária de 2001 (INE, 2004a e 2004b). O presente estudo consiste numa análise similar mas atualizada em resultado da disponibilização dos dados dos Censos de 2011 e pelo facto de considerar um modelo de análise integrado dos dois territórios metropolitanos (INE, 2014).

O objetivo principal deste estudo consistiu em elaborar um quadro de referência para a análise das características territoriais dos dois espaços metropolitanos, através da produção de uma tipologia de padrões socioeconómicos, com o intuito de identificar classes de unidades territoriais homogéneas que, do ponto de vista socioeconómico, se distingam entre si. O modelo desenvolvido permitiu reconhecer contrastes internos a cada um dos dois espaços e contribuiu também para esclarecer

---

<sup>1</sup> Instituto Nacional de Estatística, *catia.nunes@ine.pt*

<sup>2</sup> Instituto Nacional de Estatística, *francisco.vala@ine.pt*

as diferenças entre os dois territórios metropolitanos. A escala territorial de análise correspondeu à subsecção estatística. A literatura aconselha o recurso a escalas territoriais de maior detalhe para a avaliação, espacialmente integrada, de perfis e transformações socioeconómicas, como é o caso da freguesia (Ferrão, 2003) ou do bairro (Malheiros e Vala, 2004). No caso do presente estudo, dada a disponibilidade de informação estatística para microunidades censitárias, a escala territorial de análise considerou a subsecção estatística.

Este tipo de escalas, bem como os indicadores de carater socioeconómico passíveis de operacionalizar a partir da informação censitária sobre o parque habitacional e a população permitem obter uma leitura sobre os modelos de organização dos territórios. O território, entendido como o espaço em que os diversos atores e agentes sociais produzem e reproduzem práticas sociais, reflete e sofre alterações decorrentes dos diferentes processos de apropriação que ocorrem, de modo articulado, nas esferas de produção e organização económica, demográfica e social. Salgueiro (1998) demonstra, a partir do caso da cidade de Lisboa, como estudar estes processos permite perceber novas formas de organização do espaço urbano, como a “cidade fragmentada”, caracterizada pelo policentrismo, ou seja, pela existência de vários polos agregadores de atividades económicas, pela heterogeneidade na ocupação física e social do espaço e pela multiplicidade de funções.

## **2 Metodologia**

### **2.1 Âmbito territorial**

O âmbito territorial dos dois espaços metropolitanos é o definido pela Lei n.º 75/2013, de 12 de setembro, que inclui 18 municípios na Área Metropolitana de Lisboa (AML) – Alcochete, Almada, Amadora, Barreiro, Cascais, Lisboa, Loures, Mafra, Moita, Montijo, Odivelas, Oeiras, Palmela, Seixal, Sesimbra, Setúbal, Sintra e Vila Franca de Xira – e 17 municípios na Área Metropolitana do Porto (AMP) – Arouca, Espinho, Gondomar, Maia, Matosinhos, Oliveira de Azeméis, Paredes, Porto, Póvoa de Varzim, Santa Maria da Feira, Santo Tirso, São João da Madeira, Trofa, Vale de Cambra, Valongo, Vila do Conde e Vila Nova de Gaia.

Os territórios metropolitanos distinguem-se claramente do restante território nacional pela elevada densidade populacional que apresentam – em 2011, a densidade populacional era, na AML e na AMP respetivamente, oito e sete vezes superior à nacional. Contudo, os contrastes internos são também significativos: na AML, a densidade populacional por município oscilava entre 135 habitantes por km<sup>2</sup> em Palmela e 7 363 habitantes por km<sup>2</sup> na Amadora; na AMP, variava entre 68 habitantes por km<sup>2</sup> em Arouca e 5 736 habitantes por km<sup>2</sup> no Porto.

## 2.2 Informação de base

A tipologia produzida baseia-se em informação proveniente do Recenseamento Geral da População e da Habitação de 2011 (Censos 2011), disponível à escala da subsecção estatística e cobrindo a caracterização do efetivo populacional e do parque habitacional. A recolha da informação de base procurou atender sobretudo à sua relevância socioeconómica para a caracterização das duas áreas metropolitanas. De forma a expurgar o efeito da dimensão das unidades territoriais, procedeu-se à sua respetiva relativização, adotando indicadores calculados com base em proporções, médias ou índices. A área das subsecções estatísticas foi extraída da Base Geográfica de Referenciação da Informação (BGRI).

A atribuição de significado socioeconómico à informação censitária exige que sejam selecionadas as unidades territoriais relevantes para a aplicação de técnicas de análise multivariada, pelo que foram consideradas para a análise as subsecções estatísticas em que o número de edifícios, alojamentos, famílias, núcleos familiares e indivíduos fosse, simultaneamente, superior a 2. Esta opção conduziu à seleção de 48 263 (25 150 na AML e 20 113 na AMP) subsecções estatísticas relevantes em 2011, de um total de 57 636 (34 937 na AML e 22 699 na AMP). Por forma a reter a importância da dimensão de cada unidade territorial no universo estatístico, cada subsecção estatística foi ponderada pelo número de indivíduos residentes. Procurou-se abarcar diferentes domínios socioeconómicos através de um conjunto de 23 indicadores de base, suficientemente abrangentes da informação censitária [Quadro 1].

## 2.3 Métodos de análise de dados

A metodologia adotada consistiu, primeiramente, na aplicação de uma análise fatorial em componentes principais com base nos 23 indicadores censitários reportados ao ano de 2011. Posteriormente, procedeu-se a uma análise de *clusters* sobre as componentes extraídas, adotando o método das K-médias (*K-means*) com o objetivo de construir classes de subsecções estatísticas o mais homogêneas possível do ponto de vista socioeconómico.

A análise fatorial em componentes principais considerou os 23 indicadores de base e a pertinência da sua aplicação foi confirmada pelo teste de Bartlett e pela estatística de Kaiser-Meyer-Olkin (KMO). Na estatística de teste obteve-se um valor  $p < 0,001$ , sugerindo que existe correlação entre as variáveis, e na estatística KMO o valor de 0,797 confirmou que o modelo proposto tem boa adequabilidade. Foi adotado o critério de Kaiser, ou seja, extração das componentes principais cujo valor próprio associado é superior a um, e considerada a rotação *varimax*. Todos os indicadores apresentaram uma variância explicada pelo conjunto de fatores extraídos (comunalidade) superior a 0,5, com exceção do indicador referente à

proporção de edifícios não exclusivamente residenciais. Contudo, e uma vez que este indicador veicula uma ideia não refletida em nenhum outro indicador, a sua inclusão no modelo de análise foi considerada relevante.

**Quadro 1** - Indicadores censitários de base.

Indicador	Unidade
<b>Parque Habitacional</b>	
Idade média dos edifícios	Anos
Proporção de edifícios não exclusivamente residenciais	%
Proporção de edifícios com sete ou mais alojamentos	%
Proporção de alojamentos familiares clássicos de residência habitual sobrelotados	%
Proporção de alojamentos familiares clássicos de residência habitual arrendados	%
Proporção de alojamentos familiares clássicos de residência habitual sublotados	%
<b>População e Família</b>	
Proporção de núcleos familiares de casais com filhos	%
Proporção de famílias clássicas unipessoais de indivíduos com 65 ou mais anos de idade	%
Densidade populacional	N.º/km <sup>2</sup>
Média etária da população residente	Anos
Proporção de população residente de nacionalidade estrangeira	%
Média das habilitações académicas da população residente	Anos
Proporção de população residente viúva	%
Índice de diversificação religiosa	N.º
Duração média dos movimentos pendulares	Minutos
Proporção de população que trabalha ou estuda noutro município	%
Proporção de população residente empregada ou estudante cujo principal meio de transporte é o automóvel	%
Proporção de população residente empregada ou estudante cujo principal meio de transporte é coletivo	%
Proporção de população residente pertencente aos grupos socioeconómicos 14 a 18*	%
Proporção de população residente cuja principal fonte de rendimento é o trabalho	%
Taxa de atividade da população residente	%

Nota\*: Grupo 14 – Diretores e quadros dirigentes do Estado, das médias e grandes empresas; Grupo 15 – Dirigentes de pequenas empresas e organizações; Grupo 16 – Quadros intelectuais e científicos; Grupo 17 – Quadros técnicos intermédios; Grupo 18 – Quadros administrativos intermédios.

A análise de *clusters* foi aplicada recorrendo ao método das K-médias. Este método consiste na afetação de cada objeto ao *cluster* com média mais próxima, minimizando a soma do quadrado dos desvios intra-*cluster*, uma vez que o objetivo é que os elementos de cada *cluster* estejam mais próximos entre si do que em relação aos elementos pertencentes a outros *clusters*. Na especificação que foi adotada, a média é recalculada após cada afetação de um indivíduo a um *cluster*.

O método não-hierárquico das K-médias apresenta a vantagem de lidar melhor com bases de dados de dimensão significativa (no presente estudo, 48 263 subseções estatísticas) e de permitir que a afetação de um indivíduo a um *cluster* não seja definitiva enquanto o algoritmo não convergir. Contudo, exige que o número de *clusters* seja definido à partida. À semelhança de Vickers e Rees (2007), tendo em vista a definição de uma classificação socioeconómica de áreas de pequena dimensão (*output areas*) para o Reino Unido, foram testadas várias

hipóteses quanto ao número de *clusters*. De modo a assegurar um equilíbrio entre a diferenciação socioeconómica do território e a capacidade de interpretação dos dados, optou-se por reter a solução correspondente a seis classes socioeconómicas.

### 3 Resultados

#### 3.1 Identificação das dimensões socioeconómicas

A análise fatorial em componentes principais permitiu identificar cinco dimensões socioeconómicas que explicam 70,4% da variância total.

A primeira componente extraída do modelo corresponde ao *envelhecimento* e apresenta uma variância explicada de 22,2%. Esta componente encontra-se positivamente associada ao envelhecimento da população e do parque habitacional, à incidência de viúvos e de indivíduos idosos a viver sozinhos. Por oposição, apresenta uma associação negativa com a incidência de população ativa, de população cujo principal rendimento advém do trabalho e de famílias constituídas por casais com filhos [Quadro 2]. Territorialmente emergiram como mais envelhecidos os centros metropolitanos, com o envelhecimento a diminuir de intensidade à medida que aumenta o afastamento face a estes centros, sobretudo na AMP.

**Quadro 2** - Envelhecimento: principais correlações com os indicadores, 2011.

Indicadores	Peso fatorial
Média etária da população residente	0,919
Proporção de população residente viúva	0,810
Proporção de famílias clássicas unipessoais de indivíduos com 65 ou mais anos	0,807
Idade média dos edifícios	0,675
Proporção de núcleos familiares de casais com filhos	-0,735
Taxa de atividade da população residente	-0,748
Proporção de população residente cuja principal fonte de rendimento é o trabalho	-0,801

A segunda componente correspondente à *qualificação* e apresenta uma variância explicada de 18,2%. Esta componente encontra-se positivamente associada ao nível médio de habilitação da população, aos grupos socioeconómicos de maior qualificação, ao uso de automóvel nas deslocações pendulares e à expressão de habitações sublotadas. A associação é negativa com a incidência de desemprego, de alojamentos sobrelotados e de alojamentos arrendados e com o uso de meios de transporte coletivos nas deslocações pendulares [Quadro 3]. Territorialmente, a qualificação salienta, na AML, os municípios de Oeiras, Lisboa e Cascais e a parte ocidental de Sintra. Na AMP, destacam-se a parte ocidental do município do Porto e a linha de costa do território metropolitano.

**Quadro 3** - Qualificação: principais correlações com os indicadores, 2011.

Indicadores	Peso fatorial
Média das habilitações académicas da população residente	0,799
Proporção de população residente pertencente aos grupos socioeconómicos 14 a 18	0,776
Proporção de alojamentos familiares clássicos de residência habitual sublotados	0,720
Proporção de população residente empregada ou estudante cujo principal meio de transporte é o automóvel	0,613
Proporção de população residente empregada ou estudante cujo principal meio de transporte é coletivo	-0,492
Proporção de alojamentos familiares clássicos de residência habitual arrendados	-0,515
Taxa de desemprego da população residente	-0,699
Proporção de alojamentos familiares clássicos de residência habitual sobrelotados	-0,795

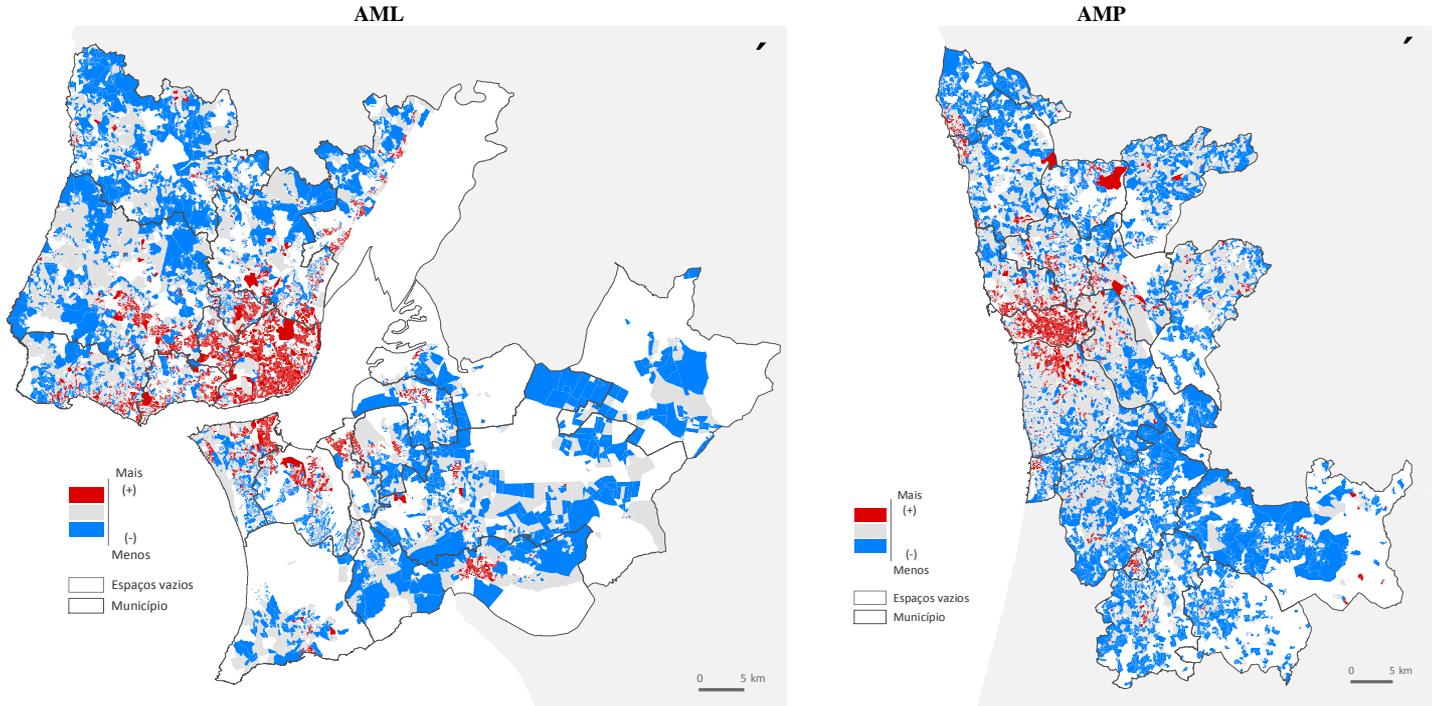
A terceira componente extraída designa-se *urbanização* e apresenta uma variância explicada de 11,9%. Encontra-se positivamente associada à presença de edifícios com sete ou mais alojamentos e não exclusivamente residenciais, à densidade populacional e à diversificação religiosa [Quadro 4].

**Quadro 4** - Urbanização: principais correlações com os indicadores, 2011.

Indicadores	Peso fatorial
Proporção de edifícios com sete ou mais alojamentos	0,834
Densidade populacional	0,782
Proporção de edifícios não exclusivamente residenciais	0,624
Índice de diversificação religiosa	0,512

Esta dimensão socioeconómica distingue claramente os municípios de Lisboa e do Porto do restante território metropolitano. Na AML, constata-se a existência de um contínuo mais urbanizado que une Lisboa a Cascais e que se estende também aos municípios da Amadora, de Odivelas e de Loures e alongando-se também a Sintra, assumindo uma disposição linear ao longo das principais vias de comunicação rododiferroviárias. Na AMP, a intensidade da urbanização vai diminuindo à medida que aumenta o afastamento face ao município do Porto. Contudo, a expressão desta componente salienta também a Norte, os centros dos municípios mais periféricos, em particular, Vila Nova de Gaia, Espinho e São João da Madeira, a Sul, e Vila do Conde, Póvoa de Varzim e Trofa [Figura 1].

A quarta componente extraída do modelo apresenta uma variância explicada de 10,1%. Esta componente foi designada de *imigração* dado que está positivamente associada à incidência de população que cinco anos antes do momento censitário residia no estrangeiro e de população residente de nacionalidade estrangeira, bem como à diversificação religiosa [Quadro 5]. Os resultados sugerem que esta dimensão socioeconómica está pouco presente na AMP e é claramente mais característica da AML, sobressaindo os municípios da Amadora, Lisboa, Cascais, Odivelas e Sintra.



**Figura 1** – Componente: Urbanização, 2011.

Nota: Para efeitos de representação os valores estandardizados da componente são apresentados em classes extremas com 30% das observações cada, e numa classe intermédia com os restantes 40% da distribuição. Uma vez que a leitura dos dados num mapa é afetada pelo impacto visual de grandes manchas que podem corresponder a espaços fracamente povoados, optou-se, para efeitos de representação em todas as figuras, considerar também como espaços vazios as subsecções fracamente povoadas, cuja área ultrapassa os 5 km<sup>2</sup>.

**Quadro 5** - Imigração: principais correlações com os indicadores, 2011.

Indicadores	Peso fatorial
Proporção de população residente que cinco anos antes do momento censitário residia no estrangeiro	0,876
Proporção de população residente de nacionalidade estrangeira	0,854
Índice de diversificação religiosa	0,546

Por último, a quinta componente extraída corresponde à *mobilidade pendular* e apresenta uma variância explicada de 8,0%. Esta dimensão está positivamente associada à duração dos movimentos pendulares e à intensidade da mobilidade pendular de indivíduos residentes por meio de transporte coletivo e para outro município que não o de residência [Quadro 6]. Territorialmente, a AML apresenta mais intensidade, em duração e extensão, dos movimentos pendulares do que a AMP, onde a mobilidade pendular é mais expressiva nos municípios de Gondomar e de Valongo. Os dois centros metropolitanos – municípios de Lisboa e do Porto – apresentam também menor expressão desta dimensão.

**Quadro 6** – Mobilidade pendular: principais correlações com os indicadores, 2011.

Indicadores	Peso fatorial
Duração média dos movimentos pendulares	0,842
Proporção de população que trabalha ou estuda noutro município	0,733
Proporção de população residente empregada ou estudante cujo principal meio de transporte é coletivo	0,462

### 3.2 Caracterização das classes socioeconómicas

Definir uma designação para cada classe socioeconómica constitui um desafio à capacidade de interpretar, combinar e dar sentido às diferentes dimensões que caracterizam uma dada classe e que, simultaneamente, a tornam diferente de outra. A atribuição de uma designação implica, por isso, avaliar e caracterizar as semelhanças e dissimelhanças *intra* e *inter* classes. A Figura 2 apresenta os valores médios de cada classe nas cinco componentes principais considerando o conjunto das duas áreas metropolitanas e a AML e AMP separadamente.

O *urbano consolidado* abrange 5 390 subsecções estatísticas (9% do total) nas quais residia, em 2011, cerca de 14% da população residente nas duas áreas metropolitanas (17% na AML e 10% na AMP). A principal característica distintiva desta classe face ao restante território metropolitano é o envelhecimento. Conjugando com os valores médios das restantes cinco componentes, conclui-se que se tratam de territórios envelhecidos, qualificados, densamente habitados e com uma mobilidade pendular pouco expressiva.

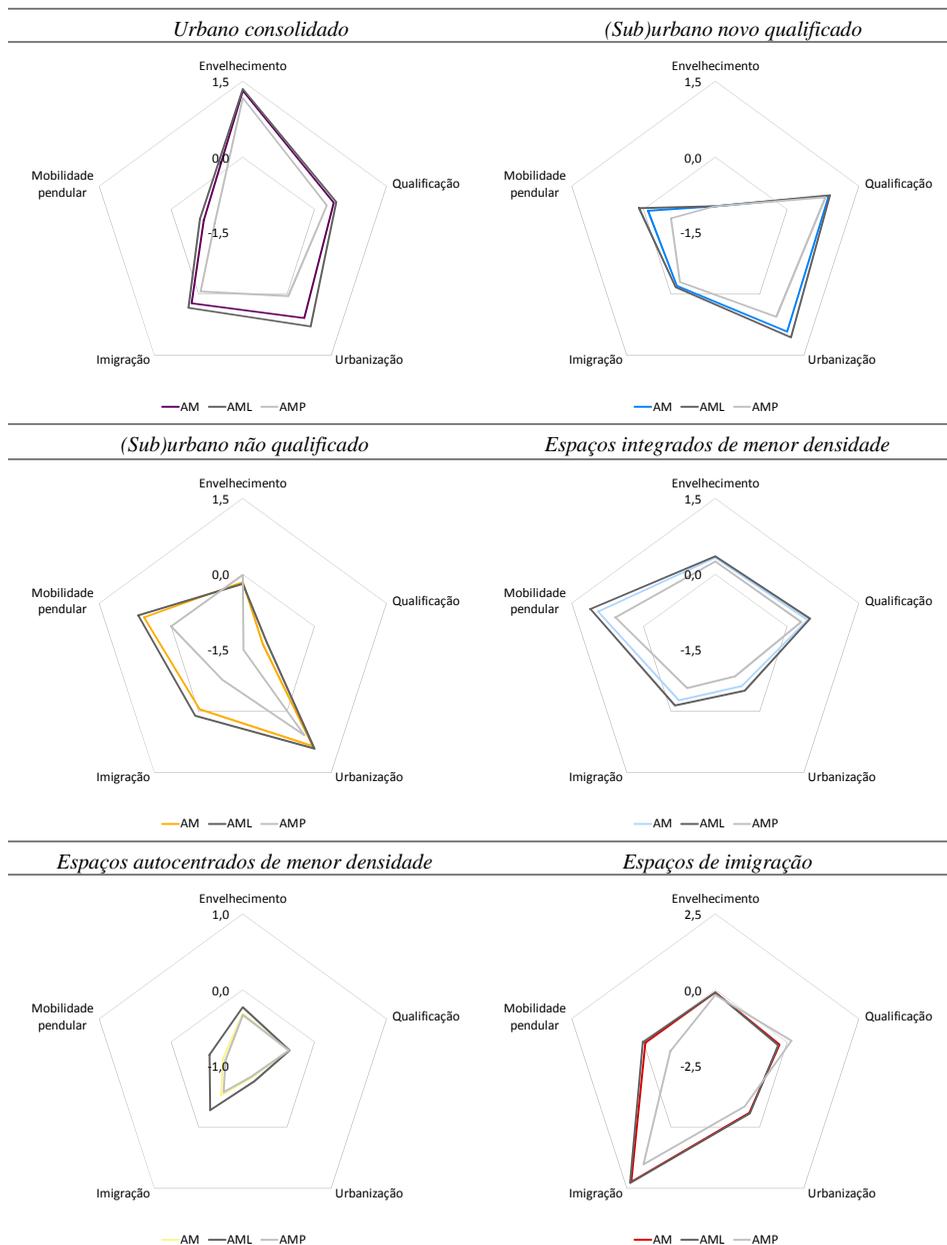
O *(sub)urbano novo qualificado* abrange 3 761 subsecções estatísticas nas quais, em 2011, residia cerca de 15% da população residente nas duas áreas metropolitanas (18% na AML e 11% na AMP). Esta classe caracteriza territórios mais jovens, qualificados e densamente habitados relativamente ao contexto dos dois espaços metropolitanos.

O *(sub)urbano não qualificado* é a classe socioeconómica que abrange o menor número de subsecções estatísticas (3 642) nas quais, em 2011, residia cerca de 16% da população residente nas subsecções estatísticas das duas áreas metropolitanas (21% na AML e 7% na AMP). A característica que mais distingue esta classe é a menor qualificação da população residente e do edificado. Paralelamente, são espaços densamente habitados cuja população residente vivencia movimentos pendulares expressivos.

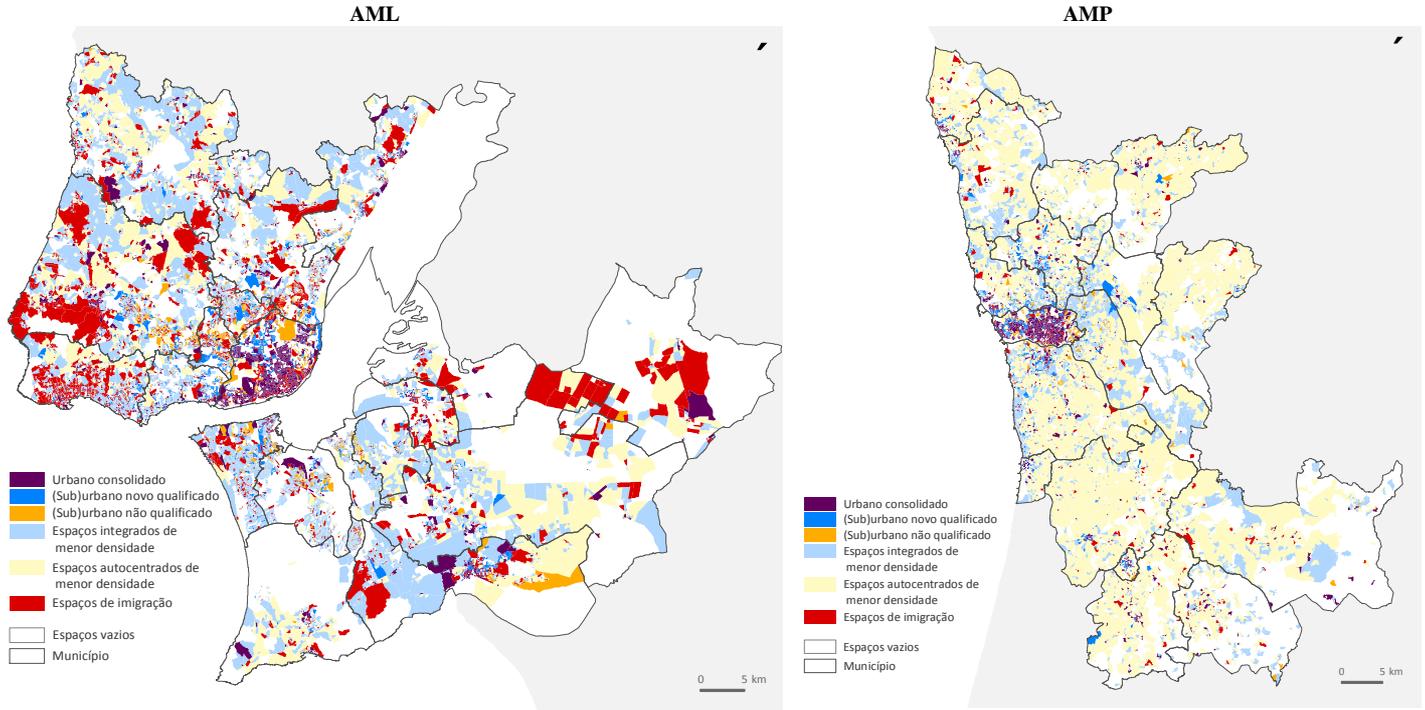
A classe referente aos *espaços integrados de menor densidade* abrange 15 009 subsecções estatísticas nas quais residia, em 2011, cerca de 21% da população residente nas duas áreas metropolitanas (24% na AML e 16% na AMP). As subsecções estatísticas classificadas nesta classe correspondem ao local de residência de uma população com mobilidade pendular acentuada, revelando territórios funcionalmente integrados. Paralelamente, constituem territórios menos urbanizados e com um grau de qualificação acima da média metropolitana.

Os *espaços autocentrados de menor densidade* constituem a classe socioeconómica com maior número de subsecções estatísticas (15 869). Nestes territórios, residia, em 2011, 24% da população residente nas duas áreas metropolitanas. Esta média é o reflexo de uma proporção de 6% para a AML e de mais de metade da população residente na AMP. O principal fator distintivo desta classe é a menor expressão da urbanização mas também o facto de serem territórios que revelam uma integração funcional menos expressiva face ao conjunto das duas áreas metropolitanas. Esta classe apresenta-se mais caracterizadora do território da AMP do que da AML.

Por fim, nas 4 592 subsecções estatísticas associadas aos *espaços de imigração* residia, em 2011, 8% da população residente nas duas áreas metropolitanas (12% na AML e apenas 2% na AMP). Esta classe sobressai pela maior expressão da dimensão relativa à imigração, revelando médias aquém da referência metropolitana para as restantes componentes.



**Figura 2** - Valores médios das componentes principais por classe socioeconómica, 2011.



**Figura 3** - Classes socioeconómicas, 2011.

## 4 Conclusões

Em síntese, as áreas metropolitanas de Lisboa e Porto constituem territórios heterogéneos e fragmentados.

A Figura 3 apresenta as seis classes socioeconómicas à escala da subsecção estatística em 2011 e revela que os territórios metropolitanos se apresentam fragmentados e heterógenos (Salgueiro, 1998, 1999) tendo em conta as características do parque habitacional e da população. A heterogeneidade apresenta-se mais acentuada na AML do que na AMP, dada a distribuição territorial e populacional mais equitativa entre as várias classes socioeconómicas na AML. A fragmentação dos espaços metropolitanos, no caso da AML, é especialmente notória em torno dos designados eixos de expansão suburbana que se formam a partir dos territórios limítrofes ao município de Lisboa. Estes eixos estendem-se, na margem Norte do Tejo, ao longo das vias de comunicação rodoviárias e ferroviárias, designadamente, nas linhas de Sintra e de Cascais e na linha da Azambuja, envolvendo territórios de Loures e de Vila Franca de Xira. Na margem Sul do Tejo, destaca-se o arco Almada-Montijo e a área centrada no município de Setúbal. Na AMP, o processo de suburbanização revela uma oposição centro-periferia que evidencia a centralidade do Porto e a dinâmica metropolitana que se estende a partir deste, formando uma coroa de expansão circunferencial que abarca Matosinhos, Maia, Valongo, Gondomar e Vila Nova de Gaia.

## Referências

- FERRÃO, J. (2003). Dinâmicas Territoriais e Trajetórias de Desenvolvimento, 1991-2001, *Revista de Estudos Demográficos*, 34: 17-25.
- INE (2014). *Tipologia Socioeconómica das Áreas Metropolitanas de Lisboa e Porto 2011*, INE, Lisboa.
- INE (2004a). *Tipologia Sócio-económica da Área Metropolitana de Lisboa 2001*, INE, Lisboa.
- INE (2004b). *Tipologia sócio-económica da Área Metropolitana do Porto 2001*, INE, Lisboa.
- INE (2000a). *Tipologia sócio-económica da Área Metropolitana de Lisboa 1991*, INE-DRLVT, Lisboa.
- INE (2000b). *Tipologia sócio-económica da Área Metropolitana do Porto à escala da subsecção estatística (censos 1991)*, INE-DRN, Lisboa.
- MALHEIROS, J. & VALA, F. (2004). A problemática da segregação residencial de base étnica – questões conceptuais e limites de operacionalização: o caso da Área Metropolitana de Lisboa, *Revista de Estudos Demográficos*, 36: 89-109.

- SALGUEIRO, T.B. (1999). Ainda em torno da fragmentação do espaço urbano, *Inforgeo*, 14: 65-76.
- SALGUEIRO, T.B. (1998). Cidade pós-moderna: Espaço fragmentado. *Território*, 4: 39-53.
- VICKERS, D. & REES, P. (2007). Creating the UK National Statistics 2001 output area classification, *Journal of the Royal Statistical Society*, 170(2): 379-403.



## Dificuldades na ANOVA com dois fatores não paramétrica com células omissas

Dulce G. Pereira<sup>1</sup> · Anabela Afonso<sup>2</sup>

**Resumo** Na análise de variância com dois fatores por vezes há combinações de níveis dos fatores que não são observadas. A inexistência destas combinações dificulta a análise dos dados, pois depende do número de células omissas e da sua localização. Já foram propostas algumas abordagens mas apenas para a ANOVA paramétrica. Uma das alternativas não paramétricas à ANOVA consiste em aplicar a abordagem paramétrica às ordens dos valores observados. Neste trabalho ilustraremos que na presença de células omissas este processo pode não ser o mais adequado, sendo necessária mais investigação nesta área.

**Palavras-chave:** ANOVA, Dados Omissos, Hipóteses de Tipo IV, Modelo de Efeitos Fixos, Modelo de Médias, Somas de Quadrados.

