

Editores

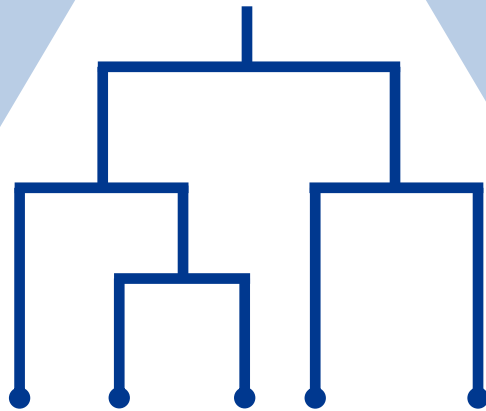
Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Ferreira

# CLASSIFICAÇÃO E ANÁLISE DE DADOS

## MÉTODOS E APLICAÇÕES



CLAD

Editores

**Helena Bacelar-Nicolau**

**Fernanda Sousa**

**Carlos Ferreira**

# **CLASSIFICAÇÃO E ANÁLISE DE DADOS**

## **MÉTODOS E APLICAÇÕES**



**Título**

Classificação e Análise de Dados – Métodos e Aplicações

**Editores**

Helena Bacelar-Nicolau (Universidade de Lisboa)

Fernanda Sousa (Universidade do Porto)

Carlos Ferreira (Universidade de Aveiro)

**Impressão**

Instituto Nacional de Estatística

Av. António José de Almeida

1000-043 LISBOA

**1.ª Edição**

**Lisboa, Abril de 2013**

**ISSN 2183-8801**

Os textos e opiniões expressos nesta publicação são da exclusiva responsabilidade dos seus autores.

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

# Índice

## CLASSIFICAÇÃO E ANÁLISE DE DADOS MÉTODOS E APLICAÇÕES

Prefácio .....v

Agradecimento aos revisores ..... vii

Sobre a Comparação de Estruturas de Classificação:  
Coeficientes e suas Distribuições .....1  
*Fernanda Sousa e Jorge Tendeiro (JOCLAD 2004)*

Application of Clustering Methods for Optimizing the Location  
of Treated Wood Remediation Units .....15  
*Helena Gomes, Victor Lobo e Alexandra B. Ribeiro (JOCLAD 2004)*

Modelos de Classe Latente de Regressão para Dados de  
Contagem. A Análise da Heterogeneidade Regional nas Eleições  
Presidenciais Portuguesas de 2006 .....23  
*Francisco Figueiredo e José G. Dias (JOCLAD 2006)*

Análise Conjunta Baseada em Preferências vs Escolhas. Selecção  
de uma Escola de Gestão .....37  
*Susana I.A. Tavares, José G. Dias e Margarida G.M.S. Cardoso  
(JOCLAD 2006)*

Geração Aleatória de Estruturas de Classificação: das  
Hierarquias às Pirâmides .....49  
*Vasco Machado e Fernanda Sousa (JOCLAD 2007)*

<b>Coefficientes de Comparação de Partições em Análise Classificatória: Abordagem Clássica vs uma Abordagem Probabilística .....</b>	<b>63</b>
--	-----------

*Oswaldo Silva, Helena Bacelar-Nicolau, Fernando C. Nicolau e Áurea  
Sousa (JOCLAD 2008)*

<b>Matriz de Dissemelhança Entrópica para Classificação Não Supervisionada.....</b>	<b>81</b>
---	-----------

*Jorge M. Santos (JOCLAD 2008)*

<b>Tipologia de Comportamentos de Crianças em Idade Pré- escolar: Aplicação de Modelos de Análise Classificatória Hierárquica .....</b>	<b>93</b>
---	-----------

*António M. Caxaria, Helena Bacelar-Nicolau e Rita M. Leal  
(JOCLAD 2008)*

<b>Estimação da Abundância e das Taxas de Incidência e Prevalência em Populações Elusivas.....</b>	<b>105</b>
--	------------

*Anabela Afonso, João F. Monteiro, Sónia Batista e Russell Alpizar-  
Jara (JOCLAD 2009)*

<b>Caracterização, Classificação e Discriminação de Doentes Atendidos no Serviço de Urgência Devido a Resultados Clínicos Negativos da Farmacoterapia.....</b>	<b>115</b>
--	------------

*Margarida Cavaco, Luís S. Dias e Fernando Fernández-Llimós  
(JOCLAD 2009)*

<b>Contributos para Previsão do Consumo de Energia Eléctrica na Ilha de São Miguel .....</b>	<b>125</b>
--	------------

*Armando B. Mendes (JOCLAD 2010)*

## Prefácio

A Associação Portuguesa de Classificação e Análise de Dados, CLAD, fundada em 1994 é uma associação científica e técnica que tem por objecto:

- a promoção e coordenação da investigação e da utilização da classificação e da análise de dados, assumindo-se como a instância nacional privilegiada da concertação e da assistência mútua nestas matérias;
- a organização, gestão e prestação de serviços e produtos relacionados com o objecto social, nomeadamente quanto a publicações, seminários, aplicações e encontros nacionais e internacionais;
- a representação dos utilizadores perante entidades nacionais, comunitárias e internacionais.

Em todas as reuniões anuais da CLAD, as JOCLAD, houve lugar à publicação de Actas, geralmente na forma de “extended abstracts”. Em 1999 e em 2001, quando a CLAD patrocinou o simpósio internacional *ASMDA’99* e o congresso europeu *EMPG’2001*, os *Proceedings* foram igualmente publicados, em inglês. Tal como o foram as lições e artigos apresentados na escola de investigação *JISS’2003*, também patrocinada pela CLAD (em conjunto com a *IFCS* e a *IASC*).

Fomos assim percebendo que havia, da parte dos participantes nas JOCLAD e dos sócios da CLAD, um interesse em reunir, prioritariamente, artigos correspondentes a alguns dos trabalhos apresentados nas JOCLAD. A presente publicação da CLAD, é o primeiro resultado desse caminho, com uma amostragem dos que foram apresentados entre as JOCLAD2004 e as JOCLAD2010. Os trabalhos aqui incluídos reflectem a diversidade de áreas que integram estas jornadas, sendo alguns de cariz marcadamente teórico ou metodológico e outros bons exemplos de aplicações, desta área do saber, a diferentes sectores da sociedade. Em qualquer caso houve uma preocupação grande com a originalidade, o rigor científico e a relevância para a dignificação e o desenvolvimento da CLAD e dos seus objectivos.

Para ordem sequencial dos artigos seguiu-se a ordem cronológica das JOCLAD em que foram apresentados e, dentro do mesmo ano, o pendor mais teórico ou mais aplicado. Uma pequena nota recomendável a uma leitura adequada desta publicação prende-se com a actualização dos trabalhos que a integram, havendo em alguns casos desenvolvimentos posteriores, dos próprios autores ou de outros investigadores, aos aqui

apresentados. Note-se enfim que alguns textos seguem o novo acordo ortográfico, outros não, já que deixámos aos autores a liberdade dessa escolha.

Acreditamos que, com a experiência entretanto adquirida e com o renovado entusiasmo de todos, a esta publicação outras se sigam, obedecendo agora a uma lei das aproximações sucessivas rapidamente convergente.

Agradecemos aos autores, aos revisores e a todos os que directa ou indirectamente nos apoiaram na criação desta publicação.

Ao INE que, desde o início, tem sido parceiro privilegiado da CLAD, cooperando nas suas actividades nacionais e internacionais, e mais especialmente na edição das publicações, o nosso agradecimento particular.

Lisboa, Abril de 2013

Helena Bacelar-Nicolau (Universidade de Lisboa)

Fernanda Sousa (Universidade do Porto)

Carlos Ferreira (Universidade de Aveiro)

## **Agradecimento aos Revisores**

Os editores agradecem aos seguintes Colegas, listados por ordem alfabética do seu apelido, pelo generoso trabalho de revisão dos artigos submetidos, que em muito valorizou o conteúdo desta publicação.

FERNANDO BAÇÃO (Universidade Nova de Lisboa)  
HELENA BACELAR-NICOLAU (Universidade de Lisboa)  
PAULA BRITO (Universidade do Porto)  
JORGE CADIMA (Universidade Técnica de Lisboa)  
MARGARIDA CARDOSO (Instituto Universitário de Lisboa)  
PEDRO COELHO (Universidade Nova de Lisboa)  
ISABEL DÓRIA (Universidade de Lisboa)  
PEDRO DUARTE SILVA (Universidade Católica Portuguesa)  
ADELAIDE FIGUEIREDO (Universidade do Porto)  
JOÃO GAMA (Universidade do Porto)  
PAULO GOMES (Instituto Nacional de Estatística)  
JOSÉ GONÇALVES DIAS (Instituto Universitário de Lisboa)  
VICTOR LOBO (Escola Naval)  
SUSANA NASCIMENTO (Universidade Nova de Lisboa)  
MANUELA NEVES (Universidade Técnica de Lisboa)  
IRENE OLIVEIRA (Universidade de Trás-os-Montes e Alto Douro)  
MARIA DE FÁTIMA SALGUEIRO (Instituto Universitário de Lisboa)  
GILDA SOROMENHO (Universidade de Lisboa)  
ANA SOUSA FERREIRA (Universidade de Lisboa)  
ÁUREA SOUSA (Universidade dos Açores)  
FERNANDA SOUSA (Universidade do Porto)  
PAULA VICENTE (Instituto Universitário de Lisboa)





# Sobre a Comparação de Estruturas de Classificação: Coeficientes e suas Distribuições

Fernanda Sousa<sup>1</sup> · Jorge Tendeiro<sup>2</sup>

© The Author(s) 2013

**Resumo** Numa Classificação Hierárquica Ascendente a função de comparação entre pares de elementos e o critério de agregação induzem relações estruturais entre os elementos do conjunto a classificar, aqui designadas por estruturas de classificação. Neste trabalho são invocadas razões para a necessidade de comparar pares destas estruturas e são introduzidos coeficientes adequados para tal comparação. As distribuições teóricas assintóticas desses coeficientes são apresentadas e é discutida a sua adequação ao problema em análise. A solução proposta passa pela dedução de distribuições empíricas, por recurso à simulação, e é ilustrada para o caso da comparação de dendrogramas sobre o mesmo conjunto de elementos, recorrendo a métodos de geração aleatória de dendrogramas.

**Palavras-chave:** Classificação Hierárquica, Dendrograma, Geração Aleatória de Dendrogramas, Coeficientes Ordinais de Comparação, Distribuição de Coeficientes.

## 1 Introdução

A Análise Classificatória tem por objectivo agrupar um conjunto de objectos num número relativamente pequeno de classes, satisfazendo a condição de que objectos de uma mesma classe sejam mais semelhantes entre si que objectos pertencentes a classes distintas. O trabalho aqui apresentado foca-se nos métodos de Classificação Hierárquica Ascendente (C.H.A.). Da aplicação de um método de

---

<sup>1</sup>Faculdade de Engenharia e CITTA, Universidade do Porto, fcsousa@fe.up.pt

<sup>2</sup>Departamento de Psicometria e Estatística, Faculdade de Psicologia, Universidade de Groningen, j.n.tendeiro@rug.nl

C.H.A. a um quadro de dados surgem várias entidades que reflectem relações de proximidade entre os elementos a classificar e informação da sua pertença a um mesmo grupo. Entre essas entidades, aqui designadas por estruturas de classificação, salientam-se as produzidas pela função de comparação entre pares de elementos e as associadas ao *output*, sendo os dendrogramas as mais usadas. É bem conhecido e aceite pela comunidade científica (Gordon, 1999) que diferentes opções na aplicação de uma C.H.A. conduzem frequentemente a resultados diferentes, não havendo escolhas reconhecidas como genericamente melhores. É atitude corrente considerar várias opções na aplicação de uma C.H.A. a um dado quadro de dados, daí resultando a necessidade de avaliar o grau de concordância das várias estruturas de classificação produzidas.

A comparação de estruturas de classificação, obtidas para o mesmo conjunto de elementos, é o tema deste trabalho. Serão apresentados coeficientes adequados a este tipo de comparação, bem como uma análise das suas distribuições.

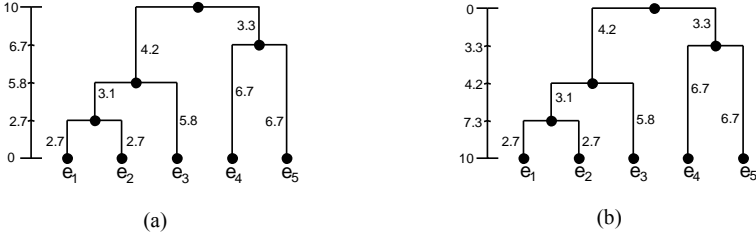
Na Secção 2 são introduzidas definições de apoio ao desenvolvimento do texto. A Secção 3 é dedicada à abordagem proposta para comparar estruturas de classificação. A Secção 4 foca-se nas distribuições assintóticas e empíricas dos coeficientes introduzidos na secção anterior. Uma apresentação e discussão dos resultados são feitas na Secção 5. A Secção 6 é dedicada a algumas conclusões.

## 2 Classificação Hierárquica Ascendente

Os métodos de C.H.A. têm por base a definição de duas funções de comparação: a **função de comparação entre elementos**,  $\gamma$ , e a **função de comparação entre classes** (critério de agregação),  $\Gamma$ . Dado um conjunto  $E = \{e_1, e_2, \dots, e_m\}$  de objectos a classificar, a função  $\gamma: E \times E \rightarrow \mathbb{R}_0^+$ , que pode ser do tipo dissemelhança ou semelhança, mede o grau de parecença entre pares de elementos. A função  $\Gamma: P(E) \times P(E) \rightarrow \mathbb{R}_0^+$ , onde  $P(E)$  é o conjunto das partes de  $E$ , mede o grau de parecença entre pares de partes de  $E$ .

Como resultado de uma C.H.A. tem-se uma **hierarquia indiciada**, um **dendrograma** ou uma **matriz ultramétrica** sobre  $E$ . Uma hierarquia indiciada ou indexada é um par  $(H, h)$ .  $H$  é uma hierarquia, isto é, uma sucessão de partições encaixadas de  $E$ . A função  $h$  associa a cada parte de  $E$ , seja  $A$ , o valor da função de comparação entre classes que deu origem ao nível em que  $A$  é formada. Um dendrograma é uma árvore ponderada com raiz, em que os nós terminais ou folhas são etiquetados e estão todos à mesma distância da raiz. Os nós internos de um dendrograma podem ser ordenados de acordo com a sua distância relativa às folhas, quando o critério de agregação é do tipo dissemelhança (Figura 1 - (a)), ou à raiz, quando o critério de agregação é do tipo semelhança (Figura 1 - (b)). Os valores associados aos nós designam-se por índices de nível e são dados pela função  $h$  introduzida acima.

Muitas vezes, mais do que trabalhar com índices de nível, trabalha-se com níveis de agregação (ou níveis de fusão), que são os valores ordinais dos índices de nível. Desta forma, dá-se mais relevância à posição relativa dos nós, em detrimento dos valores reais das distâncias entre eles. Sem prejuízo de generalização, neste trabalho consideram-se dendrogramas completamente binários.



**Figura 1** – (a) Distância relativa às folhas (b) Distância relativa à raiz.

Um dendrograma fica completamente definido por três características:

- **topologia:** diz respeito à forma, ou seja, ignora as etiquetas e os pesos atribuídos aos diferentes ramos; sob este ponto de vista, duas árvores são distintas se possuírem sistemas de bifurcação distintos (Figura 2);
- **etiquetas das folhas:** fixada uma certa topologia, há usualmente diferentes formas de etiquetar as folhas (Figura 3);
- **níveis de agregação:** dois dendrogramas que partilhem a topologia e a etiquetagem podem diferir nos níveis de agregação (Figura 4).

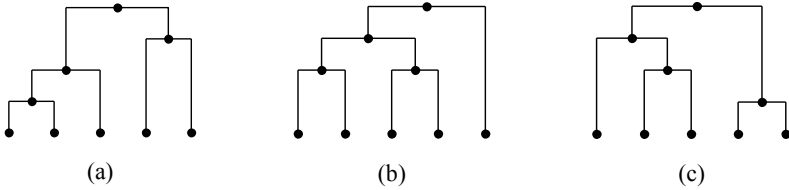
O número de dendrogramas distintos é uma função  $d(m)$  do número  $m$  de nós terminais; admitindo que não há empates nos níveis de fusão, mostra-se (por exemplo, em Podani, 2000, pg. 126) que

$$d(m) = \frac{m!(m-1)!}{2^{m-1}} \quad (1)$$

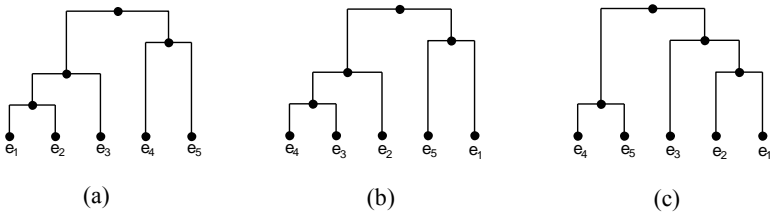
Dois dendrogramas que têm a mesma topologia, etiquetagem das folhas e níveis de fusão dizem-se **isomorfos**. Uma representação alternativa e equivalente ao dendrograma é a **matriz ultramétrica** (Figura 5). A cada par de nós terminais  $e_i, e_j \in E$  associa-se o valor  $h(\{e_i, e_j\})$ . Trata-se de uma matriz simétrica com  $M = \binom{m}{2}$  entradas relevantes, que representam todas as combinações possíveis de pares de vértices distintos. O nome desta matriz vem do facto dos seus elementos verificarem a **propriedade ultramétrica**, ou seja,

- $h(e_i, e_j) \geq \min\{h(e_i, e_k), h(e_k, e_j)\}, \forall e_i, e_j, e_k \in E$ , se a função de comparação entre elementos for do tipo semelhança.

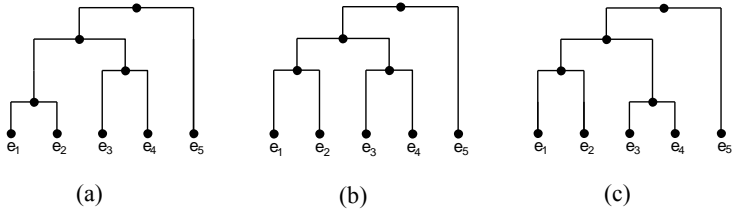
- $h(e_i, e_j) \leq \max\{h(e_i, e_k), h(e_k, e_j)\}, \forall e_i, e_j, e_k \in E$ , se a função de comparação entre elementos for do tipo dissimilaridade.



**Figura 2** – As árvores (a) e (b) não são topologicamente idênticas, mas as árvores (a) e (c) são topologicamente idênticas.



**Figura 3** – As árvores (a), (b), e (c) são todas topologicamente idênticas, mas apenas as árvores (a) e (c) são idênticas quando se consideram as etiquetas.



**Figura 4** – Os dendrogramas (a), (b) e (c) são todos distintos, embora tenham a mesma topologia e etiquetagem.



**Figura 5** – Dendrograma e respectiva matriz ultramétrica.

Uma **relação binária** sobre um conjunto  $E$  é um conjunto  $R \subseteq E \times E$ . Uma relação binária reflexiva e transitiva diz-se uma **preordem**, enquanto uma relação binária reflexiva, antisimétrica e transitiva diz-se uma **ordem**. Uma relação em que quaisquer dois elementos são comparáveis diz-se **total**.

Seja  $O$  uma relação binária sobre o conjunto  $E \times E$  que verifica as seguintes condições: (i)  $(e_i, e_i)O(e_j, e_k), \forall e_i, e_j, e_k \in E$  (ii)  $(e_i, e_j)O(e_j, e_i), \forall e_i, e_j \in E$  e (iii)  $(e_j, e_k)O(e_i, e_i) \Rightarrow e_j = e_k, \forall e_i, e_j, e_k \in E$ . Sejam ainda os conjuntos  $F$  e  $G$  definidos por  $F = \{(e_i, e_j): e_i, e_j \in E, e_i \neq e_j\}$  e  $G = \{(e_i, e_j): e_i, e_j \in E, e_i \neq e_j\}$ .

Diz-se que  $O$  é uma **ordenação (preordenação)** de  $E$  se se verificar uma das três condições (equivalentes) seguintes: (i)  $O$  é ordem (preordem) em  $E \times E$  (ii)  $O$  é ordem (preordem) em  $F$  (iii)  $O$  é ordem (preordem) total em  $G$ . Na Subsecção 3.3 será apresentado um exemplo que ilustra estas noções.

### 3 Comparação de estruturas de classificação

#### 3.1 Abordagem proposta

O conjunto de valores da função de comparação entre elementos,  $\gamma(e_i, e_j)$  com  $e_i, e_j \in E$ , a matriz ultramétrica, as sucessivas partições da hierarquia, ou o dendrograma revelam, sobre os elementos de  $E$ , relações que definem **estruturas de classificação**. À função de comparação sobre pares de elementos associa-se, em geral, uma ordenação de  $E$ , enquanto à matriz ultramétrica, a uma partição, ou a um dendrograma se associam preordenações. Escolhas diferentes de funções de comparação (entre pares de elementos ou associadas aos critérios de agregação) produzem, frequentemente, diferentes estruturas de classificação (dendrogramas, matrizes ultramétricas, hierarquias, partições). Para responder, entre outras, a questões como qual a melhor escolha a fazer para as funções de comparação, ou se o resultado da C.H.A. está de acordo com a estrutura inicial dos dados, ou qual o grau de concordância entre os resultados obtidos por dois métodos de classificação sobre um mesmo conjunto de dados, recorre-se à comparação de estruturas de classificação (Nicolau, 1984; Sousa & Nicolau, 2001). Estas preocupações inserem-se na área da Validação em Classificação e têm sido objecto de estudo nas últimas décadas (ver por exemplo Gordon, 1996; Bock, 1996 e Sousa, 2000; Halkidi *et al.*, 2001). Conforme foi já referido anteriormente, este trabalho não responde directamente a estas questões e desenvolve-se a dois níveis.

Num nível mais geral propõe-se uma abordagem ordinal para tratar o problema da comparação de estruturas de classificação associadas a uma C.H.A.. O processo consiste essencialmente em três passos:

1. Associar as estruturas de classificação a preordenações: de facto, é possível fazer corresponder dendrogramas, matrizes ultramétricas e hierarquias indiciadas a preordenações; sendo assim, o nosso propósito — comparar estruturas de classificação — é conseguido através da comparação das preordenações correspondentes.
2. Introduzir e estudar coeficientes de correlação ordinal aplicados a preordenações.
3. Deduzir distribuições adequadas para os coeficientes, tendo em vista a atribuição de significado estatístico aos seus valores.

A um segundo nível neste trabalho aplica-se esta abordagem ao caso particular de dendrogramas obtidos, a partir de um mesmo conjunto de dados, por aplicação de C.H.A. com diferentes escolhas de funções de comparação.

### 3.2 Coeficientes de comparação de preordenações

Sejam  $m$  o número de elementos a classificar e  $M = \binom{m}{2} = \text{card}(F)$  (notação introduzida na Secção 2). Consideremos duas preordenações  $\omega_1$  e  $\omega_2$ , sobre um mesmo conjunto  $E$ , de comprimento  $M$ , do tipo  $(M_1, M_2, \dots, M_k)$  e  $(N_1, N_2, \dots, N_h)$ . Isto significa que, por exemplo,  $\omega_1$  é uma preordenação em que ocorrem  $k$  valores distintos, com  $M_1$  elementos de  $F$  associados ao primeiro patamar (considerados iguais),  $M_2$  elementos associados ao segundo patamar, etc.. Tem-se portanto  $M_1 + M_2 + \dots + M_k = M = N_1 + N_2 + \dots + N_h$ . Definam-se as variáveis aleatórias  $U_l =$  "número de ordem do  $l$ -ésimo elemento de  $\omega_1$ " e  $V_l =$  "número de ordem do  $l$ -ésimo elemento de  $\omega_2$ ", com  $l = 1, 2, \dots, M$ . Aos elementos de um mesmo patamar de uma preordem atribui-se um valor que é a média das ordens que esses elementos teriam se os seus valores fossem diferentes mas consecutivos. Isto permite que a soma dos  $M$  valores associados aos elementos de  $F$  seja igual a  $1 + 2 + \dots + M = \frac{(1+M)M}{2}$ , precisamente o mesmo valor que se obteria se os valores associados aos elementos de  $F$  fossem distintos.

Considerem-se as variáveis aleatórias  $A_{ij} = \text{sgn}(U_j - U_i) \text{sgn}(V_j - V_i)$ ,  $1 \leq i < j \leq M$ , onde  $\text{sgn}$  representa a função sinal. As variáveis  $A_{ij}$  podem tomar três valores:

- $A_{ij} = 1$ , se os pares  $(U_i, V_i)$  e  $(U_j, V_j)$  são **concordantes**;
- $A_{ij} = -1$ , se os pares  $(U_i, V_i)$  e  $(U_j, V_j)$  são **discordantes**;
- $A_{ij} = 0$ , se nos pares  $(U_i, V_i)$  e  $(U_j, V_j)$  ocorre **empate**.

Definam-se ainda as variáveis aleatórias  $C$  e  $D$ .  $C$  denota o número de concordâncias entre as variáveis  $U$  e  $V$ , ou seja, o número de vezes que  $A_{ij} = 1$  (com  $1 \leq i < j \leq M$ ).  $D$  denota o número de discordâncias entre  $U$  e  $V$ , ou seja, o número de vezes que  $A_{ij} = -1$ .

Existem na literatura vários coeficientes que podem ser usados para comparar preordenações. Neste estudo, consideram-se os coeficientes de correlação ordinal de Spearman, Kendall e Goodman-Kruskal (Kendall, 1970).

O coeficiente de correlação ordinal de Spearman para as duas preordenações  $\omega_1$  e  $\omega_2$  é dado por

$$R_S = \frac{\frac{M^3-M}{6} - S(d^2) - T_1 - T_2}{\sqrt{\frac{M^3-M}{6} - 2T_1} \sqrt{\frac{M^3-M}{6} - 2T_2}} \quad (2)$$

em que

$$S(d^2) = \sum_{1 \leq i \leq M} (U_i - V_i)^2, T_1 = \frac{\sum_{1 \leq i \leq k} (M_i^3 - M_i)}{12}, T_2 = \frac{\sum_{1 \leq j \leq k} (N_j^3 - N_j)}{12}. \quad (3)$$

O coeficiente Tau de Kendall,  $T_K$ , para comparar preordenações, tem por base a noção de grafo de uma preordenação,  $gr(\omega_i)$ , em que

$$gr(\omega_i) = \{(\{x, y\}, \{z, t\}) \in F \times F : \{x, y\} \neq \{z, t\}, \{x, y\} \leq \{z, t\} \text{ e } \{z, t\} > \{x, y\}\}$$

Sendo  $\omega_1$  e  $\omega_2$  preordenações totais tem-se que

$$|gr(\omega_1)| = \binom{M}{2} - \sum_{i=1}^k \binom{M_i}{2} = \sum_{i < j} M_i M_j$$

e

$$|gr(\omega_2)| = \binom{M}{2} - \sum_{i=1}^h \binom{N_i}{2} = \sum_{i < j} N_i N_j$$

Finalmente vem

$$T_K = \frac{C-D}{|gr(\omega_1)||gr(\omega_2)|} = \frac{C-D}{\sqrt{\sum_{i < j} M_i M_j} \sqrt{\sum_{i < j} N_i N_j}}. \quad (4)$$

O coeficiente de Goodman-Kruskal,  $T_{GK}$ , para comparar preordenações é dado por

$$T_{GK} = \frac{C-D}{C+D}. \quad (5)$$

Este coeficiente é especialmente útil na comparação de preordenações que têm empates muito extensos, como os que ocorrem frequentemente nas preordenações associadas a hierarquias aquando da junção de duas classes numerosas. Outras



propriedades destes coeficientes e detalhes sobre as suas distribuições assintóticas podem ser vistas em Kendall (1970), Sousa (2000) ou Tendeiro (2005).