### 1 Introdução

Muitas experiências envolvem o estudo do efeito de dois ou mais fatores numa variável resposta. Admitindo que existem dois fatores  $R$  e  $C$ , com  $r$  e  $c$  níveis respetivamente, diz-se que o delineamento é fatorial quando cada réplica contém todas as  $rc$  combinações de níveis destes dois fatores (Montgomery, 2013). No entanto, é usual na prática encontrarem-se situações em que o número de observações por célula não é igual (delineamento desequilibrado) e até podem não ter sido observadas algumas das combinações de níveis (células omissas). Este tipo de situações por vezes ocorre devido a várias razões, que podem ser por conveniência de delineamento devido por exemplo a custos associados, mas também porque o investigador não consegue controlar em absoluto a experiência (por ex., plantas que morrem). Quando o investigador tem controlo sobre a experiência, deve ter muito cuidado para garantir que todas as combinações de níveis dos fatores sejam observadas.

Para a análise dos dados de um delineamento com células omissas foram propostos vários tipos de abordagens paramétricas. Montgomery (2013) propôs

---

<sup>1</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [dgsp@uevora.pt](mailto:dgsp@uevora.pt)

<sup>2</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

transformar o delineamento com dois fatores num delineamento a um fator com  $rc - m$ , onde  $m$  representa o número de células omissas e realizar uma análise de variância (ANOVA) a um fator. Para testar a interação entre os fatores, propôs a utilização de contrastes de interesse, linearmente independentes. Alguns *software* estatísticos, como sejam o SPSS e SAS, recomendam o uso das somas de quadrados de Tipo IV. As hipóteses de Tipo IV comparam a média de todos os níveis de cada fator com a média obtida a partir de um ou mais níveis comuns do outro fator. Estas somas coincidem com as somas de Tipo III quando não existem células omissas. Mas, vários autores têm salientado que as hipóteses de Tipo IV podem não ser únicas, pelo que o investigador deve definir as suas comparações específicas. Além disso, as conclusões dependem do número de células omissas e da posição em que surgem. Milliken e Johnson (2009) propõem que em vez de se considerar a análise através de um modelo de efeitos fixos se opte por considerar um modelo de médias (*means model*) o qual é muito simples e fácil de interpretar.

Quando as variáveis são de natureza ordinal, ou são violados os pressupostos de aplicação da ANOVA paramétrica a dois fatores, Hora e Conover (1984) propuseram um teste equivalente em que a estatística de teste  $F$  da ANOVA era obtida a partir das ordens das observações. Contudo, as conclusões referentes à possível existência de uma interação significativa não devem ser garantidas como corretas uma vez que os efeitos aditivos dos fatores nas observações originais não são extensíveis às suas ordens e vice-versa. Além disso, o estudo da interação origina probabilidades de erro de Tipo I elevadas (Shah e Madden, 2004). Em alternativa foi sugerido o uso de outras estatísticas de teste com uma distribuição aproximada a um qui-quadrado, que conduzem às mesmas conclusões relativamente aos efeitos principais dos fatores (Shah e Madden, 2004). No entanto, em nenhuma destas alternativas foi tida em conta a possível presença de células omissas.

Neste trabalho vamos aplicar, quer aos valores observados quer às respetivas ordens, as três abordagens paramétricas referidas anteriormente para a ANOVA com dois fatores com células omissas. Recorrendo a alguns exemplos de aplicação, mostraremos que este processo pode não ser o mais adequado como alternativa não paramétrica. Deste modo é necessário mais investigação nesta área.

## 2 Delineamentos com dois fatores com células omissas

As abordagens paramétricas mais usuais para analisar delineamentos com dois fatores com células omissas, e que estão disponíveis nos *software* estatísticos, são: i) calcular as somas de quadrados de Tipo IV, ii) considerar o modelo das médias, e iii) transformar o delineamento com dois fatores num delineamento com um fator.

De seguida assumiremos que os fatores  $R$  e  $C$  têm  $r$  e  $c$  níveis respetivamente e que  $n_{ij}$  representa o número de observações no nível  $i$  do fator  $R$  e nível  $j$  do fator  $C$ .

## 2.1 Somas de quadrados de tipo IV

O modelo de efeitos fixos (*effects model*) é definido por (Milliken e Johnson, 2009):

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, r; j = 1, \dots, c; k = 1, \dots, n_{ij}$$

onde  $\mu$  representa a média global,  $\tau_i$  o efeito do  $i$ -ésimo nível do fator  $R$ ,  $\beta_j$  o efeito do  $j$ -ésimo nível do fator  $C$ ,  $\gamma_{ij}$  o efeito devido à interação entre o  $i$ -ésimo nível do fator  $R$  e o  $j$ -ésimo nível do fator  $C$ , e  $\varepsilon_{ijk} \sim N(0; \sigma^2)$  e independentes. Quando a combinação  $R_i$  e  $C_j$  não é observada,  $n_{ij} = 0$ .

Quando existem células omissas Goodnight (1980) sugeriu o uso das somas de quadrados de Tipo IV. Basicamente, uma hipótese é definida como uma hipótese de Tipo IV se compara a média dos níveis de um tratamento com a média de um ou mais níveis comuns do outro tratamento. Assim, espera-se que as hipóteses sejam hipóteses de médias marginais. Quando faltam tratamentos, não se pode considerar a média em todos os níveis do outro tratamento, mas apenas em alguns dos níveis do outro fator.

Esta abordagem tem a vantagem de estar disponível em alguns *software* (por ex.: SPSS e SAS). Contudo, as hipóteses de Tipo IV podem não ser únicas e os *software* testam automaticamente algumas delas (as que são habitualmente interpretáveis), sem as especificar. Para identificar quais as hipóteses testadas é necessário analisar as funções estimadas. Caso as hipóteses não correspondam às de real interesse, recorrendo à programação o utilizador poderá definir as hipóteses a testar. No R desconhecemos a existência de algum pacote com esta abordagem. Recorrendo a alguma programação construímos algumas rotinas, que usam as funções `glm` e `glht` do R, para testar as hipóteses de interesse associadas aos contrastes definidos.

Para ilustrar, sem perda de generalidade, considere-se o delineamento com dois fatores e com duas células omissas apresentado na Tabela 1. É possível comparar  $R_1$  com  $R_2$  através da hipótese de Tipo IV  $H_0: (\mu_{11} + \mu_{13})/2 = (\mu_{21} + \mu_{23})/2$ , i.e., nos níveis  $C_1$  e  $C_3$ , ou da hipótese  $H_0: \mu_{11} = \mu_{21}$  que considera apenas o nível  $C_1$ .

**Tabela 1** – Média das células num delineamento com duas células omissas ( $\mu_{ij}$  é a média da variável resposta, ou das ordens da variável resposta, dos níveis  $i$  e  $j$  dos fatores  $R$  e  $C$ , respetivamente).

	$C_1$	$C_2$	$C_3$
$R_1$	$\mu_{11}$	$\mu_{12}$	$\mu_{13}$
$R_2$	$\mu_{21}$		$\mu_{23}$
$R_3$		$\mu_{32}$	$\mu_{33}$

Todas as possíveis hipóteses de tipo IV, associadas ao exemplo anterior, são apresentadas na Tabela 2.

**Tabela 2** – Todas as possíveis hipóteses de Tipo IV para os fatores *R* e *C*.

Para o fator <i>R</i>		Para o fator <i>C</i>	
1	$\frac{\mu_{11} + \mu_{13}}{2} = \frac{\mu_{21} + \mu_{23}}{2}$	1	$\frac{\mu_{11} + \mu_{21}}{2} = \frac{\mu_{13} + \mu_{23}}{2}$
2	$\frac{\mu_{12} + \mu_{13}}{2} = \frac{\mu_{32} + \mu_{33}}{2}$	2	$\frac{\mu_{12} + \mu_{32}}{2} = \frac{\mu_{13} + \mu_{33}}{2}$
3	$\mu_{11} = \mu_{21}$	3	$\mu_{11} = \mu_{12}$
4	$\mu_{12} = \mu_{32}$	4	$\mu_{11} = \mu_{13}$
5	$\mu_{13} = \mu_{23}$	5	$\mu_{12} = \mu_{13}$
6	$\mu_{13} = \mu_{33}$	6	$\mu_{21} = \mu_{23}$
7	$\mu_{23} = \mu_{33}$	7	$\mu_{32} = \mu_{33}$

## 2.2 Modelo das médias

Outra abordagem que muitas vezes se revela muito útil, especialmente quando existem células omissas, é abandonar o modelo de efeitos e empregar, em vez disso, o modelo das médias (*means model*), que é definido por (Milliken e Johnson, 2009):

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, r; j = 1, \dots, c; k = 1, \dots, n_{ij}$$

onde  $\mu_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij}$ ,  $\varepsilon_{ijk} \sim N(0; \sigma^2)$  e independentes, e  $n_{ij} = 0$  quando a combinação  $R_i$  e  $C_j$  não é observada.

A aplicação do modelo das médias apresenta várias vantagens: i) permite estimar cada um dos parâmetros do modelo e torna a interpretação mais fácil, ii) permite testar contrastes em simultâneo, e iii) é possível fazer inferência sobre funções com parâmetros que envolvam células omissas, desde que sejam impostas condições (por ex., a inexistência de interação) que devem estar suportadas em alguma evidência experimental (Milliken e Johnson, 2009). Contudo, é necessário recorrer à programação para definir os contrastes de interesse, o que para quem não está familiarizado com as linguagens de programação é uma desvantagem.

## 2.3 Transformar num delineamento com um fator

Montgomery (2013) propôs transformar o delineamento com dois fatores num delineamento a um fator com  $rc - m$ , onde  $m$  representa o número de células omissas, e realizar uma análise de variância (ANOVA) com um fator. Ou seja, cada nível do fator no modelo com um fator é, na verdade, uma combinação de níveis dos fatores do delineamento original. Para testar a interação entre os fatores, propôs a utilização de contrastes de interesse, linearmente independentes. Este modo de proceder tem a vantagem de ser simples, mas por outro lado a desvantagem de não permitir testar dois contrastes em simultâneo.

### 3 Exemplos

De seguida ilustra-se a aplicação das abordagens anteriores, aos dados originais e às ordens dos valores observados, em delineamentos com dois fatores e células omissas.

#### 3.1 Exemplo 1

Considere-se que para uma dada experiência foram obtidos os resultados apresentados na Tabela 3.

**Tabela 3** – Delineamento com dois fatores com células omissas (adaptado de Kleinbaum *et al.*, 1998).

	$C_1$	$C_2$	$C_3$
$R_1$	2; 7; 8; 10	2; 5	4; 5; 6
$R_2$	11; 17; 18		8; 10; 17
$R_3$		6; 9	3; 5; 8; 16

Na ANOVA paramétrica assume-se que a variável resposta é do tipo contínuo e devem-se validar os pressupostos de normalidade e homogeneidade das variâncias. Neste exemplo verificámos que estes pressupostos eram válidos (Teste Shapiro-Wilk:  $W = 0,963$ ,  $p = 0,581$ ; Teste Levene:  $F = 1,098$ ,  $p = 0,410$ ).

#### 3.1.1 Somas de quadrados de tipo IV

Recorrendo ao *software* SPSS, calcularam-se as somas de quadrados de Tipo IV associadas aos dados da Tabela 3, quer dos valores observados quer das ordens das observações, tendo-se obtido os resultados apresentados na Tabela 4.

**Tabela 4** – Resultados da ANOVA aplicada aos valores observados e às ordens.

<i>F. V.</i>	<i>g. l.</i>	Valores observados			Ordens das observações		
		<i>S. Q.</i>	<i>F</i>	<i>p</i>	<i>S. Q.</i>	<i>F</i>	<i>p</i>
<i>R</i>	2	72,8	2,348	0,132	195,7	4,182	0,038
<i>C</i>	2	31,4	1,013	0,388	45,4	1,023	0,385
<i>R × C</i>	2	3,1	0,100	0,906	10,4	0,234	0,794
Erro	14	217,1			310,8		

Neste caso, as hipóteses de Tipo IV testadas, automaticamente, foram:

- $H_0(R)$ :  $(\mu_{12} + \mu_{13})/2 = (\mu_{32} + \mu_{33})/2$  e  $\mu_{23} = \mu_{33}$
- $H_0(C)$ :  $(\mu_{11} + \mu_{21})/2 = (\mu_{13} + \mu_{23})/2$  e  $(\mu_{12} + \mu_{32})/2 = (\mu_{13} + \mu_{33})/2$
- $H_0(R \times C)$ :  $\mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} = 0$  e  $\mu_{12} - \mu_{13} - \mu_{32} + \mu_{33} = 0$ .

Ao nível de significância de 5%, a decisão relativa à existência de efeito do fator *R* difere com o tipo de abordagem considerada. Com base nos dados originais não

se rejeita a hipótese de ausência de efeito, decidindo-se o contrário com base nas ordens das observações.

Contudo, por ex., quando se troca a ordem dos níveis do fator *R* (Tabela 5), as somas de Tipo IV associadas a este fator também se alteram bem como a decisão final (Tabela 6). Neste caso, tanto com base nos dados originais como nas ordens das observações conclui-se que existe efeito do fator *R* na variável resposta. Esta alteração deve-se ao facto de terem sido testadas outras hipóteses para o fator *R*:  $H_0(R): (\mu_{12} + \mu_{13})/2 = (\mu_{32} + \mu_{33})/2$  e  $(\mu_{21} + \mu_{23})/2 = (\mu_{11} + \mu_{13})/2$ .

**Tabela 5** – Troca da ordem das linhas da Tabela 3.

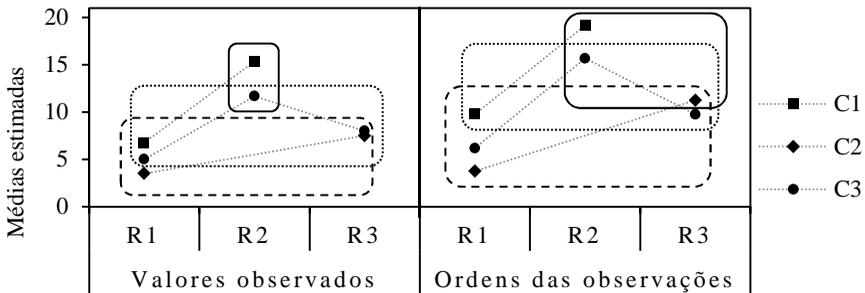
	$C_1$	$C_2$	$C_3$
$R_2$	11; 17; 18		8; 10; 17
$R_3$		6; 9	3; 5; 8; 16
$R_1$	2; 7; 8; 10	2; 5	4; 5; 6

**Tabela 6** – Resultados da ANOVA aplicada aos valores observados e às ordens.

<i>F.V.</i>	<i>g.l.</i>	Valores observados			Ordens das observações		
		<i>S.Q.</i>	<i>F</i>	<i>p</i>	<i>S.Q.</i>	<i>F</i>	<i>p</i>
<i>R</i>	2	191,8	6,185	0,012	310,7	6,998	0,008

### 3.1.2 Modelo das médias

Quando se aplica o modelo das médias aos dados originais e às ordens das observações obtêm-se conclusões similares (Figura 1). A registrar apenas que o tratamento  $R_1 \times C_3$  é posicionado em subgrupos diferentes consoante a análise tenha recaído sobre os dados originais ou sobre as ordens das observações.



**Figura 1** – Subconjuntos homogêneos das médias. (As médias localizadas no mesmo retângulo não são significativamente diferentes a 5%).

### 3.1.3 ANOVA com um fator

Na ANOVA com um fator a hipótese testada corresponde a

$$H_0: \mu_{11} = \mu_{12} = \mu_{13} = \mu_{21} = \mu_{23} = \mu_{32} = \mu_{33}$$

As conclusões da ANOVA aplicada quer aos dados originais quer às ordens das observações é a mesma, i.e., rejeitar  $H_0$  (Tabela 7).

**Tabela 7** – Resultados da ANOVA aplicada aos valores observados e às ordens.

F.V.	g.l.	Valores observados				Ordens das observações			
		S.Q.	M.Q.	F	p	S.Q.	M.Q.	F	p
Fator	6	272,1	45,3	2,924	0,046	453,3	75,5	3,403	0,028
Erro	14	217,1	15,5			310,8	22,2		

O próximo passo é formar as comparações de interesse (contrastes). A título ilustrativo, apresentam-se apenas os contrastes que permitem testar a interação e o efeito do fator  $R$  (Tabela 8). Ao nível de significância de 5%, as conclusões não são concordantes apenas na hipótese que compara o nível  $R_1$  com  $R_2$  no nível  $C_3$ . Com base nos dados originais não há evidência estatística de que os níveis  $R_1$  e  $R_2$  difiram no nível  $C_3$  concluindo-se o contrário quando se consideram as ordens das observações.

**Tabela 8** – Resultados dos contrastes.

Contrastes	g.l.	Valores observados			Ordens das observações		
		$\hat{c}$	t	p	$\hat{c}$	t	p
$H_0(R \times C)$ :							
$\mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} = 0$	14	-1,92	-0,435	0,670	0,08	0,016	0,988
$\mu_{12} - \mu_{13} - \mu_{32} + \mu_{33} = 0$	14	-1,00	-0,202	0,843	-3,92	-0,661	0,520
$H_0(R)$ :							
$(\mu_{11} + \mu_{13})/2 = (\mu_{21} + \mu_{23})/2$	14	-7,6	-3,464	0,004	-9,5	-3,591	0,003
$(\mu_{12} + \mu_{13})/2 = (\mu_{32} + \mu_{33})/2$	14	-3,5	-1,413	0,180	-5,5	-1,870	0,083
$\mu_{11} = \mu_{21}$	14	-8,6	-2,854	0,013	-9,4	-2,617	0,020
$\mu_{12} = \mu_{32}$	14	-4,0	-1,016	0,327	-7,5	-1,592	0,134
$\mu_{13} = \mu_{23}$	14	-6,7	-2,074	0,057	-9,5	-2,470	0,027
$\mu_{13} = \mu_{33}$	14	-3,0	-0,998	0,335	-3,6	-0,996	0,336
$\mu_{23} = \mu_{33}$	14	3,7	1,219	0,243	5,9	1,644	0,122

### 3.2 Exemplo 2

Como segundo exemplo, considere-se o conjunto de dados apresentado na Tabela 9. À semelhança do exemplo anterior, verificámos que os pressupostos da ANOVA paramétrica eram válidos (Teste Shapiro-Wilk:  $W = 0,967$ ,  $p = 0,676$ ; Teste Levene:  $F = 0,547$ ,  $p = 0,764$ ).

**Tabela 9** – Delineamento com dois fatores com células omissas (adaptado de Kleinbaum *et al.*, 1998).

	$C_1$	$C_2$	$C_3$
$R_1$	20; 21; 23		8; 10; 12
$R_2$		6; 9	3; 5; 8; 8
$R_3$	2; 7; 8; 11	1; 5	2; 4; 9

### 3.2.1 Somas de quadrados de Tipo IV

Na Tabela 10 apresentam-se os resultados relativos às somas de quadrados de Tipo IV associadas aos dados da Tabela 9, quer com base nos valores observados quer nas ordens dos valores observados. Estes resultados foram obtidos com auxílio do *software SPSS*.

**Tabela 10** – Resultados da ANOVA aplicada aos valores observados e às ordens.

<i>F. V.</i>	<i>g. l.</i>	Valores observados			Ordens das observações		
		<i>S. Q.</i>	<i>F</i>	<i>p</i>	<i>S. Q.</i>	<i>F</i>	<i>p</i>
<i>R</i>	2	299,1	18,834	<0,001	264,0	6,345	0,011
<i>C</i>	2	153,2	9,649	0,002	49,9	1,199	0,331
<i>R × C</i>	2	70,4	4,431	0,032	26,9	0,646	0,539
Erro	14	299,1			291,3		

Neste exemplo, as hipóteses de Tipo IV testadas, automaticamente, foram:

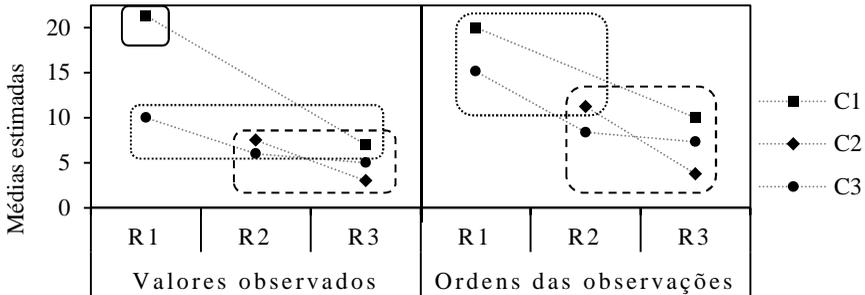
- $H_0(R)$ :  $(\mu_{11} + \mu_{13})/2 = (\mu_{31} + \mu_{33})/2$  e  $(\mu_{22} + \mu_{23})/2 = (\mu_{32} + \mu_{33})/2$
- $H_0(C)$ :  $(\mu_{11} + \mu_{31})/2 = (\mu_{13} + \mu_{33})/2$  e  $(\mu_{22} + \mu_{32})/2 = (\mu_{23} + \mu_{33})/2$
- $H_0(R \times C)$ :  $\mu_{11} - \mu_{13} - \mu_{31} + \mu_{33} = 0$  e  $\mu_{22} - \mu_{23} - \mu_{32} + \mu_{33} = 0$ .

Relativamente à interação entre os dois fatores obtêm-se conclusões contraditórias, consoante a abordagem considerada. Com base nos dados originais, conclui-se que existe interação significativa ao passo que com base nas ordens das observações não há evidência estatística dessa interação. Este resultado evidencia que existência de interação nos dados originais não implica a existência de interação nas ordens das observações, tal como já foi salientado por vários autores (e.g., Hora & Conover, 1984). Além disso, ambas as estatísticas de teste dos efeitos principais são significativas quando se consideram os dados originais, ao passo que com base nas ordens das observações a estatística associada ao efeito *C* não é significativa.

### 3.2.2 Modelo das médias

Considerando o modelo das médias, constituem-se três subconjuntos com os dados originais e apenas dois com base nas ordens das observações (Figura 2). A principal diferença reside no facto de com base nos dados originais se concluir que a média do tratamento  $R_1 \times C_1$  difere das restantes médias, mas a média das ordens

deste tratamento não difere da média das ordens dos tratamentos  $R_1 \times C_3$  e  $R_2 \times C_2$ .



**Figura 2** – Subconjuntos homogêneos das médias.

### 3.2.3 ANOVA com um fator

Na ANOVA com um fator a hipótese testada corresponde a

$$H_0: \mu_{11} = \mu_{13} = \mu_{22} = \mu_{23} = \mu_{31} = \mu_{32} = \mu_{33}.$$

Quer com base nos dados originais quer com base nas ordens das observações, conclui-se que os resultados das ANOVAS são concordantes e conduzem à mesma decisão (dados originais:  $F_{6,14} = 13,297$ ,  $p < 0,001$ ; ordens das observações:  $F_{6,14} = 3,783$ ,  $p = 0,019$ ).

Tendo em conta os resultados obtidos quando se consideraram as somas de Tipo IV, focamo-nos apenas nos contrastes que permitem testar a existência de interação.

Só concluímos que existe interação significativa ( $\alpha = 5\%$ ) se considerarmos os dados originais e usarmos a hipótese definida no primeiro contraste (Tabela 11).

Portanto, a decisão varia com o contraste considerado.

**Tabela 11** – Resultados dos contrastes.

Contrastes	<i>g.l.</i>	Valores observados			Ordens das observações		
		$\hat{c}$	<i>t</i>	<i>p</i>	$\hat{c}$	<i>t</i>	<i>p</i>
$H_0(R \times C)$ :							
$\mu_{11} - \mu_{13} - \mu_{31} + \mu_{33} = 0$	14	9,33	2,963	0,010	2,17	0,425	0,677
$\mu_{22} - \mu_{23} - \mu_{32} + \mu_{33} = 0$	14	3,50	0,987	0,340	6,46	-1,125	0,279

## 4 Conclusão

As hipóteses de Tipo IV testadas automaticamente pelos *software* dependem da posição em que surgem as células omissas e do seu número. Estas hipóteses não são únicas e não comparam exatamente os mesmos valores médios, o que pode dar

origem a decisões discordantes. Nos *software* as hipóteses a testar são dadas automaticamente sendo necessário pedir informação adicional para se saber quais foram as hipóteses testadas. Recorrendo à programação, o utilizador pode escolher as hipóteses de interesse.

Com o modelo das médias e com a ANOVA com dois fatores considerando as somas de quadrados de Tipo IV podemos testar exatamente as mesmas hipóteses. No entanto, no modelo das médias é mais simples e fácil de controlar as hipóteses de real interesse.

Quando se opta pela realização dos contrastes na ANOVA com um fator, como nem todos os contrastes geram a mesma conclusão, à semelhança do que acontece quando se usam as hipóteses de Tipo IV, é preciso saber quais as hipóteses que se pretendem testar.

Com este estudo salientámos o quão importante é saber o que se pretende testar, de forma a evitar conclusões erróneas. Como as conclusões obtidas com base nos dados originais e com base nas ordens das observações nem sempre coincidem, se os dados forem quantitativos, recomendamos que primeiro se deve tentar transformar os dados originais de forma a verificarem os pressupostos da ANOVA paramétrica.

As alternativas não paramétricas à ANOVA, de um modo geral, não permitem a existência de células omissas, pelo que é necessária mais investigação nesta área.

## Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto «UID/MAT/04674/2019(CIMA)».

## Referências

- GOODNIGHT, J. H. (1980). Test of hypotheses in fixed effects linear models. *Communications in Statistics*, 9: 167-180.
- HORA, S. & CONOVER, W. J. (1984). The F-statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, 79, 668–673.
- KLEINBAUM, D. G., KUPPER, L. L., MULLER, K. E., & NIZAM, A. (1998) *Applied Regression Analysis and Other Multivariable Methods*. Duxbury.
- MILLIKEN, G. A. & JOHNSON, D. E. (2009). *Analysis of Messy Data. Volume 1: Designed Experiments*, Second Edition, Chapman and Hall/CRC.
- MONTGOMERY, D. G. (2013). *Design and Analysis of Experiments*, 8ª Edição. John Wiley & Sons.
- SHAH, D. A. & MADDEN, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, 94, 33-43.

## Perfis de estudantes no contexto do Empreendedorismo: Análise de Correspondências Múltiplas e Análise de *Clusters*

Áurea Sousa<sup>1</sup> · Gualter Couto<sup>2</sup> · Nélia Branco<sup>3</sup> · Osvaldo Silva<sup>4</sup> · Helena Bacelar-Nicolau<sup>5</sup>

**Resumo** A promoção do empreendedorismo pelas instituições de ensino superior assume um papel preponderante. Um dos objetivos deste trabalho é o de aferir o modo como a Universidade dos Açores pode estimular o interesse dos alunos na criação de negócios. Pretende-se, ainda, relacionar as principais dificuldades subjacentes à criação de um negócio próprio com a formação científica dos estudantes. Apresentam-se as conclusões mais relevantes obtidas com base na análise dos dados recolhidos, com a aplicação da Análise de Correspondências Múltiplas e da Análise de *Clusters*. Verificou-se que a maioria dos estudantes tem uma baixa propensão empreendedora, tendo também sido encontrada uma tipologia de iniciativas académicas promotoras do empreendedorismo, com base nas perceções dos estudantes.

**Palavras-chave:** Empreendedorismo, Propensão Empreendedora, Perfil, Estudante, Análise Classificatória, Análise de Correspondências Múltiplas.

### 1 Introdução

O empreendedorismo é entendido como um fator potenciador do desenvolvimento económico e social dos países e regiões através da criação de novas empresas, de novos postos de trabalho, da promoção da competitividade e do desenvolvimento de negócios inovadores. São vários os autores que corroboram esta perspetiva,

---

<sup>1</sup> Universidade dos Açores, Faculdade de Ciências e Tecnologia; e CEEAplA, [aurea.st.sousa@uac.pt](mailto:aurea.st.sousa@uac.pt)

<sup>2</sup> Universidade dos Açores, Faculdade de Economia e Gestão; e CEEAplA, [gualter.mm.couto@uac.pt](mailto:gualter.mm.couto@uac.pt)

<sup>3</sup> CMRG, Ribeira Grande, [nelia.cavaco.branco@gmail.com](mailto:nelia.cavaco.branco@gmail.com)

<sup>4</sup> Universidade dos Açores, Faculdade de Ciências e Tecnologia; e Centro Interdisciplinar de Ciências Sociais – CICS.UAc/CICS.NOVA.UAc, [osvaldo.dl.silva@uac.pt](mailto:osvaldo.dl.silva@uac.pt)

<sup>5</sup> Universidade de Lisboa, Faculdade de Psicologia e ISAMB/FMUL, [hbacelar@psicologia.ulisboa.pt](mailto:hbacelar@psicologia.ulisboa.pt)

como se poderá consultar em Wennekers e Thurik (1999), Grossman (2009), Aghion e Howitte (2009), Cetin (2013), Galindo e Mendez-Picazo (2014), Aparicio *et al.* (2016) e Dhahri e Omri (2018), entre outros. De acordo com esta perspetiva, o papel dos empreendedores tem vindo a ser valorizado, sendo entendidos como agentes de mudança e de desenvolvimento dos contextos sociais onde se encontram inseridos.

A educação empresarial do século XXI é algo que transcende o que vem nos livros clássicos ou em casos de estudo isolados. O ensino deve estimular a capacidade de análise crítica em geral e, em particular, no contexto do empreendedorismo, a identificação de problemas e dos seus respetivos modos de resolução, de modo a capacitar os educandos para o aproveitamento das oportunidades emergentes de negócio (Sousa, 2015; Maas e Jones, 2017). Para isso, são necessários profissionais motivados e capacitados para enfrentarem os desafios cada vez mais exigentes que ocorrem ao longo da vida. O empreendedorismo é, assim, uma competência que extravasa as características pessoais que normalmente estão associadas ao processo de liderança. Os empreendedores de sucesso são os mais qualificados para gerir o risco e transformar oportunidades em negócios escaláveis, sendo de salientar que através do ensino é possível aprender e treinar competências intelectuais (e.g., conhecimentos sobre os mercados internacionais, a escalabilidade dos negócios e alguns métodos de análise de dados) que ajudam no alcance do sucesso. A importância da aquisição de competências na área de Estatística pelos cidadãos em geral, e em particular pelos estudantes que pretendem ingressar no mundo empresarial, é bem reconhecida (e.g., Silva e Sousa, 2016).

É crucial identificar quais os fatores que funcionam como obstáculos ou como potenciadores do desenvolvimento da prática empreendedora. Da sua respetiva análise, será possível delinear um conjunto de indicadores que possam servir de diretrizes para as entidades de ensino, bem como para as entidades governamentais, no sentido de estas poderem desenvolver ações e políticas equilibradas de promoção e de apoio à atividade empreendedora.

Existe uma crescente preocupação relativamente ao desenvolvimento de iniciativas e atividades que promovam o espírito e a propensão empreendedora dos estudantes do ensino superior (e.g., Maas e Jones, 2017), de modo a que surjam cada vez mais empresas produtoras de bens e serviços inovadores e com forte valor acrescentado para a sociedade. Nesse sentido, as instituições de ensino superior assumem um importante papel como impulsionadoras do empreendedorismo junto dos estudantes. Cabe a estas instituições levar a cabo diversas iniciativas e atividades que desenvolvam o espírito empreendedor dos seus estudantes, para que estes se possam tornar mais empreendedores. No entanto, nem sempre as atividades desenvolvidas pelas universidades estão ajustadas aos desafios que o empreendedorismo enfrenta hoje no mundo global, em que a inovação faz cada vez mais parte da proposta de valor. No contexto da economia global, o papel das universidades é cada vez mais o de ajudar a identificar e a transmitir as competências necessárias para o sucesso dos negócios, ao mesmo tempo que é

pertinente entender que a educação empresarial não se esgota num modelo de ensino de três ou quatro anos, constituindo um processo de aprendizagem ao longo da vida.

A inovação toma hoje em dia o papel central no ecossistema empresarial e nos processos de criação de valor (Maradana et al., 2017). Entender as necessidades dos clientes e o valor da utilidade que é retirada dos produtos/serviços, mesmo que emocional, é tão importante como a forma eficiente e célere de as satisfazer, num mundo em que a customização será a referência. Contudo, o sucesso da criação de negócios de elevado valor é mais fácil de alcançar quando estão reunidas as pessoas certas, que fazem das empresas sistemas criadores de oportunidades de forma acelerada. Isto requer boas qualificações e competências técnicas e um forte sentido de mercado, ou seja, de deteção de oportunidades à escala global dos negócios.

Um dos objetivos deste trabalho é o de aferir o modo como a Universidade dos Açores pode estimular o interesse dos alunos na criação de negócios. Pretende-se, ainda, relacionar algumas das dificuldades perspetivadas pelos estudantes em relação à inicialização de um novo negócio com a sua área científica.