### 3.3 Exemplo

Considerem-se os dendrogramas (a) e (b) da Figura 6, onde  $m = 5$  e  $M = 10$ . A preordenação associada ao dendrograma (a) é a preordem de  $F$ , do tipo (1,1,4,4), dada por  $\{e_2, e_3\} < \{e_4, e_5\} < \{e_2, e_4\} = \{e_2, e_5\} = \{e_3, e_4\} = \{e_3, e_5\} < \{e_1, e_2\} = \{e_1, e_3\} = \{e_1, e_4\} = \{e_1, e_5\}$ . Esta preordenação relaciona todos os pares de elementos por ordem crescente de semelhança entre si. Assim, por exemplo,  $\{e_4, e_5\} < \{e_1, e_3\}$  significa que os elementos  $e_4$  e  $e_5$  são mais semelhantes entre si do que os elementos  $e_1$  e  $e_3$ .

A preordenação associada ao dendrograma (b) é do tipo (1,1,2,6) e é dada por  $\{e_2, e_3\} < \{e_4, e_5\} < \{e_1, e_2\} = \{e_1, e_3\} < \{e_1, e_4\} = \{e_1, e_5\} = \{e_2, e_4\} = \{e_2, e_5\} = \{e_3, e_4\} = \{e_3, e_5\}$ .



**Figura 6** – Dendrogramas.

Escrevendo o conjunto  $F$  na forma  $F = \{\{e_1, e_2\}, \{e_1, e_3\}, \{e_1, e_4\}, \{e_1, e_5\}, \{e_2, e_3\}, \{e_2, e_4\}, \{e_2, e_5\}, \{e_3, e_4\}, \{e_3, e_5\}, \{e_4, e_5\}\}$ , obtêm-se, para as variáveis  $U$  e  $V$ , os valores que constam da Tabela 1.

**Tabela 1** – Valores das variáveis aleatórias  $U$  e  $V$  no exemplo considerado.

$\{e_i, e_j\}$	$\{e_1, e_2\}$	$\{e_1, e_3\}$	$\{e_1, e_4\}$	$\{e_1, e_5\}$	$\{e_2, e_3\}$	$\{e_2, e_4\}$	$\{e_2, e_5\}$	$\{e_3, e_4\}$	$\{e_3, e_5\}$	$\{e_4, e_5\}$
$k$	1	2	3	4	5	6	7	8	9	10
$u_k$	8.5	8.5	8.5	8.5	1	4.5	4.5	4.5	4.5	2
$v_k$	3.5	3.5	7.5	7.5	1	7.5	7.5	7.5	7.5	2

Aplicando as fórmulas (2) e (3) de cálculo de  $S(d^2)$ ,  $T_1$ ,  $T_2$  e  $R_S$  é fácil verificar que  $s(d^2) = 88$ ,  $t_1 = 10$  e  $t_2 = 18$ , donde  $r_S = 0.358$  (3 c.d.).

Atendendo a que  $c = 20$ ,  $d = 8$ ,  $\sum_{i < j} M_i M_j = 33$  e  $\sum_{i < j} N_i N_j = 29$ , substituindo em (4) e (5), obtêm-se  $t_K = 0.388$  (3 c.d.) e  $t_{GK} = 0.429$  (3 c.d.).

## 4 Distribuições dos coeficientes de comparação

Frequentemente é necessário interpretar probabilisticamente os valores fornecidos pelos coeficientes de comparação atrás definidos, por exemplo determinando a significância estatística. Alguns resultados sobre as distribuições assintóticas destes coeficientes existem (Kendall, 1970). O coeficiente de correlação ordinal de Spearman,  $R_s$ , tem distribuição assintótica normal de média zero e variância  $\frac{1}{M-1}$ . Em Kendall (1970) demonstra-se que a variável (C-D) tem também distribuição assintótica normal com:

$$\begin{aligned}
 E(C - D) &= 0 \quad \text{e} \quad V(C - D) = \\
 &= \frac{1}{18} \left[ M(M-1)(2M+5) - \sum_{1 \leq i \leq k} M_i(M_i-1)(2M_i+5) - \sum_{1 \leq l \leq h} N_l(N_l-1)(2N_l+5) \right] \\
 &\quad + \frac{1}{9M(M-1)(M-2)} \left[ \sum_{1 \leq i \leq k} M_i(M_i-1)(M_i-2) \right] \times \left[ \sum_{1 \leq l \leq h} N_l(N_l-1)(N_l-2) \right] \\
 &\quad + \frac{1}{2M(M-1)} \left[ \sum_{1 \leq i \leq k} M_i(M_i-1) \right] \times \left[ \sum_{1 \leq l \leq h} N_l(N_l-1) \right] \quad (6)
 \end{aligned}$$

Para o coeficiente de Goodman-Kruskal há também alguns resultados assintóticos para a distribuição normal, para situações particulares, que pela sua complexidade e não utilização no presente trabalho não serão aqui apresentados.

O recurso às distribuições assintóticas destes coeficientes está contudo comprometido no contexto do presente trabalho, principalmente por dois motivos:

1. As estruturas de classificação a comparar são obtidas a partir de um mesmo quadro de dados.
2. A propriedade ultramétrica induz relações entre os valores das sequências a comparar.

Assim a hipótese de “independência”, subjacente à dedução das distribuições assintóticas, não se verifica quando se trata da comparação de estruturas de classificação.

O objectivo natural seria a obtenção das correspondentes distribuições exactas para estes coeficientes, aplicados à comparação de estruturas de classificação. Foi já referido que, na aplicação de uma C.H.A. a um conjunto de dados, a informação contida no dendrograma, na matriz ultramétrica ou na hierarquia indiciada, é equivalente, e as estruturas de classificação associadas a estas entidades são coincidentes. Vamos, por isso, centrar-nos em dendrogramas. A obtenção da distribuição exacta de um coeficiente neste contexto passaria pela enumeração do

conjunto de todos os dendrogramas, fixado o número de nós terminais. Contudo tal não é exequível, tendo em conta o rápido crescimento de  $d(m)$  (ver (1)):

**Tabela 2** – Número de dendrogramas não isomorfos.

$m$	4	5	6	10	15	20	50
$d(m)$	18	180	2700	2571912000	$5.14 \times 10^{18}$	$1.53 \times 10^{29}$	$> 10^{100}$

A solução passa pela determinação de distribuições empíricas, por recurso à simulação.

Para a obtenção de amostras aleatórias de dendrogramas recorreu-se a três métodos de geração aleatória de dendrogramas propostos na literatura:

- método uniforme (Sousa, 2000)
- método RA (Podani, 2000)
- método da Permutação Dupla (Lapointe & Legendre, 1991).

Estes três métodos são uniformes no sentido de Furnas (1984), isto é, para um valor de  $m$  fixo, os métodos geram todos os dendrogramas com  $m$  nós terminais de forma equiprovável (com probabilidade  $1/d(m)$ ). Os valores de  $m$  contemplados no estudo foram: 4(1)15, 20(5)50, 75, 100(100)500. A metodologia seguida pode resumir-se em três passos:

1. gerar um par de dendrogramas
2. calcular o valor do coeficiente de comparação
3. repetir 1. e 2.  $k$  vezes.

No nosso caso fez-se  $k = 1000$ .

Sendo os três métodos de geração aleatória de dendrogramas todos uniformes no sentido de Furnas, os resultados obtidos por esta metodologia devem ser semelhantes, independentemente do método de geração aleatória que se use no passo 1.. Contudo optou-se por utilizá-los todos, meramente a título comprovativo.

Os algoritmos foram implementados em linguagem Fortran. O gerador de números pseudoaleatórios baseia-se na subrotina “random\_number”; a semente dos números aleatórios foi sempre gerada automaticamente pelo processador de acordo com o relógio do mesmo.

## 5 Discussão dos resultados

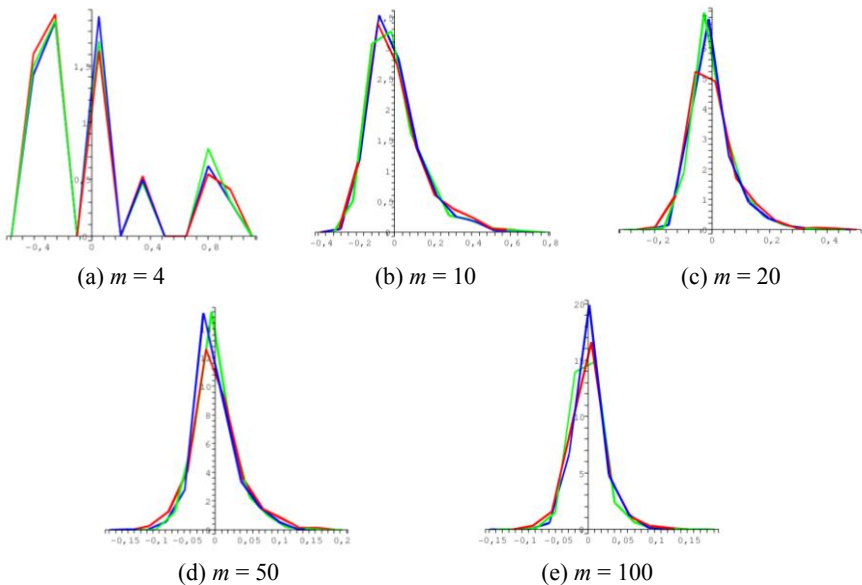
A apresentação e discussão dos resultados são, por razões de limitação de número de páginas, feitas apenas para o coeficiente Tau de Kendall. Refira-se que os

resultados encontrados para os outros coeficientes (podem ser consultados em Tendeiro, 2005) são análogos. A Figura 7 (a)-(e) apresenta, para diferentes valores de  $m$ , os polígonos de frequências dos valores do coeficiente Tau de Kendall correspondentes aos três métodos de geração. De facto, verifica-se que os polígonos são bastante semelhantes entre si, como seria de esperar.

A escolha do método de geração de dendrogramas é irrelevante. Assim, sem prejuízo das conclusões que nos propomos tirar, vamos usar a distribuição empírica obtida através do método uniforme.

Uma primeira inspecção aos resultados obtidos permite retirar algumas conclusões:

- os valores da amplitude amostral, bem como da dispersão quartal, são cada vez menores conforme  $m$  aumenta; este facto está relacionado com o rápido crescimento de  $d(m)$ ;
- para cada valor de  $m$  verifica-se que a mediana é quase sempre negativa, aproximando-se de zero conforme  $m$  aumenta;
- as distribuições são assimétricas positivas (os gráficos apoiam este facto), sendo que a assimetria vai diminuindo conforme  $m$  aumenta.



**Figura 7** – Polígonos de frequências para o coeficiente Tau de Kendall.

A comparação entre as distribuições empíricas obtidas e a distribuição assintótica é pertinente. No resultado da distribuição assintótica do coeficiente de Kendall, Secção 4, são dadas as expressões para as média e variância de C-D.

Sendo C-D uma variável cuja variação de valores depende do comprimento da preordenação e dos seus patamares de empates, a interpretação dos seus valores é mais delicada e fica inviabilizada uma comparação gráfica. No que respeita à média tem-se que  $E(C-D)=0$  e, consequentemente,  $E(T_K)=0$ . Na Tabela 3 apresentam-se, para os valores de  $m$  considerados, os valores empíricos das médias de C-D e de  $T_K$ . No que respeita à variância apresentam-se, para a variável C-D, os valores para a distribuição assintótica ( $V_t$ ), para a distribuição empírica ( $V_a$ ), bem como uma comparação relativa desses valores.

**Tabela 3** – Análise comparativa da média e dispersão das distribuições empírica e assintótica.

$m$	Média amostral		Variância de C-D		
	C-D	$T_K$	Amostral ( $V_a$ )	Teórica $V_t$	$\frac{V_a - V_t}{V_t}$
4	.2400	$.2427 \times 10^{-1}$	$.1870 \times 10^2$	$.1797 \times 10^2$	.0406
5	-.5500	$-.1702 \times 10^{-2}$	$.6881 \times 10^2$	$.8674 \times 10^2$	-.2068
6	-.1100	$-.6648 \times 10^{-3}$	$.3382 \times 10^3$	$.2960 \times 10^3$	.1426
7	-.1190x10	$-.5755 \times 10^{-2}$	$.9325 \times 10^3$	$.8284 \times 10^3$	.1256
8	-.5520x10	$-.2052 \times 10^{-1}$	$.2478 \times 10^4$	$.1989 \times 10^4$	.2459
9	$-.1631 \times 10^2$	$-.3397 \times 10^{-1}$	$.3960 \times 10^4$	$.4240 \times 10^4$	-.0658
10	$.1702 \times 10^2$	$.2278 \times 10^{-1}$	$.1217 \times 10^4$	$.8252 \times 10^4$	.4749
11	-.9140x10	$-.8627 \times 10^{-2}$	$1902 \times 10^5$	$.1494 \times 10^5$	.2728
12	.2600x10	$.1702 \times 10^{-2}$	$.2728 \times 10^5$	$.2590 \times 10^5$	.0533
13	$.2225 \times 10^2$	$.1028 \times 10^{-1}$	$.5474 \times 10^5$	$.4293 \times 10^5$	.2753
14	$.5771 \times 10^2$	$.1835 \times 10^{-1}$	$.1193 \times 10^6$	$.6886 \times 10^5$	.7330
15	$-.6442 \times 10^2$	$-.1550 \times 10^{-1}$	$.1342 \times 10^6$	$.1058 \times 10^6$	.2696
20	$-.1033 \times 10^3$	$-.7201 \times 10^{-2}$	$.1002 \times 10^7$	$.6273 \times 10^6$	.5976
25	$.1993 \times 10^3$	$.5800 \times 10^{-2}$	$.4339 \times 10^7$	$.2497 \times 10^7$	.7379
30	$-.6190 \times 10^2$	$-.9200 \times 10^{-3}$	$.1317 \times 10^8$	$.7593 \times 10^7$	.7350
35	$-.8816 \times 10^3$	$-.6083 \times 10^{-2}$	$.4605 \times 10^8$	$.1975 \times 10^8$	.1332x10
40	$-.1565 \times 10^3$	$-.5496 \times 10^{-3}$	$.1385 \times 10^9$	$.4421 \times 10^8$	.2132x10
45	$-.3524 \times 10^4$	$-.9068 \times 10^{-2}$	$.2972 \times 10^9$	$.9142 \times 10^8$	.2251x10
50	$-.1349 \times 10^3$	$-.2047 \times 10^{-3}$	$.7178 \times 10^9$	$.1709 \times 10^9$	.3201x10
75	$.2075 \times 10^5$	$.6839 \times 10^{-2}$	$.1152 \times 10^{11}$	$.1990 \times 10^{10}$	.4789x10
100	$.1010 \times 10^5$	$.9444 \times 10^{-3}$	$.7153 \times 10^{11}$	$.1142 \times 10^{11}$	.5263x10
200	$.2968 \times 10^6$	$.1846 \times 10^{-2}$	$.5946 \times 10^{13}$	$.7405 \times 10^{12}$	.7030x10
300	$.1897 \times 10^7$	$.2406 \times 10^{-2}$	$.9577 \times 10^{14}$	$.8475 \times 10^{13}$	.1030x10 <sup>2</sup>
400	$-.3218 \times 10^6$	$-.8538 \times 10^{-3}$	$.8677 \times 10^{15}$	$.4800 \times 10^{14}$	.1708x10 <sup>2</sup>
500	$.4264 \times 10^7$	$.6535 \times 10^{-3}$	$.4229 \times 10^{16}$	$.1803 \times 10^{15}$	.2246x10 <sup>2</sup>

De notar que o valor da variância teórica, dado por (6), é função do comprimento dos patamares da preordenação, não sendo, por isso, constante para cada valor de  $m$ . Os valores da variância teórica (coluna  $V_t$  na Tabela 3) foram calculados de acordo com o seguinte: (i) para cada par de dendrogramas gerado determinou-se o comprimento dos respectivos patamares e o correspondente valor de variância teórica (fórmula (6)) (ii) para cada valor de  $m$  calculou-se a média dos valores de variância teórica obtidos.

Os valores da média amostral de  $T_k$  oscilam em torno de zero, valor da média teórica, registando-se uma tendência para um decréscimo dos valores em módulo à medida que  $m$  cresce. No que respeita à dispersão verifica-se que os valores empíricos são quase sempre superiores aos valores teóricos. Tratando-se da variável C-D, os valores da variância, amostrais e teóricos, são muito elevados. Para uma análise comparativa consideram-se os valores da diferença relativa das variâncias. Constata-se que a diferença relativa aumenta com  $m$ , assumindo valores inferiores a 1 quando  $m \leq 30$  e passando a valores superiores a 1 quando  $m \geq 35$ . Estes factos parecem permitir concluir que a utilização da distribuição assintótica, no contexto da comparação de estruturas de classificação, é desadequada para qualquer valor de  $m$ .

## 6 Conclusão

Neste trabalho é apresentada uma metodologia para comparar quantitativamente pares de estruturas de classificação, que passa por associar a cada estrutura de classificação uma preordenação sobre o conjunto de elementos a classificar. São apresentados coeficientes de correlação (associação) ordinal para a comparação de pares de preordenações e discutida a adequação das suas distribuições teóricas ao problema em estudo, já que os pressupostos de independência, habitualmente aceites, não são aqui verificados.

Propõe-se a dedução de distribuições empíricas obtidas por simulação. A metodologia proposta é ilustrada com o caso em que as estruturas a comparar são dendrogramas, obtidos a partir do mesmo conjunto de dados. Para a dedução das distribuições empíricas recorreu-se a métodos de geração aleatória de dendrogramas. Os resultados obtidos parecem confirmar que os problemas teóricos observados na Secção 4 limitam, de facto, a utilização das distribuições assintóticas dos coeficientes de comparação no contexto da comparação de dendrogramas. São apresentados e discutidos resultados para o coeficiente Tau de Kendall.

Várias aplicações, como intervalos de confiança ou testes de hipóteses, podem ser levadas a cabo a partir do conhecimento das distribuições empíricas dos coeficientes aqui estudados.

## Referências

- BOCK, H. H. (1996). Probability Models and Hypotheses Testing in Partitioning Cluster Analysis, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, 377-453.
- FURNAS, G. W. (1984). The Generation of Random, Binary Unordered Trees, *Journal of Classification*, 1, 187-233.
- GORDON, A. D. (1996). Hierarchical Classification, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, 65-121.
- GORDON, A. D. (1999). *Classification*, 2nd Edition, Chapman & Hall, London.
- HALKIDI, M.; SBATISTAKIS, Y. e VAZIRGIANNIS, M. (2001). On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, 17: 2/3, 107-145.
- KENDALL, M. G. (1970). *Rank Correlation Methods*, 4ª Edição, Griffin, London.
- LAPORTE, F. J. e LEGENDRE, P. (1991). The Generation of Random Ultrametric Matrices Representing Dendrograms, *Journal of Classification*, 8, 177-200.
- NICOLAU, F. C. (1984). Problemas de Validade em Classificação Automática, *Actas do III Colóquio de Estatística e Investigação Operacional*, SPEIO, Lagos.
- PODANI, J. (2000). Simulation of Random Dendrograms and Comparison Tests: Some Comments, *Journal of Classification*, 17, 123-142.
- SOUSA, F. (2000). *Novas Metodologias em Classificação Hierárquica Ascendente*, Dissertação de Doutorado, Universidade Nova de Lisboa, Lisboa.
- SOUSA, F. e NICOLAU, F. C. (2001). Uma Abordagem ao Problema da Comparação de Estruturas Classificatórias, em *A Estatística em Movimento* (M. M. Neves, J. Cadima, M. J. Martins, F. Rosado), Sociedade Portuguesa de Estatística, 409-418.
- TENDEIRO, J. (2005). *Comparação de Dendrogramas: Obtenção de Distribuições Empíricas de Alguns Coeficientes*, Dissertação de Mestrado, Universidade do Porto, Porto.

# Application of Clustering Methods for Optimizing the Location of Treated Wood Remediation Units

Helena Gomes<sup>1</sup> · Victor Lobo<sup>1,2</sup> · Alexandra B. Ribeiro<sup>3</sup>

© The Author(s) 2013

**Resumo** Foram utilizados dois métodos de análise de *clusters* (SOM e *K-means*) para otimizar a localização de unidades de remediação de resíduos de madeira preservada com CCA (Crómio-Cobre-Arsenato) em Portugal. O software utilizado foi o SOM Toolbox para Matlab com dados dos Censos 2001, sobre a população residente segundo a freguesia, fornecidos pelo Instituto Nacional de Estatística. Embora os dois métodos utilizados tenham fornecido bons resultados, concluiu-se que, para este problema em particular, o K-means permitiu obter melhores soluções.”

**Abstract** In this paper we used two clustering methods (SOM and K-means) to optimize the location of CCA (Chromated Copper Arsenate)-treated wood waste remediation units in Portugal. We used the SOM Toolbox for Matlab implementation with the population data from the last census made by the National Statistics Institute. The two methods yield good results, but we concluded that the K-means algorithm provided better solutions than the SOM in this particular problem.

**Keywords:** Self-Organizing Map (SOM), k-means, p-median problem, CCA-treated wood waste

---

<sup>1</sup>Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, hrg@fct.unl.pt

<sup>2</sup>Escola Naval (Portuguese Naval Academy), vlobo@isegi.unl.pt

<sup>3</sup>Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, abr@fct.unl.pt



## 1 Introduction

In order to increase the service life of wood and to provide long-term resistance to attack by fungi, insects and marine organisms, wood is submitted to preservation techniques that generally apply wood preservatives using pressure. The wood preservative CCA (Chromated Copper Arsenate) is actually the most used wood preservative worldwide.

Examples of treated wood products are telephone and utility poles, railway ties, fence posts and lumbers for outdoor construction (house fronts, roof tiles, window frames, playing tools, garden houses, fences, landings, wharves, bridges, bank revetment, acoustic barriers, etc).

There is growing concern about the environmental impacts and increasing difficulty to dispose preservative treated wood products at the end of their service life. In the next decades a significant increase is expected in the amounts of treated wood removed from service. For example, in the USA the total volume of treated wood removed from service was estimated to be near 8 million m<sup>3</sup> in 1990, 9 million m<sup>3</sup> in 2000, 15 million m<sup>3</sup> in 2010 and about 18 million m<sup>3</sup> in 2020 (McQueen *et al.*, 1998).

Consumer lumber makes up the majority of current and future expected production of CCA treated wood – about 56% of 1995 production compared to about 82% of expected waste wood generated in 2010 (Cooper, 2003).

The future of the wood preservation industry depends on the solution(s) found to manage the treated wood waste problem. The recycling of these wastes, containing chromium, copper and arsenic (in the case of CCA-treated wood), can only be made after its remediation, so that planning and optimization of remediation units and its locations is of major importance.

The objective of this study is the application of clustering methods to optimize the location of remediation plants for the treatment of preserved wood waste for further recycling, minimizing costs and respecting environmental criteria. Essentially, this is a p-median problem, a classic optimization problem that involves the location of facilities in such a manner that the total weighted distance of all users to their closest facility is minimized. The p-median problem has been proved to be NP-complete (Megiddo & Supowit, 1984; Plastria, 2002), and thus a globally optimal solution is very hard to obtain. Different heuristics have been used to solve the p-median problem, such as k-means and k-medoids (Estivill-Castro & Houle, 2001); genetic algorithms (Correa *et al.*, 2001); local search heuristics (Lorena & Sene, 2002); tabu search and variable neighbourhood search (Crainic *et al.*, 2001; Mladenović *et al.*, 2003); simulated annealing (Church & Sorensen, 1994); Greedy Randomised Adaptive Search Procedure (GRASP) (Resende & Werneck, 2002); Branch-and-Cut algorithm (Avella *et al.*, 2003) and Lagrangean Relaxation (Senne & Lorena, 2000).

In this paper we will solve our particular instance of the p-median problem with a Self-Organizing Map (Kohonen, 2001), and k-means algorithm (MacQueen, 1967).

## 2 Dataset used

P-median problems are usually formalized by considering a number of demand nodes, each with a given location and weight (amount of resourced demanded), a number of facility locations that must be found, and a cost function that takes into account the relation between the demand nodes and the facility locations. In our case, the demand nodes will be the producers of waste wood, the facilities will be the remediation plants, and in a first approximation, the cost function will be the Euclidean distance between the demand nodes and the facilities, weighed by the amount of wood generated at each demand node.

In this paper we assume that all wood that needs to be sent to the remediation plants is generated by municipal solid waste, which in turn is proportional to the number of inhabitants in each area. Thus we use population instead of amount of wood needing remediation. A GIS (Geographic Information System) based location model was used to find the population of each “freguesia” of Portugal (NUTS 5). The data contained in that GIS comes from the 2001 census data, provided by the Portuguese Statistics Institute (INE), and consists of the geographical location of the centroid of the “freguesia” and its population.

The implementations of the clustering techniques that we used did not allow weighing of different data patterns. To achieve this weighing we generated one data pattern for each 1000 inhabitants of each “freguesia” (using excess rounding). We were then left with 99.146 data patterns. These data patterns were then put in random order so as not to introduce unnecessary bias.

In all our tests we assumed that 6 remediation plants would be necessary.

## 3 Proposed solutions

Two kinds of clustering methods were used: the Self-Organizing Map (SOM) and k-means algorithm. In our experiments we used the SOM Toolbox for Matlab implementation developed at Helsinki University of Technology (Vesanto *et al.*, 2000).

The SOM provides a visual representation of a vector quantification algorithm that places a number of vectors into a high-dimensional input data space in such a way that they approximate the original data patterns in an ordered fashion (Kohonen *et al.*, 1995). It’s an unsupervised neural network that tries to preserve topological relations between points in the input and output spaces. Generally,

SOMs have many neurons and these are spread in the input space proportionally to the density of data points. When we use very few neurons, as is the case in our experiments, the strong coupling between neighboring neurons will prevail over the minimization of the distances to the data patterns, thus introducing some distortions to our original problem.