## 2 Descrição dos dados e procedimentos metodológicos

A amostra é constituída por 305 estudantes (36% do sexo masculino e 64% do sexo feminino) da Universidade dos Açores, 29.5% da área de Economia e Gestão e os restantes (70.5%) de outras áreas científicas, sendo que 70.1% têm idades entre os 18 e os 24 anos, 16% estão na faixa etária dos 25-34 anos e 13.9% têm 35 ou mais anos. Quanto ao estatuto dos estudantes, apenas 23.9% dos inquiridos estão na situação de “Trabalhador Estudante”. No que concerne à experiência profissional mais recente, 18.6% dos inquiridos não a têm, 30.9% estão ou já estiveram na situação de estágio ou em outra situação e 52.2% estão ou já tiveram um trabalho a tempo inteiro.

Os dados foram recolhidos, presencialmente, em 2011, conforme descrito em Branco (2013). O questionário utilizado contém algumas variáveis sociodemográficas, tais como o “Género”, o “Grupo etário”, a “Área científica”, a “Existência de familiares empreendedores” e quatro grupos de itens (G1 a G4), apresentados em Branco (2013), que visam aferir: o conhecimento dos estudantes a nível do empreendedorismo e a sua familiaridade com este tópico; as principais dificuldades em relação à iniciação de negócios; e a sua opinião relativamente a algumas atividades e iniciativas que a Universidade pode desenvolver a nível da sensibilização e orientação para o empreendedorismo. Estes itens são avaliados numa escala de Likert: 1-*Discordo totalmente* (DT), 2-*Discordo* (D), 3-*Não discordo nem concordo* (NDNC), 4-*Concordo* (C), 5-*Concordo totalmente* (CT), sendo de referir que não foram efetuadas somas das pontuações obtidas em conjuntos de itens.

Foram utilizados métodos de Análise de Dados Multivariados, nomeadamente, de Análise Classificatória Hierárquica Ascendente (ACHA), a Análise de Correspondências Múltiplas (ACM) e o método não hierárquico das  $k$ -médias ( $k$ -means).

Para procurarmos uma tipologia dos itens referentes às iniciativas académicas promotoras do empreendedorismo, foi efetuada a ACHA desses itens com base no coeficiente de afinidade (e.g., Bacelar-Nicolau, 1980, 1988) e em critérios de agregação probabilísticos, no âmbito da metodologia VL (e.g., Bacelar-Nicolau, 1980, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1970, 1981, 2016), designadamente o Algoritmo de Validade da Ligação (AVL), o AVI e o Algoritmo de Validade da Ligação-Bacelar (AVB). A seleção da melhor partição foi efetuada com base na estatística global de níveis, *STAT* (e.g., Lerman, 1970, 1981; Bacelar-Nicolau, 1980, 1987).

A ACM (e.g., Benzécri, 1992; Greenacre e Blasius, 2006) foi utilizada com o intuito de se estudar as associações entre mais de duas variáveis categóricas e entre as categorias dessas variáveis, assim como para se obter as coordenadas das categorias e mapas perceptuais, a partir dessas coordenadas, os quais permitem percecionar visualmente essas associações. Para esse efeito, selecionaram-se as variáveis que apresentaram associações estatisticamente significativas ( $p < 0.05$ ), segundo o teste de independência do qui-quadrado, com pelo menos uma das restantes variáveis a serem consideradas na ACM. É importante salientar que diferentes combinações das características em análise, no mapa perceptual, sugerem a existência de três grupos de indivíduos com diferentes perfis. As coordenadas dos indivíduos nas três primeiras dimensões resultantes da ACM foram, então, utilizadas como novas variáveis no método não hierárquico das  $k$ -médias (MacQueen, 1967), com  $k=3$ , no contexto da articulação entre a ACM e a Análise Classificatória. A consideração das três primeiras dimensões resultantes da ACM deve-se ao facto de estas explicarem, como veremos na Secção 3.2, uma elevada quantidade de informação (83.1% da variação dos dados), reduzindo a dimensionalidade dos dados, e ao interesse na comparação com outros resultados obtidos no âmbito deste projeto.

### 3 Resultados

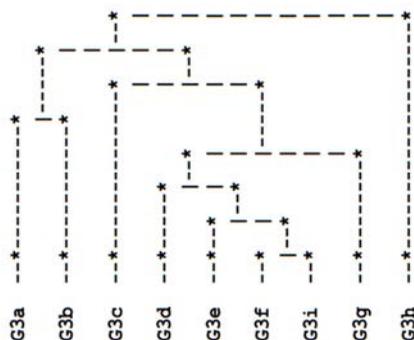
#### 3.1 ACHA de iniciativas académicas promotoras do empreendedorismo

Os itens G3a (“*Consciencializar acerca do empreendedorismo como opção de carreira*”), G3b (“*Fornecer aos alunos ideias para criarem os seus próprios negócios*”), G3c (“*Disponibilizar licenciaturas ou mestrados em empreendedorismo*”), G3d (“*Disponibilizar projetos de trabalho em*

*empreendedorismo*”), G3e (“*Organizar conferências / workshops sobre empreendedorismo*”), G3f (“*Aproximar os alunos das redes de contacto necessárias para se começar um negócio*”), G3g (“*Permitir a utilização das suas instalações por parte de empresas dirigidas por alunos*”), G3h (“*Dotar os alunos dos meios financeiros necessários para se começar um novo negócio*”) e G3i (“*Colocar os alunos empreendedores em contacto com os outros*”), respeitantes a algumas iniciativas académicas promotoras do empreendedorismo, que podem ser desenvolvidas pela Universidade, foram alvo de concordância (total ou parcial) por parte da maioria dos inquiridos, respetivamente, 54.8%, 56.5%, 55.2%, 63.2%, 67.2%, 79.9%, 57.2%, 56.4% e 70.5% dos inquiridos, sendo de realçar a importância atribuída às iniciativas subjacentes aos itens G3f e G3i. Os dendrogramas resultantes da ACHA destes nove itens apresentam no nível três uma mesma partição em seis classes, sendo esta a melhor partição, de acordo com o valor da estatística *STAT*, que atinge aqui o seu máximo ( $STAT = 3.8203$ ): C1: {G3a}; C2: {G3b}; C3: {G3c}; C4: {G3d, G3e, G3f, G3i}; C5: {G3g}; C6: {G3h}.

Esta partição faz, portanto, ressaltar uma classe forte, a Classe C4, que inclui as iniciativas relativas ao desenvolvimento de projetos de trabalho em empreendedorismo, à organização de conferências/workshops sobre empreendedorismo, a redes de contacto úteis para a criação de um negócio e à colocação dos alunos empreendedores em contacto uns com os outros. Estas iniciativas são aquelas que foram mais valorizadas pelos estudantes, conforme pode ser observado pelas percentagens supracitadas.

A Figura 1 mostra o dendrograma obtido com os métodos *AVI* e *AVB*, onde se pode observar a elevada proximidade entre os itens G3f e G3i, correspondendo à importância de, simultaneamente, aproximar os alunos, quer das redes de contacto necessárias para se começar um negócio, quer dos alunos empreendedores, a que se juntam, logo a seguir, o interesse na organização de conferências/workshops (G3e) e em disponibilizar projetos de trabalho (G3d), respeitantes ao empreendedorismo.



**Figura 1-** Dendrograma obtido pelos métodos *AVI/AVB*: iniciativas académicas promotoras do empreendedorismo.

A Figura 1 permite, ainda, observar uma segunda classe, resultante da reunião dos itens G3a (“*Consciencializar acerca do empreendedorismo como opção de carreira*”) e G3b (“*Fornecer aos alunos ideias para criarem os seus próprios negócios*”) no nível 5 do processo de agregação, constituindo um outro tipo de iniciativas conducente à mobilização empreendedora dos estudantes. Enfim, é interessante notar, no que respeita a esta tipologia de iniciativas académicas promotoras do empreendedorismo, que aquela que é mais dissimilar das outras, no que concerne à distribuição das respostas relativas ao grau de concordância (ou discordância) expressa pelos inquiridos, é a G3h, “*Dotar os alunos dos meios financeiros necessários para se começar um novo negócio*”, que se mantém isolada, até ao último nível da hierarquia.

### 3.2 Articulação entre a ACM e a Análise Classificatória dos estudantes

A ACM foi efetuada considerando sete variáveis ativas e cinco suplementares, descritas na Tabela 1, sendo de referir que os itens G1a (“*Conheço as técnicas para identificar o que o Mercado quer*”), G1b (“*Compreendo o tipo de questões subjacentes à criação de empresas*”), G4a (“*Leio regularmente livros/artigos sobre empreendedorismo e inovação*”) e G4b (“*Participo regularmente em conferências sobre empreendedorismo*”) foram previamente recodificados em três categorias (1-*Discordo* (D), 2-*Não Discordo Nem Concordo* (NDNC), 3-*Concordo* (C)), de modo a serem reunidas as condições necessárias para a aplicação do teste de independência do qui-quadrado e a se obter uma melhor perceção visual das associações entre as categorias das variáveis consideradas na ACM.

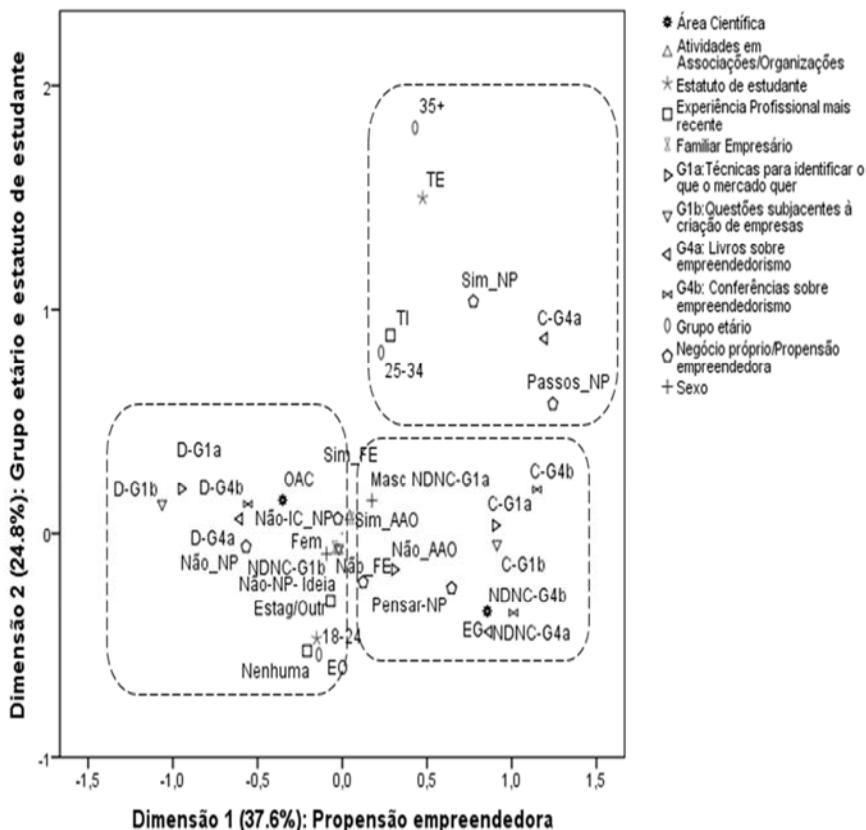
**Tabela 1** – Variáveis utilizadas na ACM e respetivas categorias.

Área Científica ( <i>Economia e Gestão (EG), Outras Áreas Científicas (OAC)</i> )
Participação em Atividades em Associações/Organizações* ( <i>Não_AAO, Sim_AAO</i> )
Estatuto de estudante ( <i>Trabalhador estudante (TE), Estudante Ordinário (EO)</i> )
Experiência profissional mais recente* ( <i>Nenhuma, Estágio/Outra (Estag/Outr), Trabalho a Tempo Inteiro (TI)</i> )
Familiar Empresário (FE)* ( <i>Não_FE, Sim_FE</i> )
G1a, G1b, G4a, G4b ( <i>Discordo (D), Não Discordo Nem Concordo (NDND), Concordo (C)</i> )
Grupo etário (18-24, 25-34, 35+)
Negócio próprio/Propensão empreendedora* ( <i>Não-NP, Não-IC-NP, Não-NP-Ideia, Pensar-NP, Passos-NP, Sim-NP</i> )
Sexo* ( <i>Masc, Fem</i> )

\*Variáveis suplementares

Os significados de algumas das siglas utilizadas para designar as categorias de variáveis apresentadas na Tabela 1 são clarificados mais à frente, neste texto.

As três dimensões resultantes da ACM explicam cerca de 83.1% (37.6% + 24.8% +20.6%) da variação dos dados. A Figura 2 mostra o mapa perceptual referente ao cruzamento das duas primeiras dimensões. As variáveis G1a, G1b, -G4a, G4b e “Área científica” estão associadas à dimensão 1, enquanto o “Grupo etário” e o “Estatuto de estudante” são as variáveis mais importantes para a dimensão 2.



**Figura 2** - Mapa perceptual resultante da ACM: (Método de normalização: *Variable Principal*).

A primeira dimensão opõe, em geral, os estudantes que frequentam cursos na área da Economia e Gestão (EG), que tendem a ser capazes de identificar as necessidades de mercado (C-G1a), a compreender o tipo de questões subjacentes à criação de empresas (C-G1b), a participar regularmente (C-G4b) ou esporadicamente (NDNC-G4b) em conferências sobre empreendedorismo e

inovação e a pensar na possibilidade de criação de um negócio próprio (Pensar-NP) ou a estar a poucos passos de o criarem (Passos\_NP); aos estudantes de outras áreas científicas (OAC), que tendem a desconhecer as técnicas para identificar as necessidades do mercado (D-G1a) e o tipo de questões referentes à criação de empresas (D-G1b), que não leem regularmente livros/artigos sobre empreendedorismo e inovação (D-G4a), que não participam regularmente em conferências sobre empreendedorismo e inovação (D-G4b) e que não têm um negócio próprio nem têm interesse em tê-lo (Não\_NP).

A segunda dimensão opõe os estudantes com 25 anos ou mais, que são trabalhadores estudantes (TE), leem regularmente livros sobre empreendedorismo e inovação (C-G4a), trabalham a tempo inteiro (TI), têm negócio próprio (Sim\_NP), ou estão a poucos passos de o criarem (Passos\_NP); aos estudantes mais jovens (18-24 anos), que são geralmente estudantes em regime ordinário (EO), sem experiência profissional (Nenhuma), ou que estão numa situação de estágio ou de trabalho a tempo parcial (Estag/Outro).

Embora não seja aqui apresentado o mapa percetual que cruza a primeira dimensão com a terceira dimensão, é de salientar que a terceira dimensão opõe os estudantes com maior investimento na área do empreendedorismo, isto é aqueles que leem regularmente livros e/ou artigos sobre empreendedorismo (C-G4a), participam regularmente em conferências sobre o tema (C-G4b), conhecem as técnicas para identificar o que o mercado quer (C-G1a) e compreendem o tipo de questões subjacentes à criação de empresas (C-G1b); aos que não concordam nem discordam no que se refere a estas questões (NDNC-G1a, NDNC-G1b, NCND-G4a, NCND-G4b).

A análise classificatória dos indivíduos utilizando o método das *k*-médias, com base nas coordenadas dos indivíduos nas três primeiras dimensões resultantes da ACM, foi efetuada considerando uma partição em três classes, a que correspondem respetivamente 55.7%, 32.1% e 12.1% dos estudantes da amostra, de forma a maximizar as diferenças entre os perfis associados às classes.

A primeira classe inclui os indivíduos com uma baixa propensão empreendedora, sendo que a maioria destes frequenta cursos que não se enquadram na área de Economia e Gestão (92.2%) e tem idades compreendidas entre os 18 e os 24 anos (71.3%), 50.9% dos indivíduos incluídos nesta classe não possuem negócio próprio mas não põem de parte essa possibilidade, 56.6% não têm empresários na família, 52.1% nunca participaram em atividades em associações/organizações, e a maioria está numa posição de discordância (total ou parcial) com as afirmações associadas aos itens G1a (54.8%), G4a (92.2) e G4b (96.4%).

A segunda classe inclui os indivíduos com algum potencial empreendedor, que frequentam maioritariamente cursos na área de Economia e Gestão (61.2%). A maioria dos indivíduos que se inserem nesta classe tem entre 18 e 24 anos (76.5%), nunca participou em atividades em associações/organizações (59.2%) e encontra-se

numa posição de indiferença (NDNC) no que se refere às afirmações relativas aos itens G1a (72.2%), G1b (50.5%), G4a (76.5%) e G4b (67.3%).

A terceira classe inclui maioritariamente estudantes com 25 ou mais anos (52.7%), que têm participado em atividades em associações/organizações (64.9%), que estão numa posição de concordância total com as afirmações referentes aos itens G1a (52.8%), G1b (86.1%), G4a (67.6%) e G4b (51.4%). A proporção de estudantes trabalhadores nesta classe é maior (29.7%), comparativamente às das classes 1 (23.4%) e 2 (22.7%).

A descrição pormenorizada dos indivíduos pertencentes a cada uma dessas classes, considerando as variáveis utilizadas na ACM, pode ser encontrada em Sousa *et al.* (2017).

Os resultados do presente estudo estão em consonância com a ideia de que a educação para o empreendedorismo ajuda a fomentar o empreendedorismo (Cardoso *et al.*, 2015; Sousa *et al.*, 2015).

## 4 Conclusão

Numa região insular, o desenvolvimento da propensão empreendedora dos estudantes poderá ter implicações a nível do desenvolvimento regional e também das oportunidades de emprego.

Os modelos de ACHA, aplicados à sub-matriz que contém a opinião dos estudantes em relação a nove iniciativas académicas promotoras do empreendedorismo, conduziram-nos a uma tipologia dessas iniciativas.

A ACM, considerando sete variáveis ativas e cinco suplementares, evidenciou três conjuntos de estudantes e permitiu-nos estudar as associações entre as categorias de algumas variáveis relevantes. A aplicação do método não hierárquico das *k*-médias, considerando as coordenadas (*scores*) dos indivíduos nas três primeiras dimensões resultantes da ACM, confirmou esses três perfis de estudantes, respetivamente, com baixa (56%), média (32%) e alta (12%) propensão empreendedora.

No contexto económico e social atual, a promoção do empreendedorismo pelas instituições de ensino superior assume um papel cada vez mais importante (facto apoiado pelos resultados do presente estudo). No entanto, o interesse pelo empreendedorismo não deve ser imposto, mas pode ser estimulado. O incremento da percentagem de diplomados no perfil com elevada propensão empreendedora poderá contribuir também para o desenvolvimento regional.

As perspetivas de trabalho futuro incluem a realização de outros estudos, com a adaptação do questionário utilizado, de forma a que sejam contempladas mais algumas questões, incluindo uma escala para a avaliação do espírito empreendedor dos estudantes. Por outro lado, o facto de não estarem representados todos os cursos existentes na Universidade, à data do estudo, constitui uma limitação deste trabalho.

## Referências

- AGHION, P. & HOWITT, P. (2009). *The economics of growth*, MIT PRESS, Cambridge.
- APARICIO, S., URBANO, D. & AUDRETSCH, D. (2016). *Institutional factors, opportunity entrepreneurship and economic growth: Panel data evidence. Technological Forecasting and Social Change*, 102, 45-61. doi:10.1016/j.techfore.2015.04.006.
- BACELAR-NICOLAU, H. (1980). *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória, Tese de Doutorado, FCL, Universidade de Lisboa.*
- 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, Springer - Verlag, New York, 73-79.
- BACELAR-NICOLAU, H. (1988). *Two probabilistic models for classification of variables in frequency tables.* In: BOCK, H.-H. (Eds.), *Classification and related methods of data analysis*, Elsevier Sciences Publishers B.V., North Holland, pp. 181-186.
- BENZÉCRI, J. P. (1992). *Correspondence Analysis Handbook*, Marcell Dekker, New York.
- BRANCO, N. (2013). *Propensão Empreendedora dos estudantes de ensino superior na Região Autónoma dos Açores*, Universidade dos Açores, Ponta Delgada.
- CARDOSO, I., SOUSA, Á. & LOPES, F. (2015). *Características empreendedoras dos técnicos de diagnóstico e terapêutica dos hospitais dos Açores.* In: CARVALHO, L.C., DOMINGUINHOS P., BALEIRAS, R.N. & DENTINHO, T.P. (Eds.) *Empreendedorismo e Desenvolvimento Regional Casos Práticos*, Edições Sílabo, Lisboa, 51-74.
- CETIN, M. (2013). The hypothesis of innovation-based economic growth: a causal relationship, *International Journal of Economic and Administrative Studies*, 6(11), 1–16.
- DHAHRI, S. & OMRI, A. (2018). Entrepreneurship contribution to the three pillars of sustainable development: what does the evidence really say? *World Development*, Elsevier, 106, 64-77. doi: 10.1016/j.worlddev.2018.01.008
- GALINDO, M. & MENDEZ-PICAZO, M. (2014). Entrepreneurship, economic growth, and innovation: are feedback effects at work, *Journal of Business Research*, 67 (5), 825–829.
- GREENACRE, M. & BLASIUS, J. (2006). *Multiple correspondence analysis and related methods*, Chapman and Hall/CRC, London, UK.
- Grossman, V. (2009). Entrepreneurial innovation and economic growth, *Journal of Macroeconomics*, 31 (4), 602–613.

- LERMAN, I.C. (1970). Sur l'analyse des données préalable à une classification automatique (Proposition d'une nouvelle mesure de similarité). *Rev. Mathématiques et Sciences Humaines*, 32 (8), 5-15.
- LERMAN, I.C. (1981). *Classification et analyse ordinale des données*, Dunod, Paris.
- LERMAN, I.C. (2016). *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering*. Series: Advanced Information and Knowledge Processing, Springer-Verlag, London. doi: 10.1007/978-1-4471-6793-8.
- MAAS, G. & JONES, P. (Eds.) (2017). *Entrepreneurship centres: Global perspectives on their contributions to Higher Education Institutions*, Palgrave MacMillan, London. doi: 10.1007/978-3-319-47892-0.
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Statistics, University of California Press, Berkeley, California, 281-297.
- MARADANA, R., PRADHAN, R., DASH, S., GAURAV, K., JAYAKUMAR, M. & CHATTERJEE, D. (2017). Does innovation promote economic growth? Evidence from European countries, *Journal of Innovation and Entrepreneurship, A Systems View Across Time and Space*, 6 (1), 1-23.
- NICOLAU, F.C. & BACELAR-NICOLAU, H. (1998). *Some trends in the classification of variables*. IN HAYASHI, C., OHSUMI, N., YAJIMA, K., TANAKA, Y., BOCK, H.-H & BABA, Y. (Eds) *Data Science, Classification, and Related Methods*, Springer-Verlag, 89-98.
- SILVA, O. & SOUSA, Á. (2016). *Estatística, uma aprendizagem para a vida*. In SOUSA, Á., MELO, H., NUNES, J., CABRAL, J., MARTINS, M.C., Silva, O., MEDEIROS, P. (Eds) *Matemática 7 perspectivas*, EGA-Empresa Gráfica Açoreana, Ponta Delgada, 15 - 39.
- SOUSA, Á., COUTO, G., BRANCO, N., SILVA, O. & BACELAR-NICOLAU, H. (2015). Entrepreneurship education: The role of the Higher Education Institutions in the Entrepreneurial Attitudes of the Students. *ICERI2015 Proceedings*, 707-714.
- SOUSA, Á., COUTO, G., BRANCO, N., SILVA, O. & BACELAR-NICOLAU, H. (2017). Entrepreneurship Promotion in Higher Education Institutions, *Journal of Entrepreneurship, Business and Economics*, 5 (1), 157-184.
- WENNEKERS, S. & THURIK, R. (1999). Linking entrepreneurship and economic growth, *Small Business Economics*, 13 (1), 27-55.



# Comparação entre métodos não paramétricos para a análise de variância com dois fatores: um estudo de simulação

Anabela Afonso<sup>1</sup> · Dulce G. Pereira<sup>2</sup>

**Resumo** Nos últimos anos têm sido propostas várias alternativas não paramétricas à Análise de Variância (ANOVA) com dois fatores. Neste trabalho, realizamos um estudo de simulação para analisar a probabilidade de erro de Tipo I e a potência de alguns desses testes alternativos, quando os dados são provenientes de distribuições discretas e os delineamentos 3×3 equilibrados. Dois testes apresentaram taxas de erro de Tipo I empíricas superiores ao nível de significância nominal e dois testes exibiram fraca potência no teste à interação, sendo por isso desaconselhada a sua utilização.

**Palavras-chave:** Estatística de Wald, Testes de Permutação, Transformação em Ordens.

## 1 Introdução

A Análise de Variância (ANOVA) foi introduzida por Fisher, com aplicações iniciais no domínio da Agronomia e Biologia, mas são várias as aplicações na área da Epidemiologia e das Ciências Sociais, por exemplo. Num delineamento fatorial pretende-se estudar a influência de um ou mais fatores numa determinada variável resposta. Neste tipo de delineamento está implícito que as amostras são aleatórias e independentes. Na ANOVA para além de se assumir que a variável resposta é do tipo contínuo, também se pressupõe que as populações têm distribuição normal e as variâncias são homogéneas.

Quando se trabalha com dados reais encontramos vários problemas: i) a natureza dos dados nem sempre é contínua; por exemplo, é muito usual nas áreas da Biologia e da Ecologia obtermos dados de contagens e na área das Ciências Sociais predominam dados ordinais; ii) a normalidade das distribuições é muitas vezes

---

<sup>1</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>2</sup> CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, [dgsp@uevora.pt](mailto:dgsp@uevora.pt)

violada devido quer à forma da distribuição quer à presença de valores atípicos; iii) vulgarmente as amostras são pequenas, o que associado a dados não normais não possibilita a aplicação do Teorema do Limite Central, o que põe em causa não só o pressuposto da normalidade como também a validade da distribuição F. Nestas situações a ANOVA paramétrica não é a técnica mais adequada o que levou ao desenvolvimento de alternativas.

Nos últimos anos têm sido propostas alternativas não paramétricas à ANOVA com dois fatores, sendo umas mais simples do que outras, não existindo uma alternativa que seja melhor que as restantes em todos os contextos. Na literatura o desempenho destas alternativas foi analisado considerando distribuições contínuas, preocupando-se com a assimetria e a presença de valores atípicos e/ou com variâncias heterogéneas, tanto para delineamentos equilibrados como desequilibrados. No entanto, não há estudos onde se utilizem distribuições discretas, que usualmente potenciam o surgimento de empates.

O principal objetivo deste trabalho é colmatar a lacuna existente ao nível das distribuições discretas, de forma a, se possível, generalizar as vantagens e desvantagens de algumas dessas alternativas: transformação em ordens, transformação normal inversa, transformação em ordens alinhadas, estatística L do teste de Puri e Sen, teste de van der Waerden, estatística de tipo Wald, estatística de tipo ANOVA, teste de permutação de tipo Wald e testes de permutação sincronizada.

## 2 ANOVA não paramétrica

As alternativas à ANOVA com dois fatores podem ser divididas em três grandes grupos cujas principais características são: 1) criam uma nova variável a partir das observações e realizam a ANOVA paramétrica; 2) propõem as suas próprias estatísticas de teste; 3) baseiam-se nas permutações das observações.

O teste de *transformação em ordens* (RT – *rank transform*), tal como o nome indica, consiste em transformar os dados originais ( $Y$ ) em ordens ( $R$ ) e posteriormente aplicar a ANOVA paramétrica à nova variável  $R$  (Conover e Iman, 1976). Este teste é muito simples, mas não deve ser utilizado para testar a interação na presença de efeitos principais significativos e vice-versa (Higgins e Tatshtoush, 1994, Beasley e Zumbo, 2009), pois pode apresentar um elevado erro de Tipo I e falta de potência (Sawilowsky, 2000).

A *transformação normal inversa* (INT – *inverse normal transformation*) consiste em transformar os dados originais ( $Y$ ) em ordens ( $R$ ), calcular os scores normais das ordens  $R$ , obtendo-se uma nova variável dependente  $Z$  e aplicar a ANOVA a essa nova variável dependente  $Z$  (van der Waerden, 1952). A salientar

que existem várias versões para calcular os scores normais (Beasley *et al.*, 2009). Mansouri e Chang (1995) mostraram que com erros normais e delineamentos equilibrados este teste é muito conservativo. Além disso, pode haver um erro de Tipo I muito elevado, se existirem outros efeitos principais significativos.

A *transformação em ordens alinhadas* (ART – *aligned rank transform*) aplicada a delineamentos fatoriais foi proposta por Higgins e Tashtoush (1994) para dados quantitativos, tendo mais tarde sido propostas alternativas que permitem a sua aplicação a dados ordinais (e.g. Peterson, 2002). Este método consiste em subtrair os efeitos de que não são de primeiro interesse antes de realizar a ANOVA, ou seja, aos resíduos adiciona-se o efeito de interesse (linha, coluna ou interação) e substituem-se os valores pelas ordens aos quais se aplica a ANOVA. Este teste é mais robusto do que o RT (Mansouri e Chang (1995) e, além disso, quando existem *outliers* ou a distribuição tem caudas pesadas este procedimento apresenta menores erros de Tipo I do que o teste F (Higgins e Tashtoush, 1994). Contudo, o erro de Tipo I aumenta com o número de observações por célula (Luepsen, 2017) e é sensível à heterocedasticidade (Leys e Schumann, 2010).

Para lidar com o problema da inflação do erro de Tipo I, Mansouri e Chang (1995) propuseram a *combinação das transformações ART com INT* (ART+INT).

Puri e Sen (1985) propuseram uma generalização do teste H de Kruskal-Wallis que consiste numa adaptação da *estatística L* do teste das ordens com distribuição qui-quadrado (teste L de PS). Na presença de efeitos nulos, este teste controla de forma adequada o erro de Tipo I tanto no teste aos efeitos principais como à interação. Mas, tem falta de potência para testar um efeito quando existirem outros efeitos não nulos no modelo (Toothaker e Newman, 1994).

O teste de van der Waerden foi generalizado para delineamentos fatoriais por Mansouri e Chang (1995). Este teste combina a transformação INT com a estatística L de PS. De um modo geral, o teste de van der Waerden apresenta um bom comportamento do erro de Tipo I e da potência. No entanto, tal como a estatística L de PS, sofre de alguma falta de potência quando existem poucas observações por célula e estamos na presença de outros efeitos não nulos.

Akritas *et al.* (1997) propuseram uma nova estatística de ordens baseada na *estatística de tipo Wald* (WTS – Wald type statistic). Apesar desta estatística ser assintoticamente exata mesmo quando é violado o pressuposto de normalidade, para amostras de dimensão pequena ou média este teste tende a dar resultados muito liberais. Quando o número de níveis dos fatores é elevado não é recomendada a inferência baseada na distribuição assintótica (Brunner *et al.*, 1997).

Para contornar o liberalismo extremo do teste WTS, Brunner *et al.* (1997) propuseram o *teste de tipo ANOVA* (ATS – ANOVA type statistic) que consiste em calcular uma estatística de teste de tipo ANOVA às ordens das observações originais. Os graus de liberdade da estatística F são corrigidos usando a

aproximação de Box. O teste ATS é mais potente que o teste WTS (Shah e Madden, 2004), mas é muito conservativo quando a distribuição dos erros é enviesada (Pauly *et al.*, 2015).

Os testes de permutação têm sido propostos como uma alternativa quando as amostras são pequenas e as distribuições dos erros não satisfazem os pressupostos. De um modo geral, estes testes assentam na condição de permutabilidade que é satisfeita quando a probabilidade dos dados observados é invariante relativamente às permutações aleatórias dos índices.

Pauly *et al.* (2015) propuseram o teste de permutação de tipo Wald (WTPS – Wald type permutation test) que é mais potente que o teste ATS. No teste WTPS os dados originais são permutados entre si, sem restrições, e é calculada uma estatística de tipo Wald com base nas observações permutadas. Este teste mostrou ter um bom controlo da taxa de erro de Tipo I mesmo com distribuições enviesadas com erros homocedásticos.

Os testes de *permutação sincronizada* (CSP – Constrained Synchronized Permutations, e USP – Unconstrained Synchronized Permutations) foram introduzidos por Pesarini (2001) e Salmaso (2003). Posteriormente foram generalizados para delineamentos equilibrados por Basso *et al.* (2007) e mais recentemente para alguns tipos de delineamentos desequilibrados por Hahn e Salmaso (2017). Ao contrário do teste WTPS, os testes CSP e USP impõem restrições à forma como os dados são permutados entre os níveis dos fatores e usam uma estatística de teste que não é studentizada (*i.e.*, não é de tipo Wald). Estes procedimentos mostraram ser potentes e ter uma boa aderência ao valor nominal  $\alpha$  (Hahn e Salmaso, 2017).