The k-means algorithm (MacQueen, 1967) is one of the classic statistical clustering methods. It generates a specific number of disjoint, flat (non-hierarchical) clusters. The k-means method is numerical, unsupervised, non-deterministic and iterative, and is basically a stochastic hill-climbing optimization technique.

One of the main drawbacks in using k-means is that it is quite sensitive to local minima, and to a certain degree we verified that in our tests. There are several ways of dealing with this problem, the most common of which is to re-initialize the k-means algorithm several times with different seeds, and then choose the best solution.

## 4 Experimental results

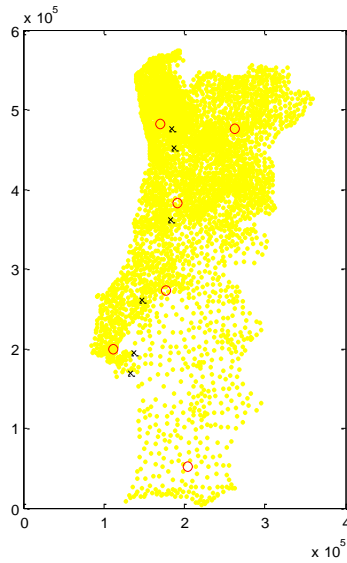
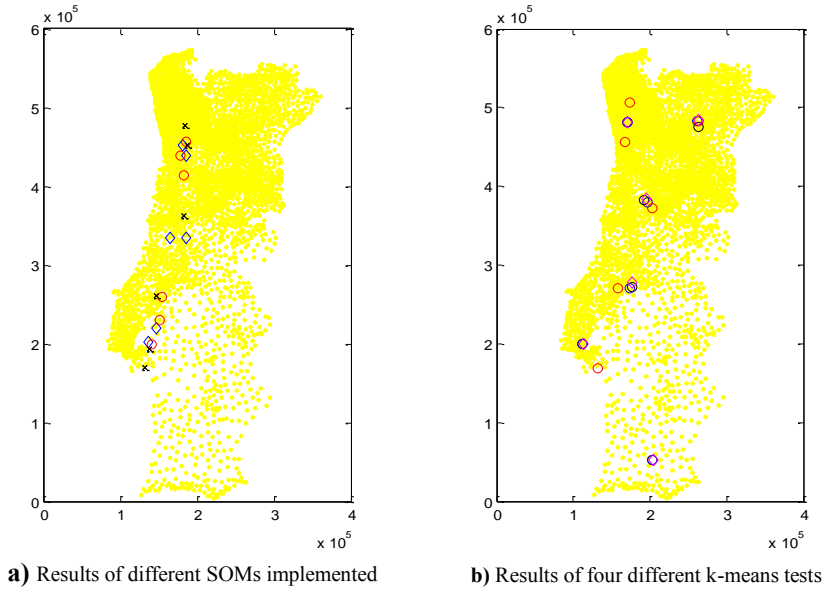
When using SOM, we have to provide an output space grid upon which the neurons are set. In our experiments we used two different grids, a 2x3 and a 1x6 grid. For the 2x3 grid we used both a hexagonal and a rectangular topology. In all three experiments, we used random initialization, and two learning phases. In the first phase we used an initial radius of 3 (linearly decreasing to 0), a learning rate of 0.3 (also decreasing linearly to 0), a sequential training algorithm, 3 epochs of training and bubble neighborhood. The second learning phase differed from the first in the initial radius (we used an initial radius of 2), in the learning rate (0.05) and in the number of training epochs (6).

For the k-means algorithm, we defined six centroids and 2400 training epochs, using a batch algorithm. We did 4 different experiments with this method, using the same parameters.

The results are presented in Figure 1.

The average distance to a remediation plant in each of the solutions is given in Table 1.

From these results we can see that the SOM introduces a tight coupling amongst the locations of the plants, forcing them to be close to each other, and thus increasing the total cost of the solution. It is also clear that the k-means algorithm is reasonably robust to initial conditions (only one solution is considerably different to the other three), and yields good solutions to the problem.



**c) Final map with SOM results (x) and k-means results (o).**

**Figure 1** – Results of the application of clustering algorithms.

**Table 1** - Average distance (in km) to a remediation plant

K-means (1st run)	32.779	SOM (2x3 hex.)	50.724
K-means (2nd run)	32.778	SOM (2x3 rect.)	51.576
K-means (3rd run)	41.510	SOM (6x1)	44.730
K-means (4th run)	32.526		

## 5 Conclusions

In this paper we show that the choice of the best locations for wood remediation plants can be found using the k-means algorithm, as well as the SOM. The k-means algorithm provided better solutions than the SOM in this particular problem. This can be explained by the fact that SOMs are better suited to problems where the number of units to be placed is large, or when there is a need to model a forced proximity between these units. As this is not the case, we suggest that k-means be used to solve the wood remediation plant location problem. The solutions obtained with our data make sense and could be used to decide on the location of these plants.

The approach used has great potential because it can easily be adapted to a more complex (and realistic) formulation of our problem. On one hand, additional sources of preserved wood waste can be added. Each source of preserved wood (population, industrial facilities, agriculture, etc) can be described as pattern composed of geographical coordinates and a quantity of treated wood that acts as weight. On the other hand, both algorithms used can use distances derived from a distance matrix with actual road distances between possible locations, thus turning the problem more realistic.

Further developments of this work include: i) the identification of the key environmental and economic factors to be considered in the optimization model; ii) the process of formulation and combination of these criteria; iii) analysis and comparison of heuristics in the p-median problem resolution; iv) production of a final map with the visualization of areas suitable for the location of remediation infrastructures.

## References

- AVELLA, P.; SASSANO, A. and VASILIEV, I. (2003). *Computational Study of Large-scale p-Median Problems*. Technical Report 08-03, DIS - Universita di Roma "La Sapienza", Rome, Italy.

- CHURCH, R.L. and SORENSEN, P. (1994). *Integrating Normative Location Models into GIS: Problem and Prospects with p-median Problem*. Technical Report 94-5, National Center for Geographic Information and Analysis. University of California, Santa Barbara, USA.
- COOPER, P.A. (2003). A Review of Issues and Technical Options for Managing Spent CCA Treated Wood. *Proceedings of the American Wood Preservation Association 99*, Granbury, TX, USA, 1-23.
- CORREA, E.S.; STEINER, M.T.A.; FREITAS, A.A. and CARNIERI, C. (2001). A Genetic Algorithm for the P-Median Problem. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, San Francisco, USA, 1268-1275.
- CRAINIC, T. G.; GENDREAU, M.; HANSEN, P.; HOEB, N. and MLADENOVIC N. (2001). Parallel Variable Neighbourhood Search for the p-Median. *Proceedings of the 4th Metaheuristics International Conference (MIC'2001)*, Porto, Portugal, 595-599.
- ESTIVILL-CASTRO, V. and Houle, M. E. (2001). Robust Distance-based Clustering with Applications to Spatial Data Mining. *Algorithmica*, 30 (2), 216-242.
- KOHONEN, T. (2001). *Self-Organizing Maps*. Springer, 3<sup>rd</sup> Ed., Berlin.
- KOHONEN, T.; HYNINEN, J.; KANGAS, J. and LAAKSONEN, J. (1996). *SOM\_PAK. The Self-Organizing Map Program Package*. SOM Programming Team of the Helsinki University of Technology. Laboratory of Computer and Information Science. Report A31.
- LORENA, L. A. N. and SENE, E. L. F. (2003). Local Search Heuristics for Capacitated p-Median Problems. *Networks and Spatial Economics*, 3 (4), 407-419.
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observation. *5th Berkeley Symposium on Mathematical Statistics and Probability, Vol.1*, University of California Press, 281-297.
- MCQUEEN, J., STEVENS, J. and KAMDEM, D. P. (1998). Recycling of CCA Treated Wood in the US. *Proceedings of the 4th International Symposium on Wood Preservation*, Cannes - Mandelieu, France, IRG Secretariat, Stockholm, Sweden, 78-93.
- MEGIDDO, N. and SUPOWIT, K (1984). On the complexity of some common geometric location problems. *SIAM Journal on Computing*, 18, 182-196.
- MLADENOVIC, N. *et al.* (2003). Solving the p-Center Problem with Tabu Search and Variable Neighborhood Search, *NETWORKS*, 42, (1) 48-64.
- PLASTRIA, F. (2002). Continuous Covering Location Problems. *Facility Location - Applications and Theory*. Z. Drezner & H. W. Hamacher, Springer-Verlag, Heidelberg.

- RESENDE, M. G. C. and WERNECK, R. F. (2002). *A GRASP with Path-relinking for the P-Median Problem*, Technical Report TD5E4QKA, AT&T Labs Research.22.
- SENNE, E. L. F. and LORENA, L. A. N. (2000). Lagrangean/surrogate heuristics for p-median problems, *Computing Tools for Modeling, Optimization and Simulation: Interfaces in Computer Science and Operations Research*. M. Laguna and J. L. Gonzalez-Velarde (eds), Kluwer Academic Publishers, 115-130.
- VESANTO, J. *et al.* (2000). *SOM Toolbox for Matlab 5*. Helsinki University of Technology.

# Modelos de Classe Latente de Regressão para Dados de Contagem. A Análise da Heterogeneidade Regional nas Eleições Presidenciais Portuguesas de 2006

Francisco Figueiredo<sup>1</sup>, José G. Dias<sup>2</sup>

© The Author(s) 2013

**Resumo** Os modelos de mistura de regressão têm vindo a ganhar expressão por permitirem em simultâneo o agrupamento das observações em diferentes grupos ou *clusters* e a estimação de um modelo de regressão para cada um dos grupos. Neste artigo procede-se à revisão do modelo de mistura de regressão Poisson utilizado na modelação de dados de contagem (valores inteiros não negativos). A sua utilização é ilustrada com uma aplicação, na área da ciência política, em que se analisam os resultados das Eleições Presidenciais de 2006 em Portugal Continental e mais particularmente a incidência de votos por concelho no candidato Manuel Alegre.

**Palavras-chave:** algoritmo EM, ciência política, modelos de mistura, regressão de Poisson.

## 1 Introdução

Os modelos de regressão de Poisson são bastante utilizados em Economia e Epidemiologia na modelação de dados de contagem (valores inteiros não-negativos), correspondentes ao número de acontecimentos ocorridos numa dada unidade espaço-temporal ou à taxa de incidência de determinado acontecimento. O objectivo deste tipo de modelos é a estimação de um modelo linear generalizado (GLM) relacionando a variável dependente com um conjunto de variáveis explicativas.

---

<sup>1</sup> ISCTE, Instituto Universitário de Lisboa, BRU-IUL, francisco.m.fig@gmail.com

<sup>2</sup> ISCTE, Instituto Universitário de Lisboa, BRU-IUL, jose.dias@iscte.pt



Frequentemente verifica-se que as observações provêm de um número (desconhecido) de grupos heterogêneos na população, tendo originado o desenvolvimento de modelos que incorporam heterogeneidade não observada, como por exemplo, o modelo de mistura de regressão de Poisson (Wedel *et al.*, 1993). Estes modelos têm vindo a ganhar expressão por permitirem, em simultâneo à estimação de um modelo de regressão para cada um dos grupos, a segmentação das observações em diferentes grupos ou *clusters*. Uma das suas possíveis aplicações é na área da ciência política e mais particularmente em estudos sobre o comportamento eleitoral.

Neste contexto, a presente investigação pretende analisar os resultados das Eleições Presidenciais de 2006 em Portugal Continental e mais particularmente o número de votos por concelho do candidato Manuel Alegre, recorrendo a um modelo de mistura de regressão de Poisson. Este estudo pretende explicar o número de votos no candidato Manuel Alegre, com base em variáveis demográficas e socio-económicas, identificar diferentes segmentos de concelhos administrativos relativamente à relação existente entre as variáveis explicativas e a incidência de votos nesse candidato e caracterizar os segmentos/grupos de concelhos identificados.

## **2 Definição do modelo**

O comportamento eleitoral tem sido teorizado sob múltiplas vertentes (Freire, 2001). Por exemplo, o modelo sociológico do voto incidia sobre as determinações estruturais do comportamento eleitoral (idade, nível de instrução, etc.). Por sua vez, os novos modelos incidem sobre o modelo económico do voto e baseiam-se na teoria da existência de novas clivagens, ou seja, diferenças apoiadas em valores e com uma base social menos definida que nas velhas clivagens. Contudo, “...quer no caso das velhas clivagens, quer no caso das novas clivagens, há uma conexão entre posições estruturais, sistemas de valores e voto (Freire, 2001)”. O modelo definido para este estudo tem como sustentação os modelos eleitorais apresentados, reflectindo as determinações estruturais, mas também as novas clivagens existentes.

## **3 Dados utilizados**

Todos os dados utilizados são provenientes de fontes secundárias. A variável dependente (o número de votos por concelho no candidato Manuel Alegre nas Presidenciais de 2006) foi recolhida *online* no *site* do STAPE (STAPE, 2006a).

Relativamente às variáveis explicativas, a percentagem de votos por concelho no Partido Socialista nas eleições legislativas de 2005 teve igualmente origem nos

dados do STAPE (STAPE, 2006b). Os indicadores da taxa de analfabetismo<sup>3</sup> e índice de envelhecimento<sup>4</sup> têm como fonte o Instituto Nacional de Estatística (INE, 2006) e resultam dos dados do Censos 2001, enquanto que o índice de desenvolvimento teve por base os *Índices de Desenvolvimento Concelhio* (Fonseca, 2002). A totalidade dos dados encontra-se desagregado ao nível concelhio.

Nenhuma das variáveis utilizadas continha dados omissos, com exceção do índice de desenvolvimento, uma vez que na altura em que foram desenvolvidos os *Índices de Desenvolvimento Concelhio* (1998), os concelhos de Odivelas, Trofa, Vizela ainda não existiam. Procedeu-se assim à sua imputação, utilizando como critério a substituição pelo índice de desenvolvimento associado ao(s) concelho(s) que lhes deram origem<sup>5</sup>.

## 4 Metodologia utilizada: Modelos de mistura de regressão de Poisson

### 4.1 Especificação do modelo

Para analisar o impacto das variáveis explicativas na incidência de voto especifica-se um modelo de regressão de Poisson. Assume-se que a amostra tem origem numa população com um número finito ( $S$ ) de grupos/segmentos de proporções desconhecidas  $(\pi_1, \dots, \pi_S)$ , não se sabendo à partida qual o segmento que deu origem a determinada observação (Dias, 2003). Essas proporções,  $\pi_s$ , correspondem à probabilidade *a priori* de uma observação pertencer ao segmento  $s$ , com:

$$\sum_{s=1}^S \pi_s = 1, \quad \pi_s > 0, \quad s = 1, \dots, S. \quad (1)$$

Em cada segmento, isto é, dado que pertence ao segmento  $s$ , cada observação é caracterizada pela seguinte função massa de probabilidade:

$$f_s(y_i; \lambda_{is}, n_i) = \frac{\exp(-\lambda_{is} n_i) (\lambda_{is} n_i)^{y_i}}{y_i!},$$

<sup>3</sup> Taxa de Analfabetismo = (População com 10 ou mais anos que não sabe ler nem escrever / População com 10 ou mais anos) \* 100.

<sup>4</sup> Índice de Envelhecimento = (População com 65 ou mais anos / população com idades entre os 0 e os 14 anos] \* 100.

<sup>5</sup> Deste modo, ao concelho de Odivelas associou-se um índice de desenvolvimento idêntico a Loures, ao concelho de Trofa um índice idêntico a Santo Tirso e a Vizela a média dos índices de desenvolvimento de Felgueiras, Guimarães e Lousada.

em que  $n_i$  e  $\lambda_{is}$  correspondem respectivamente à exposição da observação  $i$  e à taxa de incidência no segmento  $s$  e  $y_i$  à frequência observada.

Considerando uma mistura de duas ou mais distribuições de Poisson (Wedel *et al.*, 1993), o parâmetro  $\lambda_{is}$  é parametrizado como uma função das variáveis explicativas,  $\ln \lambda_{is} = \beta_{0s} + \sum_{l=1}^L x_{il} \beta_{ls}$ , em que  $\beta_{ls}$  corresponde ao impacto da variável explicativa  $l$  na taxa de incidência no segmento  $s$ ,  $\beta_{0s}$  é a constante no segmento  $s$ , e  $\ln(\cdot)$  é a função de ligação (*log-link*).  $\varphi = (\boldsymbol{\pi}, \boldsymbol{\beta})$  é o conjunto de todos os parâmetros do modelo com  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{S-1})$  e  $\boldsymbol{\beta} = (\beta_{01}, \dots, \beta_{LS})$ .

## 4.2 Estimação do modelo

A estimação dos parâmetros  $\varphi = (\boldsymbol{\pi}, \boldsymbol{\beta})$ , dados  $y_i$ ,  $\mathbf{x}_i$ ,  $n_i$  e considerando o valor de  $S$  fixo, pelo método de máxima verosimilhança, resulta da maximização da função de verosimilhança:

$$L = \prod_{i=1}^n \sum_{s=1}^S \pi_s \frac{\exp(-\lambda_{is} n_i) (\lambda_{is} n_i)^{y_i}}{y_i!}.$$

O algoritmo EM (Dempster *et al.*, 1977) permite encontrar as estimativas de máxima verosimilhança através de um processo iterativo, que se pode resumir em duas grandes etapas: etapa E (*Expectation step*) e etapa M (*Maximization step*). O algoritmo EM funciona num espaço aumentado em que se introduz uma variável latente,  $z_{is}$ , que indica se um indivíduo  $i$  pertence ao segmento latente  $s$ , assumindo assim o valor 1 caso o indivíduo  $i$  pertença ao segmento  $s$  e o valor 0, caso contrário. Admite-se que os  $\mathbf{z}_i = (z_{i1}, \dots, z_{iS})$  são independentes e identicamente distribuídos (i.i.d.) e seguem uma distribuição multinomial  $(n, \tilde{\boldsymbol{\pi}})$  com probabilidade  $\tilde{\boldsymbol{\pi}} = (\pi_1, \dots, \pi_S)$ .

A função de log-verosimilhança para os dados completos é então:

$$\ln L_c = \sum_{i=1}^n \sum_{s=1}^S z_{is} \ln f_s(y_i; \mathbf{x}_i, \boldsymbol{\beta}_s, n_i) + \sum_{i=1}^n \sum_{s=1}^S z_{is} \ln \pi_s. \quad (2)$$

A maximização da função anterior passa pela utilização do algoritmo EM, cujo funcionamento se explica de seguida.

A etapa E consiste em calcular  $E(\ln L_c)$ , o valor esperado da função  $\ln L_c$ , em ordem à distribuição da variável não observada,  $Z_{is}$ , condicionadas às variáveis observadas  $y_i$ ,  $\mathbf{x}_i$  e  $n_i$  e às estimativas dos parâmetros  $\varphi$ . O valor de  $E(\ln L_c)$  obtém-se substituindo os  $z_{is}$  da expressão (2) pelo seu respectivo valor esperado:

$$E(Z_{is} | y_i, \mathbf{x}_i, \pi, \boldsymbol{\beta}_s, n_i) = P(Z_{is} = 1) = \frac{\pi_s f_s(y_i; \mathbf{x}_i, \boldsymbol{\beta}_s, n_i)}{\sum_{r=1}^S \pi_r f_r(y_i; \mathbf{x}_i, \boldsymbol{\beta}_r, n_i)}, \quad (3)$$

que corresponde às probabilidades *a posteriori* que são designadas por  $\alpha_{is}$ , isto é, à probabilidade de pertença do indivíduo  $i$  ao segmento  $s$ .

Deste modo, em resumo, a etapa E do algoritmo EM consiste em substituir os valores não observados,  $z_{is}$ , da expressão (2), pelos valores esperados estimados correspondentes,  $\hat{\alpha}_{is}$ :

$$E(\ln L_c) = \sum_{i=1}^n \sum_{s=1}^S \hat{\alpha}_{is} \ln f_s(y_i; \mathbf{x}_i, \boldsymbol{\beta}_s, n_i) + \sum_{i=1}^n \sum_{s=1}^S \hat{\alpha}_{is} \ln \pi_s. \quad (4)$$

Por sua vez, a etapa M do algoritmo consiste em maximizar a expressão anterior (4), em ordem aos parâmetros  $\varphi$ . Assumindo que, nesta fase, as probabilidades *a posteriori*,  $\hat{\alpha}_{is}$ , são fixas, procede-se à maximização em separado da expressão (4) em ordem a  $\pi_s$  e  $\beta_{ls}$ . Maximizar em ordem a  $\pi_s$ , sujeita às restrições enunciadas em (1), corresponde a maximizar a seguinte função Lagrangiana:

$$\sum_{i=1}^n \sum_{s=1}^S \hat{\alpha}_{is} \ln \pi_s - \mu \left( \sum_{s=1}^S \pi_s - 1 \right), \quad (5)$$

em que  $\mu$  corresponde ao Multiplicador de Lagrange.

Por outro lado, maximizar a mesma expressão em ordem a  $\beta_{ls}$  é equivalente a maximizar separadamente cada uma das  $S$  expressões:

$$l_s = \sum_{i=1}^n \hat{\alpha}_{is} \ln f_s(y_i; \mathbf{x}_i, \beta_s, n_i). \quad (6)$$

Assim, na etapa M,  $l_s$  é maximizado com recurso ao algoritmo de Newton-Raphson de modo a obter as estimativas de máxima verosimilhança de  $\beta_{ls}$ <sup>6</sup>. O processo iterativo termina quando as alterações nos valores dos parâmetros estimados ( $\hat{\varphi}$ ) são suficientemente pequenas de uma iteração para a seguinte.

Resumindo, o algoritmo EM procede do seguinte modo:

1. *Inicialização*: Na primeira iteração,  $t = 0$ , são gerados aleatoriamente os valores iniciais dos parâmetros  $\varphi^{(0)}$ ;
2. *Etapa E*: Cálculo das probabilidades *a posteriori*,  $\alpha_{is}^{(t+1)}$ , de acordo com a expressão (3);
3. *Etapa M*: Condicional nos valores de  $\alpha_{is}^{(t+1)}$ , obtêm-se  $\varphi^{(t+1)}$ , através da utilização das equações (5) e (6);
4. *Teste de convergência*: caso o valor de  $\|\hat{\varphi}^{(t+1)} - \hat{\varphi}^{(t)}\|$  seja suficientemente pequeno, o algoritmo termina; caso contrário, incrementar  $t$  e voltar à etapa 2.

### 4.3 Número de segmentos na mistura

Para proceder à escolha do número de segmentos adequado, usualmente recorre-se à utilização de critérios/estatísticas da teoria de informação, cujo princípio subjacente é a parcimónia que resulta de um *trade-off* entre a complexidade do modelo (medida através do número de parâmetros) e o seu ajustamento (Dias, 2004). Dos diversos critérios existentes destacam-se pela sua popularidade e utilização, o Akaike Information Criterion (AIC) e o Bayesian Information Criterion (BIC). A selecção do modelo/número de segmentos baseia-se na minimização do seguinte critério:

$$C_S = -2 \ell_S(\hat{\varphi}; \mathbf{y}, \mathbf{x}, \mathbf{n}) + dN_S$$

em que  $\ell_S(\cdot)$  e  $N_S$  correspondem à função log-verosimilhança e ao número de parâmetros livres no modelo, respectivamente. O parâmetro  $d$  varia de acordo com o critério de informação utilizado, sendo igual a 2 no caso do AIC e a  $\ln(n)$  no caso do BIC. Valores menores destes critérios significam modelos mais parcimoniosos (Dias, 2003).

---

<sup>6</sup> Os valores iniciais das estimativas dos parâmetros podem ser gerados aleatoriamente.

4.4 Exemplo de uma mistura de regressão de Poisson

Considerou-se uma mistura de duas distribuições de Poisson para gerar observações definidas pela seguinte expressão:  $\ln \lambda_{is} = \beta_{0s} + \beta_{1s} x_i$ . Considerando a existência de dois segmentos, os parâmetros do modelo são  $\varphi = (\pi, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12})$ . Simulou-se um conjunto de dados com um total de 200 observações, geradas a partir de  $x_i \sim N(0,1)$  e  $\varphi = (0.5, -1.0, 1.5, -0.5, -1.0)$ . A exposição é igual a 1 para todas as observações ( $n_i = 1$ ).

Tabela 1 – Resultados da estimação do modelo de regressão com dados simulados.

Parâmetros estimados	Agregado	Solução com dois segmentos	
		1	2
$\hat{\beta}_{0s}$	-0.0665	-1.3595	-0.4121
$\hat{\beta}_{1s}$	0.0999	1.8838	-0.9001
$\hat{\pi}_s$	1.0000	0.4775	0.5225

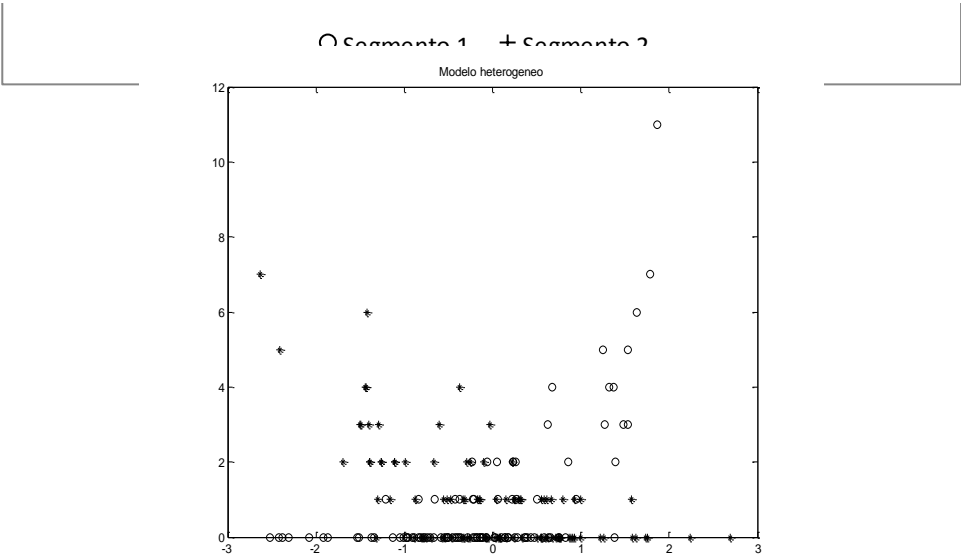


Figura 1 – Modelo de regressão heterogéneo baseado em dados simulados.

A Tabela 1 compara os resultados obtidos através da estimação de um modelo com um único segmento ( $S = 1$ ) e de um modelo com dois segmentos ( $S = 2$ ). Como seria de esperar, as estimativas obtidas na solução com dois segmentos são muito próximas dos valores dos parâmetros reais. A Figura 1 ilustra a dispersão das observações, bem como a classificação nos dois segmentos. Verifica-se que este modelo detecta a heterogeneidade existente na amostra simulada.