### 3 Simulação

No estudo de simulação levado a cabo considerou-se um modelo de ANOVA com dois fatores da forma:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

onde  $\mu$  é a média global,  $\alpha_i$  o efeito do nível  $i$  do fator  $A$ ,  $i = 1, \dots, L$ ,  $\beta_j$  o efeito do nível  $j$  do fator  $B$ ,  $j = 1, \dots, C$ ,  $\gamma_{ij}$  é o efeito da interação entre o nível  $i$  do fator  $A$  e o nível  $j$  do fator  $B$ , e  $\varepsilon_{ijk}$  é o erro aleatório,  $k = 1, \dots, n$ . Em particular, estudámos dois cenários distintos:

- Ausência de interação e de efeitos principais (modelo nulo);
- Dois efeitos principais e existência de interação (modelo completo)

Foram considerados delineamentos equilibrados com amostras de dimensão  $n = 3, 5, 10$ ,  $L = C = 3$ , e uma diversidade de cenários distribucionais com diferentes graus de dispersão e vários tipos assimetria:

- Binomial:
  - assimétrica positiva:  $B(N; 0,2)$  com  $N = 25, 50, 100$ ;
  - simétrica:  $B(N; 0,5)$  com  $N = 10, 20, 40$ ;
- Binomial Negativa:  $BN(N; 0,4)$  com  $N = 2, 4, 8$ ;
- Poisson:  $P(\lambda)$  com  $\lambda = 5, 10, 20$ ;
- Uniforme:  $U\{0, \dots, N\}$  com  $N = 10, 20, 40$ .

Considerámos várias intensidades para os efeitos:

$$\alpha_i = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & c. c. \end{cases} \quad \beta_j = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & c. c. \end{cases} \quad \gamma_{ij} = \begin{cases} c, & i = j \text{ e } i, j = 1, 2 \\ -c, & i \neq j \text{ e } i, j = 1, 2 \\ 0, & c. c. \end{cases}$$

com  $c = 0,25\sigma$ ;  $0,5\sigma$  e  $1\sigma$  e  $\sigma$  o desvio-padrão da população amostrada.

Para cada cenário distribucional foram realizadas  $M = 1000$  replicações, tendo-se registado:

- A distribuição empírica dos *valores p*,
- O número total de réplicas que ultrapassaram o nível de significância definido,  $\alpha = 1\%$ ,  $5\%$  e  $10\%$ .

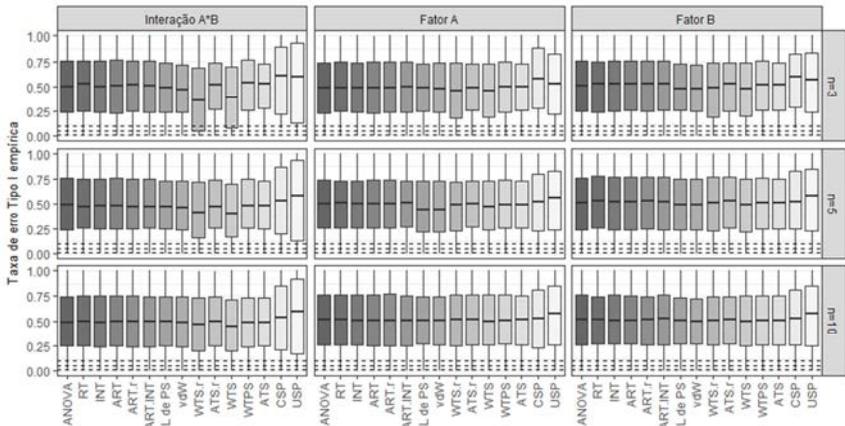
para cada um dos testes descritos na secção 2, i.e., RT, INT, ART, ART+INT, L de PS, van der Waerden (vdW), WTS, ATS, WTPS, CSP e USP, e ainda para o teste F da ANOVA. Os testes ART, WTS e ATS foram aplicados tanto aos dados originais como às ordens das observações (ART.r, WTR.r, ATS.r).

Para todos os testes foi usado o programa R project (R Core Team, 2016). Foram usadas as livrarias *ARTool*, *rankFD* e *GFD*, e as funções disponíveis em <http://www.uni-koeln.de/~luepsen/R/> e <http://static.gest.unipd.it/~salmaso/web/>.

## 4 Resultados

Na análise do erro de Tipo I, verificou-se que a distribuição empírica dos *valores p* é semelhante para todas as distribuições consideradas. Na Figura 1 apresenta-se, como exemplo, o caso particular em que os dados foram simulados com base na distribuição Poisson com média 5.

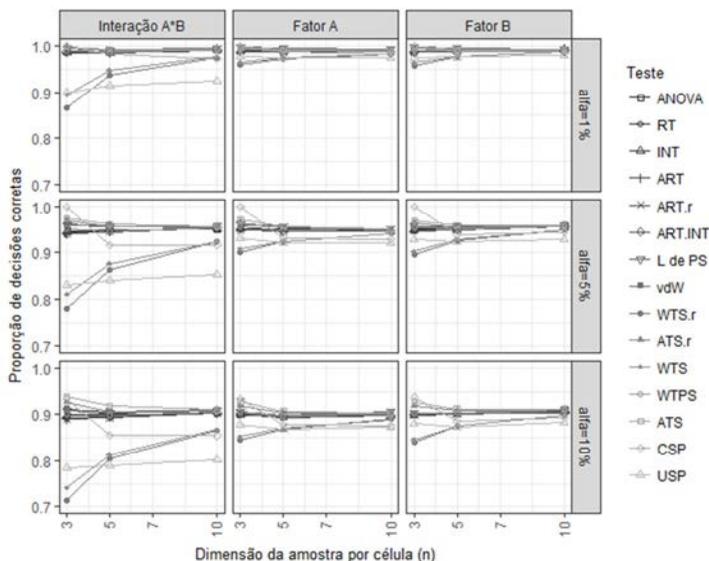
O teste WTS, aplicado quer às observações originais quer às ordens das observações, tende a apresentar menores taxas de erro de Tipo I que os restantes testes (Figura 1). Nos testes CPS e USP observou-se uma maior dispersão nas taxas de erro obtidas. Além disso, em mais de metade das vezes a taxa de erro ultrapassa o nível de significância definido.



**Figura 1** – Distribuição empírica das taxas de erro de Tipo I, quando  $c = 0$  e  $\varepsilon_{ijk} \sim P(5)$ .  
 (as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

Na análise à interação os testes RT, INT, ART e ART+INT tendem a ter desempenhos piores quando se consideram distribuições simétricas (Binomial e Uniforme), ultrapassando na maior parte das vezes o valor de alfa nominal. Os testes ART e ART+INT são sensíveis à assimetria, reagindo mal à assimetria negativa.

Na Figura 2 é possível observar a proporção de vezes que não foi cometido o erro de Tipo I tendo em conta o nível de significância definido e a dimensão da amostra em cada célula, e considerando todas as distribuições.



**Figura 2** – Proporção de cenários em que não foi cometido o erro de Tipo I, por nível de significância nominal ( $\alpha$ ) e dimensão da amostra em cada célula.

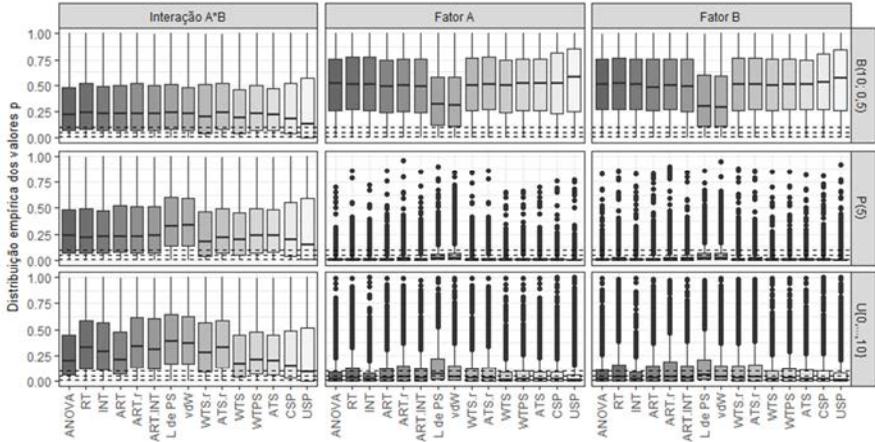
De um modo geral, à medida que aumenta a dimensão da amostra ( $n$ ), a taxa de erro de Tipo I de quase todos os testes aproxima-se do  $\alpha$  nominal (Figura 2). Os testes USP e WTS são os que apresentaram os piores desempenhos. A taxa de erro de Tipo I empírica destes dois testes foi sempre superior à do nível de significância  $\alpha$  definido, tanto no teste à interação quer nos testes aos efeitos principais.

Os resultados variam ligeiramente com o valor de  $\alpha$  e a dimensão da amostra (Figura 2). O teste ATS é o mais estável em termos de bom comportamento quando  $\alpha = 5\%$  e  $10\%$ . Quando  $\alpha = 1\%$  os testes apresentam taxas de erro de Tipo I empíricas muito similares e próximas do  $\alpha$  nominal.

Os testes RT, INT, ART e ART + INT tendem: i) a não ultrapassar o  $\alpha$  nominal com o aumento do  $n$  no teste à interação; ii) a ter desempenhos piores quando se consideram distribuições simétricas (Binomial e Uniforme), ultrapassando na maior parte das vezes o  $\alpha$  nominal, no teste à interação; iii) a ser sensíveis à assimetria, reagindo mal à assimetria negativa, nos testes aos efeitos principais.

Na análise da potência de teste, o tipo de assimetria e achatamento tem influência na distribuição empírica dos valores  $p$  dos testes aos efeitos principais (Figura 3). Estes testes parecem ser mais potentes quando se consideram distribuições assimétricas, e menos potentes quando a distribuição é simétrica

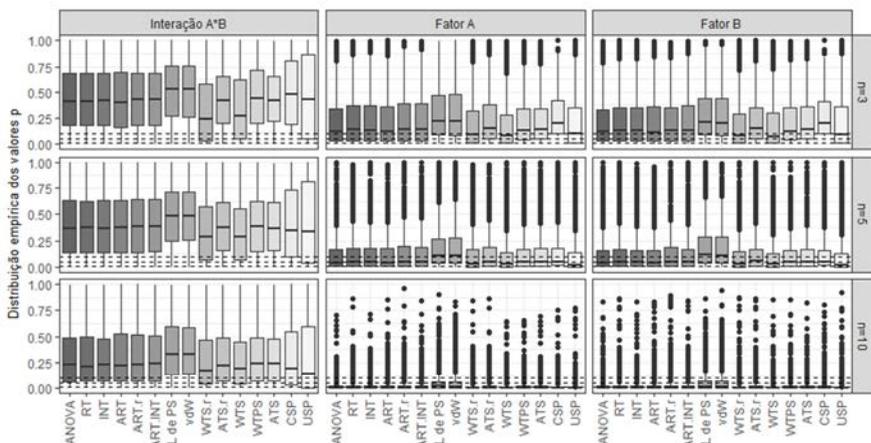
mesocúrtica. Contudo, quando se aumenta o valor do efeito  $c$  então os *valores p* diminuem e a sua distribuição aproxima-se da observada quando se considera que os erros têm distribuição Poisson.



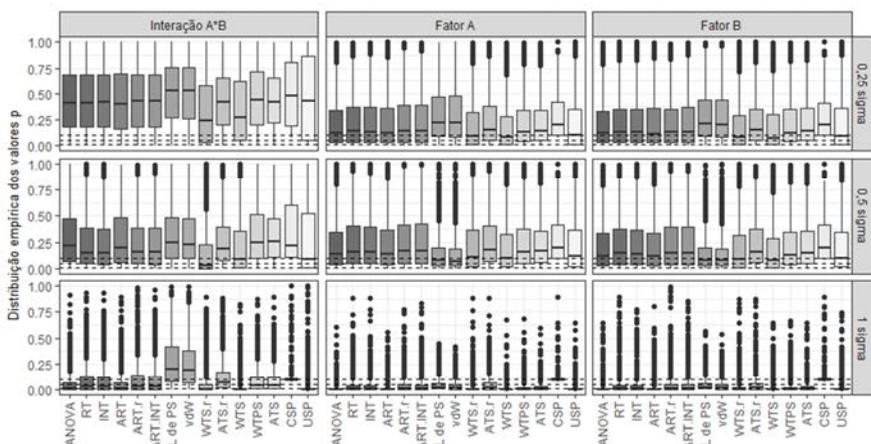
**Figura 3** – Distribuição empírica dos *valores p* quando  $n = 10$ ,  $c = 0,25\sigma$  e  $\varepsilon_{ijk} \sim B(10; 0,5)$  (cima),  $\varepsilon_{ijk} \sim P(5)$  (meio) e  $\varepsilon_{ijk} \sim U\{0, \dots, 10\}$  (baixo).  
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

A potência de teste aumenta com o aumento do  $n$ , sendo mais evidente nos testes aos efeitos principais do que no teste à interação (Figura 4). Os testes L de PS e van der Waerden são os que apresentam os piores desempenhos.

De um modo geral, à medida que aumenta a intensidade  $c$  do efeito observou-se um aumento da potência de teste (Figura 5). Todos os testes são mais potentes na avaliação dos efeitos principais do que no teste à interação. Os testes L de PS e van der Waerden são os menos potentes no teste à interação.



**Figura 4** – Distribuição empírica dos valores  $p$  quando  $n = 3, 5, 10$ ,  $c = 0,25\sigma$  e  $\varepsilon_{ijk} \sim P(5)$ .  
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)



**Figura 5** – Distribuição empírica dos valores  $p$  quando  $n = 3$ ,  $c = 0,25\sigma, 0,75\sigma, 1\sigma$  e  $\varepsilon_{ijk} \sim P(5)$ .  
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

## 5 Conclusão

Na análise da probabilidade de erro de Tipo I e da potência às alternativas não paramétricas à ANOVA com dois fatores 3×3 equilibrada, foram considerados vários cenários distribucionais que incluíram distribuições simétricas e assimétricas, bem como diferentes graus de dispersão e diferentes dimensões de amostra, para avaliar se estas características tinham influência no desempenho dos testes.

O tipo de distribuição e o número de observações em cada célula, de um modo geral, não afeta o erro de Tipo I, mas tem influência na potência dos testes tal como a intensidade dos efeitos. A potência de teste é menor quando os erros têm uma distribuição simétrica e aumenta com a dimensão das amostras.

Todos os testes são mais potentes na avaliação dos efeitos principais do que no teste à interação, sendo os testes L de PS e van der Waerden os menos potentes no teste à interação. Com o aumento do tamanho do efeito, estes dois testes aumentam substancialmente a sua potência no teste aos efeitos principais, alcançando em algumas situações desempenhos melhores que os outros testes. No entanto, estes testes são os que apresentam taxas de erro de Tipo I empírica mais próximas do  $\alpha$  definido, bem como o teste ATS.

Desaconselha-se o uso dos testes USP, WTS e CSP porque apresentam taxas de erro de Tipo I empíricas mais afastadas do nível de significância nominal. Além disso, os testes CPS e USP são os que apresentam a maior dispersão na distribuição dos *valores p*. No entanto, os testes WTS e USP mostraram ser os mais potentes no teste à interação.

A ANOVA paramétrica tem um comportamento estável e são poucos os métodos que a superam. No caso do erro de Tipo I, a ANOVA apenas tem um desempenho mais fraco que os testes ATS, L de PS, van der Waerden e WTPS, e um desempenho sempre melhor que os testes WTS e USP. No entanto, a ANOVA é mais potente que os testes L de PS e van der Waerden e menos potente que os testes USP e WTS. De um modo geral, a ANOVA comporta-se melhor no estudo dos efeitos principais do que no estudo da interação, quando comparada com as alternativas não paramétricas.

## Agradecimentos

Este trabalho é financiado por Fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto «UID/MAT/04674/2019 (CIMA)».

## Referências

- AKRITAS, M. G., ARNOLD, S. F. & BRUNNER, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of the American Statistical Association*, 92, 258-265.
- BASSO, D., CHIARANDINI, M. & SALMASO, L. (2007) Synchronized permutation tests in replicated I x J designs, *Journal of Statistical Planning and Inference*, 137, 2564-2578.
- BEASLEY, T. M. & ZUMBO, B. D. (2009). Aligned rank tests for interactions in split-plot designs: distributional assumptions and stochastic heterogeneity, *Journal of Modern Applied Statistical Methods*, 8, 16-50.
- BEASLEY, T. M., ERICKSON, S. & ALLISON, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited?, *Behavior genetics*, 39, 580–595.
- BRUNNER, E., DETTE, H. & MUNK, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association*, 92, 1494-1502.
- CONOVER, W. J., & IMAN, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs, *Communication in Statistics – Theory and Methods*, 5, 1349-1368.
- HAHN, S. & SALMASO, L. (2017). A comparison of different synchronized permutation approaches to testing effects in two-level two-factor unbalanced ANOVA designs, *Statistical Papers*, 58, 123-146.
- HIGGINS, J. J. & TASHTOUSH, S. (1994). An aligned rank transform test for interaction, *Nonlinear World*, 1, 201-211.
- LEYS, C. & SCHUMANN, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test, *Journal of Experimental Social Psychology*, 46, 684-688.
- LUEPSEN, H. (2017). The aligned rank transform and discrete variables: a warning, *Communications in Statistics - Simulation and Computation*, 46, 6923-6936.
- MANSOURI, H. & CHANG, G.-H. (1995). A comparative study of some rank tests for interaction, *Computational Statistics & Data Analysis*, 19, 85-96.
- PAULY, M., BRUNNER, E. & KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs, *Journal of the Royal Statistical Society, Series B*, 77, 461-473.
- PESARIN, F. (2001). *Multivariate permutation tests with applications in biostatistics*, Wiley & Sons, Chichester.
- PETERSON, K. (2002). Six modifications of the aligned rank transform test for interaction, *Journal of Modern Applied Statistical Methods*, 1(1), 100-109.
- PURI, M. L. & SEN, P. K. (1985). *Nonparametric methods in General Linear Models*, Wiley, New York.

- R CORE TEAM (2016). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- SALMASO, L. (2003). Synchronized permutation tests in 2k factorial designs, *Communication in Statistics*, 32, 1419-1437.
- SAWILOWSKY, S. S. (2000). Review of the rank transform in designed experiments, *Perceptual and motor skills*, 90, 489-497.
- SHAH, D. A. & MADDEN, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments, *Phytopathology*, 94,33-43.
- TOOTHAKER, L. E. & NEWMAN, D. (1994). Nonparametric competitors to the two-way ANOVA, *Journal of Educational Statistics*, 19, 237-273.
- VAN DER WAERDEN, B. L. (1952). Order tests for the two-sample problem and their power, in *Indagationes Mathematicae (Proceedings)*, 55, 453-458.

# A WebGIS-based platform for data management in water resources systems

Marta Castilho Gomes<sup>1</sup> · Reydleon Paulo<sup>2</sup> · Bernardo Varela<sup>3</sup> · Daniel Conde<sup>4</sup> · Ricardo Canelas<sup>5</sup> · Alexandre Gonçalves<sup>6</sup> · Francisco Regateiro<sup>7</sup> · Ana M. Ricardo<sup>8</sup>

**Abstract** The benefits of geographical information systems (GIS) for geocomputations in the water resources domain are countless. The INLET research project targets the development of a web platform that provides data management capabilities as well as visualization interfaces and tools to support researchers, practitioners, decision agents and policy makers in the field of water resources management. This aim was achieved by a strong articulation between a relational database and a graphical interface based on GIS. This paper describes the characteristics and development of the INLET hydro-informatics platform, which required a multidisciplinary team. Application of this customized and flexible tool is shown for a case study in Portugal.

**Key-words:** Environmental Data Management, Geospatial Information Systems, Data Visualization, Decision Support Systems.

## 1 Introduction and motivation

Nowadays, visualization of geographically-based data is nearly ubiquitous in water resources applications. Dixon & Uddameri (2016) defined geocomputation in water resources engineering and science as GIS-enabled analysis, synthesis, and design of water resources systems. The increasing volume and varying format of water resources data presents challenges in accessing, storing, managing, processing and visualizing data as well as assessing its quality. For this reason, geospatial information systems (GIS) are often incorporated into workflows for analyzing such datasets (Tsihrintzis *et al.*, 1996; Pons & Masó, 2016).

---

<sup>1-8</sup> CERIS – Civil Engineering Research and Innovation for Sustainability, Instituto Superior Técnico, Universidade de Lisboa

<sup>1</sup> [marta.gomes@tecnico.ulisboa.pt](mailto:marta.gomes@tecnico.ulisboa.pt); <sup>2</sup> [reydleon94@gmail.com](mailto:reydleon94@gmail.com); <sup>3</sup> [bernardo.f.varela@tecnico.ulisboa.pt](mailto:bernardo.f.varela@tecnico.ulisboa.pt)

<sup>4</sup> [daniel.conde@tecnico.ulisboa.pt](mailto:daniel.conde@tecnico.ulisboa.pt); <sup>5</sup> [ricardo.canelas@tecnico.ulisboa.pt](mailto:ricardo.canelas@tecnico.ulisboa.pt)

<sup>6</sup> [alexandre.goncalves@tecnico.ulisboa.pt](mailto:alexandre.goncalves@tecnico.ulisboa.pt); <sup>7</sup> [francisco.regateiro@tecnico.ulisboa.pt](mailto:francisco.regateiro@tecnico.ulisboa.pt)

<sup>8</sup> [ana.ricardo@tecnico.ulisboa.pt](mailto:ana.ricardo@tecnico.ulisboa.pt)

Jia *et al.* (2009) developed a WebGIS-based system to predict rainfall-runoff and assess water resources in real-time for the city of Beijing, China. Recently, Gomes *et al.* (2017) presented a novel WebGIS-based platform for flood risk management and demonstrated its use for the Tagus estuary inundation information system, in Portugal. In line with these developments, the focus of the INLET project is the creation of a platform that integrates Internet technologies and provides data management and data visualization interfaces and tools for researchers, professionals and policy makers in the water resources sector. The INLET web platform is a hydro-informatics platform in the sense that it provides a developing environment to deal with computerized information related to water resources systems. At its core lies a flexible and efficiently structured database coupled with a GIS-based graphical interface.

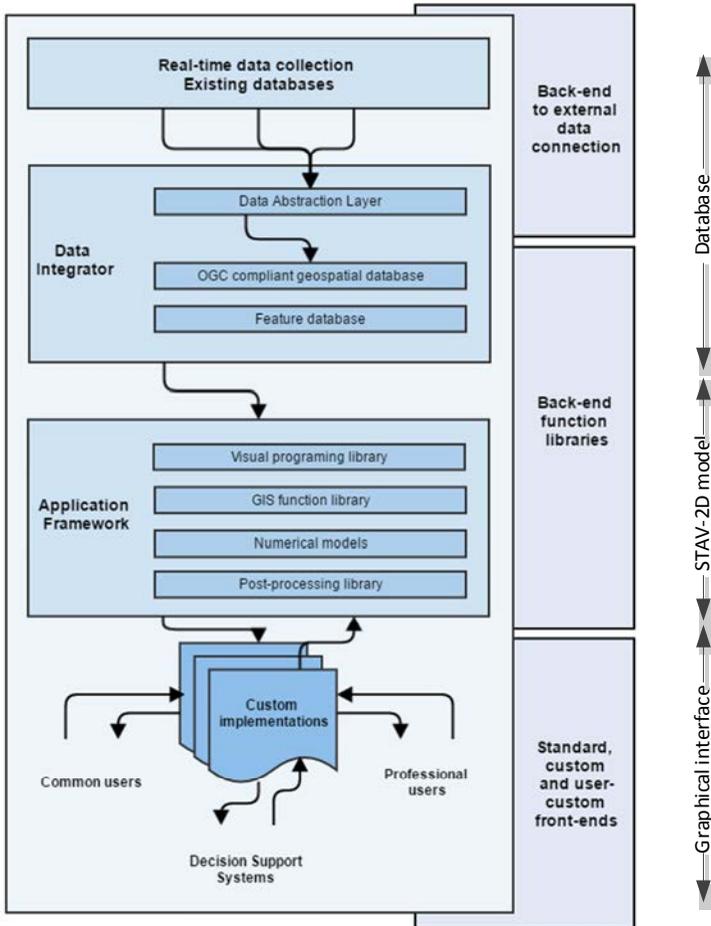
The platform and its applications are suitable for integration into decision support systems that serve water resources administration and civil protection, hence contributing to an increased resilience to floods and other natural disasters. This is ensured by a close collaboration with the Portuguese Environment Agency (APA) during project development, namely the Hydraulic Resources department. This is in charge of monitoring the hydrological network and provide public access to the collected data by means of SNIRH, the National Information System for Hydrological Resources. This service makes available a large volume of hydrological historical data, however its organization for geospatially based applications is not straightforward. Data is organized into time-series associated to the monitoring stations, and users may have to spend a long time to obtain and organize the data they need.

The main motivation for the development of the INLET platform was thus to ease these tasks for SNIRH users. Secondly, the devised system will be integrated in the future with STAV-2D, a conceptual mathematical model of fluvial hydrodynamics developed by some of the INLET project team members (Conde *et al.*, 2015).

The remainder of this paper is organized as follows. First, the platform concept is described as well as its implementation with available computational tools (Section 2). Section 3 describes an application case study for the catchment area of Mondego river, where the monthly precipitation was characterized both in time and space. Finally, conclusions are drawn and ideas for the evolution of this work discussed in section 4.

## 2 Platform concept and implementation

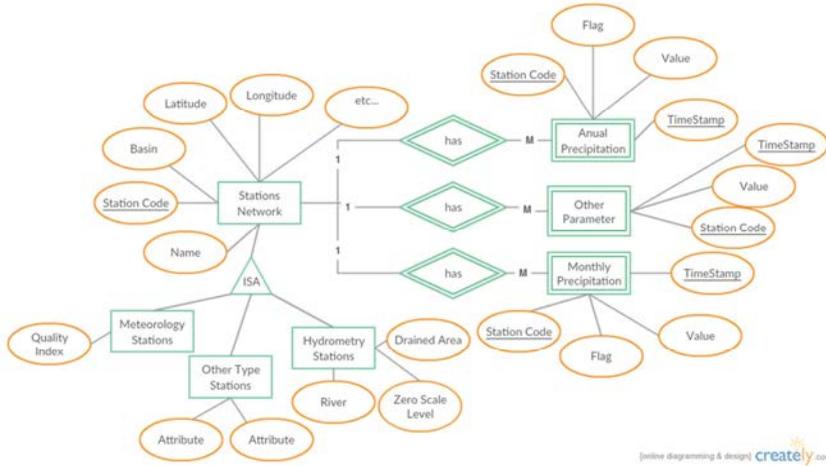
The integrated model and functionality of the INLET platform is described in Figure 1. The system is composed of a database, an application framework (STAV-2D) and a graphical interface.



**Figure 1** – Functional diagram of the INLET platform.

The database is structured according to a relational model concept detailed in Figure 2. PostgreSQL is used as the relational database management system. The

database includes time and space variables and additional metadata to maximize the efficiency of the database-interface communication. Temporal data is stored in different tables for better data management and to avoid data redundancy.



**Figure 2** – Database relational model embedded in the INLET platform.

In Portugal, hydrological and meteorological data is mainly stored following its historical design, that is to say, centered on the concept of time-series of variables collected at the monitoring stations. Following the new paradigm to organize water resources data (Dixon & Uddameri, 2016; Pons & Masó, 2016), a geospatial approach was pursued in the platform conception. PostGIS is used for managing the geospatial information.

The graphical interface of the INLET system is shown in Figure 3. This is rooted in Quantum GIS (QGIS) and tuned with customized tools and various web-services to join access to data and data analysis in a uniform environment. These tools target the ease of use and workflow simplification and were developed to find, display and process data. They were mainly designed as Python plugins. Common protocols for information transfer like Web Feature Service, Web Coverage Service, Web Map Service and Web Processing Service are also included. All the tools are coded in order to ensure compatibility with the Open Geospatial Consortium (OGC) standards.

The application framework consists of STAV-2D, a fluvial hydrodynamics model based on shallow-water equations that features dynamic bed geometries and sediment transport. A STAV-2D plugin allows access to the database, queries to import selected data, data visualization and data preparation for STAV-2D simulations.

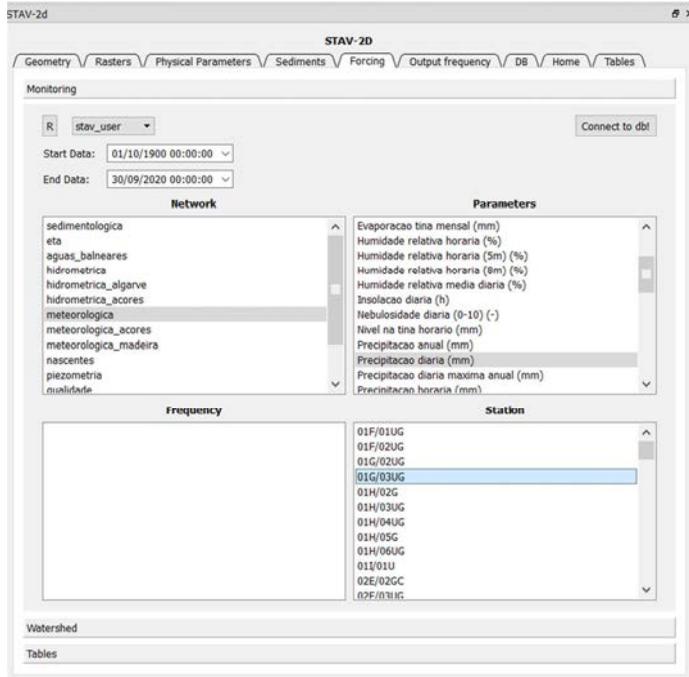
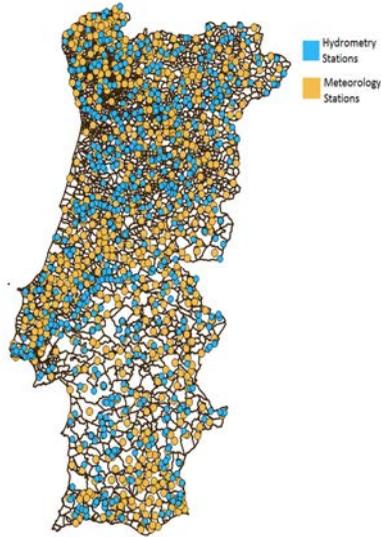


Figure 3 – Graphical interface of the INLET platform.

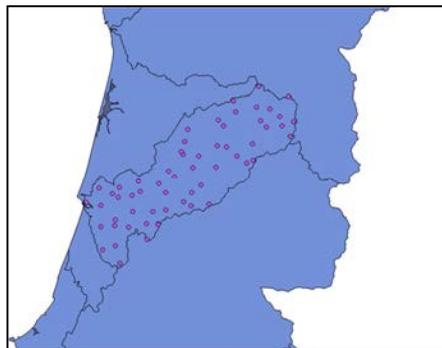
### 3 Data visualization application: the case study of Mondego watershed

The national network for monitoring water resources is composed of 620 meteorological and 311 hydrometric stations in mainland Portugal, shown in Figure 4. The meteorological stations collect data on precipitation, wind velocity and direction, evaporation, radiation, temperature and humidity while the hydrometric stations measure water level and discharge. Between 2014 and 2016, APA invested 4M€ of European Community funds to restore these stations and update the technology. Data acquisition and tele-transmission in real time is now done in two hundred stations (Saramago, 2017).

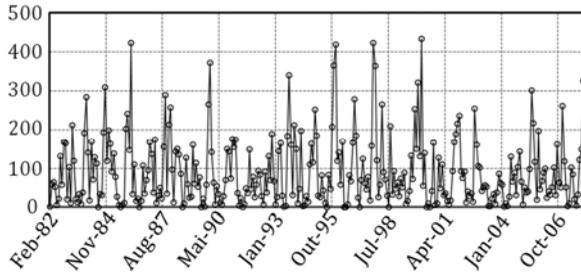


**Figure 4** – Spatial distribution of hydrometric and meteorological stations in mainland Portugal.

The Mondego watershed, in the Centre region of the country (Figure 5), was chosen to illustrate data retrieval and processing with the proposed system. At the current stage of development, the INLET database stores precipitation data and a digital elevation model (DEM) of mainland Portugal. Figure 6 shows the time-series of monthly precipitation for a monitoring station in Mondego basin.