## 5. Resultados

Tendo em conta o enquadramento técnico apresentado anteriormente, procedeu-se à estimação de um modelo de mistura de regressões de Poisson, com o objectivo de modelar a incidência de voto no candidato Manuel Alegre, com base na proporção de votos por concelho no Partido Socialista nas eleições legislativas de 2005, bem como na taxa de analfabetismo, no índice de envelhecimento e índice de desenvolvimento concelhio.

### 5.1 Decisão sobre o número de segmentos

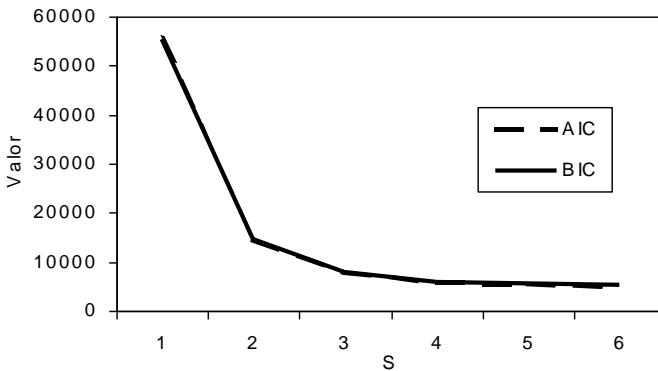
Foram utilizados diferentes valores do número de segmentos para estimar o modelo de mistura de regressão de Poisson. Para cada caso, foram utilizadas diferentes soluções iniciais conjugadas com um número de iterações elevado e um nível de tolerância reduzido (0.00001) maximizando assim as chances da obtenção de um óptimo global.

A Tabela 2 apresenta os resultados desta estimação para o diferente número de segmentos, correspondente ao valor de log-verosimilhança mais elevado das diferentes soluções obtidas.

**Tabela 2** – Critérios de informação.

S	Log-Verosimilhança	Nº de parâmetros	Critérios de	
			AIC	BIC
1	-27705.6	5	55421.	55439.33
2	-7299.7	11	14621.	14661.31
3	-3910.8	17	7855.6	7917.320
4	-2872.5	23	5791.0	5874.515
5	-2727.9	29	5513.8	5619.017
6	-2476.8	35	5023.6	5150.592

Com base nos critérios de informação, conclui-se que a melhor solução corresponde a um modelo com três segmentos (Figura 2).



**Figura 2** – Critérios de informação.

## 5.2 Resultados da estimação

Apesar da solução com três segmentos ser a mais adequada para o modelo em causa, consideram-se os resultados para o modelo agregado, para dois e três segmentos, comparando os resultados obtidos (Tabela 3). Em todas as soluções consideradas, a totalidade das variáveis incluídas são significativas ( $\text{valor-}p^7 < 0.001$ ) para explicar a incidência de voto no candidato Manuel Alegre.

Considerando a solução com um único segmento, um aumento do nível de desenvolvimento, de envelhecimento concelhio e da percentagem de votos no PS, provoca um aumento percentual esperado na votação em Manuel Alegre, enquanto a taxa de analfabetismo tem um efeito contrário.

Contudo, em termos agregados, o modelo tem uma reduzida capacidade explicativa ( $\text{Pseudo-}R^2 = 0.0494^8$ ). Quando se considera uma solução com dois segmentos (segmento 1: 58.3% dos concelhos), a capacidade explicativa melhora

<sup>7</sup> O *valor-p* num teste estatístico representa a probabilidade se observar um valor igual ou mais extremo do que o observado, assumindo que a hipótese nula é verdadeira. A hipótese nula é rejeitada quando o *valor-p* é inferior ao nível de significância. Neste caso, considerou-se um nível de significância de 0.05.

<sup>8</sup> Genericamente, o  $\text{pseudo-}R^2$  calcula-se da seguinte forma:  $1 - l(\hat{\beta}_R) / l(\hat{\beta}_U)$ , em que  $l(\hat{\beta}_R)$  é igual ao valor máximo do logaritmo da função de verosimilhança com todos os parâmetros (excepto a constante) iguais a 0 e  $l(\hat{\beta}_U)$  corresponde ao valor máximo da função de verosimilhança com todas as variáveis explicativas. Notemos que se trata de um modelo linear generalizado, enquanto o coeficiente de determinação apenas está definido para modelos lineares.



substancialmente ( $\text{Pseudo-R}^2 = 0.8132$ ) e para uma solução de três segmentos, a capacidade explicativa, traduzida pelo  $\text{Pseudo-R}^2$ , ascende aos 95.5%.

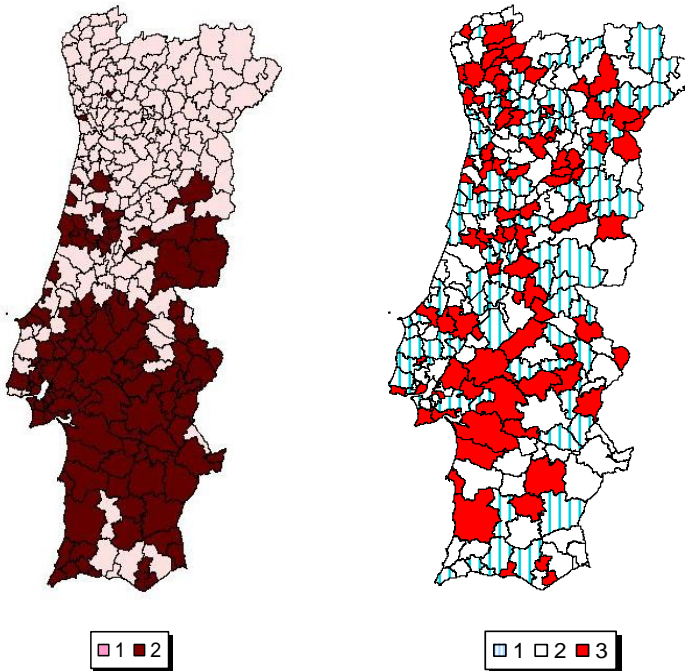
**Tabela 3** – Estimativas do modelo de regressão.

Parâmetros estimados	Agregado (S = 1)	Dois segmentos		Três segmentos		
		1	2	1	2	3
$\hat{\beta}_{0s}$	-2.4895	-3.4156	-1.5417	-3.6763	-1.4556	-2.1675
% PS nas legislativas de 2005 ( $\hat{\beta}_{1s}$ )	2.0098	1.5460	0.8712	1.5283	0.7492	1.1946
Índice de desenvolvimento ( $\hat{\beta}_{2s}$ )	0.0006	0.0097	-0.0013	0.0124	-0.0017	-0.0005
Índice de envelhecimento ( $\hat{\beta}_{3s}$ )	0.0013	0.0022	-0.0004	0.0016	0.0002	0.0008
Taxa de analfabetismo ( $\hat{\beta}_{4s}$ )	-0.0221	-0.0143	-0.0041	-0.0104	-0.0081	-0.0064
Peso do segmento ( $\hat{\pi}_s$ )	1.0000	0.5831	0.4169	0.3779	0.3157	0.3064
Pseudo- $R^2$	0.0494	0.8132		0.9545		

Interpretando os resultados obtidos para o modelo com dois segmentos, destaca-se a diferente relação existente entre as variáveis índice de desenvolvimento e envelhecimento e a incidência de votos nos dois segmentos em análise. Com efeito, no segmento 1, um aumento do índice de desenvolvimento e de envelhecimento aumenta percentualmente a incidência de voto Manuel Alegre, enquanto no segmento 2 o efeito é contrário. Constata-se também que uma maior percentagem de votos no PS reflecte uma maior incidência de votos no candidato Manuel Alegre, quer considerando apenas um segmento, quer considerando ambos os segmentos em separado, enquanto que uma maior taxa de analfabetismo traduz uma menor incidência de votos no candidato Manuel Alegre.

Na solução com três segmentos e independentemente do segmento em consideração, uma maior percentagem de votos no PS e um maior índice de envelhecimento traduzem uma maior incidência de votos no candidato Manuel Alegre, enquanto que uma maior taxa de analfabetismo traduz uma menor percentagem de votos neste candidato. Por sua vez, a relação entre o índice de desenvolvimento e a incidência de votos em Manuel Alegre é distinta se considerarmos o primeiro segmento e os restantes: no primeiro segmento, um aumento do índice de desenvolvimento provoca um aumento percentual na incidência de voto, enquanto nos restantes segmentos o efeito é contrário.

Como se observa na Figura 3, numa solução com dois segmentos, parece existir, em termos espaciais, uma clara separação Norte-Sul na distribuição dos concelhos pelos dois segmentos. Os concelhos do segmento 1 concentram-se predominantemente na zona norte do país, enquanto que a zona sul de Portugal é composta na sua esmagadora maioria por concelhos do segmento 2.



**Figura 3** – Distribuição dos concelhos pelos segmentos definidos.

Contudo, quando se analisa a solução de três segmentos, esta separação desvanece-se, não se conseguindo identificar um padrão claro quanto à distribuição regional dos segmentos, em termos de grandes linhas de demarcação tradicionais do espaço português (*e.g.*, Norte-Sul, Litoral-Interior). Não obstante, a tabela seguinte identifica as características de cada um dos segmentos.

**Tabela 4** – Caracterização dos segmentos.

	Agregado	Solução com três segmentos		
		1	2	3
Incidência de votos em Manuel Alegre	0.202	0.142	0.272	0.204
Proporção de votos no PS nas legislativas de 2005	0.457	0.414	0.496	0.468
Índice de desenvolvimento	80.3	76.3	86.1	79.2
Índice de envelhecimento	161.4	164.9	151.5	167.0
Taxa de analfabetismo	13.5	13.9	13.5	13.1

Observa-se que o segmento 2 apresenta uma maior incidência de votos no candidato Manuel Alegre e no PS, bem como um maior índice de desenvolvimento e uma população mais jovem. Por sua vez, o segmento 1 é o oposto, caracterizando-se por uma menor incidência de votos em Manuel Alegre e no PS e um menor índice de desenvolvimento, apresentando, contudo, uma maior taxa de analfabetismo. Por último, o segmento 3 caracteriza-se por ter uma população mais idosa, mas uma menor taxa de analfabetismo.

## 6 Conclusões

A aplicação de um modelo de mistura de regressão de Poisson permite verificar a existência de grupos heterogêneos na população, no que diz respeito à relação existente entre a incidência de voto no candidato Manuel Alegre, e a taxa de envelhecimento da população, a taxa de analfabetismo, o nível de desenvolvimento concelhio e a percentagem de votos no PS nas eleições legislativas. A solução de três segmentos permite concluir que independentemente do segmento em consideração, uma maior percentagem de votos no PS e um maior índice de envelhecimento traduzem uma maior incidência de votos no candidato Manuel Alegre, enquanto que uma maior taxa de analfabetismo traduz uma menor percentagem de votos neste candidato. Por sua vez, a relação entre o índice de desenvolvimento e a incidência de votos em Manuel Alegre é distinta se considerarmos o primeiro segmento e os restantes.

Em termos regionais não se consegue identificar um padrão claro quanto à distribuição dos concelhos para três segmentos. Esta aplicação mostra que nem sempre as medidas de selecção do número de segmentos na mistura finita identificam uma solução facilmente interpretável e com forte suporte substantivo. Por vezes, como no presente caso, a selecção do número de segmentos deverá ter em atenção outros factores de natureza qualitativa, pois mais segmentos não significa necessariamente maior riqueza de interpretação dos dados.

## Referências

- DEMPSTER, A.P.; LAIRD, N.M. e RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1–38.
- DIAS, J.G. (2003). Introdução aos modelos de mistura finita. In: REIS, E. e M. HILL, *Temas em Métodos Quantitativos 3*, Lisboa, Edições Sílabo, 101-115.
- DIAS, J.G. (2004). *Finite Mixture Models: Review, Applications, and Computer-Intensive Methods*, University of Groningen, Holanda, 206 p.
- FONSECA, P. (2002). Índices de Desenvolvimento Concelhio, *Revista de Estatística*, 2º Quadrimestre de 2002, INE.
- FREIRE, A. (2001). *Modelos do Comportamento Eleitoral – Uma breve introdução crítica*. Celta Editora: Oeiras.
- INE – Instituto Nacional de Estatística (2006). Dados dos Censos 2001 (<http://www.ine.pt>).
- STAPE – Secretariado Técnico dos Assuntos para o Processo Eleitoral (2006a). Resultados das Eleições Presidenciais de 2006 (<http://www.eleicoes.mj.pt/Presidenciais2006>, data de download: 02/02/2006).
- STAPE – Secretariado Técnico dos Assuntos para o Processo Eleitoral (2006b). Resultados das Eleições Legislativas de 2005 (<http://www.stape.pt/resultref/ar2005.htm>, data de download: 17/03/2006).
- WEDEL, M.; DESARBO, W.S.; BULT, J.R. e RAMASWAMY, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8(4), 397-411.



## Análise Conjunta Baseada em Preferências vs Escolhas. Selecção de uma Escola de Gestão

Susana I.A. Tavares<sup>1</sup>, José G. Dias<sup>2</sup>, Margarida G.M.S. Cardoso<sup>3</sup>

© The Author(s) 2013

**Resumo** Existem diferentes formas de medir a importância relativa de atributos considerados num processo de tomada de decisão (escolha entre alternativas possíveis). Na Análise Conjunta Tradicional, também designada por Análise Conjunta de Preferências (ACp), é normalmente pedido aos inquiridos para expressarem as suas preferências através de uma pontuação (*rate*) ou ordenação (*rank*) de diversas alternativas ou perfis, caracterizadas por um número limitado de níveis de atributos definidos mediante um desenho experimental. A Análise Conjunta baseada em escolhas (ACe) é uma técnica de experimentação mais realista. Perante cada conjunto de alternativas que são apresentados aos inquiridos é pedida uma escolha, podendo estes não optar por qualquer alternativa. O objectivo principal deste artigo consiste em apresentar as principais diferenças entre as duas metodologias de Análise Conjunta: uma baseada em dados relativos a preferências e outra a escolhas. Ilustra-se uma comparação entre os dois processos numa aplicação referida aos atributos que estudantes do secundário consideram mais importantes na escolha de uma Escola de Gestão.

**Palavras-chave:** Análise Conjunta Baseada em Preferências (ACp), Análise Conjunta Baseada em Escolhas (ACe), Desenho Experimental, Educação.

---

<sup>1</sup> ISCTE, Instituto Universitário de Lisboa, BRU-IUL, s.tavares@iscte.pt

<sup>2</sup> ISCTE, Instituto Universitário de Lisboa, BRU-IUL, jose.dias@iscte.pt

<sup>3</sup> ISCTE, Instituto Universitário de Lisboa, BRU-IUL, margarida.cardoso@iscte.pt

## 1 Introdução

A Análise Conjunta é uma técnica estatística usada para analisar as preferências dos indivíduos, na qual é fornecido aos inquiridos um conjunto de perfis que são, normalmente, produtos definidos por combinações de categorias (níveis) dos seus atributos. A partir da preferência que cada inquirido manifesta pelos vários perfis, estima-se a importância dos atributos e a valorização atribuída aos seus níveis. Estes valores são normalmente referidos como *utilidades parciais* (*part-worth utilities*).

A Análise Conjunta baseada em preferências (ACp) foi dominante durante décadas. Nesta abordagem, também designada por Análise Conjunta Tradicional, é pedido aos inquiridos uma pontuação (*rate*) ou ordenação (*rank*) de diversas alternativas ou perfis, caracterizadas por um número limitado de níveis de atributos, definidos mediante um desenho experimental. Utilizando um ou outro tipo de medida de preferência (*rate* ou *rank*), esta técnica fornece estimativas das utilidades parciais para os níveis dos atributos, mediante os coeficientes de uma regressão linear múltipla aplicada à modelação das mesmas preferências (é unânime na literatura que a regressão é uma alternativa robusta para analisar dados de ordenação, quando o número de postos é superior a 10). As quotas de determinados perfis (*share*) (combinações específicas dos níveis dos atributos que condicionam a preferência) podem ser obtidas através de simulações, sendo os métodos mais usuais a abordagem segundo a utilidade máxima, BTL (Bradley-Terry-Luce) e a modelação logit. A regressão associada à ACp pode realizar-se ao nível do indivíduo. Torna-se, assim, possível a utilização do *output* da ACp como *input* para outras técnicas, como por exemplo, algoritmos de agrupamento para segmentação de mercado – constituindo segmentos com preferências semelhantes – utilizando como variáveis os resultados imediatos da ACp (coeficientes dos modelos de regressão individuais).

A Análise Conjunta baseada em escolhas (ACe) é uma técnica de experimentação mais realista. Os perfis a testar são divididos em vários conjuntos. Perante cada conjunto é pedida aos inquiridos uma escolha, podendo estes optar por não escolher qualquer perfil. A ACe viabiliza a previsão de quotas de perfis directamente, mediante o modelo que propõe. Os resultados da ACe são obtidos de forma agregada. No entanto, adoptando um modelo de segmentos latentes, podem obter-se segmentos homogêneos na formação das suas escolhas.

Este artigo apresenta as principais diferenças entre as duas metodologias de Análise Conjunta: uma baseada em preferências e outra em escolhas (Gustafsson *et al.*, 2000; Louviere *et al.*, 2000).

No sentido de estudar as características de cada técnica (ACp e ACe), utilizam-se, neste trabalho, duas amostras de estudantes finalistas do ensino secundário, respondentes a um inquérito relativo a escolhas/preferências associada a Escolas de Gestão. Sobre estas duas amostras comparam-se as estimativas de utilidades parciais e importância dos atributos, resultados de segmentação e, ainda,

estimativas de quotas de perfis (simulações), não contemplados nos respectivos desenhos experimentais.

## **2 Principais diferenças entre a análise conjunta baseada em preferências (ACp) e análise conjunta baseada em escolhas (ACE)**

Este ponto destaca as principais diferenças entre as duas metodologias de Análise Conjunta – Análise Conjunta baseada em Preferências e Análise Conjunta baseada em Escolhas.

### **2.1 Desenho experimental**

O desenho experimental utilizado na ACp baseia-se essencialmente na propriedade da ortogonalidade – inexistência de correlação entre os atributos. A ortogonalidade de um desenho factorial garante a independência entre as estimativas dos parâmetros e, geralmente, estimadores eficientes.

Na ACE, o desenho experimental baseia-se em três propriedades que são as mais usadas para gerar desenhos de escolha eficientes (Huber e Zwerina, 1996):

- Balanceamento: os níveis de cada atributo ocorrem com igual frequência (distribuição proporcional dos níveis dos atributos);
- Ortogonalidade: inexistência de correlação entre os atributos;
- Sobreposição mínima: em cada conjunto de alternativas deve ser mínima a duplicação de um nível de um determinado atributo.

Existe, contudo, alguma discussão sobre os desenhos factoriais fraccionários ortogonais e não ortogonais. Nos últimos anos tem crescido a utilização de desenhos não ortogonais eficientes (Kuhfeld *et al.*, 1994). Considera-se mesmo que a ortogonalidade pode não ser desejável em algumas situações em que (Kuhfeld, 1997):

- o número de níveis é diferente para a maior parte dos atributos;
- é predefinido o número de cartões desejado;
- existe uma correlação considerável entre alguns atributos, um desenho ortogonal pode produzir alguns perfis não realistas (Green e Srinivasan, 1978). Neste caso é necessário eliminar esses perfis e incluir outros considerados aceitáveis.

Para avaliar a proximidade do desenho experimental face a um desenho ortogonal hipotético, medindo assim a eficiência de um desenho experimental, pode utilizar-se o índice *d-efficiency* (Kuhfeld *et al.*, 1994).



## 2.2 Recolha da Informação

No modelo de Análise Conjunta baseada em preferências, todos os perfis são apresentados aos respondentes e a tarefa de cada um consiste em ordenar ou avaliar cada um dos perfis, enquanto que, no modelo de Análise Conjunta baseada em escolhas são apresentados vários conjuntos de alternativas e a tarefa consiste em escolher a alternativa preferida em cada conjunto.

## 2.3 Modelação/ estimação

Na ACp a variável dependente é, no mínimo, ordinal. Para a sua modelação é habitual utilizar-se um modelo de regressão linear múltipla com estimação feita através do Método dos Mínimos Quadrados. Na ACE a variável dependente é categorizada, utilizando-se um Modelo de Regressão Logística com estimação feita através do Método da Máxima Verosimilhança. Para uma exposição mais detalhada sobre este tipo de modelos, *vide* Dias (1997), Ben-Akiva e Lerman (1985) e Haaijer e Wedel (2000).

## 2.4 Resultados

No caso da Análise Conjunta baseada em preferências podem obter-se, como resultado imediato, as utilidades desagregadas ao nível do indivíduo.

No caso da Análise Conjunta baseada em escolhas, os resultados são obtidos para a amostra (resultados agregados), sendo feita *a posteriori* a sua desagregação. Esta é a maior desvantagem deste tipo de modelos, pois não existem observações suficientes para estimar o modelo de escolha ao nível do indivíduo, o que faz com que seja impossível estimar modelos individuais. Para proceder à desagregação pode, contudo, estimar-se um modelo com segmentos latentes.

## 2.5 Segmentação

O processo de segmentação mais utilizado na ACp é tradicionalmente feito em duas etapas, em que primeiramente se determinam as utilidades parciais dos níveis dos atributos, e, numa segunda etapa, se agrupam os inquiridos utilizando um algoritmo baseado nas semelhanças entre os perfis de benefício<sup>4</sup>.

---

<sup>4</sup> Note-se que o processo de segmentação em duas etapas é o mais utilizado na ACp. No entanto, as duas abordagens de segmentação (uma etapa ou duas etapas) podem ser utilizadas tanto na ACp como na Ace.

No caso da Análise Conjunta baseada em escolhas (ACe) utilizam-se modelos de misturas ou de regressão *clusterwise* que simultaneamente estimam as utilidades parciais e os respectivos segmentos de pertença dos inquiridos.

## 2.6 Simulação (Quotas de escolha)

A ACe produz directamente probabilidades de escolha, ao contrário da ACp que necessita de uma conversão para estimar essas probabilidades. Os métodos mais utilizados neste caso são: modelo de máxima utilidade ou de primeira escolha, modelo BTL ou proporcional e o modelo logit.

A Tabela 1 sintetiza as principais diferenças entre as duas técnicas (ACp e ACe) nos vários passos necessários para a aplicação de uma Análise Conjunta.

**Tabela 1** - Quadro comparativo entre as técnicas ACp e ACe.

	<b>ACp (Preferências)</b>	<b>ACe (Escolhas)</b>
<b>Desenho experimental</b>	Ortogonal/ Quasi ortogonal	Baseado na ortogonalidade, balanceamento dos níveis e sobreposição mínima dos mesmos
<b>Recolha da informação</b>	Pontuação ( <i>rate</i> ) ou ordenação ( <i>rank</i> )	Escolha de um perfil em vários conjuntos de alternativas
<b>Variável dependente</b>	Mínimo ordinal	Categorizada
<b>Modelo</b>	Regressão Linear Múltipla	Modelo Logístico Multinomial
<b>Estimação</b>	Método dos Mínimos Quadrados	Método da Máxima Verosimilhança
<b>Resultados</b>	Nível do indivíduo	Agregada. Para desagregar pode utilizar-se um Modelo de Classes Latentes
<b>Segmentação de mercado</b>	2 etapas (caminho tradicional): Análise de Agrupamento sobre os resultados da ACp	1 etapa: Modelo de Classes Latentes que simultaneamente estima as utilidades parciais e o respectivo segmento
<b>Quotas de escolha</b>	Conversão para estimar as probabilidades de escolha (Métodos BTL, Máxima utilidade e Logit)	Produce directamente as quotas de escolha

### **3 Aplicação prática – Selecção de uma escola de gestão**

Nesta secção ilustra-se a aplicação das duas técnicas, ACp e ACe, com o objectivo de comparar os principais resultados entre elas. Para tal realizou-se um estudo sobre os critérios que os estudantes do 12º ano mais valorizam na escolha de uma Escola de Gestão.

#### **3.1 Selecção de atributos**

Os atributos mais importantes para a formação de uma preferência ou escolha de uma Escola de Gestão foram identificados através de um estudo prévio (McKinsey, 2003). Esses atributos são: *Prestígio*, *Qualidade da Formação Prática*, *Ligação ao Mundo Empresarial* e *Facilidade em Arranjar Emprego*. Neste estudo a cada um destes quatro atributos foram associados três níveis ordinais: acima da média, na média e abaixo da média.

#### **3.2 Amostra**

Foram constituídas duas amostras de alunos do 12º ano da área de Gestão/Economia no Distrito de Lisboa. Uma delas destinou-se à aplicação da ACp e a outra à Ace, respectivamente com 134 e 130 indivíduos. Os 264 alunos estudam em 22 escolas, sendo 141 do sexo masculino.

#### **3.3 Método de recolha de informação**

Foi utilizado o método de auto-preenchimento para os dois questionários, um para cada amostra/ técnica. Cada turma seleccionada foi dividida em duas partes, de forma a que cada parte respondesse a um tipo de questionário (preferências ou escolhas).

#### **3.4 Desenho experimental**

Para a ACp, foi utilizado um desenho experimental próximo de um desenho ortogonal<sup>5</sup> com 10 perfis. Para a Ace, foi utilizado um desenho experimental

---

<sup>5</sup> O desenho para a ACp foi gerado pelo programa ORTHOPLAN do SPSS (SPSS, 2005).

eficiente<sup>6</sup> (próximo da ortogonalidade, com balanceamento dos níveis e sobreposição mínima dos mesmos em cada conjunto de perfis) com 6 conjuntos de quatro perfis.

3.5 Modelação/ Estimação

Para a ACp foi utilizado um Modelo de Regressão Linear Múltipla, enquanto que na Ace foi utilizado um Modelo de Regressão Logística. Na Ace, para desagregar a informação ao nível do indivíduo foi utilizado um Modelo de Segmentos Latentes.

3.6 Interpretação dos resultados

Os pontos seguintes referem-se à análise comparativa dos resultados obtidos pelas duas técnicas, nomeadamente no que respeita à importância dos atributos, utilidade dos perfis e simulação de quotas de alguns perfis não contemplados nos desenhos experimentais. Apresenta-se ainda um sumário dos resultados de segmentação.

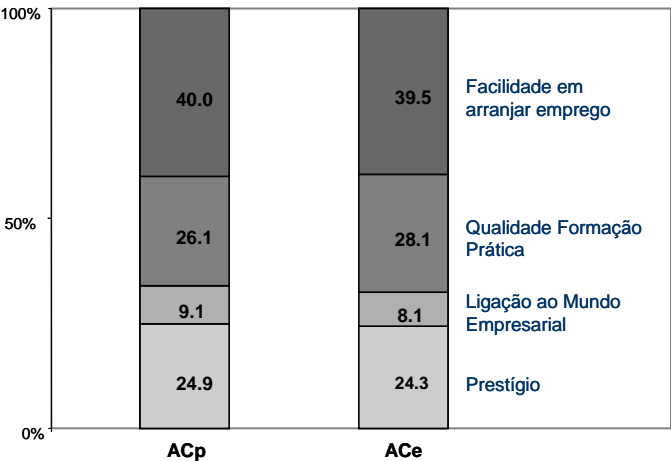


Figura 1 – Importância dos atributos.

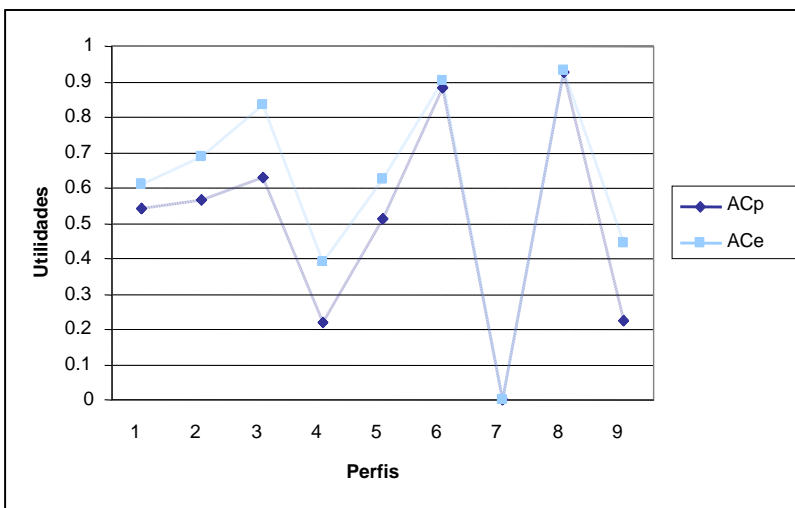
<sup>6</sup> O desenho para a Ace foi gerado pelo programa JMP (SAS Institute, 2002), Módulo DOE (*Design of Experiments*).

- **Importância dos atributos**

Comparando as duas técnicas (ACp e ACe), pode observar-se pela Figura 1 que a importância estimada para os atributos é muito semelhante. Ambas as técnicas indicam que o atributo que os alunos consideram mais importante na escolha de uma Faculdade de Gestão é a *Facilidade em arranjar emprego*, seguido dos atributos *Qualidade da Formação Prática* e *Prestígio*. O atributo *Ligação ao mundo empresarial* foi o menos valorizado pelos alunos.

- **Utilidades de perfis**

Comparando os valores das utilidades obtidas para os perfis do desenho experimental da ACp com as utilidades dos mesmos perfis geradas pela ACe, depois de reescaladas entre 0 e 1, foram obtidos valores muito semelhantes (Fig. 2).



**Figura 2** – Utilidades dos perfis para as duas técnicas.

- **Simulações de quotas para perfis**

Consideraram-se três opções de selecção de uma Faculdade de Gestão, não contempladas em nenhum dos desenhos experimentais (Tab. 2).

Os resultados da simulação de quota para as três opções com base nos modelos estimados e utilizando a regra logit são apresentados na Figura 3. Conclui-se que os resultados utilizando as duas técnicas são semelhantes, sendo a *Opção 3* claramente a mais desejada.

Tabela 2 – Novo conjunto de perfis.

Atributo	Opção 1	Opção 2	Opção 3
Prestígio/ Reputação	Acima da média	Na média	Abaixo da média
Ligação ao Mundo Empresarial	Abaixo da média	Na média	Abaixo da média
Qualidade da Formação Prática	Na média	Acima da média	Acima da média
Facilidade em arranjar emprego	Na média	Abaixo da média	Acima da média

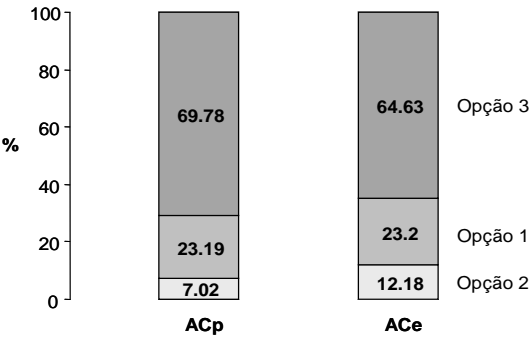


Figura 3 – Simulação de perfis hipotéticos.

• Segmentação

A estrutura de segmentos que resulta da aplicação do Método Ward sobre as importâncias em ACp e do Modelo de Segmentos Latentes em ACe é constituída por três segmentos. As características que se podem associar a estes segmentos não são coincidentes, embora, no geral sejam concordantes, como se pode verificar na Tabela 3.

Tabela 3 – Segmentação resultante da ACp e ACe.

ACp	ACe
<b>Pró-Emprego:</b> (Facilidade em arranjar emprego). Escolas públicas, sexo feminino e rendimentos do agregado médios; <b>Pró-Prestígio:</b> (Prestígio). Ensino privado, sexo masculino, Universidade “Nova”, a instrução máxima dos pais é licenciatura e rendimentos elevados. <b>Pró-Formação Prática:</b> (Qualidade da Formação Prática). Escolas públicas, instrução máxima dos pais o 9º ano e rendimentos baixos.	<b>Pró-Emprego:</b> (Facilidade em arranjar emprego e Qualidade da formação prática). Escolas públicas, sexo feminino, rendimentos do agregado familiar médios; <b>Pró-Prestígio:</b> (Prestígio). Preferem a Universidade “Nova”, a instrução máxima dos pais é licenciatura, ensino privado. <b>Indecisos:</b> (Não diferenciam os atributos). Ensino público, sexo masculino e baixa instrução dos pais.

## 4 Conclusão

Este artigo abordou duas metodologias distintas de Análise Conjunta – Análise Conjunta baseada em preferências (ACp) e Análise Conjunta baseada em escolhas (ACe).

Existem algumas diferenças entre estas duas metodologias. A tarefa de escolha é mais próxima da realidade conduzindo a uma maior validade externa (Elrod *et al.*, 1992), no entanto implica mais tempo para a sua realização (Huber *et al.*, 2002). Por outro lado, é mais fácil para o inquirido fazer uma escolha sem se preocupar em ordenar todos os perfis.

A ACe produz directamente quotas de escolha dos perfis contemplados no desenho experimental, ao contrário da ACp que necessita de uma conversão para estimar essas probabilidades (Métodos BTL, Máxima Utilidade, Logit). Por outro lado, a ACp dá uma maior informação acerca das preferências de cada indivíduo. A informação que nos dá a ACe é sempre agregada. Para desagregar informação ao nível do indivíduo utilizou-se um Modelo de Segmentos Latentes.

Fez-se uma análise comparativa das duas técnicas através de uma aplicação prática. Esta aplicação teve como objectivo identificar os atributos determinantes na escolha de uma Escola de Gestão. Ambas as técnicas indicam que o atributo que os alunos consideram mais importante na escolha de uma Faculdade de Gestão é a *Facilidade em arranjar emprego*, seguido dos atributos *Qualidade da formação prática* e *Prestígio*. O atributo *Ligação ao mundo empresarial* foi o menos valorizado pelos alunos, o que permite concluir que as duas técnicas valorizaram de forma idêntica os atributos considerados na formação de preferências ou escolha de uma Faculdade de Gestão. As utilidades geradas pelas duas técnicas são equivalentes. Pondo em confronto vários perfis hipotéticos, obtiveram-se os mesmos valores de quota nas duas técnicas. A conversão de preferências em probabilidade de escolhas feita através da regra logit gerou valores idênticos aos obtidos na ACe. A estrutura de segmentos que resulta da aplicação do Método Ward sobre as importâncias da ACp e do Modelo de Segmentos Latentes em ACe é constituída por três segmentos. As características que se podem associar a estes segmentos não são coincidentes, embora, no geral sejam concordantes.

No que respeita à aplicação das duas técnicas detectaram-se algumas diferenças nos tempos e na dificuldade de preenchimento dos inquéritos. Os entrevistados que utilizaram a técnica ACp registaram tempos de preenchimento do inquérito sensivelmente inferiores, no entanto, houve uma maior dificuldade no preenchimento da parte da ACp. A percentagem de inquéritos com ordenação incompleta foi superior à percentagem de inquéritos com conjuntos de perfis não respondidos, provavelmente porque a técnica baseada em escolhas se assemelha mais a situações reais.

Embora não se possam tirar conclusões definitivas apenas com um estudo, aparentemente nenhuma das duas técnicas se sobrepõe à outra.

Em análises posteriores seria interessante:

- comparar as duas técnicas (ACe e ACp) com outros atributos e noutras áreas e verificar se mantêm a consistência de resultados;
- comparar a ACp com utilidades geradas ao nível do indivíduo e a ACp feita com segmentos latentes como forma de desagregação individual das utilidades;
- realizar estudos de simulação em que se poderá controlar um conjunto de factores e comparar de forma mais adequada as técnicas estudadas.

## Referências

- BEN-AKIVA, M. e LERMAN, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press: Cambridge.
- DIAS, J. (1997). *Análise Conjunta. Aplicação ao Processo de Escolha de um Curso Superior*. Tese de Mestrado em Ciências Empresariais: INDEG/ ISCTE, Lisboa.
- ELROD, T.; LOUVIERE, J. e DAVEY, K. (1992). An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models, *Journal of Marketing Research*, 29, 368-377.
- GREEN, P. e SRINIVASAN, V. (1978). Conjoint Analysis in Consumer Research: Issues and Outlook, *Journal of Consumer Research*, 5, 103-123.
- GUSTAFSSON, A.; HERRMANN, A. e HUBER, F. (2000). *Conjoint Measurement, Methods and Applications*, Springer: Berlim.
- HAAIJER, M. e WEDEL, M. (2000). Conjoint Choice Experiments: General Characteristics and Alternative Model Specifications, em GUSTAFSSON, A.; HERRMANN, A. e HUBER, F. (eds.): *Conjoint Measurement, Methods and Applications*, 319-360 (Springer, Berlim).
- HUBER, J. e ZWERINA, K. (1996). The Importance of Utility Balance in Efficient Choice Designs, *Journal of Marketing Research*, 33, 307-317.
- HUBER, J.; ARIELY, D. e FISCHER, G. (2002). *Expressing Preferences in a Principal-Agent Task: A Comparison of Choice, Rating, and Matching*, *Organizational Behavior and Human Decision Processes*, 87(1), 66-90.
- JMP (2002). *Design of Experiments*. USA: SAS Institute Inc..
- KUHFELD, W. (1997). *Efficient Experimental Designs Using Computerized Searches*, SAS Institute, Inc.



- KUHFELD, W.; TOBIAS, R. e GARRATT, M. (1994). Efficient Experimental Design with Marketing Research Applications, *Journal of Marketing Research*, 31, 545-557.
- LOUVIERE, J.; HENSHER, D. e SWAIT, J. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University Press: Cambridge.
- Mckinsey (2003). *Definir Prioridades Estratégicas da Escola de Gestão do ISCTE*, Estudo de Consultoria Realizado para o ISCTE.
- SPSS (2005). *SPSS Conjoint 14.0 (Manual)*. Chicago: SPSS.

# Geração Aleatória de Estruturas de Classificação: das Hierarquias às Pirâmides

Vasco Machado<sup>1</sup> · Fernanda Sousa<sup>2</sup>

© The Author(s) 2013

**Resumo** Nas últimas décadas vários algoritmos de geração de estruturas de classificação foram propostos. Neste trabalho é analisado e discutido o desempenho do método *RAP* - Random Generation Algorithm of Pyramids, método de geração aleatória de uma estrutura piramidal. Sendo o modelo de classificação piramidal uma generalização do modelo de classificação hierárquica, o método em análise surge como uma extensão de trabalhos anteriores de geração aleatória de dendrogramas. A avaliação do método *RAP* inclui uma abordagem teórica e de simulação. Para um número fixo de nós terminais identificam-se os diferentes tipos topológicos de pirâmides e o respectivo número de pirâmides não isomórficas. A complexidade deste estudo limitou-o a pirâmides com um número reduzido de nós terminais.

**Palavras-chave:** Análise Classificatória, Classificação Piramidal, Pirâmide, Geração Aleatória de Pirâmides, Simulação.

## 1 Introdução

De uma forma geral pode afirmar-se que o objectivo de uma classificação é encontrar uma estrutura de classificação, sobre o conjunto de elementos a classificar, que reflecta as relações de semelhança e/ou dissemelhança entre esses elementos. Existem diferentes tipos de estruturas de classificação, com distintos graus de complexidade, de acordo com as propriedades que lhes são subjacentes. De entre os vários tipos dessas estruturas as mais conhecidas são as partições e as hierarquias.

---

<sup>1</sup>Departamento de Saúde Pública, ARS Norte, I.P., [vmpmachado@gmail.com](mailto:vmpmachado@gmail.com)

<sup>2</sup>Faculdade de Engenharia e CITTA, Universidade do Porto, [fcsousa@fe.up.pt](mailto:fcsousa@fe.up.pt)

Vários trabalhos têm sido desenvolvidos no contexto da geração aleatória de árvores de classificação. O interesse por esta temática é em grande parte justificado pelas aplicações de técnicas de Monte Carlo em classificação. Como trabalhos importantes para o presente artigo e para uma lista mais completa de referências sobre o assunto, refere-se Furnas (1984), Lapointe e Legendre (1991), Podani (2000) e Sousa (2000). A comparação de estruturas de classificação é uma exigência frequente em classificação, várias medidas têm vindo a ser desenvolvidas e/ou adaptadas para comparar os diferentes tipos de estruturas. A interpretação probabilística dos valores resultantes da comparação de estruturas necessita do conhecimento das distribuições estatísticas destas medidas. Estas distribuições estatísticas teóricas são desconhecidas ou desadaptadas neste contexto, já que as hipóteses de independência, usualmente aceites, são aqui fortemente postas em causa. A dedução das distribuições teóricas destas medidas, tendo em conta as particularidades do contexto, só é conseguida para dimensões muito reduzidas. Atendendo a que as estruturas a comparar em classificação têm habitualmente dimensões elevadas, o recurso à simulação de Monte Carlo para a obtenção de distribuições empíricas surge como uma solução, tornando a geração aleatória de estruturas de classificação uma ferramenta fundamental.

Um dendrograma é o resultado mais comum de uma Classificação Hierárquica Ascendente (C.H.A.) e a sua geração aleatória tem ganho uma atenção crescente com vários algoritmos a serem propostos. O método de Permutação Dupla, proposto por Lapointe e Legendre (1991), o método de Geração Uniforme, apresentado por Sousa (2000), e o método *RA* - Random Agglomeration, proposto por Podani (2000), permitem gerar aleatória e uniformemente dendrogramas ponderados no sentido de Furnas (1984). Sousa (2000) propôs ainda um método de geração aleatória com um Parâmetro de Forma que permite gerar dendrogramas predominantemente de um certo tipo topológico prefixado.

O trabalho aqui apresentado foca-se na geração aleatória de um tipo mais complexo de estruturas de classificação, as pirâmides. A aplicação de um método de Classificação Piramidal Ascendente (C.P.A.), a um conjunto de dados multivariados, produz uma pirâmide como resultado principal. Numa sucessão natural de investigação, Machado e Sousa (2006) propuseram um método de geração aleatória de pirâmides, designado *RAP* - Random Generation Algorithm of Pyramids, que gera pirâmides para um número fixo de nós terminais. No essencial a linha de actuação do algoritmo de geração *RAP* é semelhante à subjacente ao método de C.P.A. proposto por Bertrand (1986).

Na secção seguinte, Secção 2, descrevem-se algumas propriedades que distinguem o modelo hierárquico do modelo piramidal, apresentam-se algumas definições e propriedades de Classificação Piramidal necessárias a uma melhor compreensão das secções seguintes, e referem-se os principais métodos de C.P.A.. Na Secção 3 é detalhado o algoritmo proposto para a geração aleatória de pirâmides e são tecidas algumas considerações sobre a sua performance e limitações. A análise do desempenho do método de geração de pirâmides constitui a Secção 4 e é feita em duas vertentes: i) por enumeração exaustiva dos diferentes

tipos topológicos de pirâmides, para um número fixo de nós terminais, e do respectivo número de pirâmides não isomórficas. Dada a sua complexidade este estudo só foi conseguido para as pirâmides com 3 ou 4 nós terminais. ii) por recurso à simulação, abordagem para pirâmides com qualquer número de nós terminais.

## 2 Classificação Piramidal Ascendente (CPA)

As pirâmides, resultado de uma Classificação Piramidal (Diday (1984) e Bertrand (1986)), são uma generalização das hierarquias, resultado de uma Classificação Hierárquica. Uma pirâmide é uma colecção de classes de um conjunto de elementos a classificar, com propriedades mais específicas e complexas do que uma hierarquia. Pode afirmar-se que duas características principais distinguem estes dois procedimentos de classificação. Numa classificação piramidal um elemento, num mesmo patamar, pode pertencer no máximo a duas classes, já numa classificação hierárquica pertence a uma só classe. Outra característica relevante de uma classificação piramidal é a capacidade de produzir um número pequeno de ordens, sobre os elementos a classificar (Bertrand (1986) e Mfoumoune (1998)).

Métodos associados a outros tipos de estruturas de classificação têm sido desenvolvidos. Entre estes referem-se os que dão origem às hierarquias 2-3 (Bertrand (2002)) e às hierarquias fracas (Bandelt e Dress (1989)).

### 2.1 Modelo Piramidal como uma extensão do Modelo Hierárquico

Seja  $E$  um conjunto de  $n$  elementos a classificar. Como resultado da aplicação de um método de Classificação Hierárquica Ascendente tem-se uma hierarquia indiciada, um dendrograma (Figura 1(a)) ou uma matriz ultramétrica. Uma hierarquia é um conjunto de partições encaixadas. Formalmente, diz-se que  $H$ ,  $H \in P(E)$  (conjunto de partes não vazias de  $E$ ), é uma **hierarquia** sobre  $E$  se:

- (i)  $E \in H$ ;
- (ii)  $\{a\} \in H, \forall a \in E$ ;
- (iii)  $\forall h, h' \in H, h \cap h' = \{\emptyset, h, h'\}$ .

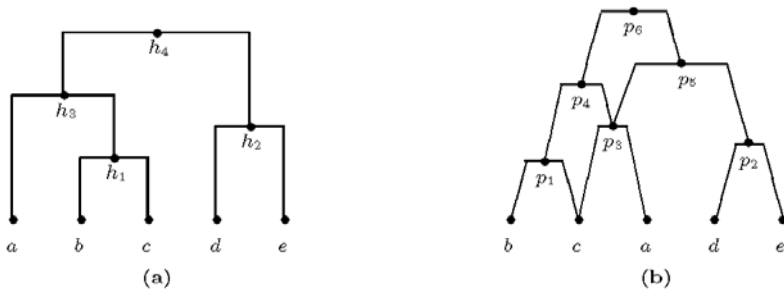
Define-se  $(H, f)$  como uma **hierarquia indiciada** sobre  $E$  se  $H$  é uma hierarquia e  $f: H \rightarrow \mathbb{R}_0^+$  é uma aplicação em que  $\forall h, h' \in H$  se tem:

$$f(h) = 0 \Leftrightarrow \text{card}(h) = 1 \text{ e } \forall h, h' \in H, h \subset h' \Rightarrow f(h) \leq f(h').$$

Dendrograma é a representação mais comum de uma classificação hierárquica e será aqui entendido como uma árvore ponderada com raiz, terminalmente

etiquetada (a cada nó terminal está associado um elemento do conjunto a classificar) em que todos os nós terminais estão à mesma distância da raiz. A cada nó interno de um dendrograma ponderado está associado um valor numérico de uma variável contínua, designada índice de nível. A distribuição desta variável é, em geral, desconhecida e depende das duas escolhas subjacentes ao processo de classificação hierárquica: a função de comparação entre pares de elementos do conjunto a classificar e a função de comparação entre classes associada ao método de agregação. A utilização dos valores dos índices de nível tem sido frequentemente questionada, sendo apontado como preferencial o uso dos respectivos valores ordinais. Esta opção corresponde a trabalhar com dendrogramas completamente ordenados ou dendrogramas invariantes de ordem global (*GOI* – Global Order Invariant).

A matriz ultramétrica é uma matriz quadrada, simétrica, de dimensão  $n = \text{card}(E)$ , em que o seu elemento  $(i, j)$  é dado pelo valor da função  $f$  para o menor subconjunto (classe) que contém os elementos de ordem  $i$  e  $j$  de  $E$ . A hierarquia indiciada, a matriz ultramétrica e o dendrograma contêm a mesma informação.



**Figura 1** – Hierarquia vs pirâmide.

Analogamente três entidades caracterizam o resultado de um método de classificação piramidal: uma pirâmide indiciada, a representação piramidal (Figura 1(b)) e a matriz de Robinson. Uma cobertura de  $E$  é um conjunto de classes não vazias cuja reunião é  $E$ . Uma **pirâmide** é uma sucessão de coberturas encaixadas. Diz-se que  $P, P \in P(E)$ , é uma pirâmide sobre  $E$  se:

- (i)  $E \in P$ ;
- (ii)  $\{a\} \in P, \forall a \in E$ ;
- (iii)  $\forall p, p' \in P, p \cap p' = \emptyset$  ou  $p \cap p' \in P$ ;
- (iv) existe uma ordem  $\theta$ , compatível com  $P$ .

Define-se  $p \in P$  como um **sucessor** de  $p' \in P$  ( $p'$  diz-se um **predecessor** de  $p$ ) se  $p \subseteq p'$  e não existe  $p'' \in P: p \subseteq p'' \subseteq p'$ . Na Figura 1 (b) pode verificar-se que  $p_3$  tem dois predecessores  $p_4$  e  $p_5$ . Uma das propriedades que distingue uma hierarquia de uma pirâmide é que numa pirâmide uma classe admite, no máximo,

dois predecessores enquanto que uma hierarquia admite, no máximo, um predecessor. Note-se também que uma partição é um caso particular de uma cobertura, o que permite concluir que o conjunto das hierarquias está incluído no conjunto das pirâmides. Uma **pirâmide indiciada** sobre  $E$  é um par  $(P, f)$  onde  $P$  é uma pirâmide e  $f: P \rightarrow \mathbb{R}_0^+$  tal que:

$$\forall p, p' \in P, f(p) = 0 \Leftrightarrow \exists a \in E: p = \{a\} \text{ e } p \subset p' \Rightarrow f(p) \leq f(p').$$

Os valores de  $f$  são designados **índices de nível** e os valores ordinais correspondentes **níveis de agregação**.

Uma ordem  $\theta$  definida sobre  $E$  diz-se compatível com um índice de dissemelhança  $d$ , dito piramidal, se

$$\forall x, y, z \in E, x <_{\theta} y <_{\theta} z \Rightarrow d(x, z) \geq \max\{d(x, y), d(y, z)\}.$$

Quando associada uma ordem compatível sobre os elementos uma classificação piramidal define uma matriz, denominada matriz de Robinson. Nesta matriz verifica-se a seguinte propriedade: os termos das linhas e das colunas nunca diminuem quando nos afastamos da diagonal principal, em qualquer das direcções.

## 2.2 Algumas definições

A Classificação Piramidal tem por base um vasto conjunto de definições, cuja apresentação formal não é aqui possível tendo em conta as limitações na extensão do trabalho. Assim optou-se por uma apresentação mais leve, ilustrando a partir de exemplos da Figura 2., remetendo para Bertrand (1986) e Mfoumoune (1998) a consulta dessas definições matemáticas.

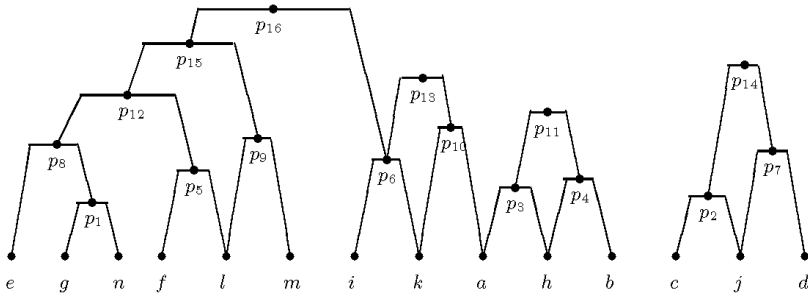
Sejam  $E = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n\}$  um conjunto de 14 elementos a classificar,  $P$  uma pirâmide incompleta (uma vez que  $E \notin P$ ),  $p, q \in P$  e  $\theta$  uma ordem compatível com  $P$ , neste exemplo  $\theta = (e, g, n, f, l, m, i, k, a, h, b, c, j, d)$ .

Designa-se por  $C_p$  a **componente conexa** de  $p$ , por  $\min(p)$  o **menor elemento** da classe  $p$  e por  $\max(p)$  o **maior elemento** da classe  $p$ , segundo a ordem compatível  $\theta$ . Na Figura 2, para a classe  $p_{13}$ , tem-se que  $C_{p_{13}}$  é a componente conexa constituída pelos elementos  $\{e, g, n, f, l, m, i, k, a, h, b\}$ ,  $\min(p_{13}) = i$  e  $\max(p_{13}) = a$ .

Uma classe diz-se **singular** ou **terminal** se não possui nenhum sucessor e **maximal** se não possui nenhum predecessor. Na Figura 2,  $\{a\}, \dots, \{n\}$  são as classes singulares e  $p_{16}$ ,  $p_{13}$ ,  $p_{11}$  e  $p_{14}$  são as classes maximais de  $P$ .

Diz-se que uma classe  $p$  é um **descendente** da classe  $q$  se e só se  $p \subset q$ . O conjunto dos descendentes de  $p$  designa-se por  $D_p$ . Uma classe  $p$  é um **ascendente** da classe  $q$  se e só se  $p \supset q$ . O conjunto dos ascendentes de  $p$  designa-se por  $A_p$ . Na Figura 2,  $D_{p_{13}} = \{p_6, p_{10}, \{i\}, \{k\}, \{a\}\}$  e  $A_{p_5} = \{p_{12}, p_{15}, p_{16}\}$ .

Diz-se que  $p$  **está antes** de  $q$  se  $(\min(p) <_{\theta} \min(q) \text{ e } \max(p) <_{\theta} \max(q))$  ou  $p = q$ . Se  $p$  está antes de  $q$ , diz-se também que  $q$  **está depois** de  $p$ . Diz-se ainda que  $p$  é **interior** a  $q$  se  $\min(q) <_{\theta} \min(p) \text{ e } \max(p) <_{\theta} \max(q)$ . Na Figura 2, a classe  $p_5$  é interior a  $p_{15}$ .



**Figura 2** – Pirâmide incompleta para ilustração de conceitos.

Uma classe  $p \in P$  é definida como **uma classe extrema** se e só se  $\min(p) = \min(C_p)$  ou  $\max(p) = \max(C_p)$ . Por exemplo, na pirâmide da Figura 2  $\{e\}, p_8, p_{12}, p_{15}, p_{16}, p_{11}, p_4, \{b\}, \{c\}, p_2, p_{14}, p_7, \{d\}$  são classes extremas.