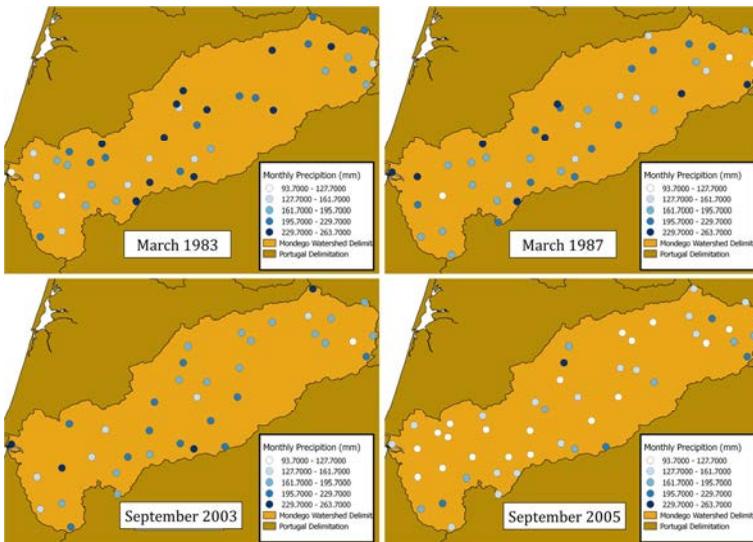


**Figure 5** – Location of Mondego watershed and the corresponding hydrometric and meteorological stations considered in the case study.



**Figure 6** – Monthly precipitation (mm) in recent decades at Aguiar da Beira station, in Mondego watershed.

The TimeManager plugin (available at the QGIS Python plugin repository) is employed for visualizing the contents of the database, allowing a flexible and simple management of geospatially referenced hydrological data. Figure 7 displays an example of an animation of spatial and temporal data for the case study.



**Figure 7** – Visualization of the monthly precipitation in Mondego watershed with the INLET platform.

## 4 Conclusion and future work

INLET is a GIS-based web platform that results from a multidisciplinary effort to create a novel, flexible, efficient and customized tool for water resources management and risk assessment. Developing this work required strong cooperation with APA, the Environment Agency in charge of the network of monitoring stations in Portugal. The platform is fed with data available in databases such as the one of SNIRH, a service deployed by APA to allow public access to water resources data.

Currently, the INLET platform integrates precipitation and topographic data, as shown in the application case study of Mondego hydrographic basin. In the future, other environmental datasets, of meteorological, hydrometric, sedimentological and geometric nature, will be included in the platform. Another direction of work is to provide GIS typical functionalities (like interpolation) in the supporting GIS software, enhancing the flexibility of the platform.

The final stage to be achieved is the total integration of the developed system and the existing STAV-2D mathematical model (Conde *et al.*, 2015), so as to obtain a real-time, operational fluvial hydrodynamics modelling tool able to simulate solid material transport of debris, sediments and pollutants.

## Acknowledgements

This work was funded by CERIS (Civil Engineering Research and Innovation for Sustainability) through the 2016 exploratory actions program (RD0475/01).

## References

- CONDE, D. A. S.; TELHADO, M. J.; VIANA BAPTISTA, M. A. & FERREIRA, R. M. L. (2015). Severity and exposure associated with tsunami actions in urban waterfronts: the case of Lisbon, Portugal. *Natural Hazards*, 79, 2125-2144.
- DIXON, B., UDDAMERI, V., & RAY, C. (2015). *GIS and Geocomputation for Water Resource Science and Engineering*. John Wiley & Sons.
- GOMES, J. L.; JESUS, G.; ROGEIRO, J.; OLIVEIRA, A.; DA COSTA, R. T.; FORTUNATO, A. B. (2017). An innovative web platform for flood risk management. In: Grzenda M., Awad A., Furtak J., Legierski J. (Eds) *Advances in Network Systems*. Advances in Intelligent Systems and Computing, Springer, 461, 217-231.

- JIA, Y.; ZHAO, H.; NIU, C.; JIANG, Y.; GAN, H.; XING, Z.; ZHAO, X.; ZHAO, Z. (2009). A WebGIS-based system for rainfall-runoff prediction and real-time water resources assessment for Beijing. *Computers & Geosciences*, 35(7), 1517-1528.
- PONS, X. & MASÓ, J. (2016). A comprehensive open package format for preservation and distribution of geospatial data and metadata. *Computers & Geosciences*, 97, 89-97.
- SARAMAGO, M. (2017). Redes de monitorização hidrometeorológicas. *Recursos Hídricos*, 38(1), 33-39.
- TSIHRINTZIS, V. A., HAMID, R., & FUENTES, H. R. (1996). Use of geographic information systems (GIS) in water resources: a review. *Water Resources Management* 10(4), 251-277.



## O uso do web scraping nas Estatísticas Oficiais

**Maria José Fernandes<sup>1</sup>**

**Resumo** No contexto de modernização dos processos de recolha o Instituto Nacional de Estatística tem vindo a explorar o potencial da web como fonte de informação complementar ou alternativa às fontes convencionais. Este trabalho pretende partilhar e divulgar o modelo de recolha desenvolvido para o Índice de Preços no Consumidor, baseado em técnicas de web scraping. A infraestrutura criada usa o Python como linguagem base de programação e assenta na utilização de ferramentas “free and open source”.

**Palavras-chave:** Free and Open Source, Índice de Preços no Consumidor, Jupyter Notebook, MongoDB, Python, Web Scraping.

### 1 Introdução

Todos os meses o Instituto Nacional de Estatística, IP (INE) produz o Índice de Preços no Consumidor (IPC) que permite medir a inflação através da evolução temporal dos preços de um conjunto de bens e serviços representativos da estrutura de despesa de consumo da população residente em Portugal (INE, 2009). Durante uma semana predefinida são recolhidos aproximadamente 128.000 preços correspondentes a 1.200 bens e serviços observados; cerca de 95% destes preços são recolhidos por 130 entrevistadores, em 10.000 lojas físicas distribuídas por todo o País (recolha tradicional); os restantes preços são recolhidos, centralmente, em catálogos, por telefone e através da web.

A recolha na web é feita manualmente, visitando os websites relevantes (mobiliário para o lar, telemóveis, hotéis, viagens, automóveis usados, farmácias, etc.) e transferindo a informação (copy/paste) para a base de dados do IPC; não obstante os procedimentos definidos que asseguram a consistência da recolha, trata-se de um processo sujeito a erros, pouco estimulante e pouco eficiente.

---

<sup>1</sup> Instituto Nacional de Estatística, [mjose.fernandes@ine.pt](mailto:mjose.fernandes@ine.pt)

A importância da Internet como canal preferencial de comunicação entre instituições, empresas e famílias e o progressivo crescimento do comércio eletrônico constituem uma oportunidade para os INE's enquanto fonte de informação e para o IPC, em particular: mais produtos e serviços disponíveis na web significam mais preços observáveis com custos de recolha potencialmente mais reduzidos.

Perante este cenário e em linha com a estratégia nacional de modernização dos processos de recolha de dados e com as recomendações do Eurostat para incorporação da fonte comércio eletrônico no IPC (Eurostat, 2013), o INE decidiu, à semelhança de outros Estados Membros, explorar o potencial da recolha automática na web - web scraping - enquanto novo método de recolha no IPC (Boettcher, 2015; Griffioen *et al.*, 2016; Hoekstra *et al.*, 2010; Polidoro *et al.*, 2015; Swier, 2015). O projeto foi cofinanciado pelo Eurostat no âmbito da Grant "Price statistics - methodological and practical improvements and developments, Objective F: Web scraping as a source for HICP".

Qualquer que seja o futuro alcance desta mudança, substituição da recolha manual na web, complemento ou alternativa à recolha tradicional, os benefícios potenciais desta "evolução" são óbvios: potencial redução de custos, aumento da cobertura (mais preços recolhidos), aumento da frequência de observação, diminuição dos tempos de recolha e processamento, indicadores com mais qualidade, indicadores derivados, novos indicadores e...mais tempo livre para novos desafios.

Este trabalho tem como objetivo partilhar e divulgar o modelo de web scraping desenvolvido para o IPC, em particular, a opção tecnológica que suporta a extração, armazenamento, processamento e análise dos dados.

## **2 Web scraping – conceito, utilização, funcionamento, opções tecnológicas, conhecimentos necessários**

A quantidade de informação disponibilizada na web tem crescido exponencialmente o que configura um desafio quando pensamos na sua recolha (Cavallo & Rigobon, 2016; IDC & EMC, 2012; Internet Live Stats, 2018). A recolha manual é enfadonha, lenta, falfível, cara e impraticável quando dirigida a grandes volumes de informação.

Torna-se por isso crucial simular a interação humana com a web, ou seja, automatizar a web. E é isso mesmo que faz o web scraping, automatiza o processo de recolha através de ferramentas (aqui chamados web scrapers) que navegam e extraem a informação semiestruturada de páginas web e a armazenam em bases de dados. De uma forma simples mas explícita podemos afirmar que o web scraping transforma dados de páginas web em estruturas de dados.

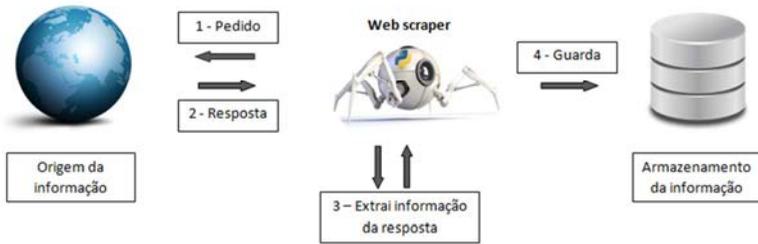
O web scraping é usado numa infinidade de áreas de estudo, sempre que se pretende extrair informação da web; da investigação, aos mercados financeiros, passando pelo jornalismo, pela recolha de contactos, pela procura/oferta de emprego, pela análise de sentimentos até à monitorização, comparação e recolha de preços.

E é exatamente na recolha de preços que o web scraping se torna relevante para o IPC. Um exemplo bem conhecido da utilização desta técnica computacional no cálculo do IPC é o “The Billion Prices Project” (BPP) , desenvolvido no MIT Sloan School of Management; neste estudo o professor Alberto Cavallo e a sua equipa usaram técnicas de web scraping para recolher os preços de centenas de lojas online e mostraram que é possível calcular estatísticas de preços a partir de dados recolhidos na web (Cavallo & Rigobon, 2016).

O web scraping é uma técnica computacional que, como o próprio nome indica, usa a world wide web (web) como fonte de informação. Por isso o entendimento do funcionamento da Internet e da web, dois termos tantas vezes confundidos, é fundamental para quem deseja utilizar esta técnica. A Internet é uma infraestrutura de redes, que por meio de complexas estruturas de comunicação, liga biliões de dispositivos à escala global, através de satélites, cabos ou sem fios. Vários serviços funcionam sobre a infraestrutura da Internet, tais como serviços de voz, mensagens instantâneas, correio eletrónico, transferência de arquivos, etc. A web, também é construída sobre a infraestrutura da Internet e é um modelo de partilha de informação constituído por um conjunto de páginas interligadas por hyperlinks.

Quando visitamos uma página web o conteúdo da resposta enviada pelo servidor onde a página está alojada, é um documento html (hypertext markup language) que diz ao navegador/browser (google chrome, firefox, safari, Internet explorer, etc.) do cliente como exibir a página. O navegador analisa o conteúdo do documento html e realiza outros pedidos para obter todos os outros recursos descritos no html que também compõe a página web, tais como formatações, scripts, imagens, etc.; ou seja, o navegador, não faz apenas um, mas diversos pedidos para obter todo o conteúdo da página web solicitada. Este processo de troca de informação entre cliente e servidor é conhecido como pedido/resposta.

Tecnicamente um web scraper é um programa/script que envia, de forma autónoma, um pedido http (hypertext transfer protocol) para o servidor onde está alojada a página web e interpreta a resposta enviada pelo servidor identificando, extraindo e armazenando a informação pretendida que pode ser texto, imagem ou links para outras páginas que servirão de base a novos pedido http (Figura 1).



**Figura 1** – Arquitetura do Web Scraping.

Para que o web scraping seja efetuado com sucesso e de uma forma eficiente é fundamental não só dominar a estrutura de uma página web mas também entender os processos de comunicação e os desenvolvimentos tecnológicos associados à web. A complexidade associada ao web scraping está, no nosso ponto de vista, mais associada a estes conhecimentos do que à técnica ou ao software utilizado para o operacionalizar.

Existem várias técnicas para implementar o web scraping que implicam diferentes graus de conhecimento/especialização em IT e que proporcionam diferentes níveis de controlo sobre o processo:

- **Ferramentas “point and click”** - versões simplificadas de web scrapers também conhecidas como “visual scrapers” que são capazes de identificar elementos de páginas que queremos extrair e transforma-los em formato tabular apenas com um click; são direcionadas a utilizadores sem conhecimentos de programação ou de funcionamento da web. Três dos produtos comerciais mais conhecidos são import.io, dexti.io e octoparse. Fizemos algumas experiências com estas ferramentas mas as dificuldades identificadas em lidar com conteúdos dinâmicos (paginação, “infinite scrolling”, etc.), a falta de qualquer controlo sobre o processo de extração, a possibilidade de serem descontinuadas e o facto de serem serviços pagos constituíram para nós demasiados fatores limitantes;
- **Ferramentas baseadas na automatização dos navegadores/browsers** – neste grupo encontram-se dois tipos distintos:
  - **plugins muito simples** como o Web Scraper, Data Miner ou Scraper, grátis, virados para a extração de informação de páginas pouco evoluídas tecnologicamente; à semelhança dos “visual scrapers” não exigem qualquer conhecimento de programação ou tecnologias da web;
  - **softwares dedicados**, pagos, que combinam automatização do browser com extensões programáveis e que exigem conhecimento das tecnologias da web e de programação (imacros é um dos softwares mais conhecido);

- **Ferramentas Ad-hoc** – programas específicos desenvolvidos com linguagens de programação de uso geral e que por isso podem cobrir outras fases do processo para além da extração pura, tais como, limpeza e armazenamento. Python (Mckinney, 2012; Python Software Foundation, 2018) é uma das linguagens mais usadas neste âmbito oferecendo recursos dedicados (packages, frameworks) que facilitam muitas operações envolvidas em todo o ciclo de web scraping. Requer conhecimentos especializados para o desenvolvimento e manutenção mas oferece um controle total, mesmo nos sites mais complexos e evoluídos.

Não obstante as diversas soluções de web scraping oferecidas pelo mercado, o INE optou por desenvolver o seu próprio modelo, ou seja, uma ferramenta Ad-hoc. Esta opção implica um elevado nível de conhecimentos técnicos mas proporciona um total controlo do processo. É ainda a abordagem que se considera mais importante porque potencia o uso de novas linguagens/ferramentas estimulando a investigação por parte dos intervenientes e incentivando o debate sobre os processos tecnológicos em uso. Acresce ainda a tudo isto uma vontade em desenvolver uma infraestrutura de captura, armazenamento, processamento e análise totalmente assente em ferramentas “free and open source”.

Para implementar uma solução Ad-hoc é necessário não só conhecer uma linguagem de programação (Python, Java, C#, Javascript, R, Ruby, etc.), de preferência de uso geral e com fortes facilidades na web, mas também dominar e acompanhar o desenvolvimento dos processos de comunicação na web; destacam-se os conhecimentos em Http (“linguagem de comunicação” entre cliente e servidor web), Html (“linguagem base da Internet” para estruturar páginas web), Cascading Style Sheets (css - linguagem para formatar conteúdos estruturados), Javascript (linguagem de programação para tornar as páginas web dinâmicas e interativas), Xml Path Language (xpath - linguagem de consulta para selecionar nós num documento xml), Expressões regulares (regex – define um padrão a ser usado para procurar palavras ou grupos de palavras).

### **3 Solução tecnológica adotada para extração, armazenamento, processamento, classificação e análise dos dados**

A infraestrutura de extração, armazenamento, processamento, classificação e análise foi construída tendo como base a linguagem de programação Python. A escolha do Python reside obviamente nas suas características, que se descrevem de seguida, mas também na exploração do seu potencial enquanto agente de mudança no INE:

- É uma linguagem de uso geral, ou seja, cobre as várias fases do processo no caso específico do web scraping mas é capaz de muito mais, desde o desenvolvimento web, à visualização ou ao machine learning;

- É simples de entender (extremamente legível), muito intuitiva e produtiva (a escrita é virada para a estruturação dos scripts sendo por isso muito fácil traduzir o raciocínio em algoritmos);
- Frequentemente escolhida para soluções pedagógicas que visam ensinar a programar e com uma curva de aprendizagem curta é a escolha perfeita para quem nunca programou e pretende adquirir rapidamente conhecimentos e competências em IT;
- É uma linguagem multiplataforma, free e open source, com uma comunidade muito grande e ativa, sempre pronta a esclarecer dúvidas e que produz constantemente packages e frameworks para executar todo o tipo de tarefas; o código-fonte distribuído pode ser testado, alterado e melhorado o que contribui em muito para a sua robustez;
- Linguagem muito bem documentada, com um site oficial bem estruturado e atualizado, com uma oferta infindável quer de cursos online (na sua maioria grátis) quer de outras ferramentas de aprendizagem, tais como ambientes de desenvolvimento integrado (IDE's) online para um primeiro contacto com a linguagem, sem necessidade de qualquer instalação;
- De fácil instalação através da distribuição gratuita Anaconda (Anaconda, 2018) que, para além do Python, agrega a maioria dos packages científicos de Python e uma série de outros recursos tais como o Jupyter Notebook (Project Jupyter, 2018) também utilizado neste projeto.

Para a extração dos dados foram usadas duas abordagens distintas com base em duas ferramentas Python:

- A **library Selenium** que cria instâncias automáticas de browsers, simulando a interação do utilizador com o website; ferramenta de eleição sempre que é necessário interagir com as páginas web (carregar em botões, imputar valores em campos de pesquisa, etc.) e portanto muito boa a lidar com Javascript;
- A **framework Scrapy** que permite fazer vários pedidos ao mesmo tempo possibilitando extrações mais rápidas; assenta em mecanismos pré-construídos que são customizáveis e numa bateria de regras que permite seguir links com um determinado padrão; não consegue lidar com Javascript.

A automatização da execução dos scripts para extração da informação foi efetuada através do Task Scheduler (Windows Task Scheduler, 2018) do sistema operacional Windows.

Tradicionalmente o INE trabalha com bases de dados relacionais (SQL - Structured Query Language) possuindo a infraestrutura e know-how necessários. Neste projeto quisemos juntar mais um desafio e optámos por testar e utilizar pela primeira vez uma base de dados não relacional (NoSQL).

Não obstante a vontade de experimentar a nova geração de bases de dados percebemos que uma nova abordagem seria necessária, uma vez que:

- sabíamos que íamos recolher grandes quantidades de informação mas desconhecíamos, à priori, o seu volume; poderíamos ou não vir a precisar de processamento em várias máquinas;
- a escrita e a leitura teriam que ser rápidas, para que não fossem perturbados os processos de extração e análise;
- finalmente e muito importante, precisávamos de uma base de dados sem um esquema definido à priori ou seja, queríamos armazenar informação sem nos preocuparmos com a estrutura dos dados, nº de campos, tipo de campos, etc..

A juntar a estas necessidades pretendíamos uma solução: a) simples de instalar e de consultar, b) com suporte para a linguagem Python, c) multiplataforma e free and open source.

A escolha recaiu sobre o MongoDB (MongoDB, 2018) que é uma base de dados NoSQL orientada a documentos, ou seja a informação é guardada não em tabelas mas em documentos com esquema dinâmico. Os documentos de MongoDB (conjuntos de pares chave-valor) são semelhantes aos objetos Json (javaScript object notation) e a sua estrutura pode ser alterada, adicionando novos campos ou excluindo os já existentes. O MongoDB oferece alto desempenho e fácil escalabilidade.

Conjuntamente com o MongoDB é utilizado o Robomongo/Robo 3T (Robo 3T, 2018) que é um software grátis, multiplataforma, que funciona como uma interface gráfica (Gui) para gestão/manipulação de dados armazenados no MongoDB; substitui a necessidade de utilização da linha de comandos (consola) possibilitando uma experiencia ao utilizador muito melhor que a Mongo Shell padrão do MongoDB.

Para o processamento, classificação e análise foi escolhida a solução Jupyter Notebook que é um ambiente computacional interativo no qual se pode combinar execução de código, texto explicativo, visualização gráfica, imagem e vídeo (“both human and machine readable”). Trata-se de uma aplicação web, free and open source, incluída na distribuição python Anaconda, que permite editar e executar documentos, chamados “notebooks”, através de um browser e que pode ser acedida num PC sem ligação à Internet ou num servidor remoto através da Internet. É uma ferramenta que permite, ao mesmo tempo, explorar/analisar e partilhar informação pois os inputs/outputs podem ser guardados e distribuídos no formato de notebook (ficheiro .ipynb). Exporta em diversos formatos (Python, Html, Latex, Markdown, Slides, etc.) e suporta várias linguagens (Python, R, Julia, Scala, Javascript, C++, etc.).

O desenvolvimento e implementação do protótipo foi efetuado num PC com Windows uma vez que é o sistema operativo usado no INE; contudo, de forma a tornar completamente real a premissa “free and open source” e a comprovar a característica “multiplataforma” comum a todas as ferramentas escolhidas, foram recentemente realizados testes em Linux que se revelaram um sucesso em todas as

fases da infraestrutura; os testes foram efetuados com as distribuições de Linux “Ubuntu” (numa máquina virtual) e “Raspbian” (num Raspberry Pi 3).

#### 4 Teste e implementação do protótipo

A seleção do grupo de produtos do IPC e correspondente website a utilizar no protótipo obedeceu aos seguintes critérios: a) website já utilizado na recolha online manual do IPC, b) website com grande quantidade de produtos recolhidos na recolha manual, c) loja representativa do respetivo grupo de produtos.

A escolha recaiu sobre o website português do IKEA (multinacional especializada em desenho e venda de móveis, utensílios de cozinha e acessórios para a casa) onde mensalmente são recolhidos, de forma manual, os preços de 109 produtos pertencentes a 11 classes da COICOP - Classificação do consumo individual por objetivo - (INE, 2012) relativas ao mobiliário para o lar.

Durante a fase de testes não eram esperadas (e não foram observadas), entre as duas recolhas (manual e automatizada), diferenças nas variáveis recolhidas para os 109 produtos, uma vez que se tratava da mesma fonte de informação (website). Contudo foi possível identificar diferenças em relação ao número de produtos recolhidos decorrentes das distintas frequências de observação, entre recolhas (um momento de observação na recolha manual e cinco, um por cada dia da semana, na recolha automatizada).

O website demonstrou ser muito estável; as falhas pontuais dos web scrapers foram sobretudo originadas por problemas de comunicação, que foram entretanto solucionados.

Na Tabela 1 pode observar-se a acentuada redução no tempo de recolha/armazenamento associada ao web scraping, em comparação com a recolha manual. Para além de mais intensiva em horas de trabalho e pouco estimulante para quem a executa, a recolha manual está mais sujeita a erros no processo de “copy/paste” do website para a base de dados; a qualidade dos indicadores obtidos pode ainda aumentar através da incorporação de mais produtos observados (aumento da cobertura) e maiores frequências de observação, ou seja, são perceptíveis os potenciais ganhos de eficiência e de qualidade.

**Tabela 1** – Recolha manual *versus* web scraping (website do IKEA).

método de recolha	nº de produtos recolhidos	duração da recolha	observações
recolha manual	109	≈ 5 horas	
web scraping <sup>a</sup>	109	≈ 35 minutos <sup>c</sup>	com “input” no website das referências dos 109 produtos (selenium)
web scraping <sup>b</sup>	≈ 7000	≈ 60 minutos <sup>c</sup>	todos os produtos (scrapy)

a – em produção no IPC; b - incorporação no IPC em estudo; c – inclui “data cleaning” e armazenamento.

Face aos resultados obtidos entendeu-se que a substituição da recolha manual era viável e podia ser efetuada de imediato. A recolha mensal passou a ser realizada pelos web scrapers desenvolvidos, de acordo com a seguinte metodologia de recolha: dois momentos de observação (dois dias da semana de recolha) sempre que no primeiro momento os 109 produtos não estejam disponíveis no website; recolha paralela de todos os produtos disponíveis, por defeito (sem interação/customização) no website (cerca de 7000 produtos).

A classificação COICOP é efetuada com base em técnicas de “string matching” combinando os descritivos detalhados dos produtos recolhidos e a classificação utilizada no próprio website.

Mensalmente são calculadas e incorporados no índice, para cada uma das 11 classes COICOP observadas, as médias geométricas e as variações de preço para os 109 produtos da lista IPC. Num futuro próximo pretende-se que o cálculo inclua: a) todos os produtos disponíveis no website pertencentes às 11 classes COICOP observadas (cerca de 1000 produtos); b) outros produtos disponíveis no website (cerca de 6000 produtos) pertencentes a outras classes de produtos observadas no IPC mas não obtidas através do IKEA.

O INE informou previamente o IKEA sobre os objetivos do projeto tendo obtido total cooperação por parte da empresa; desta forma foram evitados bloqueios ao web scraping e minimizados os efeitos de um dos maiores desafios/problemas deste processo que é a alteração, sem aviso prévio, de um site. Não obstante a legalidade do processo (Lei do Sistema Estatístico Nacional, 2008) e a relação profissional estabelecida com o IKEA, o INE assegura que os processos automatizados de extração de informação não causam impactos negativos na performance do site, cumprindo as regras da net ou “netiquette”: minimização dos pedidos de página efetuados pelos web scrapers, temporização da extração das variáveis e identificação dos web scrapers.

## 5 Conclusões e desenvolvimentos futuros

Desta primeira experiência de recolha automática de dados da web testada pelo INE, podemos retirar as seguintes conclusões:

- A recolha através de web scraping constitui uma clara alternativa à recolha manual na web, tratando-se de um processo mais eficiente (potencial redução de custos e indicadores com mais qualidade);
- O seu potencial enquanto substituto/complemento da recolha tradicional (nas lojas físicas) deve ser explorado e amplamente estudado;
- A solução Ad-hoc desenvolvida neste projeto é eficiente e flexível, proporcionando um nível de controlo sobre todo o processo impossível de atingir com soluções “point and click”;

- Implementar uma solução Ad-hoc exige muitos conhecimentos e competências específicas em IT mas também representa uma oportunidade imensa de aprendizagem e preparação para os novos desafios e contantes mudanças;
- A aquisição de conhecimentos e competências de IT foi efetuada através de cursos e materiais online gratuitos (MOOC - Massive Open Online Course -, Youtube, sites oficiais, etc.). Desafios que têm a ver com aquisição de competências e conhecimentos específicos sempre existiram no INE, mas uma abordagem baseada em soluções adaptadas ao tempo/horários disponíveis, grátis e colaborativas devem ser uma opção a considerar no futuro;
- O uso de software free and open source não constituiu qualquer obstáculo ao desenvolvimento do projeto e por isso sugerimos que a sua utilização seja promovida.

Quanto aos trabalhos e desenvolvimentos futuros, pretende-se:

- Continuar com a progressiva substituição da recolha manual na web;
- Melhorar o atual modelo da classificação COICOP e testar a possibilidade de utilização de técnicas de machine learning;
- Promover estudos metodológicos que avaliem a possibilidade de substituição da recolha tradicional para alguns produtos;
- Apostar na requalificação dos técnicos existentes no INE, privilegiando a utilização das plataformas online de aprendizagem (MOOC) e o uso de recursos “free and open source”.

## Referências

- ANACONDA (2018), <https://www.anaconda.com/> (acedido em 17 de Maio 2018).
- BOETTCHER, I. (2015) Automatic price collection on the Internet (web scraping). NNTS Conference, 10-12 March 2015, Brussels.
- CAVALLO, A., RIGOBON R. (2016) The Billion Prices Project: Using Online Prices for Inflation Measurement and Research. *Journal of Economic Perspectives*, Vol 30(2):151-178.
- EUROSTAT (2013). Compendium of HICP Reference Documents - 2013 edition, <http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-017> (acedido em 16 de Maio 2018).
- GRIFFIOEN,R., TEN BOSCH, O. AND HOOGTEIJLING, E. (2016) Challenges and solutions to the use of Internet data in the Dutch cpi. UNECE, Workshop on Statistical Data Collection ‘Visions on Future Surveying’.
- HOEKSTRA, R., TEN BOSCH, O. AND HARTEVELD, F. (2010) Automated Data Collection from Web Sources for Official Statistics: First Experiences. Statistics Netherlands.

- IDC & EMC (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, [www.emc.com/leadership/digital-universe/2012iview/executive-summary-a-universe-of.htm](http://www.emc.com/leadership/digital-universe/2012iview/executive-summary-a-universe-of.htm) (acedido em 16 de Maio 2018).
- INE (2012). Classificação do consumo individual por objetivo, adaptada às necessidades do índice de preços no consumidor, <http://smi.ine.pt> (acedido em 17 de Maio 2018).
- INE (2009). Índice de Preços no Consumidor. Documento Metodológico, <http://smi.ine.pt> (acedido em 17 de Maio 2018).
- INTERNET LIVE STATS (2018). Total number of Websites, [www.internetlivestats.com/total-number-of-websites](http://www.internetlivestats.com/total-number-of-websites) (acedido em 16 de Maio 2018).
- LEI DO SISTEMA ESTATÍSTICO NACIONAL (2008). Lei nº 22/2008 (D.R. nº 92 1ª Série, de 2008-05-13).
- MCKINNEY, W. (2012) Python for Data Analysis. Data Wrangling with Pandas, Numpy, and Ipython. O'Reilly Media.
- MONGODB (2018), <https://www.mongodb.com/> (acedido em 16 de Maio 2018).
- POLIDORO F., GIANNINI R., LO CONTE R., MOSCA S., ROSSETTI F. (2015) Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. Statistical Journal of the IAOS 31, 165-176.
- PROJECT JUPYTER (2018), <http://jupyter.org/> (acedido em 17 de Maio 2018).
- PYTHON SOFTWARE FOUNDATION (2018), <https://www.python.org/> (acedido em 17 de Maio 2018).
- ROBO 3T (2018), <https://robomongo.org/> (acedido em 16 de Maio 2018).
- SWIER, N. (2015) Using Web Scraped Data to Construct Consumer Price Indices. NNTS Conference, 10-12 March 2015, Brussels.
- WINDOWS TASK SCHEDULER (2018), [https://msdn.microsoft.com/en-us/library/windows/desktop/aa383614\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/windows/desktop/aa383614(v=vs.85).aspx) (acedido em 17 de Maio 2018).



## Avaliação estatística da eficácia do projeto AdolesSer

Paulo Infante<sup>1</sup> · Gonçalo Jacinto<sup>2</sup> · Anabela Afonso<sup>3</sup> · Ana Carla Coelho<sup>4</sup> ·  
Ana Gabriela Pontes<sup>5</sup> · Isabel Fernandes<sup>6</sup> · Edgar Palminhas<sup>7</sup>

**Resumo:** O projeto AdolesSer é dirigido a alunos dos 6º e 9º anos de Escolas de Évora e pretende reforçar/promover a informação sobre as dimensões da sexualidade humana, crescimento e relações interpessoais. Neste estudo faz-se uma breve caracterização dos alunos do Concelho de Évora em termos de conhecimentos, comportamentos e atitudes sobre a sexualidade e avalia-se a eficácia das intervenções realizadas com base em questionários submetidos antes e após essas intervenções.

**Palavras-chave:** Análise Classificatória, Análise de Correspondências Múltiplas, Modelos Lineares Generalizados.

### 1 Introdução

A Adolescência é o resultado da interligação entre os fatores biológicos, sociais, comportamentais e culturais, que ocorrem na transição da infância para a idade adulta (Cole *et al.*, 2004). Nesta etapa ocorrem as principais transformações psicológicas e fisiológicas que alteram a qualidade dos processos cognitivos e emocionais. Enquanto as transformações físicas da puberdade não afetam diretamente os estados psicológicos dos adolescentes, os efeitos psicológicos são medidos pelas respostas às alterações físicas emitidas pelos pais, colegas e pelos próprios adolescentes (Sprinthall *et al.*, 1999). A educação sexual foi integrada, por lei, na educação para a saúde, com vista à promoção da saúde física, psicológica e

---

<sup>1</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [pinfante@uevora.pt](mailto:pinfante@uevora.pt)

<sup>2</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [gjcj@uevora.pt](mailto:gjcj@uevora.pt)

<sup>3</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

<sup>4</sup> ARSA/ACES AC/ UCC Évora, [AnaCarla.Coelho@alentejocentral.min-saude.pt](mailto:AnaCarla.Coelho@alentejocentral.min-saude.pt)

<sup>5</sup> ARSA/ ACES AC/ UCC Évora, [Ana.Pontes@alentejocentral.min-saude.pt](mailto:Ana.Pontes@alentejocentral.min-saude.pt)

<sup>6</sup> ARSA/ ACES AC/ UCC Évora, [Isabel.Fernandes@alentejocentral.min-saude.pt](mailto:Isabel.Fernandes@alentejocentral.min-saude.pt)

<sup>7</sup> Porta Mágica, [edpalminhas@hotmail.com](mailto:edpalminhas@hotmail.com)

social dos indivíduos. O projeto AdoslesSer – Projeto de educação para a saúde sexual e reprodutiva - desenvolvido pela equipa de saúde escolar da Unidade de Cuidados na Comunidade de Évora, do Agrupamento de Centros de Saúde do Alentejo Central (ACES), é dirigido a alunos dos 6º e 9º anos de escolas de Évora e pretende reforçar/promover a informação sobre as dimensões da sexualidade humana, crescimento e relações interpessoais, assim como a aquisição de competências.