Para **classe livre** considera-se a definição dada por Mfoumoune (1998), na qual uma classe é livre se e só se: (i)  $p$  tem um só predecessor; (ii) todas as classes ascendentes de  $p$  têm, no máximo, um só predecessor; (iii) todas as classes descendentes cujo menor ou maior elemento coincide com o da classe  $p$  têm um só predecessor; (iv) seja  $s_p$  o mais pequeno ascendente extremo de  $p$  (no sentido de inclusão), todos os descendentes de  $s_p$  com o mesmo elemento mínimo ou máximo têm um só predecessor. Na Figura 2, a classe  $p_9$  é livre e a classe  $p_5$  não.

Seja  $B(p)$  o conjunto das classes com o mesmo elemento mínimo ou (exclusivamente) máximo que  $p$  e que não tenham ascendente possuindo dois predecessores. Designa-se **vale** de duas classes  $p$  e  $q$  o conjunto  $B(p) \cup B(q)$  tal que  $\forall x \in B(p) \text{ e } \forall y \in B(q) \text{ se tem } x \cap y = p \cap q$ . Nenhum ascendente de  $p \cap q$  possui mais do que um predecessor.  $V(p, q)$  representa o vale das classes  $p$  e  $q$ . Na Figura 2  $V(p_3, p_{10}) = \{p_3, p_{10}, p_{11}, p_{13}\}$ .

Para **classe activa** apresenta-se uma definição não formal, que pode ser compreendida de uma forma intuitiva. Em classificação ascendente, entende-se por classe activa uma classe que pode ser agregada em iterações seguintes.

## 2.3 Algoritmos de Classificação Piramidal Ascendente (CPA)

Nesta subsecção faz-se referência à evolução de alguns algoritmos de Classificação Piramidal Ascendente (CPA). Diday (1984) propôs um princípio fundamental para

a construção de pirâmides, que serviu de ponto de partida aos outros algoritmos do tipo ascendente aplicados à Classificação Piramidal. Bertrand (1986) formalizou as condições de agregação das classes, que se apoiam nas condições físicas dos elementos das classes, não tendo contudo em conta as classes livres na construção de uma pirâmide. Mfoumoune (1998) propôs uma nova implementação do algoritmo de CPA, designado por *QuikCAP*, que integra dados numéricos e simbólicos (Brito (1991)) e utiliza uma metodologia ligeiramente diferente, assente na *Seleção – Agregação – Eliminação*. Mostra que desta maneira consegue-se, por um lado, diminuir a complexidade algorítmica e, por outro lado, obter resultados mais satisfatórios.

## 2.4 Inversão parcial e total de componentes conexas

A construção de uma pirâmide implica a resolução de numerosos problemas algorítmicos, em particular, os algoritmos de inversão parcial ou total de componentes conexas. Com efeito, a agregação de duas classes ainda não agregadas atribui uma ordem arbitrária entre os elementos. Esta ordem pode ser revista numa segunda agregação de uma das classes ou de um *parente* desta (ascendente ou descendente). A revisão desta ordem consiste em inverter parcial ou globalmente uma, ou outra, ou as duas componentes conexas da pirâmide em construção. Como exemplo, considere-se a agregação das classes  $p_1$  e  $\{c\}$  na pirâmide incompleta  $P$  da Figura 2. A sua agregação implica a inversão total da componente conexa  $C_{\{c\}}$  e a inversão parcial de  $C_{p_1}$ , isto é, dos elementos da classe  $p_8$ , colocando os elementos de  $C_{\{c\}}$  antes de  $C_{p_1}$ .

## 3 Geração aleatória de pirâmides

A geração aleatória de estruturas de classificação é uma ferramenta de grande utilidade na avaliação do desempenho de métodos classificatórios, temática que se enquadra na Validação em Classificação.

No que respeita aos dendrogramas, vários trabalhos propondo métodos para a geração deste tipo de estruturas foram desenvolvidos: o método de Permutação Dupla, proposto por Lapointe e Legendre (1991), o método de Geração Uniforme, apresentado por Sousa (2000), e o método *RA* (*Random Agglomeration*), desenvolvido por Podani (2000). Estes métodos geram dendrogramas aleatória e uniformemente no sentido de Furnas (1984). Este conceito significa que, sendo  $d_n$  o número de dendrogramas distintos com  $n$  nós terminais, cada um destes  $d_n$  dendrogramas é gerado equiprovavelmente, isto é, com probabilidade  $\frac{1}{d_n}$ . Estudos analíticos e de simulação, Sousa (2000) e Tendeiro (2005), mostraram que estes três métodos têm desempenhos muito semelhantes.



O método de geração aleatória de pirâmides proposto por Machado e Sousa (2006), denominado *Random Generation Algorithm of Pyramids (RAP)* actua de modo muito semelhante ao algoritmo de C.P.A. proposto por Bertrand (1986), em que o par de classes a agregar é gerado aleatoriamente. Pode ser entendido como uma extensão, para as pirâmides, do método *RA* proposto por Podani (2000) para dendrogramas. Para definições formais de *condições de agregação* e *classes activas*, utilizadas no método proposto, remete-se para Machado (2007). O algoritmo *RAP* foi implementado em linguagem Matlab e gera pirâmides para qualquer número de nós terminais.

### 3.1 O algoritmo *RAP*

Apresentam-se, de uma forma simplificada, os passos principais do algoritmo *RAP* de acordo com a seguinte notação:

- $c_{act}^0$  - vector de classes com  $n$  nós terminais ( $n \geq 1$ ), numerados de 1 até  $n$ ;
- $c_{act}^i$  - vector das classes activas na iteração  $i$  ;
- $c_j^i$  ( $j = 1, 2$ ) - classe  $j$  gerada na iteração  $i$  ;
- $c_{agr}^i$  - vector de possíveis classes agregáveis com  $c_1^i$  ;
- $c^i$  - nova classe formada na iteração  $i$ ;
- $\theta$  - vector de dimensão  $n$  associado à ordem compatível sobre os elementos (por abuso de linguagem está a usar-se a mesma letra para a ordem e o vector associado);
- $CC$  - matriz associada às componentes conexas;
- $P$  - matriz associada à pirâmide. Formada uma classe  $k$ , com  $k \geq n + 1$ , é acrescentada uma nova linha à matriz  $P$ ;
- $M$  - matriz  $n \times n$ , inicialmente com todas as entradas nulas, que vai sendo preenchida passo a passo e, no final, quando associada a ordem  $\theta$ , será uma matriz de Robinson.

**Iteração 0:** é introduzido o número de nós terminais,  $n$ , e são inicializadas as variáveis  $c_{act}^0$ ,  $CC$ ,  $\theta$ ,  $P$  e  $M$ .

Para  $i$  de 1 até à paragem do algoritmo (o algoritmo pára quando a classe que contém todos os elementos é formada):

**Iteração  $i$ :**

1. geração aleatória e uniforme do par  $(c_1^i, c_2^i)$  de classes a agregar:
  - geração aleatória e uniforme de um elemento do vector  $c_{act}^{i-1}$ , a classe  $c_1^i$ ;

- formação do vector de classes agregáveis,  $c_{agr}^i$ : é construído um vector com as classes que são agregáveis com  $c_1^i$ , verificando as condições de agregação (ver Bertrand (1986) e Machado (2007));
- geração aleatória e uniforme de um elemento do vector  $c_{agr}^i$ , a classe  $c_2^i$ .

2. actualizações: ordem  $\theta$ , matriz  $P$ , matriz associada à pirâmide  $P$ , matriz  $M$  e vector das classes activas  $c_{act}^i$ .

**Dados de saída:** ordem  $\theta$ , matriz  $P$  associada à pirâmide e matriz de Robinson  $MR = M(\theta)$ .

### 3.2 Algumas limitações do algoritmo *RAP*

O algoritmo *RAP*, no seu processo iterativo (da base até ao topo), apresenta fundamentalmente duas limitações que estão identificadas em Machado (2007): a geração do par de classes não é obtida de forma *completamente aleatória* e uniforme do conjunto de todos os pares de classes agregáveis e as classes livres são eliminadas das classes activas por serem nós interiores.

No algoritmo *RAP* o par de classes a agregar não é produzido de uma forma uniforme, isto é, os pares possíveis de classes a agregar não têm igual probabilidade. Por exemplo, na Figura 3, os pares de classes possíveis para agregar são (1,3), (1,4), (2,3), (2,4), (3,4), (3,5) e (4,5). Portanto, numa distribuição uniforme, o par (3,5) tem probabilidade  $\frac{1}{7}$  de ser gerado. Com o algoritmo *RAP* a probabilidade deste par ser obtido é de  $\frac{3}{20}$ , que resulta da primeira classe a ser gerada ser a 3 e a segunda a 5, ou a primeira ser a 5 e a segunda a 3. Então, a probabilidade de gerar o par (3,5), usando o algoritmo *RAP*, é  $\frac{1}{5} \times \frac{1}{4} \times \frac{1}{5} \times \frac{1}{2} = \frac{3}{20}$ .

A agregação de duas classes atribui normalmente uma ordem arbitrária entre os elementos, que pode ser revista em agregações futuras. Uma inversão parcial na ordem dos elementos não é efectuada no algoritmo *RAP* e, por isso, classe livres não são contempladas como activas. Por exemplo, na Figura 3, a agregação das classes 3 e 5 forma a classe 6. Com o algoritmo *RAP*, após este momento, a classe 2 é excluída por ser uma classe interior. No entanto, uma inversão parcial na nova componente conexa formada (constituída pelo elementos 1, 2 e 3), isto é, uma inversão na ordem dos elementos da classe 5, permite que a classe 2 esteja activa para agregação nas iterações seguintes. Seria então desejável que o algoritmo considerasse classes activas, para além das classes extremas, das classes maximais e das classes pertencentes a um vale de duas classes maximais consecutivas (da mesma componente conexa), as classes livres que se transformam em classes extremas com uma inversão parcial da componente conexa à qual pertence.

Um novo algoritmo, adaptado do algoritmo *RAP*, deverá ser desenvolvido com o objectivo de remover algumas das suas limitações, de modo a obter resultados

mais satisfatórios. Machado (2007) lança alguns dos princípios para a implementação de um novo algoritmo de geração aleatória de pirâmides que permita implementar melhorias, à semelhança dos progressivos melhoramentos que se observaram nos algoritmos de C.P.A., nomeadamente com o método *QuikCAP* proposto por Mfoumoune (1998).



**Figura 3** – Hierarquia vs pirâmide.

Dois aspectos principais deverão caracterizar este novo algoritmo de geração aleatória de pirâmides: uma identificação integral de todos os pares de classes agregáveis e uma gestão eficaz e rigorosa da matriz que contém todos os pares de classes agregáveis. Desta forma, o objectivo será acelerar o procedimento de procura do par de classes a agregar em cada etapa da construção ascendente da pirâmide. Há assim necessidade de eliminar todos os pares de classes que se tornam não agregáveis à medida que se formam novas classes e ainda deduzir as classes agregáveis com a nova classe formada.

## 4 Simulação e discussão

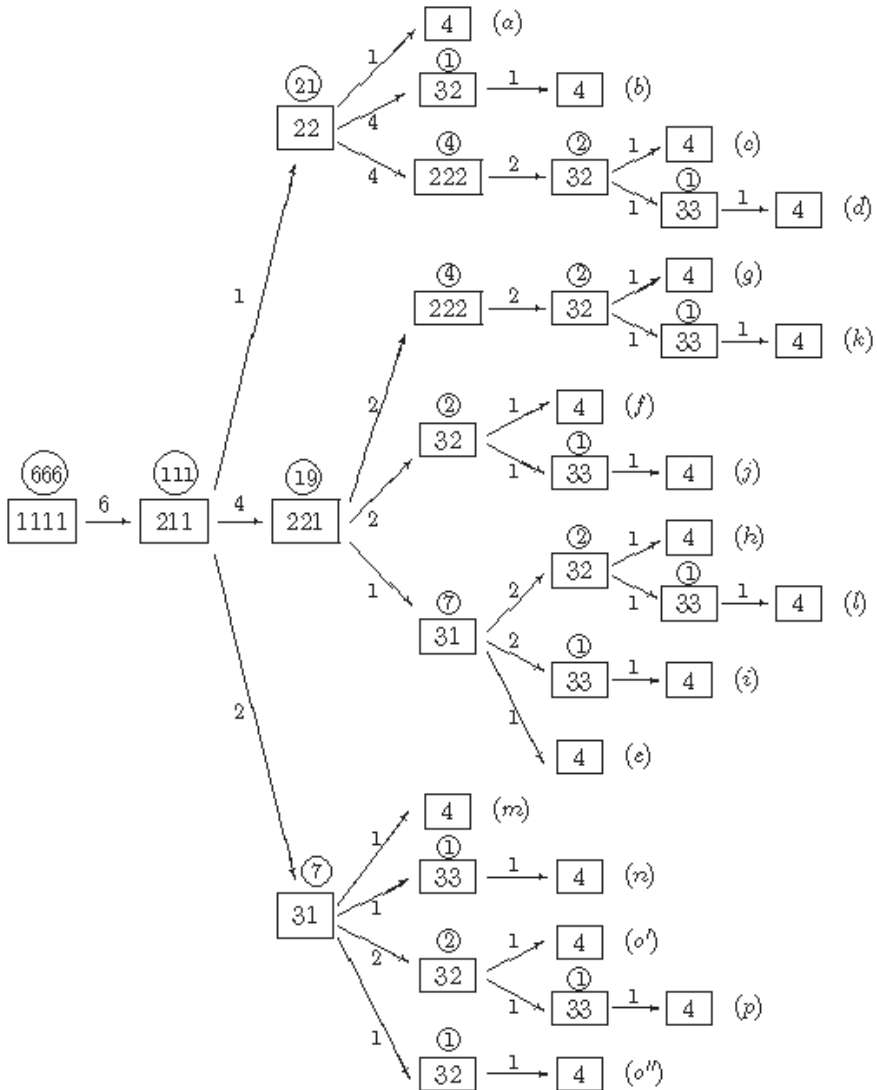
Para avaliar o desempenho do algoritmo desenvolvido realizaram-se estudos teóricos e de simulação. Para um número de nós terminais,  $n$ , os estudos teóricos incluem: a determinação do número de diferentes tipos topológicos de uma pirâmide; o número de pirâmides não isomórficas; a distribuição do número de pirâmides não isomórficas pelos diferentes tipos topológicos; a probabilidade de obter uma pirâmide para cada tipo topológico.

No caso particular dos dendrogramas, a determinação do número de tipos topológicos e de dendrogramas não isomórficos para qualquer número de nós terminais está facilitada pela existência de fórmulas matemáticas. No caso das pirâmides estas fórmulas não existem e o seu estudo é bastante mais complexo, pelo que se procurou concretizá-lo, para pequenos valores de  $n$ , por recurso a ferramentas de cálculo combinatório. Para pirâmides com 4 nós terminais ( $n = 3$  é um caso trivial) este estudo é já bastante complexo, a sua extensão a valores de  $n$  superiores poderá ser conseguida, implicando contudo um procedimento moroso.

Na Figura 4 apresenta-se um esquema de contagem do número de pirâmides distintas que é possível ter com 4 nós terminais.

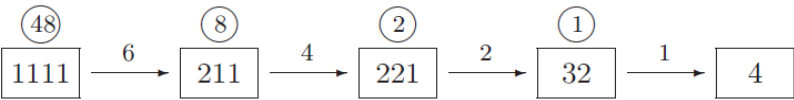
Inicialmente tem-se quatro classes singulares, representadas por  $[1111]$ . A primeira agregação pode ser feita de 6 formas distintas (combinações de 4 a 2),

dando origem a uma nova classe com dois elementos. Usar-se-á a notação  $\boxed{211}$  para indicar que existe agora uma classe com dois elementos e duas classes singulares ainda não agregadas. A explicação dos vários caminhos é idêntica, pelo que será apenas exemplificado para um deles, o caminho  $(f)$  que se encontra esquematizado na Figura 5.



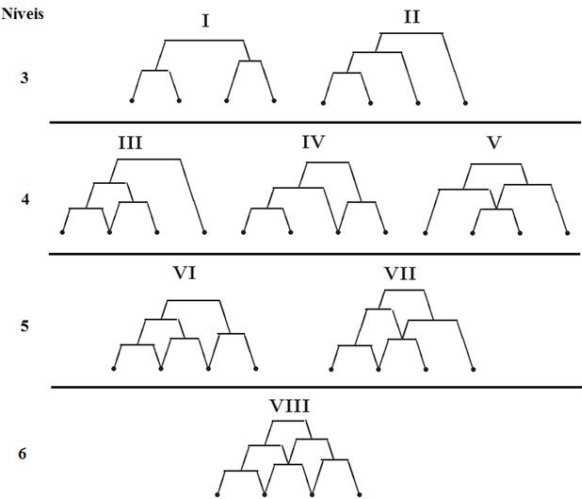
**Figura 4** – Representação esquemática do número de pirâmides para  $n = 4$ .

De [211] para [221] há 4 formas distintas de o fazer. De [221] para [32] existem 2 agregações possíveis e, por fim, de [32] para [4] a agregação é feita de forma única. O número de pirâmides binárias não isomórficas distintas para este caminho é agora dado pelo produto  $1 \times 2 \times 4 \times 6 = 48$ . Em Machado (2007) este estudo é mais detalhado e, para  $n = 4$ , contabilizaram-se 666 pirâmides não isomórficas e oito tipos topológicos distintos foram identificados.



**Figura 5** – Esquema de contagem do número de pirâmides para o caminho (f) da Figura 4.

Na Figura 6 pode observar-se a representação piramidal dos oito tipos topológicos, por níveis de agregação, das pirâmides com 4 nós terminais e na Tabela 1 o número de pirâmides não isomórficas por tipo topológico.



**Figura 6** – Topologias das pirâmides para  $n = 4$ .

**Tabela 1** – Número de pirâmides para  $n = 4$ , por tipo topológico.

Tipos topológicos	I	II	III	IV	V	VI	VII	VIII
Número de pirâmides	6	12	24	108	12	192	120	192

Na Tabela 2 pode observar-se a probabilidade de obter uma pirâmide com 4 nós terminais para cada tipo topológico. O algoritmo *RAP* foi usado para simular pirâmides aleatoriamente, identificando o seu tipo topológico. Na Tabela 2 apresentam-se as frequências observadas para as oito topologias, com base na simulação de 100000 pirâmides.

**Tabela 2** – Distribuições teórica e empírica das pirâmides para  $n = 4$ , por tipo topológico.

Tipo topológico	I	II	III	IV	V	VI	VII	VIII
Probabilidade	0.016	0.057	0.023	0.292	0.057	0.169	0.217	0.169
Frequência <sup>(*)</sup>	0.011	0.076	0.024	0.245	0.077	0.195	0.175	0.197

(\*) Algoritmo *RAP*

Analisando os resultados teóricos e de simulação conclui-se que, para alguns tipos topológicos, os valores de probabilidade e de frequência observada com o algoritmo *RAP* se afastam. Este facto é explicado pelas limitações já referidas do algoritmo e torna pertinente o desenvolvimento de novos algoritmos de geração aleatória de pirâmides. Um estudo sobre a identificação topológica das estruturas piramidais para valores um pouco superiores de  $n$  surge também como um desafio futuro.

## Referências

- BANDELT, H., DRESS, A., (1989). Weak hierarchies associated with similarity measures: an additive clustering technique, *Bull. Math. Biol.*, 51, 133–166.
- BERTRAND, P. (1986). *Etude de la représentation pyramidale*. 3th Cycle Thesis, Université Paris IX-Dauphine.
- BERTRAND, P. (2002). Les 2-3 Hiérarchies: une structure de classification pyramidale parcimonieuse. *Actes du IXème Congrès de la Société Francophone de Classification*, 16-18 Septembre, Toulouse, France.
- BRITO, P. (1991). *Analyse de données symboliques: Pyramides d'heritage*. 3th Cycle Thesis, Université Paris IX-Dauphine.

- DIDAY, E. (1984). Une représentation visuelle des classes empiétantes: les pyramides. *Rapport de recherche I.N.R.I.A.* n. 291, Rocquencourt, France.
- FURNAS, G.W. (1984). The generation of random, binary unordered trees. *Journal of Classification* 1, 187-233.
- LAPOINTE, F., LEGENDRE, P. (1991). The generation of random ultrametric matrices representing dendrograms. *Journal of Classification* 8, 177-200.
- MACHADO, V., SOUSA, F. (2006). Geração Aleatória de uma Estrutura Classificatória Piramidal. *Actas do XIII Congresso Anual da SPE*.
- MACHADO, V. (2007): *Geração aleatória de estruturas classificatórias*. Master Thesis, Faculdade de Engenharia, Universidade do Porto.
- MFOUMOUNE, E. (1998). *Les aspects algorithmiques de la Classification Ascendante Pyramidale et Incrémentale*. PhD Thesis, Université Paris IX-Dauphine.
- PODANI, J. (2000). Simulation of random dendrograms and comparison tests: some comments. *Journal of Classification* 17, 123-142.
- SOUSA, F. (2000). *Novas metodologias e validação em classificação hierárquica ascendente*. PhD Thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.
- TENDEIRO, J. (2005). *Comparação de dendrogramas: obtenção de distribuições empíricas de alguns coeficientes*. Master Thesis, Faculdade de Engenharia, Universidade do Porto.

# Coeficientes de Comparação de Partições em Análise Classificatória: Abordagem Clássica vs uma Abordagem Probabilística

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

© The Author(s) 2013

**Resumo** Existem diversos índices para a comparação de partições, o que dificulta a tomada de decisão, dado que diferentes índices põem geralmente em evidência diferentes peculiaridades das partições a comparar. Com o intuito de auxiliar nessa avaliação, é apresentada e exemplificada uma abordagem para a comparação de partições, com base na semelhança *VL* (*Validade da Ligação*), a qual tem, entre outras, a vantagem de uniformizar a escala de medida. A comparação dos resultados obtidos com a aplicação de índices de comparação de partições, clássicos e probabilísticos do tipo *VL*, é feita sobre um conjunto de dados reais.

**Palavras-chave:** Análise classificatória hierárquica, comparação de partições, coeficiente de afinidade, metodologia *VL*.

## 1 Introdução

A Análise Classificatória (*Cluster Analysis*) tem como objetivo identificar grupos (classes ou *clusters*) de entidades (indivíduos, objetos, etc.), relativamente homogêneos e bem separados, com base nas semelhanças ou dissimilaridades entre essas entidades.

Existem diversos índices para a comparação de partições, o que dificulta a tomada de decisão, dado que diferentes índices avaliam geralmente diferentes peculiaridades das partições a comparar. Por outro lado, existe uma grande diversidade de técnicas de validação capazes de auxiliar na escolha da melhor

---

<sup>1</sup> Departamento de Matemática e CMATI, Universidade dos Açores, osilva@uac.pt

<sup>2</sup> Faculdade de Psicologia e LEAD, Universidade de Lisboa, hbacelar@fp.ul.pt

<sup>3</sup> Departamento de Matemática, FCT, Universidade Nova de Lisboa, geral@datascience.org

<sup>4</sup> Departamento de Matemática, CEEAplA e CMATI, Universidade dos Açores, aurea@uac.pt



partição dos elementos a classificar, mas, em geral, cada uma apresenta tendência para favorecer um determinado tipo de algoritmo. Assim, é imprescindível encontrar formas de comparar os resultados obtidos usando diferentes abordagens.

Na Secção 2 faz-se uma introdução aos índices para a comparação de partições, com recurso a coeficientes clássicos. A Secção 3 é dedicada à comparação de partições com recurso a coeficientes probabilísticos do tipo *VL*. Na Secção 4 são comparados os resultados obtidos com a aplicação das duas abordagens, clássica e probabilística, a um conjunto de dados reais, no âmbito de um trabalho mais vasto de validação em Análise Classificatória, com recurso a métodos de reamostragem. Finalmente, na Secção 5, são apresentadas as principais conclusões.

## 2 Coeficientes para a comparação de pares de partições

A comparação de duas partições, no âmbito da Análise Classificatória, pode ser efetuada usando diversos índices ou coeficientes clássicos no contexto de três abordagens (com base, respetivamente, na contagem de pares, no emparelhamento das classes e na variação da informação). No entanto, cada um desses coeficientes assume um determinado valor, em função da sua expressão analítica, e alguns apresentam intervalos de variação diferentes ou não variam no intervalo previsto mas somente num subintervalo desse intervalo. Para que esses coeficientes sejam mais facilmente comparáveis, deve-se ter em atenção as suas características intrínsecas, categorizando-os em grupos com características similares.

Para comparar duas partições,  $P$  e  $P'$ , de um mesmo conjunto de dados, de cardinal  $n$ , com base na contagem de pares, pode-se começar por construir uma tabela de contingência  $2 \times 2$  associada, tal como a Tabela 1.

**Tabela 1** - Tabela de contingência com base na contagem de pares.

Partição $P$	Partição $P'$	
	$a$	$b$
	$c$	$d$

A Tabela 1 refere os pares de elementos que existem nas duas partições, onde “ $a$ ” é o número de pares de elementos que estão nas mesmas classes em ambas as partições, “ $b$ ” é o número de pares de elementos que pertencem às mesmas classes na partição  $P$  mas a diferentes classes na partição  $P'$ , “ $c$ ” é o número de pares que pertencem a diferentes classes na partição  $P$  e às mesmas classes na partição  $P'$  e “ $d$ ” é o número de pares de elementos que pertencem a classes diferentes em ambas as partições. O número total de pares de objetos é  $a+b+c+d=n \times (n-1)/2$ .

O Anexo B contém uma lista de índices para a comparação de dados binários, os quais são funções dos quatro valores da Tabela 1 e são também usados para a comparação de partições. Nessa lista foi considerada a subdivisão desses índices em coeficientes de semelhança que consideram a ausência conjunta “ $d$ ”, coeficientes de semelhança que não consideram a ausência conjunta “ $d$ ” e outros coeficientes de associação. Para cada um dos coeficientes, é apresentada a respetiva fórmula, o símbolo com que é designado, o seu intervalo de variação e o(s) autor(es).

Estes índices devem ser avaliados, tendo por base propriedades comuns, e podem ser sensíveis ao número de classes nas partições. Alguns dos índices (por exemplo, os de Hubbert e Rand) têm tendência para tomar valores elevados no caso de partições com mais classes, outros no caso de partições com um pequeno número de classes (por exemplo o de Jaccard). O índice de Rand ajustado não tem nenhuma destas características indesejáveis (Milligan e Cooper, 1985; Jain e Dubes, 1988), razão pela qual este é um dos índices englobados na metodologia utilizada no âmbito deste trabalho. Também o índice de Ochiai (caso particular do coeficiente de afinidade) centrado e reduzido tem sido usado, com bons resultados, no âmbito da comparação de partições (Silva, 2004).

Como referimos atrás, a avaliação dos índices de comparação de partições, com base na contagem de pares, deve ter em atenção a escala de variação e a relação que se pode estabelecer entre os diversos índices a partir da sua fórmula de cálculo. Vários estudos de comparação e classificação destes coeficientes têm sido propostos por diversos autores, desde Sneath e Sokal (1963). Sibson (1972) fez o agrupamento dos coeficientes em classes monotónicas, estabelecendo uma relação de equivalência no conjunto dos coeficientes de comparação para dados binários. Bacelar-Nicolau (1980, 1987) determinou classes de coeficientes “*distribucionalmente equivalentes*”, conceito que vamos utilizar neste trabalho, conforme se refere na próxima secção.

### **3 Comparação de pares de partições com recurso a coeficientes probabilísticos**

Lerman (1970) propôs a utilização de um coeficiente de semelhança de natureza probabilística entre variáveis binárias, que depois generalizou a coeficientes de proximidade entre estruturas do mesmo tipo (Lerman, 1981). Bacelar-Nicolau (1980, 1987) desenvolveu um estudo distribucional dos coeficientes de comparação para dados binários, tendo verificado e comprovado a equivalência distribucional de uma vasta classe de coeficientes, sob a hipótese de margens fixas da tabela de contingência  $2 \times 2$  associada a cada par de elementos do conjunto a classificar. Para outros coeficientes, bem como na hipótese de margens livres, embora não se verifique a equivalência distribucional exata, podemos encontrar classes de coeficientes equivalentes no que respeita à sua distribuição assintótica, e tomar sempre, como informação associada a um coeficiente, a sua função de

distribuição limite (Bacelar-Nicolau, 1980, 1987; Lerman, 1981), que é um coeficiente de semelhança probabilístico  $\gamma$  ou da *Validade da Ligação*, *VL*. Tem-se então, para um coeficiente de semelhança  $S$ :

$$\gamma = F_s(s) = Prob_{H_0}(S \leq s) \equiv Prob_{H_0}(S^* \leq s^*) \equiv \phi(s^*)$$

onde  $H_0$ , é uma hipótese de referência adequada,  $F_s$  é a função de distribuição de  $S$ ,  $S^* = (S - E(S))/\sigma_S$ ,  $s^*$  é uma realização de  $S^*$ ,  $\phi$  é a função de distribuição da lei normal reduzida e  $E(S)$  e  $\sigma_s$  são, respetivamente, o valor médio e o desvio padrão de  $S$ , geralmente assintóticos. O coeficiente probabilístico assume valores em  $[0,1]$  (segue a distribuição Uniforme  $(0, 1)$ ) e, em geral, é calculado assintoticamente, porque a função de distribuição exata de  $S$  pode não ser conhecida. O coeficiente *VL* foi posteriormente estendido a outros tipos de dados e a misturas de dados de diferentes tipos (e.g. Bacelar-Nicolau, 1988, Nicolau, 1983; Nicolau e Bacelar-Nicolau, 1998; Bacelar-Nicolau *et al*, 2009, 2010).

A abordagem à comparação de partições, com recurso a coeficientes probabilísticos do tipo *VL*, apoia-se nos estudos relativos aos coeficientes de comparação para dados binários de Bacelar-Nicolau e processa-se do seguinte modo:

i) Parte-se de um índice de semelhança  $S$  para a comparação de duas partições  $P$  e  $P'$ , tendo por base a contagem de pares de elementos que existem nas duas partições.

ii) Calcula-se o valor de  $\gamma_{PP'}$  da função de distribuição do índice de semelhança utilizado  $S$  no ponto  $s$ , sob a hipótese de referência considerada:

$$\gamma_{PP'} = F_S(s) = Prob_{H_0}(S \leq s) \equiv Prob_{H_0}(S^* \leq s^*) \equiv \phi(s^*).$$

Duas partições,  $P$  e  $P'$ , serão consideradas tanto mais concordantes quanto maior for o valor de  $F_s(s)$ , ou seja, quanto mais improvável for ultrapassar a realização  $s$  de  $S$  na hipótese de referência.

Como tem sido sublinhado por vários autores (e.g., Lerman, 1973, 1981; Bacelar-Nicolau, 1980, 1987; Jain e Dubes, 1988), os diferentes índices não apresentam todos valores em  $[0, 1]$  e uma parte da semelhança entre as duas partições é atribuída ao acaso. No entanto, demonstra-se que os índices mais utilizados são equivalentes do ponto de vista distribucional (Bacelar-Nicolau, 1980, 1987). A aplicação da metodologia *VL* a estes coeficientes permite obter índices de comparação de partições cujos valores seguem a distribuição Uniforme  $(0, 1)$  e podem ser interpretados numa escala probabilística. Assim, utilizando um coeficiente probabilístico podemos escolher, para efeitos de comparação de partições, unicamente um coeficiente clássico em cada uma das classes de coeficientes (assintoticamente) distribucionalmente equivalentes.

## 4 Comparação de resultados obtidos pelas abordagens clássica e probabilística sobre um conjunto de dados reais

Os dados (amostra de 164 alunos) foram obtidos através de um questionário contendo vinte e duas questões relativas às atitudes/crenças dos estudantes dos cursos da Área de Ciências Sociais e Humanas do Ensino Superior em relação à disciplina de Estatística (Silva *et al.*, 2007). Cada aluno seleccionou uma e uma só de sete possibilidades de resposta para cada uma das questões (1-*discordo totalmente*,..., 4- *não concordo nem discordo*,..., 7- *concordo totalmente*).

Foi efectuada a Análise Classificatória Hierárquica Ascendente (*ACHA*) utilizando o coeficiente de afinidade (e.g., Bacelar-Nicolau, 1985) entre variáveis e os critérios de agregação probabilísticos *AVL*, *AVI* e *AVB*, da *Validade da Ligação* - respectivamente, funções do produto e das médias aritmética e geométrica dos cardinais das classes a reunir (e.g., Lerman, 1981; Nicolau, 1983; Bacelar-Nicolau, 1985; Nicolau e Bacelar-Nicolau, 1998). Os dendrogramas, os quadros contendo os valores dos índices de validação utilizados para a seleção da partição mais significativa, obtida a partir da matriz inicial de dados e a interpretação das classes correspondentes a essa partição, em quatro classes, podem ser encontradas em Silva *et al.* (2007). Verificou-se que a partição mais significativa é a mesma para os três critérios de agregação.

Os resultados, que apresentamos abaixo, foram obtidos com vista à avaliação e comparação de partições com recurso à reamostragem. No presente estudo, tratou-se de avaliar a partição mais significativa fornecida pela *ACHA* dos dados, baseada no coeficiente de afinidade e nos critérios de agregação mencionados acima. A parte do procedimento que aqui nos interessa, pode descrever-se, brevemente, do modo seguinte: 1) a partir dos dados originais foram geradas 50 subamostras, com uma taxa de amostragem definida *a priori* (80%), utilizando a amostragem aleatória simples; 2) aplicou-se o mesmo modelo de classificação hierárquica ascendente às matrizes de dados (subamostras) geradas aleatoriamente, pelo método de simulação de Monte Carlo, para determinar as partições com o mesmo número de classes que a partição mais significativa obtida a partir dos dados originais; 3) comparou-se esta partição com cada uma das partições obtidas em 2), com base na contagem de pares, utilizando cada um dos coeficientes clássicos da lista do Anexo B, ou o coeficiente probabilístico *VL* associado; foram também calculadas estatísticas de localização e de dispersão associadas a cada um dos índices, para analisar o comportamento dos mesmos.

Na Tabela 2 encontram-se os valores de estatísticas sumárias (medidas de tendência central, dispersão e quantis) referentes a coeficientes clássicos (Tabela 2-a) e a coeficientes probabilísticos (Tabela 2-b) na situação onde não é considerada a ausência conjunta “*d*”. No Anexo A são apresentadas tabelas com os valores das mesmas estatísticas, que descrevem as distribuições de amostragem dos coeficientes onde é considerada a ausência conjunta “*d*”, bem como de outros coeficientes de associação, obtidas nas 50 reamostragens. As tabelas A1, A3 e A5

são referentes aos coeficientes de comparação clássicos, enquanto que as tabelas A2, A4 e A6 se reportam aos coeficientes probabilísticos.

**Tabela 2-a)** - Valores de estatísticas sumárias referentes a coeficientes clássicos que não consideram a ausência conjunta “d”.

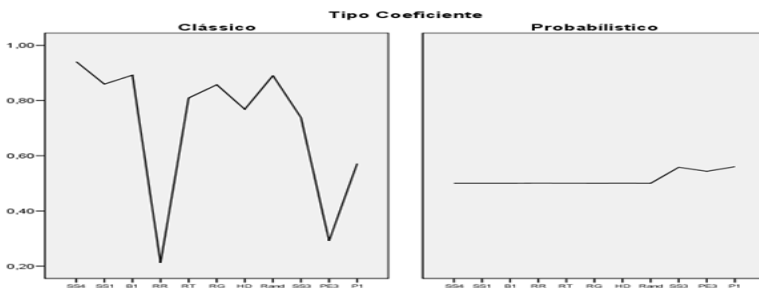
	S	J	O	CZ	K1	DW1	DW2	SS2	BB1	BB2	SO	JO	K2	FMG1
<i>Min</i>	.609	.432	.607	.603	.609	.547	.609	.276	.547	.354	.299	1.219	.761	.544
<i>Max</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938
<i>Media</i>	.955	.908	.941	.941	.942	.934	.950	.870	.929	.479	.891	1.884	2.491	.879
<i>DP</i>	.095	.177	.117	.119	.116	.138	.097	.238	.139	.043	.222	.233	3.769	.118
<i>Centro</i>	.804	.716	.803	.801	.804	.773	.804	.638	.773	.427	.650	1.609	5.505	.741
<i>.005</i>	.609	.432	.607	.603	.609	.547	.609	.276	.547	.354	.299	1.219	.761	.544
<i>.01</i>	.609	.432	.607	.603	.609	.547	.609	.276	.547	.000	.299	1.219	.761	.544
<i>.025</i>	.673	.438	.609	.609	.610	.609	.673	.281	.609	.379	.371	1.220	.761	.547
<i>.05</i>	.765	.513	.683	.678	.687	.609	.765	.345	.609	.379	.371	1.374	.761	.620
<i>.1</i>	.765	.513	.683	.678	.687	.609	.765	.345	.609	.379	.371	1.374	.770	.620
<i>.25</i>	.969	.912	.954	.954	.954	.969	.939	.838	.939	.484	.938	1.908	1.054	.891
<i>.5</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	2.491	.938
<i>.75</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	5.798	.938
<i>.9</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938
<i>.95</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938
<i>.975</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938
<i>.990</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938
<i>.995</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.500	1.000	2.000	10.333	.938

**Tabela 2-b)** - Valores de estatísticas sumárias referentes a coeficientes probabilísticos que não consideram a ausência conjunta “d”.

	S	J	O	CZ	K1	DW1	DW2	SS2	BB1	BB2	SO	JO	K2	FMG1
<i>Min</i>	.000	.004	.002	.002	.002	.003	.000	.006	.003	.002	.004	.374	.194	.002
<i>Max</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
<i>Media</i>	.559	.557	.559	.559	.559	.561	.555	.554	.558	.560	.560	.438	.463	.559
<i>DP</i>	.233	.249	.242	.242	.242	.238	.249	.261	.247	.241	.242	.152	.305	.242
<i>Centro</i>	.341	.351	.347	.347	.347	.343	.349	.357	.349	.346	.346	.687	.779	.347
<i>.005</i>	.000	.004	.002	.002	.002	.003	.000	.006	.003	.002	.004	.374	.194	.002
<i>.01</i>	.000	.004	.002	.002	.002	.003	.000	.006	.003	.000	.004	.374	.194	.002
<i>.025</i>	.002	.004	.002	.003	.002	.009	.002	.007	.011	.010	.009	.374	.194	.002
<i>.05</i>	.023	.013	.014	.013	.014	.009	.029	.014	.011	.010	.009	.374	.194	.014
<i>.1</i>	.023	.013	.014	.013	.014	.009	.029	.014	.011	.010	.009	.374	.195	.014
<i>.25</i>	.559	.509	.543	.544	.541	.600	.456	.447	.529	.550	.583	.374	.217	.540
<i>.5</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.374	.360	.691
<i>.75</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.479	.365	.691
<i>.9</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.530	.702	.691
<i>.95</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
<i>.975</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
<i>.990</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
<i>.995</i>	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691

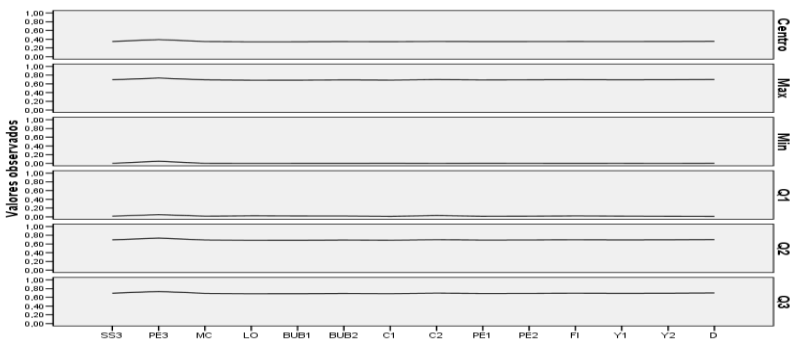
Verifica-se, na Tabela 2-a) e nas Tabelas A1, A3 e A5 do Anexo A, que a maior parte dos coeficientes de comparação clássicos toma valores no intervalo

[0,1]. No entanto, o intervalo entre o mínimo e o máximo valores da distribuição de amostragem é muito variável, mesmo nestes coeficientes. O valor máximo da distribuição atinge o limite superior 1 do intervalo em muitos dos coeficientes, situando-se o mínimo frequentemente acima de 0.5, para as duas primeiras classes consideradas de coeficientes, mas não para a terceira classe, dos outros coeficientes de associação. Analogamente, as medidas de localização e de dispersão dos vários coeficientes clássicos mostram variação elevada. Entretanto as distribuições de amostragem dos coeficientes probabilísticos associados apresentam intervalos de variação de amplitude semelhante, com valores mínimos e valores máximos aproximados, bem como as restantes estatísticas de localização e de dispersão.



**Figura 1** - Variação das médias obtidas para alguns coeficientes clássicos e probabilísticos.

Na Figura 1 é ilustrada a variação dos valores das médias de alguns dos coeficientes, clássicos (tomando valores no intervalo  $[0,1]$ ) e probabilísticos. Como se pode observar, os valores das médias dos índices clássicos, considerando os valores obtidos em 50 reamostragens, variam muito de índice para índice.



**Figura 2** - Variação dos valores de algumas estatísticas sumárias para os coeficientes probabilísticos *VL* associados a coeficientes de associação clássicos.

No contexto da abordagem *VL* verifica-se que, contrariamente aos respetivos índices básicos, os valores obtidos para as médias e outras medidas de localização das distribuições de amostragem dos coeficientes probabilísticos apresentam-se muito próximos, como se pode observar nas Figuras 1 e 2, bem como na Tabela 2-b) e nas Tabelas A1, A3 e A5 do Anexo A.

Estes resultados estão de acordo com a teoria que demonstra a propriedade da equivalência distribucional (exata ou assintótica) entre coeficientes de comparação para dados binários (Bacelar-Nicolau, 1980, 1987), mencionada na Secção 3.

A comparação de partições com recurso a coeficientes probabilísticos do tipo *VL* é, portanto, uma abordagem mais simples e mais robusta do que a comparação baseada nos coeficientes clássicos: em vez de se determinarem vários destes índices, bastará escolher um só índice em cada uma das classes (exacta ou assintoticamente) distribucionalmente equivalentes e usar o coeficiente probabilístico *VL* a este associado, o que tem, ainda, a vantagem de uniformizar a escala de medida numa mesma escala probabilística. Finalmente, os intervalos de variação e as outras estatísticas fornecidas pelas tabelas dos coeficientes *VL*, permitem-nos ainda atribuir uma avaliação média à qualidade da partição mais significativa fornecida pelos três modelos de classificação probabilísticos. Essa avaliação foi corroborada por um conjunto adequado de coeficientes de validação, que já não foram incluídos neste trabalho.

## 5 Conclusões

Neste trabalho foram referidas algumas das dificuldades relativas à utilização de índices clássicos para a comparação de partições, entre as quais a de poderem apresentar intervalos de variação distintos ou não tomarem valores em todo o intervalo de variação mas somente numa parte desse intervalo. Em contrapartida, a abordagem probabilística para a comparação de partições com recurso a coeficientes probabilísticos do tipo *VL*, tem, entre outras, a vantagem de que todos os índices clássicos utilizados conduzem, exata ou assintoticamente, a valores do índice probabilístico muito próximos (teoricamente, ao mesmo valor, na hipótese de referência aqui considerada), e numa escala probabilística (0, 1). Assim, em vez de se determinarem vários índices, poderemos aplicar a abordagem *VL* a um qualquer dos índices, pertencentes a uma dada classe de índices distribucionalmente equivalentes, para se proceder à comparação de pares de partições, com o mesmo número de classes.

A metodologia de reamostragem descrita faz parte de um trabalho desenvolvido sobre a avaliação da estabilidade das classificações obtidas numa *ACHA*. No exemplo apresentado, foi considerada a hipótese de margens fixas da tabela  $2 \times 2$  associada ao par de partições, no que concerne ao cálculo dos valores dos coeficientes probabilísticos para a comparação das partições.

Desenvolvimentos futuros poderão incluir a utilização da hipótese de margens livres.

## Referências

- BACELAR-NICOLAU, H. (1980). Contribuições ao estudo dos coeficientes de comparação em análise classificatória, Tese de Doutoramento, FC-UL.
- BACELAR-NICOLAU, H. (1985). *The affinity coefficient in cluster analysis*. Meth. Oper. Research, vol. 53, M. J. Beckmann *et al.* (ed.), Verlag Anton Hain, Munchen, 507-512.
- BACELAR-NICOLAU, H. (1987). *On the distribution equivalence in cluster analysis*. In: Pattern Recognition Theory and Applications, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, P. A. Devijver, J. Kittler (edit), Springer-Verlag, New York, 73-79.
- BACELAR-NICOLAU, H. (1988). *Two probabilistic models for classification of variables in frequency tables*. In: Classification and Related Methods of Data Analysis, H.H.Bock (ed.), North Holland, 181-186.
- BACELAR-NICOLAU, H.; NICOLAU, F. C.; SOUSA, A.; BACELAR-NICOLAU, L. (2009). *Measuring similarity of complex and heterogeneous data in clustering of large data sets*. Biocybernetics and Biomedical Engineering, Vol. 29, nº 2: 9-18.
- BACELAR-NICOLAU, H.; NICOLAU, F. C.; SOUSA, A.; BACELAR-NICOLAU, L. (2010). *Clustering complex heterogeneous data using a probabilistic approach*. Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010), Chania Crete Greece, 8-11 June 2010 – published on the CD Proceedings of SMTDA2010 (*electronic publication*).
- JAIN, A. K.; DUBES, R. C. (1988). *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ.
- LERMAN, I. C. (1970). *Les Bases de la classification automatique*. Paris, Gauthier -Villars.
- LERMAN, I. C. (1973). Étude distributionnelle de statistiques de proximité entre structures algébriques finies de même type – application à la classification automatique. In: Cahiers du B.U.R.O., N<sup>o</sup>. 19, Paris.
- LERMAN, I. C. (1981). *Classification et analyse ordinaire des données*. Paris, Dunod.
- MILLIGAN, G. W.; COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159-179.
- 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: Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, Y. Baba (Eds.), Springer-Verlag, 89-98.



SIBSON, R. (1972). *Multidimensional scalling in theory and pratice*. In: Les Méthodes Mathématiques de l'Archéologie, Centre d'Analyse Documentaire pour l'Archéologie, Marseille, 43-73.

SILVA, A., SAPORTA, G. e BACELAR-NICOLAU, H. (2004). *Missing data and imputation methods in partition of variables*. Classification, Clustering and Data Mining Applications, Springer, 631-637.

SILVA O.; BACELAR-NICOLAU, H. e NICOLAU, F. C. (2007). Utilização da análise classificatória para avaliar as atitudes/crenças em relação à estatística de alunos da área de ciências sociais e humanas. In: Ferrão, M. et al. Eds. Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística, 2006, Edições S.P.E., 751-759.

SNEATH, P. H.; SOKAL, R. R. (1963). *Principles of numerical taxonomy*. Freeman, San Francisco.

## Anexo A - Tabelas de estatísticas sumárias para as distribuições de amostragem dos coeficientes de comparação clássicos e probabilísticos *VL*.

**Tabela A1** - Valores de estatísticas sumárias para coeficientes clássicos que consideram a ausência conjunta “d”.

	SS4	SS1	B1	RR	RT	RG	HD	SMR	SS5	H
<i>Min</i>	.879	.730	.784	.152	.644	.730	.589	.784	3.000	.460
<i>Max</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000
<i>Media</i>	.983	.961	.970	.259	.947	.960	.935	.969	14.678	.921
<i>DP</i>	.033	.078	.060	.038	.103	.080	.125	.061	13.253	.157
<i>Centro</i>	.939	.865	.892	.215	.822	.865	.794	.892	20.250	.730
<i>.005</i>	.879	.730	.784	.152	.644	.730	.589	.784	3.000	.460
<i>.01</i>	.879	.730	.784	.152	.644	.730	.589	.784	3.000	.460
<i>.025</i>	.889	.739	.804	.169	.668	.735	.599	.801	3.000	.477
<i>.05</i>	.913	.791	.843	.169	.724	.786	.660	.840	3.000	.580
<i>.1</i>	.913	.791	.843	.169	.724	.786	.660	.840	3.500	.580
<i>.25</i>	.987	.968	.974	.268	.949	.968	.938	.974	5.000	.936
<i>.5</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	10.000	1.000
<i>.75</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	21.375	1.000
<i>.9</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000
<i>.95</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000
<i>.975</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000
<i>.990</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000
<i>.995</i>	1.000	1.000	1.000	.277	1.000	1.000	1.000	1.000	37.500	1.000

**Tabela A2** - Valores de estatísticas sumárias para coeficientes probabilísticos *VL* que consideram a ausência conjunta “d”.

	SS4	SS1	B1	RR	RT	RG	HD	SMR	SS5	H
<i>Min</i>	.458	.406	.424	.341	.375	.406	.357	.424	.185	.310
<i>Max</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534
<i>Media</i>	.500	.500	.500	.500	.500	.500	.500	.500	.463	.501
<i>DP</i>	.013	.032	.024	.057	.043	.033	.052	.025	.305	.066
<i>Centro</i>	.482	.461	.468	.435	.449	.461	.442	.469	.737	.422
<i>.005</i>	.458	.406	.424	.341	.375	.406	.357	.424	.185	.310
<i>.01</i>	.458	.406	.424	.341	.375	.406	.357	.424	.185	.310
<i>.025</i>	.462	.409	.432	.365	.385	.408	.361	.431	.185	.317
<i>.05</i>	.472	.430	.448	.365	.408	.428	.385	.447	.185	.357
<i>.1</i>	.472	.430	.448	.365	.408	.428	.385	.447	.189	.357
<i>.25</i>	.501	.503	.502	.514	.501	.503	.501	.502	.221	.506
<i>.5</i>	.507	.516	.512	.528	.522	.517	.527	.513	.369	.534
<i>.75</i>	.507	.516	.512	.528	.522	.517	.527	.513	.658	.534
<i>.9</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534
<i>.95</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534
<i>.975</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534
<i>.990</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534
<i>.995</i>	.507	.516	.512	.528	.522	.517	.527	.513	.962	.534

**Tabela A3** - Valores de estatísticas sumárias para outros coeficientes clássicos de associação (I).

	SS3	PE3	MC	LO	BUB1	BUB2	C1	C2	PE1	PE2	F1	Y1	Y2
<i>Min</i>	.518	.083	.219	.460	.416	.560	.415	.460	.445	.460	.460	.797	.491
<i>Max</i>	1.000	.917	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>Media</i>	.927	.516	.884	.937	.683	.701	.914	.931	.917	.926	.921	.982	.928
<i>DP</i>	.144	.337	.233	.132	.083	.045	.179	.134	.168	.146	.157	.045	.142
<i>Centro</i>	.759	.501	.609	.730	.569	.641	.707	.730	.722	.730	.730	.898	.746
<i>.005</i>	.518	.083	.219	.460	.416	.560	.415	.460	.445	.460	.460	.797	.491
<i>.01</i>	.518	.083	.219	.460	.416	.560	.415	.460	.445	.000	.460	.797	.491
<i>.025</i>	.526	.083	.220	.548	.431	.570	.460	.548	.460	.511	.477	.828	.502
<i>.05</i>	.610	.083	.374	.675	.512	.606	.499	.675	.538	.626	.580	.905	.589
<i>.1</i>	.610	.083	.374	.675	.512	.606	.499	.675	.538	.626	.580	.905	.636
<i>.25</i>	.937	.115	.908	.956	.697	.705	.956	.916	.945	.927	.936	.998	.945
<i>.5</i>	1.000	.712	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.75</i>	1.000	.777	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.9</i>	1.000	.869	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.95</i>	1.000	.917	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.975</i>	1.000	.917	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.990</i>	1.000	.917	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>.995</i>	1.000	.917	1.000	1.000	.723	.723	1.000	1.000	1.000	1.000	1.000	1.000	1.000

**Tabela A4** - Valores de estatísticas sumárias para coeficientes probabilísticos *VL* associados aos coeficientes de associação (I).

	SS3	PE3	MC	LO	BUB1	BUB2	C1	C2	PE1	PE2	F1	Y1	Y2
<i>Min</i>	.002	.049	.002	.000	.001	.001	.003	.000	.003	.002	.001	.002	.000
<i>Max</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>Media</i>	.558	.543	.559	.559	.560	.559	.561	.555	.560	.559	.557	.559	.557
<i>DP</i>	.245	.305	.242	.233	.234	.240	.239	.249	.241	.242	.245	.243	.250
<i>Centro</i>	.348	.391	.347	.341	.342	.346	.344	.349	.346	.347	.348	.347	.348
<i>.005</i>	.002	.049	.002	.000	.001	.001	.003	.000	.003	.002	.001	.002	.000
<i>.01</i>	.002	.049	.002	.000	.001	.001	.003	.000	.003	.002	.000	.002	.000
<i>.025</i>	.003	.049	.002	.002	.001	.002	.006	.002	.003	.002	.002	.002	.001
<i>.05</i>	.014	.050	.014	.023	.019	.018	.010	.029	.012	.014	.020	.015	.012
<i>.1</i>	.014	.050	.014	.023	.019	.018	.010	.029	.012	.014	.020	.015	.019
<i>.25</i>	.529	.058	.541	.557	.565	.538	.593	.457	.566	.540	.503	.538	.547
<i>.5</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.75</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.9</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.95</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.975</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.990</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694
<i>.995</i>	.694	.734	.691	.683	.684	.690	.685	.697	.689	.691	.695	.692	.694

**Tabela A5** - Valores de estatísticas sumárias para outros coeficientes clássicos de associação (II).

	D	M	K	T	SC	MP	F	FO1	FO2	B2	P1	P2	HA	MO
<i>Min</i>	.211	.524	.000	.375	.460	.460	.460	2.199	.460	.089	.418	48.810	.460	.024
<i>Max</i>	1.000	.801	.781	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320
<i>Media</i>	.871	.767	.111	.545	.920	.921	.922	3.428	.937	.183	.669	201.297	.920	.123
<i>DP</i>	.246	.078	.219	.043	.159	.158	.155	.351	.132	.034	.079	56.493	.159	.115
<i>Centro</i>	.605	.663	.391	.470	.730	.730	.730	2.904	.730	.145	.563	139.905	.730	.172
<i>.005</i>	.211	.524	.000	.375	.460	.460	.460	2.199	.460	.089	.418	48.810	.460	.024
<i>.01</i>	.211	.524	.000	.375	.460	.460	.460	2.199	.460	.089	.000	48.810	.460	.024
<i>.025</i>	.215	.557	.000	.417	.471	.476	.482	2.429	.548	.092	.430	52.544	.472	.024
<i>.05</i>	.287	.589	.000	.468	.572	.578	.587	2.760	.675	.108	.502	77.706	.573	.024
<i>.1</i>	.336	.589	.000	.468	.572	.578	.587	2.760	.675	.108	.502	77.706	.573	.025
<i>.25</i>	.876	.797	.000	.544	.936	.936	.936	3.391	.956	.189	.683	202.376	.936	.380
<i>.5</i>	1.000	.801	.000	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.083
<i>.75</i>	1.000	.801	.094	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.185
<i>.9</i>	1.000	.801	.578	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320
<i>.95</i>	1.000	.801	.578	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320
<i>.975</i>	1.000	.801	.719	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320
<i>.990</i>	1.000	.801	.781	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320
<i>.995</i>	1.000	.801	.781	.566	1.000	1.000	1.000	3.609	1.000	.200	.707	231.000	1.000	.320

**Tabela A6** - Valores de estatísticas sumárias para coeficientes probabilísticos *VL* associados aos coeficientes de associação (II).