A avaliação da eficácia do projeto, bem como das melhorias a implementar, é concretizada pela aplicação de um questionário inicial de validação de conceitos e conhecimentos (pré-intervenção) antes da realização da primeira sessão de intervenção e a sua reaplicação imediatamente após a realização da última sessão (pós-intervenção).

Neste trabalho faz-se uma breve caracterização dos alunos do Concelho de Évora no ano letivo de 2016/2017 em termos de conhecimentos, comportamentos e atitudes sobre a sexualidade e, com base em questionários submetidos antes e após intervenções realizadas junto dos alunos, avalia-se a eficácia do projeto AdoslesSer. Este projeto teve início no ano letivo de 2013/14 e só este ano é possível comparar o desempenho dos alunos que participaram previamente no projeto (estavam na altura no 6º ano) com os alunos que participaram no projeto pela primeira vez apenas no 9º ano, permitindo também fazer uma avaliação longitudinal do projeto AdoslesSer.

## **2 Métodos**

### **2.1 Recolha da informação**

A população alvo deste estudo é constituída por alunos do 6º e 9º ano de escolaridade das escolas do concelho de Évora.

Em cada ano de escolaridade foi adotado um esquema de amostragem multietápico, estratificado por escola e com seleção aleatória total de 10 turmas respeitando a proporcionalidade do número de turmas por escolas. Responderam ao questionário todos os alunos que estavam presentes na sala de aula no dia do questionário e que tinham o consentimento de participação assinado pelos pais ou encarregados de educação.

Aos alunos das turmas selecionadas foi solicitada a resposta anónima aos questionários na sala de aula em dois instantes: 1) antes da realização da primeira sessão de intervenção (pré-intervenção), e 2) imediatamente após a realização da última sessão (pós-intervenção).

Os questionários eram compostos por diversas questões de caracterização do aluno, participação neste projeto em anos anteriores e fontes de informação sobre a

temática. Além disso continha outras questões que visavam avaliar os conhecimentos sobre as temáticas abordadas nas intervenções, que se podem agrupar em: corpo, sistema reprodutor, grupo e o próprio, para alunos do 6º ano; contraceção, infeções sexualmente transmissíveis (IST), violência, orientação sexual, comportamento na relação e pressão de pares, para alunos de 9º ano. Em algumas questões a resposta era do tipo Verdadeiro ou Falso (V/F), conforme o aluno acertou ou não na resposta. Noutras questões o aluno indicava se Concordava ou Discordava (C/D) da afirmação, tendo-se analisado se o aluno deu ou não a resposta esperada.

Este projeto está a ser realizado desde o ano letivo 2013/2014 e no ano letivo 2016/17 a amostra final do 6º ano é composta por 155 alunos na pré-intervenção e 150 na pós-intervenção, enquanto no 9º ano a amostra final é composta por 184 alunos na pré-intervenção e 217 na pós-intervenção. Para manter a confidencialidade não foi possível fazer o emparelhamento dos questionários na pré e pós intervenção, sendo as amostras analisadas como se fossem independentes.

## 2.2 Análise estatística

Para averiguar associações entre as respostas dadas às questões (categorizadas apenas em duas categorias, resposta certa ou de acordo com o esperado *vs.* caso contrário), por ano de escolaridade, realizou-se uma análise de correspondências múltiplas (Greenacre, 2007). Para a escolha do número de dimensões a reter foram considerados os seguintes critérios: i) análise da representação gráfica dos valores próprios; ii) número de valores próprios superiores a  $1/Q$ , sendo  $Q$  o número de variáveis; iii) Percentagem total da variabilidade explicada. Tendo em conta as categorias mais discriminantes das variáveis que mais contribuíam para cada uma das dimensões consideradas, foram identificados alguns perfis.

Foi usado o teste  $t$  para amostras independentes para comparar o desempenho médio dos alunos por período de intervenção, sexo, freguesia de residência, existência de experiência sexual e participação prévia neste projeto.

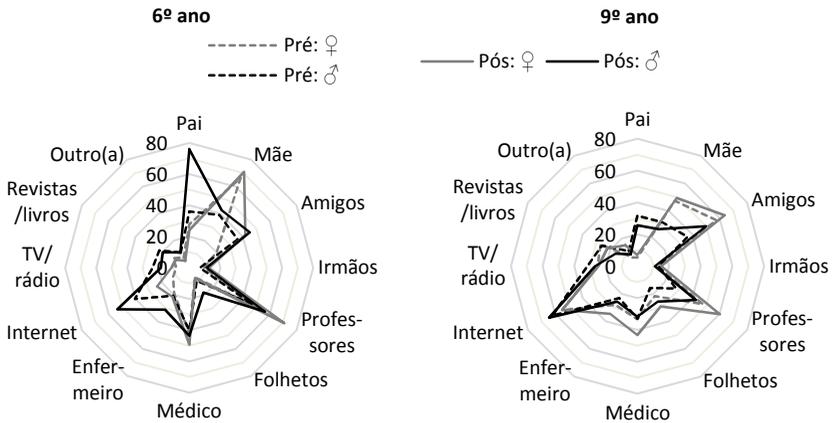
Foram usados testes de associação do qui-quadrado, correlação ponto-bisserial e modelos de regressão logística para as questões de V/F para identificar algumas características dos alunos que afetam o seu desempenho. Para a análise longitudinal, com base nos resultados medidos na pré-intervenção em 2016/2017, comparou-se o desempenho dos alunos que já tinham sido intervencionados em 2013/2014 com os que nunca tiveram a formação do projeto AdolesSer. Ajustou-se, ainda, um modelo de regressão logística para as questões C/D (Hosmer *et al.*, 2013) para identificar os fatores potenciadores para um aluno ter mais de 70% de respostas de acordo com o esperado, pois foi nesta parte do questionário que se registaram as maiores diferenças entre os dois grupos de alunos.

### 3 Resultados

#### 3.1 Caracterização geral

Em ambos os anos de escolaridade obteve-se mais respostas de alunos do sexo feminino (54% no 6º ano e 58% no 9º ano na pré-intervenção; 53% no 6º ano e 58% no 9º ano na pós-intervenção).

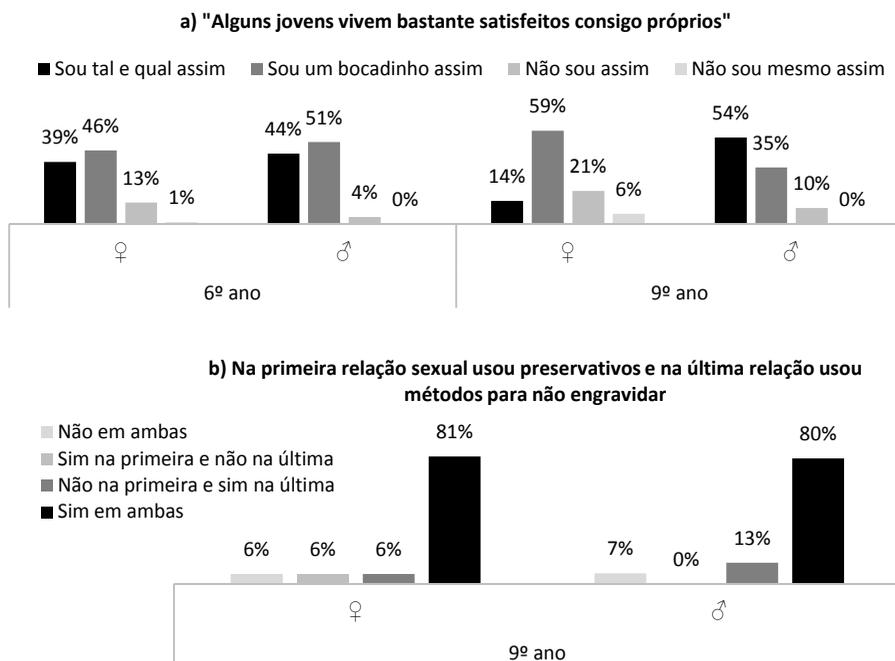
Relativamente às fontes de informação sobre a sexualidade, nas raparigas destaca-se como maior fonte de informação na pós-intervenção os professores em ambos os anos e a mãe no 6º ano, enquanto nos rapazes se destaca a *internet* em ambos os anos e o pai no 6º ano (Figura 1).



**Figura 1** – Fontes de informação sobre sexualidade.

A auto estima parece ser mais elevada entre os rapazes do que entre as raparigas (Figura 2a) e entre os alunos do 6º ano do que entre os alunos do 9º ano, o que pode estar relacionado com a fase da adolescência em que os alunos se encontram.

Cerca de 3 em cada 4 alunos (rapazes e raparigas) indicaram ter tido pela primeira vez relações sexuais entre os 14 e os 15 anos. Os alunos indicaram ter usado preservativo na sua primeira relação sexual e ter usado métodos para não engravidar na última relação sexual (Figura 2b).



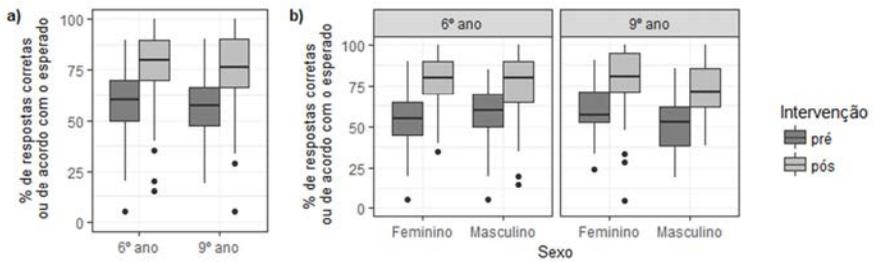
**Figura 2** – Algumas características dos alunos na pré-intervenção: a) autoavaliação da autoestima; b) utilização de métodos contraceptivos na 1ª e na última relação sexual.

### 3.2 Análise de desempenho das intervenções

Analisando as respostas dadas pelos alunos nas duas passagens do questionário (pré e pós-intervenção), podemos constatar que a percentagem média de respostas corretas/dadas de acordo com o esperado aumentou após a intervenção, e diminuiu a percentagem de respostas “não sei”, em ambos os anos de escolaridade ( $p < 0,001$ , Figura 3a).

Não foram detetadas diferenças no desempenho dos alunos por tipo de freguesia de residência ( $p > 0,26$ ) nem por já terem ou não tido experiência sexual ( $p > 0,77$ ).

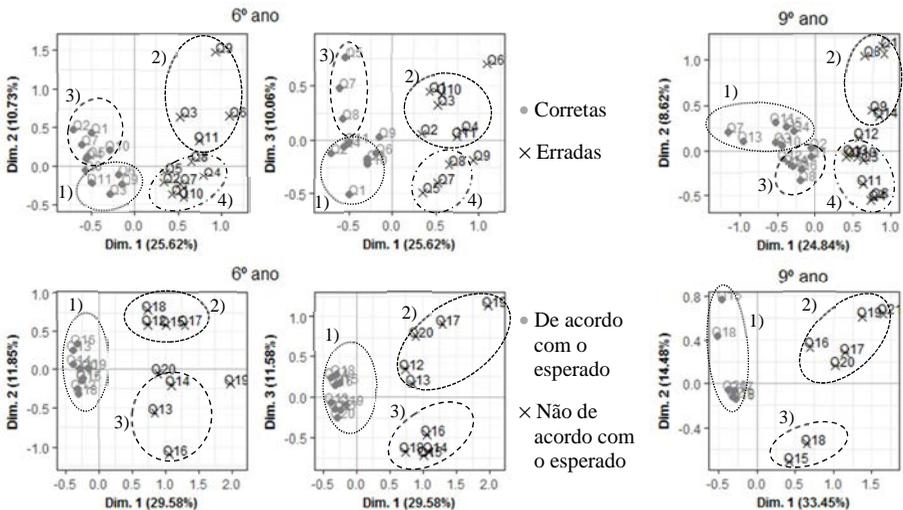
No 6º ano não existem diferenças significativas entre sexos em cada uma das passagens ( $p > 0,17$ ). No 9º ano as raparigas tiveram, em média, um desempenho superior ao dos rapazes em ambas as passagens ( $p < 0,02$ , Figura 2b). Observou-se uma correlação (Spearman) positiva entre o desempenho dos alunos e a quantidade de fontes de informação a que recorrem ( $r_s > 0,18$  e  $p < 0,02$ ).



**Figura 3** – Distribuição da percentagem de respostas corretas por período de intervenção e: a) ano de escolaridade; b) sexo do aluno e ano de escolaridade.

### 3.3 Associações entre as respostas dadas

Através da análise multivariada das respostas dadas nas questões foram identificados entre 3 e 4 perfis: 1) associa alunos que tendem de um modo geral a acertar ou a dar a resposta esperada a quase todas as questões; 2) associa alunos que tendem a não ter acertado ou a não ter dado a resposta esperada à maior parte das questões; 3) e 4) associa alunos que se distinguem dos restantes ora porque acertam, ou não, ou dão a resposta esperada, ou não, num conjunto particular de questões (Figura 4).



**Figura 4** – Representação das primeiras dimensões da análise multivariada das respostas às questões dadas pelos alunos respondentes, divididas pelo tipo de resposta.

### 3.4 Algumas relações mais relevantes

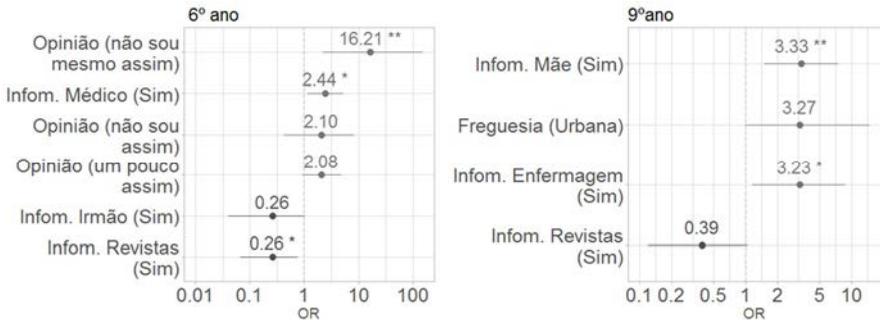
A autoestima é a faceta mais importante do autoconceito, encontrando-se associada aos aspetos avaliativos que o sujeito elabora a seu respeito (Romano *et al.*, 2007). Neste trabalho concluímos que a autoestima está associada ao ano ( $p < 0,001$ ), sendo maior nos alunos do 6º ano. Esta diferença acentua-se na pós intervenção.

Na pré-intervenção junto dos alunos do 9º ano, o melhor desempenho dos alunos nas questões sobre IST (infecções sexualmente transmissíveis) está correlacionado positivamente com os alunos terem usado preservativo na primeira relação sexual (corr. biss. pont.=0,41,  $p=0,02$ ). Esse desempenho não está correlacionado com a não utilização de métodos contraceptivos na última relação sexual (corr. biss. pont.=0,08,  $p=0,35$ ). Ainda para este grupo de alunos, os que não acham que *ao ter um comportamento igual os outros têm maior aceitação*, parecem ter propensão para que a primeira relação sexual ocorra na altura desejada ( $p=0,09$ ).

### 3.5 Alguns fatores determinantes para o bom e mau desempenho

Para o 6º ano ajustou-se um modelo de regressão logística (Hosmer *et al.*, 2013) para o mau desempenho dos alunos, tendo como evento de interesse o ter acertado no máximo em 2 respostas (menos de 20% de respostas corretas), nas questões do tipo V/F na pré-intervenção. Este modelo revelou um bom ajustamento aos dados ( $p=0,49$  no teste da Deviance) e uma boa capacidade discriminativa (AUC=0,74). Com base nos resultados do modelo, podemos concluir que procurar informação junto do médico mas não em revistas, não falar com os irmãos sobre a temática e ter baixa autoestima são fatores potenciadores de um mau desempenho (Figura 5a).

No 9º ano considerou-se que alunos que responderam corretamente a mais de 80% das questões do tipo V/F tiveram um desempenho muito bom. Com base no modelo de regressão logística ajustado (valor  $p=0,17$  no teste da Deviance e AUC=0,73), podemos concluir que conversar com as mães e com o pessoal de enfermagem, não procurar informações em revistas e viver numa freguesia urbana são fatores potenciadores de um bom desempenho (Figura 5b). Uma nota para referir que no 9º ano se optou por definir o evento de interesse no modelo de regressão logística como o bom desempenho, porque não se conseguiu ajustar nenhum modelo com boa capacidade discriminativa para o mau desempenho (como se conseguiu para os resultados dos alunos do 6º).



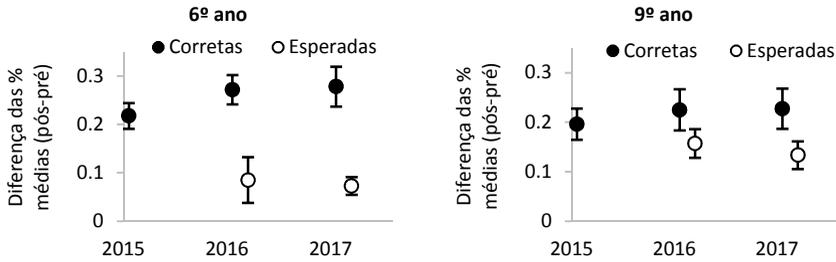
**Figura 5** – Razão de chances (OR) e respetivos intervalos de confiança para alunos com: a) mau desempenho no 6º ano (esq.), b) muito bom desempenho no 9º ano (direita). (\* sig. a 5% e \*\* sig. a 1%).

### 3.6 Análise longitudinal

Nesta secção iremos comparar o desempenho dos alunos do 9º ano que já participaram no projeto (em 2013/14 estavam no 6º ano) com o dos alunos do 9º ano que estão a participar pela primeira vez no projeto em 2016/17, com base nas respostas dadas na pré intervenção. Cerca de 51% dos rapazes afirmou já ter participado no projeto contra 82% das raparigas. Observamos que 52% dos que participaram previamente no projeto indicaram quatro ou mais fontes de informação contra 40% entre os que não participaram no projeto. Quer tenham ou não participado anteriormente no programa, as principais fontes privilegiadas são a *internet*, os amigos e a mãe. Entre os que não participaram destaca-se ainda os professores, ao passo que entre os que participaram destaca-se o médico e no caso dos rapazes também a televisão/rádio.

Podemos concluir que há indícios de um melhor desempenho de quem já participou relativamente aos que não participaram previamente no projeto (valor  $p=0,06$ ). Verificou-se uma diferença significativa nas respostas às questões do tipo C/D ( $p=0,002$ ), não havendo diferenças significativas nas questões V/F ( $p=0,80$ ). Nas questões do tipo C/D registaram-se diferenças significativas nas temáticas de orientação sexual ( $p=0,036$ ) e nas do comportamento na relação ( $p=0,007$ ).

As raparigas deram mais respostas de acordo com o esperado do que os rapazes (todas as questões com diferenças absolutas entre 12% e 32%). O padrão da diferença entre sexos é semelhante quer tenham ou não participado anteriormente no projeto. De um modo geral, as diferenças entre sexos são menores entre os que já participaram no projeto do que entre os que não participaram. Relativamente a 2015, entre as duas passagens do questionário, registou-se nos anos seguintes uma melhoria no desempenho dos alunos nas questões do tipo V/F (Figura 6).



**Figura 6** – Diferença entre as percentagens médias de respostas corretas ou de acordo com o esperado entre a pós e a pré-intervenção, por ano letivo e ano de escolaridade (as barras representam os limites do respetivo intervalo de confiança a 95%).

Ajustou-se um modelo de regressão logística ao desempenho nas questões C/D da pré-intervenção dos alunos do 9º ano, considerando como evento de interesse na variável resposta o aluno ter mais de 70% de respostas de acordo com o esperado. Este modelo ajustou-se bem aos dados ( $p=0,10$  no teste Deviance) e apresentou uma excelente capacidade discriminativa ( $AUC=0,80$ ). Neste modelo, registou-se uma interação significativa entre o sexo e a fonte de informação ser a internet, tendo sido possível identificar como fatores potenciadores de ter mais de 70% de respostas de acordo com o esperado: ser do sexo feminino e não obter informação sobre a temática na *internet* ( $OR=18,01$ ,  $IC_{95\%}(OR)=(4,89; 66,36)$ ); ser do sexo masculino e obter informações na *internet* ( $OR=1,12$ ;  $IC_{95\%}(OR)=(0,43; 2,89)$ ); procurar informação junto das mães ( $OR=3,23$ ;  $IC_{95\%}(OR)=(1,54; 7,08)$ ) e não junto dos médicos ( $OR=2,14$ ,  $IC_{95\%}(OR)=(1,00; 4,57)$ ) nem em revistas ( $OR=2,34$ ;  $IC_{95\%}(OR)=(1,05; 5,22)$ ); e ter participado no projeto AdolesSer quando estavam no 6º ano ( $OR=2,63$ ,  $IC_{95\%}(OR)=(1,22; 5,69)$ ).

## 4 Conclusão

Os resultados sugerem que a intervenção da equipa do projeto AdolesSer contribuiu para a aquisição de conhecimentos sobre a sexualidade e relações interpessoais dos alunos do 6º e do 9º ano do concelho de Évora. Observam-se diferenças significativas nas percentagens de respostas corretas entre a pré e a pós-intervenção. Esta significância é observada tanto nos rapazes como nas raparigas.

Quando comparamos o desempenho nas respostas dos alunos do 9º ano que já participaram no projeto (estavam na altura no 6º ano) com os alunos do 9º ano que nunca participaram no projeto, parece haver um melhor desempenho entre quem já participou no projeto. Além de ter participado no projeto ser um fator potenciador de um melhor desempenho nas questões do tipo C/D, também o ser do sexo masculino para os que obtêm informação na internet ou o ser do sexo feminino para

os que não obtêm informação na internet e o procurar informação junto das mães em vez de a procurar em revistas ou nos médicos parecem aumentar a probabilidade de dar respostas de acordo com o esperado. Já nas questões V/F a probabilidade de um bom desempenho para os alunos do 9º ano é mais elevada para os que vivem numa freguesia urbana e procuram informação sobre a temática junto das mães e do pessoal de enfermagem e não em revistas.

Finalmente, conclui-se que o projeto AdolesSer contribui para a aquisição de conhecimentos e competências sobre sexualidade humana, crescimento e relações interpessoais, quer ao longo do mesmo ano letivo, quer mesmo após um horizonte temporal de 3 anos. Estas competências adquiridas são relevantes na educação para a saúde física, psicológica e social dos adolescentes e, portanto, têm um impacto direto na sociedade.

## **Agradecimentos**

Anabela Afonso, Gonçalo Jacinto e Paulo Infante são membros do CIMA, centro de investigação financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal, pelo projeto PEst-OE/MAT/UI0117/2019.

## **Referências**

- COLE, M. & COLE, S. R. (2004). *O Desenvolvimento da Criança e do Adolescente*, 4ª Edição, Artmed, Brasil.
- GREENACRE, M. (2007). *Correspondence analysis in practice*, 2nd edition, Chapman & Hall/ CRC Press, Boca Raton.
- HOSMER, D., LEMESHOW, S. & STURDIVANT, R. (2013). *Applied Logistic Regression*, 3rd Edition, Wiley New York.
- JOHNSON, R. & WICHERN, D. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.
- Sprinthall, N. A., & Collins, W. A. (1999). *Psicologia do adolescente: uma abordagem desenvolvimentista* (2ª ed.).
- ROMANO, A., NEGREIROS, J., & MARTINS, T. (2007). Contributos para a Validação da Escala de Auto-estima de Rosenberg numa Amostra de Adolescentes da Região Interior Norte do País, *Psicologia, Saúde e Doenças*, 8(1), 107-114.

## Labour market attractiveness in the EU at country-level

João Sollari Lopes<sup>1</sup> · Marco Folgado Moura<sup>2</sup> · Sónia Quaresma Gonçalves<sup>3</sup>

**Abstract** In this work we present a preliminary exploratory study to characterize EU-28 countries according to their Labour Market Attractiveness. This concept was developed by combining 76 variables from 6 Eurostat datasets on different subjects (i.e., Demographics; Earnings structure; Education and training; Life conditions; Employment and unemployment; and National accounts). Using data mining techniques, we showed that this set consistently captures the country-level heterogeneity in EU. Although preliminary, this study suggests that this characterization can help to understand different aspects of the EU labour market (e.g., Skills mismatch, Labour Market Mobility).

**Key-words:** Labour Market Attractiveness, EU-28, Social Networks, Partition-Around-Medoids, Model Selection.

### 1 Introduction

This work was developed for the European Big Data Hackathon 2017, whose aim was to answer the pressing policy question “How to tackle the mismatch between jobs and skills at regional level in Europe?” (European Commission, 2017). Indeed, the mismatch between the available skills of the labour force and the skills required by the labour market entail significant economic and social costs for individuals and firms (European Commission, 2016). Nonetheless, skills mismatch in European Union (EU) remains at 45% (CEDEFOP, 2015). There is a need for developing skills efficiently and in a coordinated manner, particularly, in the emerging new economy with a fast-changing labour market (European Commission, 2016).

We developed a framework focused on the concept of Labour Market Attractiveness. This concept has to be considered carefully and our approach

---

<sup>1</sup> Instituto Nacional de Estatística, [joao.lobes@ine.pt](mailto:joao.lobes@ine.pt)

<sup>2</sup> Instituto Nacional de Estatística, [marco.moura@ine.pt](mailto:marco.moura@ine.pt)

<sup>3</sup> Instituto Nacional de Estatística, [sonia.quaresma@ine.pt](mailto:sonia.quaresma@ine.pt)

should be seen as a first-step towards a more mature definition. As such, we took a comprehensive approach by considering most of the Eurostat labour-related datasets with regional information, comprising of 17 variables with information on Demography, Earnings structure, Education and training, Life conditions, Employment and unemployment, and National accounts. These variables were broken by several categorical levels (e.g. *age groups*, *level of education*, *qualifications*, *occupations*) originating more than 70 variables. Several data mining techniques were then considered to analyse the compiled Labour Market Attractiveness set: distances between regions were calculated and visualized using networks; the regions were clustered using these distances; and the clusters were characterized using over-representation analysis and model selection.

Our data product entailed two parts: an exploration part, in which we aimed to better understand the EU global labour market and capture its heterogeneity; and a more inferential part, where we established associations between characteristics of the EU labour market and indicators designed to capture important aspects of the labour market, such as Labour market mobility, Migration, or the previously mentioned Skills mismatch.

In the present work, we are going to present the first exploration part of our framework. Nevertheless, the follow-up association study is invaluable to fully understand the influences on the complex dynamics of the EU labour market. Indeed, the final goal is to use this understanding to help setting policies to tackle such problems as localized excess or deficit of available labour force and/or of specific labour skills. Moreover, although we are going to focus on country-level analyses, our framework can be easily applied to datasets with detailed regional information.

## 2 Material and methods

### 2.1 Datasets

For the Labour Market Attractiveness set, we considered Eurostat datasets to capture different aspects of the attractiveness of the labour market. Thus, we chose 6 main datasets: *reg\_demo* for Demographics data; *earn* for Earnings structure data; *edtr* for data on Education and training; *ilc* for Life conditions information; *employ* for Employment and unemployment data; and *na10* for National accounts data. These main datasets comprised of 17 smaller datasets from which we considered data mostly from 2014. The exceptions were *educ\_uoe\_fine06* and *nama\_10r\_2hhinc* from which we used information from 2013 in order to obtain a more complete dataset (see Table 1).

For each of the 17 Eurostat datasets composing the Labour Market Attractiveness dataset, we extracted one variable. These 17 main variables were then broken by several categorical levels [i.e. type of contract, age groups, level of education (ISCED), economic activity (NACE) and occupation title (ISCO)] originating 76 variables. The entire set of these variables composed the Labour Market Attractiveness set (Table 1).

**Table 1** – Variables from labour market attractiveness set.

variable	description	dataset	units
ARPR	At-risk-of-poverty	ilc_li41	PC_POP
ARPR_socexcl	At-risk-of-poverty or social excl.	ilc_peps11	PC_POP
low_work	Very low work intensity	ilc_lvhl21	PC_POP_YLE60
mat_depriv	Severe material deprivation	ilc_mddd21	PC_POP
rooms_pp	Number of rooms per person	ilc_lvho04n	AVG
earn_OC[...] <sub>Nace[...]</sub>	Earning by ISCO and NACE	earn_ses_hourly	MN_PPS
emp_T[...]	Employment by work contract	lfst_r_lfe2eftpt	PC_POP_YGE15
emp_Y[...]	Employment by age	lfst_r_lfe2emp	PC_POP_Y[...]
emp_Y[...] <sub>ED[...]</sub>	Employment by age and ISCED	lfst_r_lfe2eedu	PC_EMP_Y[...]
emp_Y[...] <sub>Nace[...]</sub>	Employment by age and NACE 2	lfst_r_lfe2en2	PC_EMP_Y[...]
unemp_Y[...]	Unemployment by age	lfst_r_lfu3pers	PC_POP_Y[...]
expend_ED5-8	Public expenditure on education	educ_uoe_fine06	PC_GDP
disp_income	Disposable income	nama_10r_2hhinc	PPCS_HAB
GDP	Gross Domestic Product	nama_10r_2gdp	PPS_HAB
GVAgr	Gross Value Added growth	nama_10r_2gvagr	PCH_PRE
pop_Total	Population	demo_r_d2jan	NR
pop_Y[...]	Population by age	demo_r_d2jan	PC_POP
training	Participation in education/training	trng_lfse_0	PC_POP_Y25-64

PC\_POP, Percentage of Population; AVG, Average; MN\_PPS, Mean by group in Purchasing Power Standard; PC\_GDP, Percentage of Gross Domestic Product; PPCS\_HAB, Purchasing Power Consumption Standard per inhabitant; PPS\_HAB, Purchasing Power Standard per inhabitant; PCH\_PRE, Percentage change on previous period; NR, Number.

## 2.2 Methods

Several data mining techniques were considered to analyse the compiled Labour Market Attractiveness set: Social network analysis [SNA, see Fruchterman-Reingold (1991) and references therein]; Partition-around-medoids [PAM, Kaufman & Rousseeuw (1990); see Reynolds et al. (1992) for a comparison of clustering methods]; Over-representation analysis (ORA) via hypergeometric tests [i.e., Fisher's exact test, Fisher (1922)]; and Model selection using multinomial logistic regression (Calcagno & Mazancourt, 2010). This data mining framework was chosen to allow an easy rescaling of the analyses to detailed regional data.

**SNA:** A social network was constructed using similarity values calculated as the additive inverse of weighted Euclidean distances between countries. These distances were calculated using the 76 variables weighted in such a way that each set of variables originated from one of the 17 main variables had a weight of 1. This weighting scheme was employed to assure that there was no bias towards main variables broken in many secondary variables. Prior to the calculation of the distances, the variables were made dimensionless using a min-max transformation. Finally, the network was constructed on the similarities above 0.65 using the Fruchterman–Reingold algorithm as implemented in the R package *sna*.

**PAM analysis:** The countries were clustered using a PAM analysis on the previously calculated weighted Euclidean distances. The number of clusters  $k$  was chosen by running the analysis with all possible number of clusters ( $k = 2, 3, \dots, n - 1$ , where  $n$  was the number of subjects) and examining their average silhouette width. The PAM analysis was implemented in the R package *cluster*.

**ORA:** In order to describe the clusters created, we performed an ORA on each variable of the Labour Market Attractiveness set. Before performing this analysis the variables were discretized by defining a cut-off on the 90th percentile (over-representation of the higher values) and on the 10th percentile (over-representation of the lower values). The significance of these analyses was set to  $p$ -value = 0.05.

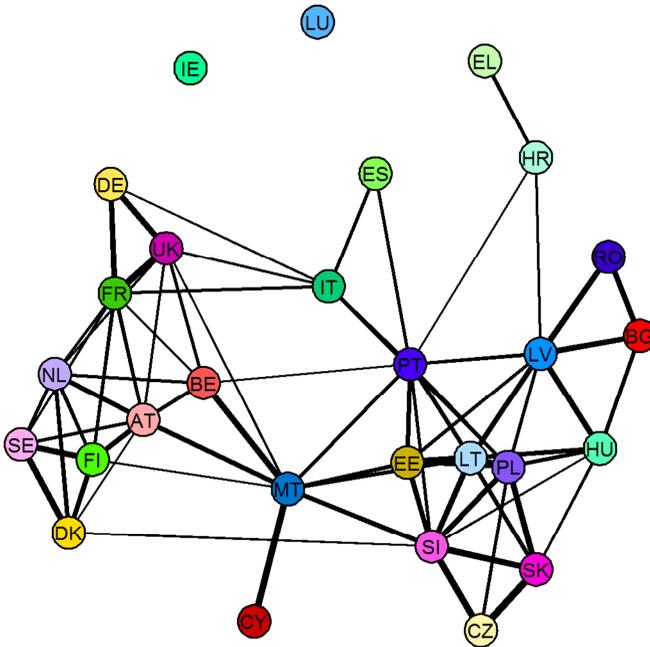
**Model selection:** The clustering classification was further studied by fitting multinomial logistic regressions using the clustering structure as dependent variable. This was performed using a single-layer feed-forward neural network as implemented in R package *nnet* [Chapter 8.10 in Venables & Ripley (2002)]. Using a model selection heuristic from the R package *glmulti*, we defined which model best predicted the clustering created. The model selection was performed using an exhaustive screening on all possible models and storing the 100 models with the lowest AIC. The maximum number of predictors for the models was chosen to be the hard threshold [i.e., maximum predictors is  $m$  such that  $(k - 1) * (1 + m) < (n - 1)$ , where  $k$  is the number of clusters, and  $n$  is the number of subjects]. Before performing model-choice, the dataset was pre-processed by removing variables until obtaining a set of 30 using the following steps: 1) any variable with missing data; 2) variables highly correlated among them ( $r < 0.90$ ); variables little correlated with the clustering classification (i.e., correlation between observed values and predicted values, where the predicted values are just the mean values of the cluster to which the subject belongs to). The selected variables were also made dimensionless using a min-max transformation.

See Lopes et al. (2018) for a more detailed description of the methods used. All the analysis can be reproduced using the R scripts in the repository <https://github.com/jsollari/EUhackathon2017>.

### 3 Results

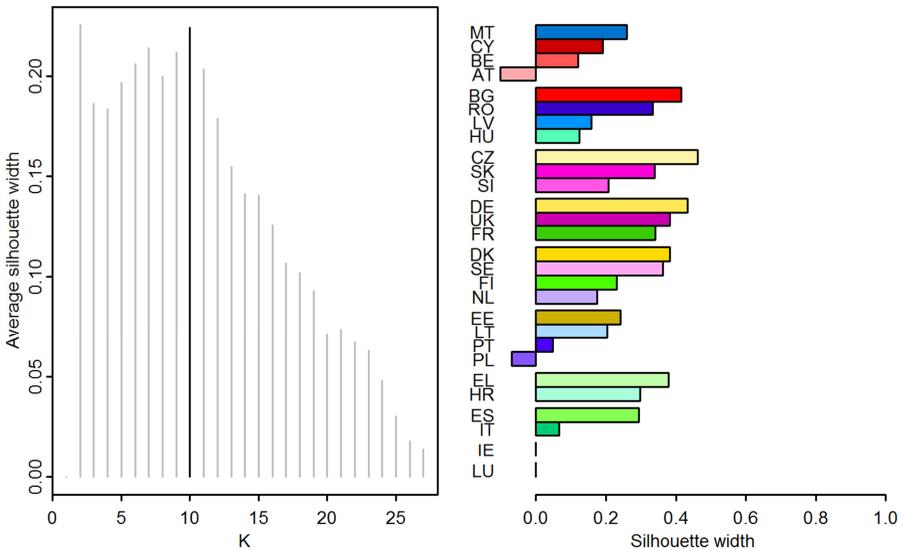
Our first approach was to visualize the entire Labour Market Attractiveness set by constructing a social network based on distances between countries (Figure 1).

From the social network analysis we saw a clear separation between Northern and Western European countries and Eastern European countries, while Southern European countries seemed to somewhat make the bridge between these two main blocks. More in detail, we observed a group formed by Denmark (DK), Finland (FI), Sweden (SE), Netherlands (NL) and Austria (AT), and another by France (FR), Germany (DE) and the United Kingdom (UK). On the other hand, Czech Republic (CZ), Estonia (EE), Lithuania (LT), Poland (PL), Slovenia (SI) and Slovakia (SK) seemed to form a group, while Bulgaria (BG), Hungary (HU), Latvia (LV) and Romania (RO) seemed to form another one. Belgium (BE), Malta (MT) and Portugal (PT) were positioned in the centre of the network. We also observed strong links between pairs of countries, i.e., Greece (EL) and Croatia (HR), Italy (IT) and Spain (ES), and Cyprus (CY) and MT. Finally, both Luxemburg (LU) and Ireland (IE) seemed to be far apart from all the other countries.



**Figure 1** – Country network based on the labour market attractiveness set.

Following the construction of a country network, we performed a clustering analysis on the data. This analysis gave a more detailed overview of the grouping, suggesting 10 groups as a good clustering structure (Figure 2, left panel). The Northern and Western European countries were further divided in Cluster 4 (DE, UK, FR) and Cluster 5 (DK, SE, FI, NL), while the Eastern European countries were divided in Cluster 2 (BG, RO, LV, HU) and Cluster 3 (CZ, SK, SI). As in the previous analysis, some countries were paired, such as in Cluster 7 (EL, HR) and Cluster 8 (ES, IT), while LU and IE remained separated in Clusters 9 and 10, respectively. Interestingly, we observed two quite heterogeneous groupings Cluster 1 (MT, CY, BE, AT) and Cluster 6 (EE, LT, PT, PL) (Figure 2, right panel). Overall, the SNA and the clustering analysis gave very consistent results.



**Figure 2** – Clustering of countries based on the labour market attractiveness set.

In order to characterize the obtained clusters, we performed an ORA on discretized variables (Table 2). Thus, we found that Cluster 1 (AT, BE, CY, MT) was characterized as having small-sized populations, high proportions of youngsters and high proportions of employed youngsters with higher education. Cluster 2 (BG, HU, LV, RO) was mostly characterized as having high rates of poverty, low disposable incomes and low expenditures in higher-education. Cluster 3 (CZ, SK, SI) was characterized as having high proportions of adults in the population, of which the employed ones had high proportions achieving secondary education and low proportions achieving only primary education. Cluster 4 (DE,

FR, UK) was characterized as having large-sized populations with high proportions of children and high disposable incomes. Cluster 5 (DK, FI, NL, SE) was characterized as having high rates of employed youngsters and adults, high expenditures in higher education and low rates of poverty. Cluster 6 (EE, LT, PL, PT) was characterized as having high proportions of employed elders. Cluster 7 (EL, HR) was characterized as having high unemployment rates of youngsters and low employment rates of adults. Finally, Cluster 8 (ES, IT) was characterized as having both low proportions of youngsters in the population and of employed youngsters. Regarding the single-element clusters 9 and 10, we were unable to perform ORA since obtaining significant statistics in these cases is quite difficult. Nonetheless, we note that IE (Cluster 9) had the highest proportion of children and the highest GVAgr of the whole EU-28 countries, while LU (Cluster 10) had the highest proportion of employed adults with higher education and the highest GDP.

**Table 2** – Characterization of the country clusters.

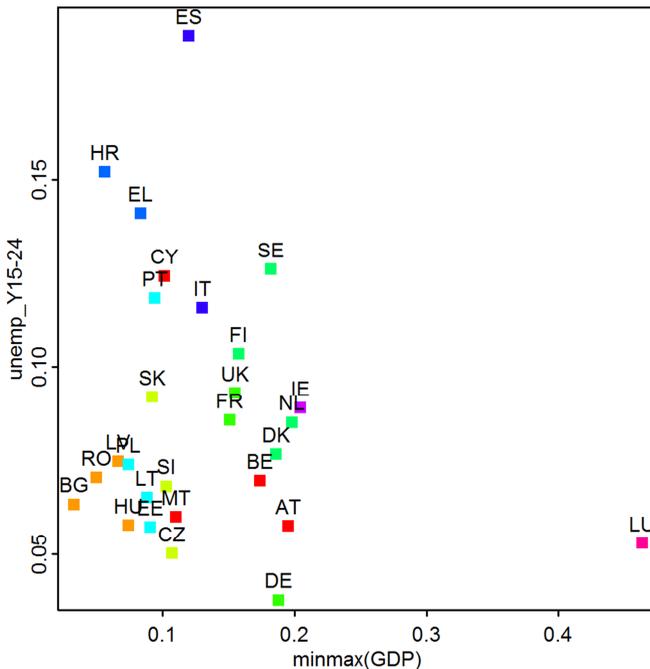
#	label	Over-represented variables
1	AT-BE-CY-MT	> <b>P90</b> : emp_Y[15-24]_ED[5-8]; emp_Y[15-24]_Nace[F]; pop_Y[15-24]; rooms_pp. < <b>P10</b> : pop_Total.
2	BG-HU-LV-RO	> <b>P90</b> : ARPR_socexcl; emp_Y[15-24]_Nace[J]; mat_depriv. < <b>P10</b> : disp_income; earn_OC[1-9]_Nace[B-F]; earn_OC[2-5,7-9]_Nace[G-N]; earn_OC[1-9]_Nace[P-S]; emp_Y[15-24,25-64]_Nace[O-Q]; expend_ED[5-8]; GDP; training.
3	CZ-SK-SI	> <b>P90</b> : emp_Y[15-24,25-64]_ED[3-4]; emp_Y[15-24]_Nace[B-E]; emp_Y[25-64]_Nace[B-E,F]; pop_Y[25-64]. < <b>P10</b> : emp_Y[25-64]_ED[0-2].
4	DE-FR-UK	> <b>P90</b> : disp_income; pop_Total; pop_Y[0-14]. < <b>P10</b> : emp_Y[15-24]_Nace[A].
5	DK-FI-NL-SE	> <b>P90</b> : earn_OC[7-9]_Nace[B-F]; earn_OC[5,9]_Nace[G-N]; earn_OC[6,8]_Nace[P-S]; emp_Y[15-24,25-64]; emp_Y[15-24]_ED[0-2]; emp_Y[15-24]_Nace[M-N]; emp_Y[25-64]_Nace[J,M-N,O-Q]; expend_ED[5-8]; pop_Y[65-74]; training. < <b>P10</b> : ARPR; ARPR_socexcl; emp_Y[15-24]_ED[3-4,5-8]; emp_Y[25-64]_Nace[G-I]; mat_depriv; pop_Y[25-64].
6	EE-LT-PL-PT	> <b>P90</b> : emp_Y[GE65]. < <b>P10</b> : earn_OC[1]_Nace[G-N]; emp_Y[25-64]_Nace[J,K].
7	EL-HR	> <b>P90</b> : unemp_Y[15-24]. < <b>P10</b> : emp_Y[25-64]; emp_Y[25-64]_Nace[L].
8	ES-IT	> <b>P90</b> : - < <b>P10</b> : emp_Y[15-24]; pop_Y[15-24].
9	IE	> <b>P90</b> : - < <b>P10</b> : -
10	LU	> <b>P90</b> : - < <b>P10</b> : -

Finally, we performed model selection to find which group of at most 2 variables from the Labour Market Attractiveness set could define better the clustering scheme obtained. The model which obtained the lowest AIC value included both GDP (i.e., gross domestic product) and unemp\_Y15-24 (i.e., unemployment rates among youngsters). Using this model, we were able to correctly identify 82% of the clusters to which the countries belonged (Table 3).

**Table 3** – Analysis of Deviance Table (Type II tests) for best multinomial model.

Variables	Likelihood ratio	Df	Pr(> $\chi^2$ )
<b>GDP</b>	80.18	9	<0.001
<b>unemp_Y15-24</b>	36.13	9	<0.001
EU_groups ~ GDP + unemp_Y15-24			
$R^2_{McFadden} = 0.84$ ( $p$ -value < 0.001)			
$R^2_{count} = 0.82$			

As a visual help to analyse the model selected, we plotted the EU-28 countries considering the two selected predictors. And, indeed, we were able to roughly differentiate the plot areas corresponding to each of the 10 clusters (Figure 3).



**Figure 3** – Best multinomial model for the country clusters.

## 4 Discussion and conclusion

The use of the Labour Market Attractiveness set aimed to capture the heterogeneity of the EU global labour market. For this purpose, we chose a broad collection of socio-demographic and economic information. The chosen set of variables seemed to successfully describe the EU space as the country network built showed complex interactions between the regions. Indeed, there seems to be a main separation between Northern and Western Europe and Eastern Europe, which at a finer detail revealed a far more complex net of connections. Thus, we observed a strongly connected group of Scandinavian countries (DK, FI and SE), a group of the most populous countries (FR, DE and UK), two distinct groups of Eastern European countries (CZ, EE, LT, PL, SI and SK; and BG, HU, LV and RO), several pairings (EL and HR; IT and ES; and CY and MT), and two single-element groups (LU; and IE).

The social network approach was complemented using a clustering analysis coupled with ORA to extract the clusters defining characteristics. These characteristics can reflect different degrees of attractiveness depending on the stakeholder. Thus, Cluster 3 (CZ, SI and SK) can be more attractive when looking for labour forces composed mostly by adults, while Cluster 1 (AT, BE, CY and MT) can be attractive when looking for younger labour forces. Similarly, Cluster 2 (BG, LV, HU and RO) can be attractive for enterprises looking for low-waged economies, while Cluster 5 (DK, SE, FI and NL) and Cluster 4 (FR, DE and UK) can be attractive for job seekers looking for high-paid jobs. Stakeholders interested in labour markets fairly open to elders can look at Cluster 6 (EE, LT, PL and PT). Nevertheless, characteristics such as high poverty rates, high unemployment, low-level of education and low public expenditure in education are generally unattractive.

Interestingly, using a model-selection approach, we could roughly capture the dynamics of the Labour Market Attractiveness set using only 2 variables: GDP and young unemployment. Nevertheless, although recovering the clustering structure fairly well (82% correct clustering prediction), this approach greatly reduces countries information.

The main goal for developing this exploration framework was to extract useful information from data to help policy-makers define strategies to tackle labour market problems. In fact, we argue that the characteristics extracted from the data can be invaluable to define strategies more suitable for each particular regional case. For example, taking into account countries characteristics such as ageing labour forces (Cluster 6: EE, LT, PL and PT), ageing population (Cluster 5: DK, FI, NL and SE; and Cluster 8: ES and IT) or highly skilled labour forces (Cluster 1: AT, BE, CY and MT; and Cluster 10: LU) can be of utmost importance when implementing policies to reduce Skills mismatch. Moreover, knowing of high

unemployment (Cluster 7: EL and HR) and low access to employment (Cluster 8: ES and IT) can be important to diagnose potential labour market problems.

A natural follow-up to the work presented here is to define relevant indicators of labour market (e.g. Labour market mobility, Skills mismatch) and of the population in general (e.g., Migration), and test their association to the characteristic of EU labour market extracted in the exploration part. These association studies entail the inferential second part of our proposed data product, and can be important to better understand the dynamics between socio-economic and demographic characteristics and important aspects of the labour market [see Lopes et al. (2018) for a follow-up association study]. These studies can not only provide a better understanding, but also provide possible strategies to tackle labour market related problems such as mismatch between local supply and demand of labour force and/or of specific labour skills, hence providing invaluable tools for helping policy-makers defining strategies to shape labour force to the needs of a fast-changing labour market.

## References

- CALCAGNO, V. & DE MAZANCOURT, C. (2010). glmulti: an R package for easy automated model selection with (generalized) linear models, *Journal of Statistical Software*, 34.12, 1-29.
- CEDEFOP (2015) *Skills, qualifications and jobs in the EU: the making of a perfect match?* Publications Office of the European Union, Luxembourg.
- EUROPEAN COMMISSION (2016) Skills, education and lifelong learning – European Pillar of Social Rights, <https://ec.europa.eu> (accessed at 2<sup>nd</sup> October 2017).
- EUROPEAN COMMISSION (2017) European Big Data Hackathon, [https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon\\_en](https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon_en) (accessed at 2<sup>nd</sup> October 2017).
- FISHER, R.A. (1922). On the Interpretation of  $\chi^2$  from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1): 87-94.
- FRUCHTERMAN, T.M. & REINGOLD, E.M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11), 1129-1164.
- KAUFMAN, L. & ROUSSEEUW, P.J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 68-125.
- LOPES, J.S., MOURA, M. & QUARESMA S. (2018) Labour market attractiveness in the EU. *Statistical Working Papers*. Luxembourg: Eurostat.
- REYNOLDS, A.P., RICHARDS, G., DE LA IGLESIA, B. & RAYWARD-SMITH, V.J. (1992). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms, *Journal of Mathematical Modelling and Algorithms*, 5.4, 475-504.
- VENABLES, W. N. & RIPLEY, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

# Avaliação de métodos de estimação da variância em amostras complexas

Eládio Muianga<sup>1</sup> · Anabela Afonso<sup>2</sup>

**Resumo** Em amostras complexas nem sempre é possível obter uma expressão analítica para o estimador da variância dos estimadores, existindo na literatura alguns métodos para obter aproximações para esse estimador. Neste trabalho estudou-se o desempenho de alguns desses métodos. São simulados dados, a partir da informação real da atividade económica Moçambicana, e usam-se diferentes esquemas de amostragem com o objetivo de contribuir para as investigações por amostragem em Moçambique. O estimador *Taylor* apresentou o melhor desempenho e o estimador *bootstrap* foi o menos preciso.

**Palavras-chave:** Amostras Complexas, Enviesamento, Erro Quadrático Médio, Inferência, Estimadores da Variância.

## 1 Introdução

A necessidade de conhecer uma população impulsiona o processo de recolha e análise de informação. Usualmente, é muito difícil, ou impossível, estudar a totalidade da população, daí a importância do seu estudo com recurso a amostras. Conceber um estudo por amostragem é um processo complexo, desde antes da recolha dos dados até a fase de análise dos mesmos.

As amostras complexas combinam um conjunto de métodos probabilísticos de amostragem para a seleção de uma amostra representativa da população (Szwarcwald e Damacena, 2008). Estas amostras têm pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção

---

<sup>1</sup> Mestrado em Modelação Estatística e Análise de Dados, Instituto Nacional de Estatística de Moçambique, [eladio.muianga@ine.gov.mz](mailto:eladio.muianga@ine.gov.mz)

<sup>2</sup> CIMA/IIFA e DMAT/ECT, Universidade de Évora, [aafonso@uevora.pt](mailto:aafonso@uevora.pt)

desiguais, ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008). Comparando com a amostragem aleatória simples, sabe-se que a amostragem estratificada, quando usada de forma conveniente, em geral, produz estimativas mais precisas e que a amostragem por conglomerados acarreta uma perda de precisão das estimativas. Para medir o efeito do delineamento de amostragem, Kish (1965) propôs a utilização da medida *deff* (*design effect*). Esta medida consiste na razão entre a variância do estimador do delineamento de amostragem complexo e a variância do estimador considerando uma amostragem aleatória simples. Esta medida para além de quantificar a perda, ou o ganho, de precisão da estimativa também pode ser usada para determinar a dimensão da amostra complexa com base na dimensão da amostra aleatória simples.

Contudo, para as amostras complexas nem sempre é possível obter uma expressão analítica para o estimador da variância dos estimadores dos parâmetros. Ao longo dos anos foram propostas aproximações para este estimador, ajustadas à natureza complexa do plano da amostra, sendo as mais utilizadas para estimar a variância dos totais e médias estimados: o método de linearização *Taylor* e as técnicas de reamostragem e replicação (*Jackknife* e *bootstrap*).

Neste trabalho avalia-se o desempenho, e respetivas propriedades, dos estimadores mais usuais da variância do totais e médias amostrais em amostras complexas. Este estudo tem como objetivo contribuir para as investigações de amostragem em Moçambique. Para tal, usam-se dados simulados a partir da realidade da atividade económica Moçambicana e consideram-se diferentes esquemas de amostragem.

## 2 Estimação da variância

Um requisito básico em todas as formas de análise, senão a principal exigência nas investigações práticas, é que uma medida de precisão deve ser fornecida para cada estimativa derivada dos dados de investigação. A medida de precisão mais usada é a variância do estimador (Wolter, 2007).

Os totais e as médias da população podem ser facilmente estimados a partir dos pesos de amostragem. Estimar variâncias é um processo mais complexo pois, em amostras complexas com várias etapas de estratificação e conglomerados, a variância do estimador dos totais e médias deve ser calculada para cada nível e, em seguida, deve ser combinada segundo o delineamento de amostragem (Lohr, 2010).

Estimar a variância dos estimadores em amostras complexas é um tema atual e bastante complexo em termos gerais, uma vez que, de um modo geral, não existe uma expressão analítica para um estimador centrado e eficiente da variância do

estimador. Para estimar a variância dos totais e médias estimados existem, basicamente, duas abordagens:

- Analítica, usando o método de linearização de *Taylor*;
- Métodos de reamostragem ou replicação (*Jackknife*, réplicas equilibradas repetidas (*BRR*) e *bootstrap*).

Nos métodos de reamostragem duas ou mais subamostras são selecionadas de uma dada população, ou eventualmente, de uma amostra. Com base em cada uma das amostras estima-se o parâmetro de interesse. É a partir da combinação das estimativas obtidas que se obtém uma estimativa da variância. Estes métodos diferem entre si na forma de gerar réplicas de amostras.

A escolha de um método para estimar a variância envolve um equilíbrio de fatores tais como a precisão e o custo. Nenhum dos métodos para estimar a variância do estimador é o melhor no geral (Wolter, 2007, pág. 366). Por isso, num bom julgamento, no qual está envolvida a escolha de um método para estimar a variância, não será surpresa se o estatístico recomendar métodos diferentes para diferentes aplicações da investigação.

Seja  $\hat{\theta}$  um estimador para o parâmetro populacional  $\theta$  e  $Var(\hat{\theta})$  a sua variância. De seguida, apresentam-se os três métodos de estimação que são usualmente mais utilizados para estimar  $Var(\hat{\theta})$ , quando o parâmetro é uma média ( $\mu$ ) ou um total ( $\tau$ ): o método de *Linearização em série de Taylor* e os métodos de replicação (reamostragem) *Jackknife* e *bootstrap*.

## 2.1 Método de linearização de *Taylor*

Este método de linearização baseia-se na expansão em série de *Taylor* a qual permite obter uma aproximação linear para a estatística de interesse.

Sejam  $\hat{\theta}_j$ ,  $j = 1, \dots, k$ , estimadores não enviesados para os  $\theta_j$ ,  $j = 1, \dots, k$  parâmetros populacionais. Seja  $\hat{\theta} = \sum_{j=1}^k \hat{\theta}_j$  e  $a_j$  e  $a_l$  constantes de linearização; então pelo método de linearização de *Taylor* a variância do estimador pode ser estimada por (Lohr, 2010):

$$\widehat{Var}_{\tau}(\hat{\theta}) = \sum_{j=1}^k a_j \widehat{Var}(\hat{\theta}_j) + \sum_{j=1}^{k-1} \sum_{l=j+1}^k a_j a_l \widehat{Cov}(\hat{\theta}_j, \hat{\theta}_l).$$

A precisão da aproximação da linearização depende do tamanho da amostra, o que origina a que a variância do estimador por vezes seja subestimada se a amostra não for grande o suficiente (Lohr, 2010).

Quando o estimador  $\hat{\theta}$  é uma função não linear das observações, o estimador *Taylor* da  $Var(\hat{\theta})$  é enviesado, mas tipicamente consistente (Wolter, 2007).

## 2.2 Método de *Jackknife*

Este método foi proposto por Quenouille (1956) como um método para reduzir o enviesamento dos estimadores, num contexto da Estatística Clássica. Posteriormente, Tukey (1958) propôs usá-lo para estimar variâncias e calcular intervalos de confiança.

A ideia geral deste método consiste em tomar como ponto de partida uma amostra de dimensão  $n$  a partir da qual serão consideradas todas as subamostras possíveis em que se eliminam  $k$  elementos de cada vez. Com base em cada uma destas subamostras é obtida uma estimativa para o parâmetro de interesse e a variância é estimada usando a informação de todas estas estimativas.

No caso do método *delete-1*, considera-se  $\hat{\theta}_{(j)}$  um estimador com a mesma forma que  $\hat{\theta}$  sem a observação  $j$  e  $n-1$  a dimensão da amostra, sendo o estimador de *Jackknife* dado por (Lohr, 2010):

$$\widehat{var}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}_{(j)} - \hat{\theta})^2.$$

Portanto, a variância do estimador é estimada com base na variabilidade entre as estimativas obtidas a partir das subamostras constituídas.

No caso de uma amostra por grupos, de forma a não perder a estrutura dos grupos, em vez de se eliminar uma observação de cada vez na formação das subamostras, eliminam-se todas as observações pertencentes ao mesmo grupo.

## 2.3 Método de *bootstrap*

O método de *bootstrap* foi proposto por Efron (1979). É um método de reamostragem de computação intensiva que tem sido bastante aplicado em muito devido aos avanços computacionais.

Tal como no método de *Jackknife*, neste método a variância pretendida será estimada a partir de subamostras, designadas por réplicas, que serão extraídas de uma amostra inicial.

Seja  $R$  o número de réplicas *bootstrap* e  $\hat{\theta}_r^*$  um estimador de  $\theta$  calculado usando o vetor dos pesos replicados. A variância estimada pelo método de *bootstrap* é dada por (Lohr, 2010):

$$\widehat{var}_B(\hat{\theta}) = \frac{1}{B-1} \sum_{l=1}^R (\hat{\theta}_l^* - \hat{\theta})^2.$$

Para algumas configurações, o método de *bootstrap* pode exigir mais cálculos do que o método de *Jackknife*, uma vez que tipicamente  $R$  representa um número muito grande de réplicas. No entanto, em outras investigações de larga escala se, por exemplo, for retirada uma amostra aleatória estratificada, o método de *bootstrap* pode exigir menos cálculos do que o *Jackknife*. De notar que, a estimativa de variância *bootstrap* difere quando é retirado um conjunto diferente de amostras *bootstrap*.

### 3 Métodos

Devido à confidencialidade no acesso aos dados reais, foi adotada uma metodologia de geração de números pseudoaleatórios para a criação do universo de estudo, “Empresas do sector do comércio”, com base na informação de uma amostra representativa da estrutura deste sector fornecida pelo Instituto Nacional de Estatística de Moçambique (INE-M).

Foram simulados três conjuntos de dados populacionais de empresas por região, classificador de atividade económica CAE Rev-2, número de trabalhadores ( $NT$  e  $NTCat$ ) e volume de negócio ( $VN$ ). Assumiu-se que a variável de interesse  $VN$  dependia linearmente de  $NT$ , i.e.,  $VN = \alpha + \beta NT + \varepsilon$ , com os valores de  $\alpha$  e  $\beta$  a variarem segundo a população e  $\varepsilon \sim N(0; \sigma NT)$ , com  $\sigma > 0$ .

As principais características das populações geradas são (Tabela 1):

- I. Com características semelhantes às da população real;
- II. Considerou-se uma igual distribuição de empresas por Região e CAE;
- III. Introduziu-se uma maior dispersão na distribuição da variável de interesse.

**Tabela 1** – Características das populações geradas.

Variável	Categorias	População		
		I	II	III
Região	Norte	14%	33%	14%
	Centro	47%	34%	47%
	Sul	39%	33%	39%
CAE	45	6%	32%	6%
	46	5%	34%	5%
	47	89%	34%	89%
NTCat*	Pequena	69%	69%	40%
	Média	20%	20%	42%
	Grande	10%	10%	18%

\* Dimensão das empresas: pequena ( $NT < 50$ ), média ( $50 \leq NT < 250$ ), e grande ( $NT \geq 250$ ).

Com vista a estimar o parâmetro de interesse, i.e., a média do  $VN$ , de cada um dos universos populacionais foram sorteadas 10 000 réplicas de amostras de acordo com quatro delineamentos de amostragem:

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

Para analisar a qualidade dos estimadores recorreu-se a duas propriedades fundamentais: o não enviesamento e a precisão. A precisão foi medida pelo erro quadrático médio ( $EQM$ ) e pela raiz do erro quadrático médio escalado ( $REQME$ ):

$$REQME(\hat{\theta}) = \frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})} \times 100.$$

O  $EQM$  é uma medida de precisão que depende da escala da variável em estudo enquanto o  $REQME$  é uma medida cujo resultado é expresso em percentagem. Deste modo, esta medida de desempenho escalado, que combina o enviesamento e a precisão, permite comparar o desempenho do estimador quando as populações têm características diferentes (Walther & Moore, 2005).

Todo o estudo foi realizado com auxílio do *software R Project*.

## 4 Resultados

### 4.1 Distribuição dos estimadores da variância da média estimada

As distribuições de amostragem obtidas para os estimadores *Taylor* ( $\widehat{Var}_T$ ), *Jackknife* ( $\widehat{Var}_{JK}$ ) e *bootstrap* ( $\widehat{Var}_B$ ) para a variância de  $\hat{\mu}$ , para cada uma das três populações, são apresentadas na Figura 1. Além disso, quando possível, também é apresentada a distribuição de amostragem do estimador usualmente indicado na literatura ( $\widehat{Var}_L$ ) para os delineamentos de amostragem considerados neste trabalho e cuja expressão é apresentada, por ex., em Lohr (2010).

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

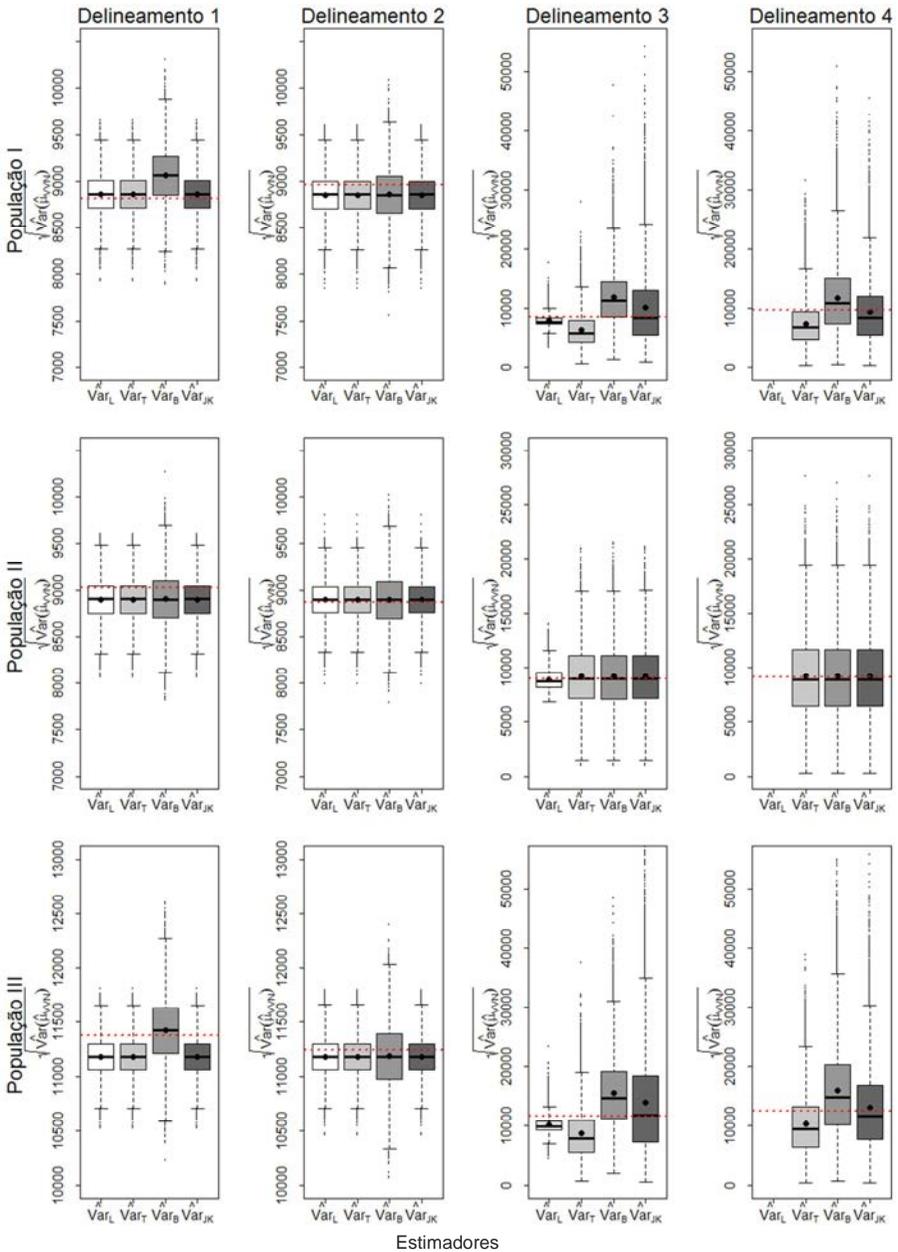


Figura 1 – Distribuição empírica de  $\sqrt{\widehat{Var}(\hat{\mu}_{VN})}$ .

Nos delineamentos de amostragem estratificados (1 e 2) as distribuições são aproximadamente simétricas e as distribuições dos estimadores *Taylor*, *Jackknife* e o indicado na literatura são similares. Resultados diferentes podem ser visualizados nos planos multietápicos (delineamentos 3 e 4), que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais. A distribuição dos estimadores varia nas três populações em estudo. A assimetria da distribuição depende da população e, em todas as populações, predominam os valores atípicos superiores.

Nas populações I e III os estimadores *bootstrap* e *Jackknife* são enviesados no delineamentos multietápicos, sendo o estimador *bootstrap* o mais enviesado. Na população II, a distribuição dos estimadores é simétrica apresentando o estimadores *Taylor* e *Jackknife* distribuições similares.

### 4.2 Análise comparativa dos estimadores da variância

Nos delineamentos de amostragem estratificados (1 e 2) os estimadores apresentam enviesamentos idênticos nas três populações (Tabela 2), mas o estimador *bootstrap* é o menos preciso, apresentando os maiores valores de *REQME* (Tabela 3). Os restantes três estimadores apresentam valores iguais de *REQME*.

Nos delineamentos multietápicos (3 e 4), o enviesamento do estimador indicado na literatura depende das características da população, mas é o estimador mais preciso (menor *REQME*). O estimador *bootstrap* é, de uma forma geral, o mais enviesado. No delineamento a duas etapas (3) o estimador *Jackknife* é o menos preciso enquanto no delineamento 4 o estimador *bootstrap* foi o mais impreciso.

**Tabela 2** – Enviesamento dos estimadores.

População	Delineamento	Env ( $\widehat{Var}(\hat{\mu}_{VVN})$ ) %			
		$\widehat{Var}_L$	$\widehat{Var}_T$	$\widehat{Var}_B$	$\widehat{Var}_{JK}$
I	1	0,96	0,96	5,71	0,96
	2	-2,41	-2,41	-2,28	-2,41
	3	-10,96	-31,10	125,96	96,68
	4	-	-25,73	87,51	25,89
II	1	-2,65	-2,65	-2,56	-2,65
	2	0,68	0,68	0,77	0,68
	3	0,96	16,03	16,06	16,36
	4	-	18,82	18,86	18,84
III	1	-3,43	-3,43	0,93	-3,43
	2	-1,00	-1,00	-0,88	-1,00
	3	-20,04	-29,70	109,32	102,27
	4	-	-14,98	100,54	43,12

**Tabela 3** – Precisão dos estimadores.

População	Delineamento	$REQME (\widehat{Var}(\hat{\mu}_{VVN})) \%$			
		$\widehat{Var}_L$	$\widehat{Var}_T$	$\widehat{Var}_B$	$\widehat{Var}_{JK}$
I	1	5,10	5,10	9,11	5,10
	2	5,45	5,45	6,91	5,45
	3	27,87	75,84	225,13	296,43
	4	-	86,72	223,89	162,75
II	1	5,44	5,44	6,95	5,44
	2	4,85	4,85	6,61	4,85
	3	24,11	73,45	73,71	73,88
	4	-	96,38	96,62	96,39
III	1	4,59	4,59	5,60	4,59
	2	3,30	3,30	5,55	3,30
	3	33,81	80,88	202,19	296,82
	4	-	91,26	227,57	177,61

## 5 Conclusão

Nos delineamentos de amostragem estratificados os resultados dos estimadores *Taylor* e *Jackknife* coincidiram com o indicado na literatura. Apresentaram melhores resultados quanto ao enviesamento e à precisão das estimativas do que o estimador *bootstrap*. Verificou-se, à semelhança do referido por Pessoa & Silva (1998), que neste tipo de delineamentos as estimativas *Jackknife* são iguais às estimativas *Taylor*.

Nos planos multietápicas, que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais, o estimador *Taylor* mostrou ser menos enviesado e mais preciso do que os estimadores *Jackknife* e *bootstrap*. Estes dois últimos estimadores são enviesados, dependendo o enviesamento do tipo de população, e, além disso, muito imprecisos.

Os estimadores indicados na literatura para os delineamentos estratificados e por grupos em duas etapas foram os que garantiram maior confiabilidade nas estimativas. Seguiu-se o estimador *Taylor* com resultados mais fiáveis para as aproximações.

Em Moçambique grande parte dos estudos realizados pelo Instituto Nacional de Estatística são baseados em amostras, havendo a necessidade de garantir um elevado grau de precisão para as estimativas geradas. Com este trabalho pretendeu-se dar um contributo no sentido de minorar essa necessidade, auxiliando na escolha do estimador a usar na produção de estatísticas oficiais no país.

## Agradecimentos

Anabela Afonso é membro do CIMA, centro de investigação financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal, no âmbito do projeto «UID/MAT/04674/2019 (CIMA)».

## Referências

- EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7(1), 1-26.
- LAVRAKAS, P. J. (2008). *Encyclopedia of Survey Research Methods*, SAGE Publications, Thousand Oaks.
- LORH, S. L. (2010). *Sampling: Design and Analysis*, Second Edition, Michelle Julet, Boston.
- PESSOA, D. G. C. & SILVA P. L. N. (1998). *Análise de Dados Amostrais Complexos*, Associação Brasileira de Estatística, São Paulo.
- QUENOUILLE, M. H. (1956). Notes on Bias in Estimation, *Biometrika*, 43, 353–360.
- SZWARCWALD, C. L. & DAMACENA, G. N. (2008). Amostras Complexas em Inquéritos Populacionais: Planeamento e Implicações na Análise Estatística dos Dados. *Revista Brasileira de Epidemiologia*, 11(11), 38–45.
- TUKEY, J. W. (1958). Bias and Confidence in Not-quite large Samples, *Annals of Mathematical Statistics*, 29, 614.
- WALTHER, B. A. & MOORE, J. L. (2005). The Concepts of bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance, *Ecography*, 28, 815–829.
- WOLTER K. M. (2007). *Introduction to Variance Estimation*, Springer-Verlag, New York.

# O papel mediador da avaliação cognitiva na relação entre a mútua interferência trabalho-família e o *Burnout*

Jéssica Rodrigues<sup>1</sup> · A. Manuela Gonçalves<sup>2</sup> · Susana Faria<sup>2</sup> · Clara Simões<sup>3</sup>  
· A. Rui Gomes<sup>4</sup>

**Resumo** Aplicando Modelos com Equações Estruturais pretende-se avaliar o papel mediador da avaliação cognitiva na relação entre a mútua interferência trabalho-família e a experiência de *burnout*, em professores portugueses que lecionam desde o jardim de infância até ao secundário. Os resultados evidenciam o efeito mediador da avaliação cognitiva nessa relação. A avaliação cognitiva é assim um importante mecanismo para explicar a adaptação no trabalho.

**Palavras-chave:** Avaliação Cognitiva, *Burnout*, Mediação, Modelo com Equações Estruturais, Mútua Interferência Trabalho-Família.

## 1 Introdução

As profissões de ensino têm sido associadas pela investigação em saúde ocupacional a elevados níveis de *stress* crónico e de *burnout* (Gomes, Faria, & Gonçalves, 2013). A investigação neste domínio tem vindo a destacar a mútua interferência trabalho-família como uma importante fonte de *stress* ocupacional. A dificuldade de conciliação da vida laboral com as responsabilidades familiares tem demonstrado significativos efeitos negativos sobre a saúde física, mental e social dos sujeitos (e.g., *distress*, depressão, esgotamento, fadiga crónica, insatisfação laboral e com a vida em geral), justificando a importância de analisar os

---

<sup>1</sup>DMAT – Departamento de Matemática e Aplicações, Universidade do Minho, Portugal, [pg32814@alunos.uminho.pt](mailto:pg32814@alunos.uminho.pt)

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

<sup>3</sup>Escola Superior de Enfermagem, Universidade do Minho, Portugal, [csimaes@ese.uminho.pt](mailto:csimaes@ese.uminho.pt)

<sup>4</sup>Departamento de Psicologia Aplicada, CIPsi, Centro de Investigação em Psicologia, Universidade do Minho, Portugal, [rgomes@psi.uminho.pt](mailto:rgomes@psi.uminho.pt)

mecanismos que podem explicar o aparecimento e desenvolvimento do *stress* ocupacional. Um destes mecanismos, refere-se aos processos de avaliação cognitiva (Lazarus & Folkman, 1984), que nos indicam até que ponto uma dada situação de *stress* pode ser avaliada pelo indivíduo como mais ameaçadora ou mais desafiadora.

Neste trabalho pretende-se aplicar modelos com equações estruturais a um conjunto de dados reais, com o objetivo de averiguar se a avaliação cognitiva que os professores fazem da sua atividade laboral (Glaser & Tracy, 2013) desempenha um papel mediador na relação entre a perceção de mútua interferência trabalho-família (e.g., conflito trabalho-família e conflito família-trabalho) e a experiência de *burnout* (e.g., fadiga física, fadiga cognitiva e exaustão emocional).

## 2 Metodologia

O modelo com equações estruturais (*Structural Equation Modeling*, SEM) permite estimar, simultaneamente um conjunto de relações múltiplas de dependência entre variáveis, algumas das quais podem ser variáveis latentes (fatores), i.e., variáveis não diretamente mensuráveis mas medidas a partir de um conjunto de indicadores (variáveis manifestas ou observadas) (Salgueiro, 2012). Esta metodologia de análise estatística é uma técnica de modelação multivariada, usada para testar a validade de modelos teóricos que incluem variáveis latentes e variáveis manifestas.

O modelo com equações estruturais é um modelo linear e pode ser dividido em dois submodelos: o submodelo de medida (que especifica os indicadores de medida de cada variável latente) e o submodelo estrutural (que estabelece o conjunto de relações estruturais entre as variáveis latentes).

O modelo com equações estruturais é definido pelas três equações (as duas primeiras constituem o modelo de medida e a última constitui o modelo estrutural),

$$\begin{aligned} X &= \Lambda_X \xi + \delta \\ Y &= \Lambda_Y \eta + \epsilon \\ \eta &= B\eta + \Gamma\xi + \zeta \end{aligned}$$

onde  $X$  é o vetor das  $q$  variáveis manifestas ou observadas independentes,  $\xi$  é o vetor das  $n$  variáveis latentes independentes,  $\delta$  é o vetor dos erros de medida das  $q$  variáveis manifestas independentes,  $Y$  é o vetor das  $p$  variáveis manifestas ou observadas dependentes,  $\eta$  é o vetor das  $m$  variáveis latentes dependentes,  $\epsilon$  é o vetor dos erros de medida das  $p$  variáveis manifestas dependentes,  $\zeta$  é o vetor dos  $m$  termos residuais das relações estruturais e  $B, \Lambda_X, \Lambda_Y$  e  $\Gamma$  são os parâmetros do modelo. Este modelo tem os seguintes pressupostos i) os erros dos submodelos de medidas têm valor esperado nulo, ii)  $\epsilon$  e  $\eta$  são independentes, iii)  $\delta$  e  $\xi$  são

independentes, iv)  $\zeta$  e  $\xi$  são independentes e v)  $\epsilon$ ,  $\zeta$  e  $\delta$  são mutuamente independentes, vi)  $\mathbf{B}$  é uma matriz cujos valores da diagonal principal são nulos e vii)  $(I - \mathbf{B})$  é uma matriz invertível. Tem-se ainda que  $E[\xi \xi^T] = \Phi$  e  $E[\zeta \zeta^T] = \Psi$ , sendo  $\Phi$  e  $\Psi$ , respectivamente, as matrizes de variância-covariância das variáveis latentes independentes  $\xi$  e dos termos residuais das relações estruturais  $\zeta$ .

No ajustamento do modelo, utiliza-se uma estratégia em duas etapas (*two-step*): na primeira etapa especifica-se e ajusta-se o submodelo de medida e na segunda etapa especifica-se e ajusta-se o submodelo estrutural. Na estimação dos parâmetros do modelo (pesos fatoriais, coeficientes de regressão, covariâncias,...) aplica-se o método da máxima verosimilhança. Na avaliação da qualidade de ajustamento dos modelos, em cada uma das etapas, utilizam-se os seguintes índices: a estatística qui-quadrado  $\chi^2$ , *Root Mean Square Error of Approximation* (RMSEA), *Tucker-Lewis Index* (TLI), *Comparative Fit Index* (CFI) e *Normed Fit Index* (NFI). Para estimar os erros-padrão e os intervalos de confiança das estimativas de todos os efeitos recorre-se ao método de reamostragem *Bootstrap* (Mackinnon, Lockwood & Williams, 2004).

Todas as análises são realizadas no *software* AMOS 24.0 (Byrne, 2010).

### 3 Base de dados e modelo

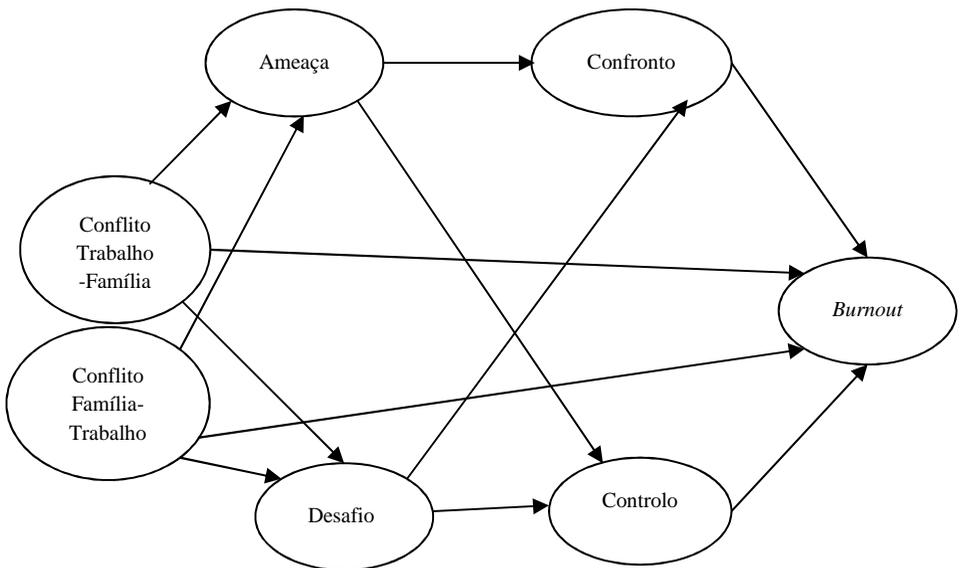
Neste estudo participaram 438 professores portugueses que lecionam desde o jardim de infância até ao ensino secundário. Todos os professores responderam a um questionário sociodemográfico e a um protocolo de avaliação constituído pelos seguintes instrumentos de autorrelato: a versão portuguesa do *Work-Family Conflict & Family-Work Conflict Scales* (Netemeyer, Boles, & McMurrian, 1996); o Questionário de Avaliação Cognitiva (Gomes, 2008); e a Medida de *Burnout de Shirom-Melamed* (Gomes, 2012). O questionário sociodemográfico avalia a informação pessoal e profissional dos professores. A versão portuguesa do *Work-Family Conflict & Family-Work Conflict Scales* avalia a mútua interferência trabalho-família e é constituída por 10 itens distribuídos pelos seguintes fatores: Conflito Trabalho-Família e Conflito Família-Trabalho. O questionário de avaliação cognitiva avalia de que forma um indivíduo classifica uma situação de *stress* relacionada com a sua atividade profissional e é constituído por 12 itens distribuídos pelos seguintes fatores: Ameaça, Desafio, Confronto e Controlo. A medida de *Burnout de Shirom-Melamed* avalia a experiência de *Burnout* (que se caracteriza pelo estado de exaustão emocional, mental e física de um indivíduo) e é constituída por 14 itens distribuídos pelos seguintes fatores: Fadiga Física, Fadiga Cognitiva e Exaustão Emocional.

Dos 438 professores, 304 são do género feminino (69,41%), 129 do género masculino (29,45%) e 5 não forneceram informação sobre o seu género (1,14%). A

idade dos professores varia entre os 28 e os 67 anos, sendo a idade média 46,85 anos e o desvio-padrão 7,88 anos. A maioria dos professores é casada (70,78%), 13,47% são solteiros, 10,96% divorciados e 2,74% pertencem a um outro estado civil (2,05% dos inquiridos, o que corresponde a 9 professores, não forneceram informação sobre o seu estado civil). O número de anos de serviço varia de 2 a 44 anos, sendo a média 23,24 anos e o desvio-padrão 8,33 anos. Relativamente ao número de anos de serviço enquanto docente, este varia de 2 a 42 anos, sendo a média 22,63 anos e o desvio-padrão 8,39 anos. A maioria dos professores não tirou baixa nos 12 meses que antecederam a recolha dos dados (94,52%).

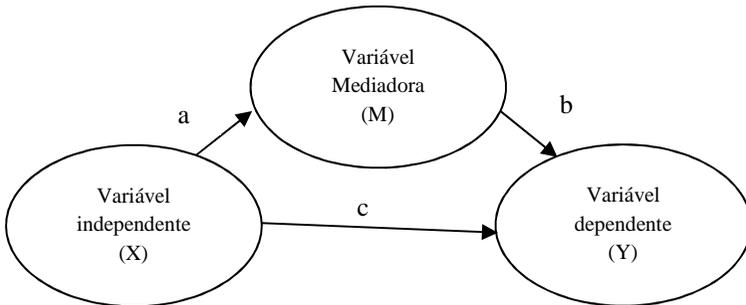
Os conflitos entre o trabalho e a família são uma fonte de *stress* para muitos trabalhadores, sendo a mútua interferência trabalho-família definida como uma situação em que as exigências laborais e os papéis familiares são mutuamente incompatíveis. A vida laboral interfere na vida familiar (trabalho-família) e a vida familiar interfere na vida laboral (conflito família-trabalho) o que dá origem a uma doença ocupacional, o síndrome de *Burnout*. A avaliação cognitiva surge como um mecanismo para explicar o aparecimento e desenvolvimento do *stress* ocupacional e indica de que forma uma dada situação de *stress* é avaliada por um indivíduo.

Na Figura 1 é apresentado o modelo utilizado para testar se a avaliação cognitiva dos professores desempenha um papel mediador na relação entre a mútua interferência trabalho-família e a experiência de *Burnout*.



**Figura 1** - Modelo teórico em estudo.

As relações estruturais, que refletem hipóteses causais sobre efeitos diretos e efeitos indiretos de variáveis latentes, são particularmente apropriados para testar hipóteses de mediação entre variáveis. Essas relações estruturais podem ser representadas usando um diagrama de caminhos (Figura 2).



**Figura 2** - Representação gráfica de um modelo de mediação.

A análise de caminhos permite estimar conjuntamente efeitos diretos, indiretos e totais entre variáveis. Neste modelo, o efeito direto da variável X sobre a variável Y é medido pelo coeficiente da regressão estrutural que relaciona as duas variáveis (c). O efeito indireto da variável X sobre a variável Y, potencialmente mediado pela variável mediadora M, é medido pelo produto dos coeficientes de regressão ao longo da cadeia de caminhos que liga as duas variáveis ( $a*b$ ). A soma dos efeitos diretos com os efeitos indiretos designa-se por efeito estrutural total ( $a*b+c$ ).

## 4 Resultados

Na Tabela 1 apresentam-se os alfa de Cronbach ( $\alpha$ ) das variáveis latentes (fatores) do modelo e a matriz de correlação de Spearman entre essas variáveis. As variáveis Conflito Trabalho-Família (CTF) e Conflito Família-Trabalho (CFT) estão positivamente correlacionadas entre si e com as variáveis Ameaça (AM), Fadiga Física (FF), Fadiga Cognitiva (FC) e a Exaustão Emocional (EE) e negativamente correlacionadas com as variáveis Desafio (DF) e Confronto (CF). A variável Ameaça (AM) está negativamente correlacionada com as variáveis Desafio (DF), Confronto (CF) e Controlo (CL) e positivamente correlacionada com as variáveis Fadiga Física (FF), Fadiga Cognitiva (FC) e a Exaustão Emocional (EE). As variáveis Desafio (DF) e Confronto (CF) e Controlo (CL) encontram-se positivamente correlacionados entre si e negativamente correlacionados com as variáveis Fadiga Física (FF), Fadiga Cognitiva (FC) e a Exaustão Emocional (EE). Estas três últimas variáveis encontram-se positivamente correlacionadas entre si.

**Tabela 1** - Alfa de Cronbach e correlações de Spearman entre as variáveis latentes.

Variáveis	$\alpha$	1	2	3	4	5	6	7	8
1. CTF	0,955								
2. CFT	0,901	0,398**							
3. AM	0,825	0,435**	0,290**						
4. DF	0,856	-0,125**	-0,084	-0,356**					
5. CF	0,851	-0,104*	-0,127**	-0,376**	0,397**				
6. CL	0,790	-0,105*	0,003	-0,281**	0,376**	0,438**			
7. FF	0,950	0,559**	0,281**	0,501**	-0,344**	-0,292**	-0,216**		
8. FC	0,968	0,455**	0,344**	0,447**	-0,264**	-0,374**	-0,216**	0,769**	
9. EE	0,914	0,185**	0,421**	0,311**	-0,133**	-0,233**	-0,089	0,335**	0,520**

Notas: \*  $p < 0,05$ ; \*\*  $p < 0,01$

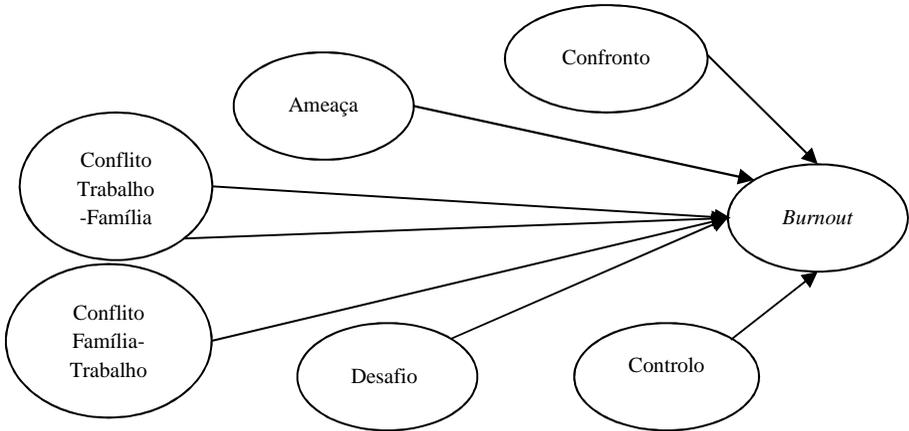
Inicialmente realiza-se a análise fatorial confirmatória dos instrumentos de autorrelato: mútua interferência trabalho-família, avaliação cognitiva e *Burnout*. Todos os itens da escala mútua interferência trabalho-família, da escala de avaliação cognitiva e do *Burnout* apresentam pesos fatoriais adequados e os modelos dos instrumentos de autorrelato apresentam índices de qualidade de ajustamento considerados bons.

Após a análise fatorial confirmatória dos instrumentos de autorrelato, estabeleceu-se um submodelo de medida adequado. Ajustam-se inicialmente dois modelos: um modelo com nove fatores (Conflito Trabalho-Família, Conflito Família-Trabalho, Ameaça, Desafio, Confronto, Controlo, Fadiga Física, Fadiga Cognitiva e Exaustão Emocional), e um modelo com apenas um único fator, ou seja, um modelo em que todas as variáveis manifestas em estudo definem apenas uma variável latente.

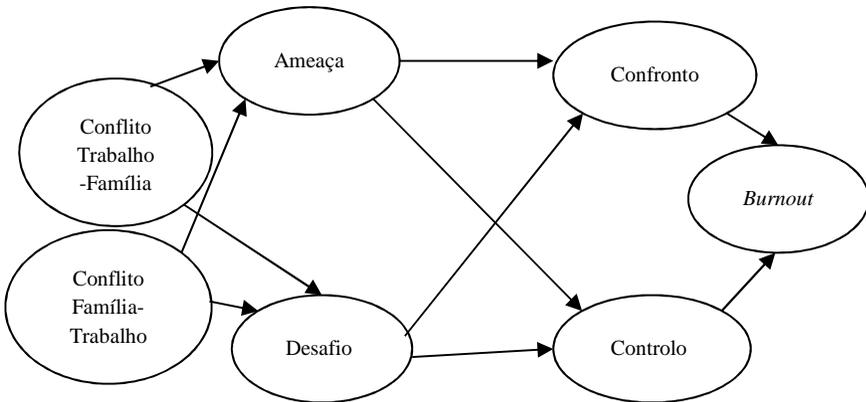
De acordo com os índices de qualidade de ajustamento, o modelo com os nove fatores apresenta um bom ajustamento ( $\chi^2(554)=949,160$ ; RMSEA= 0,041; TLI=0,970; CFI=0,974 e NFI=0,939) e melhor do que o ajustamento do modelo com apenas um único fator ( $\chi^2(590)=7612,278$ ; RMSEA=0,168; TLI=0,502; CFI=0,534 e NFI=0,515). A diferença da estatística  $\chi^2$  dos dois modelos ( $\Delta\chi^2(36)=6663,118$ ;  $p < 0,001$ ) permite concluir que os dois modelos diferem significativamente na qualidade do ajustamento, o modelo com nove fatores é preferível e é considerado como submodelo de medida. Os pesos fatoriais deste modelo são todos estatisticamente significativos e variam entre 0,626 e 0,975.

Depois de se estabelecer um submodelo de medida, estabelece-se um submodelo estrutural. Diferentes modelos são ajustados: o modelo de efeitos diretos (Figura 3), o modelo de mediação parcial (Figura 1), e o modelo de mediação total (Figura 4). O modelo de efeitos diretos assume caminhos diretos dos fatores Conflito Trabalho-Família, Conflito Família-Trabalho, Ameaça, Desafio, Confronto, Controlo para o *Burnout*. O modelo de mediação total assume caminhos diretos dos fatores Conflito Trabalho-Família e Conflito Família-Trabalho para os fatores Ameaça, Desafio, Confronto, Controlo e destes fatores para o *Burnout*. Por último,

o modelo de mediação parcial acrescenta ao modelo de mediação total, caminhos diretos dos fatores Conflito Trabalho-Família e Conflito Família-Trabalho para o *Burnout*. Os índices de qualidade de ajustamento destes modelos encontram-se resumidos na Tabela 2.



**Figura 3** - Modelo de efeitos diretos.



**Figura 4** - Modelo de mediação total.

**Tabela 2** - Índices de ajustamento dos modelos estruturais.

Modelo	$\chi^2$	gl	RMSEA	PCLOSE	TLI	CFI	NFI
Efeitos Diretos	1585,156	581	0,064	0,000	0,928	0,933	0,891
Mediação Total	1338,404	576	0,056	0,070	0,945	0,949	0,915
Mediação Parcial	1209,977	574	0,051	0,319	0,954	0,958	0,923

De acordo com os índices de qualidade de ajustamento, o modelo de mediação total apresenta um bom ajustamento ( $\chi^2(576)=1338,404$ ; RMSEA= 0,056; TLI=0,945; CFI=0,949 e NFI=0,915) e melhor do que o ajustamento do modelo de efeitos diretos ( $\chi^2(581)=1585,156$ ; RMSEA=0,064; TLI=0,928; CFI=0,933 e NFI=0,899). A diferença da estatística  $\chi^2$  dos dois modelos ( $\Delta\chi^2(7)= 375,179$ ;  $p < 0,001$ ) permite concluir que os dois modelos diferem significativamente na qualidade do ajustamento e por isso os efeitos de mediação não podem ser ignorados. No entanto, o modelo de mediação parcial apresenta um bom ajustamento ( $\chi^2(574)=1209,977$ ; RMSEA=0,051; TLI=0,954; CFI=0,958 e NFI=0,923) e melhor do que o ajustamento do modelo de mediação total. A diferença da estatística  $\chi^2$  dos dois modelos ( $\Delta\chi^2(2)=128,427$ ;  $p < 0,001$ ) permite concluir que os dois modelos diferem significativamente na qualidade do ajustamento e por isso os efeitos diretos dos fatores Conflito Trabalho-Família e Conflito Família-Trabalho sobre o *Burnout* não devem ser ignorados. O modelo de mediação parcial é considerado o submodelo estrutural.

Recorrendo à reamostragem *bootstrap*, obtiveram-se as estimativas dos efeitos diretos e dos efeitos indiretos do submodelo estrutural assim como, os correspondentes intervalos de confiança a 95% com correcção de enviesamento (*bias-corrected bootstrap*), tal como recomendado por MacKinnon et al. (2004). Estas estimativas são apresentadas na Tabela 3.

O fator Conflito Trabalho – Família (CTF) apresenta um efeito direto estatisticamente significativo sobre o *Burnout* ( $\hat{\beta}=0,474$ ) e um efeito indireto (efeito mediado pelos fatores Ameaça, Confronto e Controlo) estatisticamente significativo sobre o *Burnout* ( $\hat{\beta}=0,070$ ). No entanto, o fator Conflito Família – Trabalho (CFT) não apresenta nem um efeito direto, nem um efeito indireto estatisticamente significativos sobre o *Burnout*.

A percentagem da variância do *Burnout* explicada pelo modelo é 45,5%.

**Tabela 3** - Estimativas estandarizadas dos efeitos diretos e indiretos e intervalos de confiança de 95% no modelo de mediação parcial.

	Variáveis Dependentes				<i>Burnout</i>	
	AM	DF	CF	CL	Efeitos Diretos	Efeitos Indiretos
	CTF	0,448** (0,329;0,558)	-0,158* (-0,278;-0,036)			0,474** (0,363;0,599)
CFT	0,113 (ns) (0,002;0,237)	-0,029 (ns) (-0,155;0,098)			0,101 (ns) (-0,008;0,208)	0,016 (ns) (-0,012;0,046)
AM			-0,304** (-0,423;-0,188)	-0,223** (-0,346;-0,106)		
DF			0,420** (0,305;0,522)	0,385** (0,262;0,499)		
CF					-0,271** (-0,373;-0,167)	
CL					-0,093 (ns) (-0,189;0,003)	
R <sup>2</sup>	0,258** (0,174;0,347)	0,030** (0,005;0,069)	0,291** (0,198;0,374)	0,213** (0,130;0,291)	0,455** (0,350;0,546)	

Notas: \*  $p < 0,05$ ; \*\*  $p < 0,01$ ; ns: não significativo

## 5 Resultados

Este estudo confirmou a importância da avaliação cognitiva na relação entre a percepção de mútua interferência trabalho-família (e.g., conflito trabalho-família e conflito família-trabalho) e a experiência de *burnout* (e.g., fadiga física, fadiga cognitiva e exaustão emocional) dos professores portugueses.

O modelo de mediação da avaliação cognitiva em que os efeitos diretos dos fatores Conflito Trabalho-Família e Conflito Família-Trabalho sobre o *Burnout* foi mantido, confirma a influência da avaliação cognitiva na relação entre a percepção de mútua interferência trabalho-família e a experiência de *burnout*; no entanto, também reforçou a influência da percepção de mútua interferência trabalho-família na experiência de *burnout*.

## Agradecimentos

Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos Nacionais através da FCT - “Fundação para a Ciência e a Tecnologia”, no âmbito do projeto PEstOE/MAT/UI0013/2017.

## Referências

- BYRNE, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, New York: Routledge.
- GLASER, W. & Tracy, D. H. (2013). Work-family conflicts, threat-appraisal, self-efficacy and emotional exhaustion. *Journal of Managerial Psychology*, 28(2), 164-182.
- GOMES, A. R. (2008). *Escala de Avaliação Cognitiva*. Relatório técnico não publicado. Braga: Instituto de Educação e Psicologia, Universidade do Minho.
- GOMES, A. R. (2012). *Medida de “Burnout” de Shirom-Melamed (MBSM)*. Relatório técnico não publicado. Braga: Escola de Psicologia, Universidade do Minho.
- GOMES, A. R., FARIA, S. & GONÇALVES, A. M. (2013). Cognitive appraisal as a mediator in the relationship between stress and burnout, *Work & Stress*, 27(4), 351-367.
- LAZARUS, R. S. & FOLKMAN, S. (1984). *Stress, appraisal, and coping*. New York: Springer.
- MACKINNON, D., LOCKWOOD, C. & WILLIAMS, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99-128.
- NETEMEYER, R.G., MCMURRIAN, R. & BOLES, J.S. (1996). Development and validation of work-family conflict and family-work conflict scales. *Journal of Applied Psychology*, 81, 400-410.
- SALGUEIRO, M.F. (2012). *Modelos com Equações Estruturais*. Lisboa / Portugal. Sociedade Portuguesa de Estatística.

**Editores:** Helena Bacelar-Nicolau, Fernanda Sousa,  
Carlos Marcelo, Ana Sousa Ferreira, Paulo Infante,  
Adelaide Figueiredo

O volume CLADMAp III da série CLASSIFICAÇÃO E ANÁLISE DE DADOS – Métodos e Aplicações, vem continuar a dar resposta à vontade expressa pelos associados e participantes nas sucessivas JOCLAD – Jornadas de Classificação e Análise de Dados, de que a CLAD divulgasse os trabalhos nelas apresentados, dando-lhes visibilidade e valorizando-os. Os artigos incluídos neste CLADMAp III, após processo de revisão interpares, são desenvolvimentos de trabalhos apresentados nas JOCLAD 2014-2017 e mostram, mais uma vez, a interdisciplinaridade e a diversidade das áreas que integram estas Jornadas.



Associação Portuguesa de  
Classificação e Análise de Dados