	D	M	K	T	SC	MP	F	FO1	FO2	B2	P1	P2	HA	MO
<i>Min</i>	.003	.001	.307	.000	.002	.002	.001	.000	.000	.002	.001	.003	.002	.184
<i>Max</i>	.700	.669	.999	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962
<i>Media</i>	.556	.564	.442	.555	.559	.559	.559	.555	.559	.560	.560	.556	.559	.463
<i>DP</i>	.255	.229	.244	.239	.243	.243	.243	.249	.233	.240	.235	.254	.243	.305
<i>Centro</i>	.352	.335	.653	.344	.347	.347	.347	.349	.341	.346	.342	.352	.347	.573
<i>.005</i>	.003	.001	.307	.000	.002	.002	.001	.000	.000	.002	.001	.003	.002	.184
<i>.01</i>	.003	.001	.307	.000	.002	.002	.001	.000	.000	.002	.000	.003	.002	.184
<i>.025</i>	.004	.004	.307	.002	.002	.002	.002	.002	.002	.003	.001	.004	.002	.184
<i>.05</i>	.009	.012	.307	.038	.014	.015	.015	.029	.023	.012	.017	.014	.014	.184
<i>.1</i>	.014	.012	.307	.038	.014	.015	.015	.029	.023	.012	.017	.014	.014	.188
<i>.25</i>	.507	.650	.307	.493	.540	.538	.536	.458	.557	.566	.570	.508	.539	.220
<i>.5</i>	.700	.669	.307	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.368
<i>.75</i>	.700	.669	.470	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.659
<i>.9</i>	.700	.669	.984	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962
<i>.95</i>	.700	.669	.984	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962
<i>.975</i>	.700	.669	.997	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962
<i>.990</i>	.700	.669	.999	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962
<i>.995</i>	.700	.669	.999	.688	.692	.692	.692	.697	.683	.689	.684	.700	.692	.962

## Anexo B - Índices para a comparação de partições com base na Contagem de Pares.

<i>Símbolo</i>	<i>Fórmula</i>	<i>Amplitude variação</i>	<i>Autor</i>
<i>Coeficientes de semelhança que não consideram a ausência conjunta “d”</i>			
<i>S</i>	$\frac{a}{\min(a+b, a+c)}$	[0,1]	Simpson (1943)
<i>J</i>	$\frac{a}{(a+b+c)}$	[0,1]	Jaccard (1901)
<i>O</i>	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]	Ochiai (1957) Fowles e Mallows (1983)
<i>CZ</i>	$\frac{2a}{(2a+b+c)}$	[0,1]	Czekanowski (1932) Dice (1945)
<i>K1</i>	$\frac{1}{2} \left( \frac{a}{a+b} + \frac{a}{a+c} \right)$	[0,1]	Kuleczynski (1927)
<i>W1</i>	$\frac{a}{a+b}$	[0,1]	Dice (1945) Wallace (1983)
<i>W2</i>	$\frac{a}{a+c}$	[0,1]	Dice (1945) Wallace (1983)
<i>SS2</i>	$\frac{a}{a+2(b+c)}$	[0,1]	Sneath e Sokal (1963)
<i>BB1</i>	$\frac{a}{a+\max(b,c)}$	[0,1]	Braun-Blanquet (1928)
<i>BB2</i>	$\frac{a}{a+\max\{(a+b), (a+c)\}}$	[0,1]	Braun-Blanquet (1928)
<i>SO</i>	$\frac{a^2}{(a+b)(a+c)}$	[0,1]	Sorgenfrei (1958)
<i>JO</i>	$\frac{a}{a+b} + \frac{a}{a+c}$	[0,2]	Johnson (1967)

$K2$	$\frac{a}{b+c}$	$[0, \infty[$	Kulczynski (1927)
$FMG1$	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{(a+b)}}$	$\left[-\frac{1}{2}, 1\right]$	Fager e McGowan (1963)
<b><i>Coefficientes de semelhança que consideram a ausência conjunta “d”</i></b>			
$SS4$	$\frac{2(a+d)}{2(a+d)+b+c}$	$[0, 1]$	Sneath e Sokal (1963)
$SSI$	$\frac{1}{4} \left( \frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c} \right)$	$[0, 1]$	Sneath e Sokal (1963)
$BI$	$\frac{(a+b+c+d)^2 - (a+b+c+d)(b+c) + (b-c)^2}{(a+b+c+d)^2}$	$[0, 1]$	Baulieu (1989)
$RR$	$\frac{a}{a+b+c+d}$	$[0, 1]$	Russel e Rao (1940)
$RT$	$\frac{a+d}{a+2(b+c)+d}$	$[0, 1]$	Rogers e Tanimoto (1960)
$RG$	$\frac{a}{(a+b)+(a+c)} + \frac{d}{(c+d)+(b+d)}$	$[0, 1]$	Rogot e Goldberg (1966)
$HD$	$\frac{1}{2} \left( \frac{a}{a+b+c} + \frac{d}{b+c+d} \right)$	$[0, 1]$	Hawkins e Dotson (1968)
$KSM$	$\frac{(a+d)}{(a+b+c+d)}$	$[0, 1]$	Sokal e Michener (1958) Rand (1971)
$SS5$	$\frac{a+d}{b+c}$	$[0, \infty[$	Sneath e Sokal (1963)
$H$	$\frac{(a+d)-(b+c)}{a+b+c+d}$	$[-1, 1]$	Hamann (1961) Hubert (1977)
<b><i>Outros coeficientes de associação</i></b>			
$SS3$	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[0, 1]$	Sneath e Sokal (1963)

<i>PE3</i>	$\frac{(ab+bc)}{(ab+2bc+cd)}$	$[0,1]$	Peirce (1884)
<i>MC</i>	$\frac{a^2-bc}{(a+b)(a+c)}$	$[-1,1]$	McConnaughey (1964)
<i>LO</i>	$\frac{ad-bc}{\min((a+b)(b+d), (a+c)(c+d))}$	$[-1,1]$	Loevinger (1947)
<i>BUB1</i>	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	$[-1,1]$	Baroni-Urbani e Buser (1976)
<i>BUB2</i>	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	$[-1,1]$	Baroni-Urbani e Buser (1976)
<i>C1</i>	$\frac{ad-bc}{(a+b)(b+d)}$	$[-1,1]$	Cole (1949)
<i>C2</i>	$\frac{ad-bc}{(a+c)(c+d)}$	$[-1,1]$	Cole (1949)
<i>PE1</i>	$\frac{ad-bc}{(a+b)(c+d)}$	$[-1,1]$	Peirce (1884)
<i>PE2</i>	$\frac{ad-bc}{(a+c)(b+d)}$	$[-1,1]$	Peirce (1884)
<i>FI</i>	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	$[-1,1]$	Yule (1912)
<i>Y1</i>	$\frac{ad-bc}{ad+bc}$	$[-1,1]$	Yule (1900) Goodman e Kruskal (1954)
<i>Y2</i>	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$[-1,1]$	Yule (1912)
<i>D</i>	$\frac{(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$	$[-1,1]$	Doolittle (1885) Pearson (1926)

$M$	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$[-1, 1]$	Michael (1920)
$K$	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	$[-1, 1]$	Cohen (1960)
$T$	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	$[-1, 1]$	Tarwid ((1960)
$SC$	$\frac{4ad-(b+c)^2}{((a+b)+(a+c))((c+d)+(b+d))}$	$[-1, 1]$	Scott (1955)
$MP$	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	$[-1, 1]$	Maxwell e Piliner (1968)
$F$	$\frac{(ad-bc)[(a+b)(b+d)+(a+c)(c+d)]}{2(a+b)(a+c)(c+d)(b+d)}$	$[-1, 1]$	Fleiss (1975)
$FO2$	$\frac{[na-(a+b)(a+c)]}{[n^* \min\{(a+b), (a+c)\} - (a+b)(a+c)]}$	$[-1, 1]$	Forbes (1925)
$B2$	$\frac{ad-bc}{(a+b+c+d)^2}$	$\left[-\frac{1}{4}, \frac{1}{4}\right]$	Baulieu (1989)
$PI$	$\frac{\chi^2}{\sqrt{n+\chi^2}}$	$[0, 1[$	Pearson (1905)
$HA$	$\frac{2(ad-bc)}{(a+b)(b+d)(a+c)(c+d)}$	$[-1, 1]$	Hubert e Arabie (1985)
$MO$	$\frac{2a}{2bc+ab+ac}$	$[0, \infty[$	Mountford (1962)
$FO1$	$\frac{na}{(a+b)(a+c)}$	$[0, \infty[$	Forbes (1907)
$P2$	$\frac{n(ad-bc)^2}{[(a+b)(c+d)(a+c)(b+d)]} = \chi^2$	$[0, \infty[$	Pearson (1905)





# Matriz de Dissemelhança Entrópica para Classificação Não Supervisionada

Jorge M. Santos<sup>1</sup>

© The Author(s) 2013

**Resumo** A classificação hierárquica não supervisionada é um processo usualmente baseado em medidas de semelhança ou dissemelhança entre objetos ou conjuntos de objetos de um determinado conjunto de dados. A medida de dissemelhança mais comum é a métrica pesada  $l_p$  (a distância Euclidiana é um caso particular da métrica não pesada  $l_p$ ) que serve de suporte para a construção de matrizes de dissemelhança, elemento base dos algoritmos de classificação hierárquica. A medida de dissemelhança “ideal” para um algoritmo de agrupamento deveria providenciar informação sobre a estrutura dos dados de forma a facilitar a obtenção de soluções óptimas, o que pode não suceder com a distância Euclideana. Neste trabalho mostramos como podemos obter uma medida de dissemelhança com as características referidas usando uma medida de dissemelhança entrópica.

**Palavras-chave:** entropia, matriz de dissemelhança, análise classificatória, agrupamento.

## 1 Introdução

Os métodos de agrupamento são processos que tentam encontrar diferentes grupos de entidades num determinado conjunto de dados com base na semelhança ou diferença (dissemelhança) entre os seus objetos. Questões como “Quais os critérios de semelhança ou dissemelhança?” ou “O que distingue os vários grupos e como os podemos encontrar?” são exemplos de perguntas aparentemente simples que se colocam no âmbito de um problema de agrupamento mas para as quais não temos resposta única.

---

<sup>1</sup> ISEP e LEMA, Instituto Politécnico do Porto e Instituto de Engenharia Biomédica, jms@isep.ipp.pt

Os algoritmos hierárquicos são geralmente baseados numa matriz de semelhança/dissemelhança. Frequentemente esta matriz é construída usando uma medida de distância, em particular a Euclidiana, pelo que os resultados obtidos poderão não refletir a estrutura dos dados. Neste trabalho propõe-se uma nova matriz de dissemelhança baseada numa medida entrópica com a qual as ligações/pesos entre objetos são sensíveis à estrutura local e direcional dos dados. Usando esta matriz os grupos formados refletirão essa mesma estrutura.

Na secção seguinte introduzimos alguns conceitos e notação que servirá de suporte ao trabalho apresentado. Na Secção 3 apresentamos a nova matriz de dissemelhança e apresentamos algumas experiências com dados artificiais que mostram a validade da medida proposta. Na última secção tiramos as conclusões.

## 2 Conceitos básicos

### 2.1 Medidas de proximidade

Seja  $X = \{x_i\}$ ,  $i=1,2,\dots,N$  um conjunto de dados com  $N$  objetos, em que  $x_i$  é um vector  $m$ -dimensional representativo de cada objeto. Define-se  $S$ , um  $s$ -agrupamento de  $X$ , como uma partição de  $X$  em  $s$  grupos,  $C_1, C_2, \dots, C_s$ , que

obedece às seguintes condições:  $C_i \neq \emptyset$ ,  $i=1,2,\dots,s$  ;  $\bigcup_{i=1}^s C_i = X$  e

$C_i \cap C_j = \emptyset$ ,  $i \neq j$ ,  $i, j=1,2,\dots,s$ . Cada vetor, definidas as condições anteriores, pertence a um só grupo (ao contrário do que acontece nos algoritmos fuzzy em que a cada elemento é atribuído um grau de pertença a cada grupo). Elementos pertencentes a um determinado grupo possuem um grau de semelhança maior entre si do que com qualquer um dos outros elementos pertencentes a outros grupos. Este grau de semelhança é usualmente definido com recurso a uma medida de semelhança/dissemelhança.

A medida de dissemelhança mais comum entre dois vetores  $x$  e  $y$  é a métrica pesada  $l_p$ ,

$$d_p(x, y) = \left( \sum_{i=1}^l w_i |x_i - y_i|^p \right)^{\frac{1}{p}}, \quad (1)$$

em que  $x_i$  e  $y_i$  são as componentes  $i$  dos vetores  $x$  e  $y$ ,  $i=1,2,\dots,m$  e  $w_i \geq 0$  é o coeficiente associado ao peso índice  $i$ . A métrica não pesada  $l_p$ ,  $w=1$ , é também conhecida como a distância de Minkowski de ordem  $p$  ( $p \geq 1$ ). Exemplos desta distância são as conhecidas distância Euclidiana, que se obtém fazendo  $p=2$ , a distância de Manhattan,  $p=1$ , e a  $l_\infty$  ou distância de Chebyshev.

## 2.2 Algoritmos de agrupamento

Os algoritmos de agrupamento do tipo aglomerativo são frequentemente usados nas áreas de Biologia, Medicina e também nas Ciências de Computadores e Engenharia (Jain, 1988; Jain, 1999; Jain, 2004; Berkhin, 2002). Os algoritmos hierárquicos aglomerativos começam por considerar um número de grupos igual ao número de elementos do conjunto de dados e posteriormente, usando uma medida de semelhança/dissemelhança, em cada passo do algoritmo efetuam a junção ou aglomeração de elementos em grupos maiores acabando, no último passo, num grupo único com todos os elementos do conjunto. A árvore hierárquica resultante define os níveis de agrupamento e posteriores técnicas de corte produzem o número de grupos final. Temos como exemplos de algoritmos que usam processos hierárquicos aglomerativos os algoritmos CURE (Guha, 1998), ROCK (Guha, 2000), AGNES (Kaufman, 1990), BIRCH (Zhang, 1996; Zhang, 1997) e Chameleon (Karypis, 1999).

A junção de grupos nos diferentes algoritmos aglomerativos produz frequentemente resultados diferentes dependendo da medida usada para avaliar a semelhança/dissemelhança entre grupos. Os métodos clássicos mais comuns para efetuar a junção dos grupos nas várias etapas são o método do vizinho mais próximo, o do vizinho mais distante, o método dos centróides e o método de Ward ou da inércia mínima. Em (Kamvar, 2002) podemos encontrar uma interpretação probabilística destes métodos aglomerativos.

## 2.3 Entropia quadrática de Rényi

Desde a sua introdução, por Shannon (Shannon, 1948), que a entropia e os conceitos de teoria de informação são usados em sistemas de aprendizagem. A entropia de Shannon,

$$H_s(X) = - \sum_{i=1}^N p_i \log p_i, \quad (2)$$

mede a quantidade média de informação que é conduzida pelos eventos  $X = x_i$  que ocorrem com probabilidade  $p_i$ . A entropia também pode ser vista como a quantidade de incerteza associada a uma variável aleatória. Quanto mais incertos são os eventos de  $X$ , maior a informação, com valor máximo para eventos equiprováveis. A extensão da entropia de Shannon para variáveis aleatórias contínuas é:

$$H_s(X) = - \int_C f(x) \log f(x) dx, \quad (3)$$

em que  $X \in C$  e  $f(x)$  é a função densidade de probabilidade (fdp) da variável  $X$ .

Alfred Rényi generalizou o conceito de entropia (Rényi, 1976) e definiu a entropia de Rényi de ordem  $\alpha$ , de uma distribuição discreta, como

$$H_{R\alpha}(X) = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^N p_i^\alpha \right), \alpha > 0, \alpha \neq 1, \quad (4)$$

que engloba a entropia de Shannon quando  $\alpha \rightarrow 1$ . Para distribuições contínuas e  $\alpha=2$  obtém-se a Entropia Quadrática de Rényi (Rényi, 1976):

$$H_{R2}(X) = -\log \left( \int_C [f(x)]^2 dx \right), \quad (5)$$

A função densidade de probabilidade,  $f(x)$ , pode ser estimada usando o método de Parzen o que nos permite estimar a entropia de Rényi de uma forma muito eficiente e não-paramétrica. O método de Parzen estima a fdp como (Parzen, 1962)

$$f(x) = \frac{1}{Nh^m} \sum_{i=1}^N G\left(\frac{x-x_i}{h}\right), \quad (6)$$

em que  $N$  é o número de elementos da amostra,  $m$  é a dimensão do vector  $x$ ,  $h$  é o tamanho da janela ou parâmetro de suavização e  $G$  é uma função Kernel. Neste caso usamos o Kernel Gaussiano simétrico com média nula e matriz de covariância diagonal

$$G(x;0,I) = \frac{1}{(2\pi)^{m/2} |I|^{1/2}} \sum_{i=1}^N \exp\left(-\frac{1}{2} x^T I x\right), \quad (7)$$

em que  $I$  é a matriz identidade.

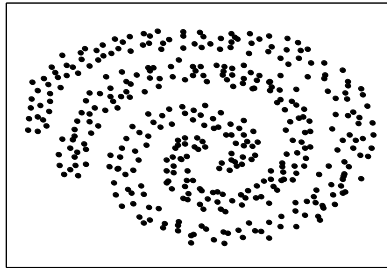
Substituindo (7) e (6) em (5) e fazendo a integração dos kernel gaussianos a Entropia Quadrática de Rényi é estimada por (Xu, 1999)

$$\begin{aligned} \hat{H}_{R2} &= -\log \left[ \int_{-\infty}^{+\infty} \left( \frac{1}{Nh^m} \sum_{i=1}^N G\left(\frac{x}{h}, \frac{x_i}{h}, I\right) \right)^2 dx \right] = \\ &= -\log \left( \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G(x_i - x_j; 0, 2h^2 I) \right), \end{aligned} \quad (8)$$

No cálculo da matriz de dissemelhança que apresentaremos na secção seguinte usamos a entropia quadrática de Rényi devido à sua relativa simplicidade de cálculo quando comparado com o cálculo de outras entropias.

### 3 Matriz de dissemelhança entrópica

Os métodos de agrupamento baseados numa matriz de dissemelhança, conforme referido na Secção 2., não produzem frequentemente grupos que reflitam a estrutura espacial dos dados. Um conjunto de dados como o representado na Figura 1 possui uma forte estrutura direcional, isto é, os dados estão dispostos localmente segundo uma determinada direção. Esta estrutura pode também aparecer, noutros conjuntos de dados, a um nível mais global. Os habituais algoritmos de agrupamento não produzem, para o caso representado na Figura 1, os grupos relativos a cada uma das espirais.



**Figura 1** – Conjunto de dados com uma forte estrutura direcional (dupla espiral).

Como já salientamos anteriormente, não existe uma solução única para um problema de agrupamento, pelo que uma das dificuldades que se apresenta é o da validação dos resultados. Nos problemas bidimensionais como o representado na Figura 1, o resultado de um algoritmo de agrupamento pode ser verificado visualmente o que não é possível para conjuntos de dados de elevada dimensão.

Seja  $X = \{\mathbf{x}_i\}$ ,  $i=1,2,\dots,N$ ,  $\mathbf{x}_i \in \mathbb{R}^m$ , um conjunto de vetores correspondentes a um conjunto de  $N$  objectos. Cada elemento de uma matriz de dissemelhança  $A$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $n=N$ , é calculado usando a medida de dissemelhança  $A_{ij}=d(\mathbf{x}_i, \mathbf{x}_j)$ . Usando esta matriz de dissemelhança, elaboremos primeiro uma matriz de vizinhança,  $L$ , em que cada linha  $i$  representa um ponto do conjunto de dados, cada coluna  $j$  a ordem de proximidade (1ª coluna=ponto mais próximo, ..., última coluna=ponto mais afastado) e cada elemento representa o ponto que está na posição de vizinhança  $j$  relativamente ao ponto da respectiva linha  $i$ . Um exemplo de uma matriz de dissemelhança e da respectiva matriz de vizinhança está representado na Tabela 3. Os pontos da primeira coluna da matriz de vizinhança são aqueles que têm menor dissemelhança (maior semelhança) com o respetivo elemento de cada linha.

Na Figura 2 estão representados, para o conjunto de dados da Figura 1, os sub-grafos correspondentes à primeira coluna de uma matriz de vizinhança. Na Figura 2a, cada ligação corresponde à relação entre cada um dos elementos e o elemento

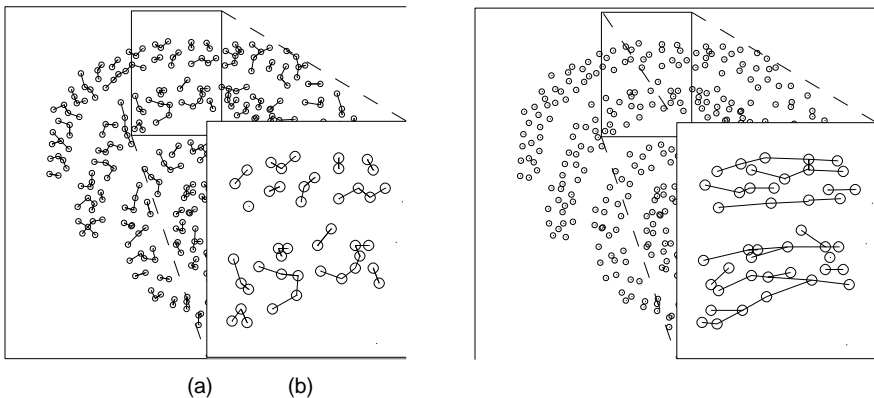
referenciado na primeira coluna usando a distância Euclidiana. Pode-se facilmente verificar que estas ligações não têm qualquer relação com a estrutura do conjunto de dados. Na Figura 2b representa-se o que consideramos serem as ligações “ideais”, as que refletem a estrutura local dos dados.

As medidas de distância clássicas não nos permitem obter este comportamento “ideal” em que as ligações estão relacionadas com a estrutura direcional local dos dados. Veremos como conseguir este comportamento através do uso de uma medida de dissemelhança entrópica.

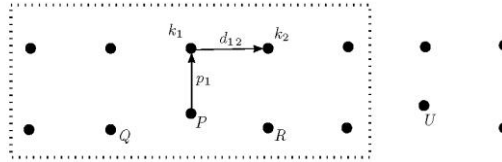
Consideremos o conjunto de pontos representado na Figura 3. Todos os pontos estão situados sobre uma grelha quadrada, com excepção dos pontos  $P$  e  $U$ . Consideremos que:

- $K = \{k_i\}$ ,  $i=1,2,...,M$  é o conjunto dos  $M$  vizinhos mais próximos de  $P$ .
- $d_{ij}$  é o vetor diferença entre os pontos  $k_i$  e  $k_j$ ,  $i,j=1,2,...,M$ ,  $i \neq j$  a que chamaremos vetor de ligação entre esses pontos.
- $p^i$  é o vetor diferença entre o ponto  $P$  e cada um dos  $M$  vizinhos mais próximos  $k_i$ .

Pretendemos encontrar a ligação entre  $P$  e um dos seus vizinhos que melhor reflita a estrutura local dos dados. Sem efetuar qualquer tipo de cálculo, olhando somente para a figura, podemos dizer que, apesar de a ligação mais curta ser  $p^1$ , as candidatas a “ligação ideal” são as ligações de  $P$  a  $Q$  e de  $P$  a  $R$  pois são aquelas que melhor refletem a estrutura direcional dos pontos.



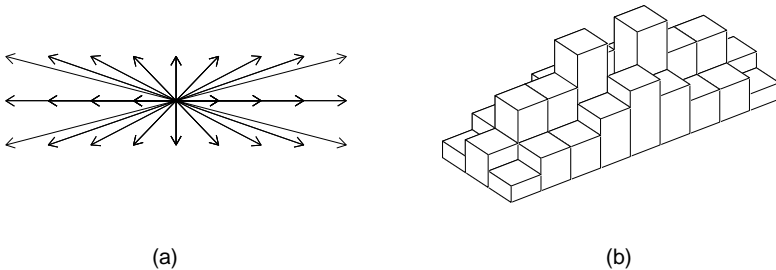
**Figura 2** – As ligações relativas à primeira coluna usando distância Euclidiana (a) e as ligações “ideais” (b).



**Figura 3** – Um simples exemplo com os  $M$  vizinhos mais próximos de  $P$ , ( $M=9$ ).

Representemos todos os vetores de ligação  $d_{ij}$  trasladados para uma origem comum conforme representado na Figura 4a. Este será o campo de vetores dos  $M$  vizinhos de  $P$  (vetores de ligação com  $P$  não estão incluídos neste campo). Este campo de vetores pode ser interpretado como uma função densidade de probabilidade com correspondência com o histograma bidimensional mostrado na Figura 4b, onde em cada caixa contaremos o número de ocorrências das extremidades dos vetores  $d_{ij}$ . Este histograma estima a função densidade de probabilidade das ligações  $d_{ij}$  e pode também ser interpretado como uma estimação por janela de Parzen usando um núcleo (kernel) rectangular.

A função densidade de probabilidade associada ao ponto  $P$  reflete, neste caso, uma estrutura horizontal e, portanto, devemos escolher uma ligação para  $P$  que siga esta direção. Sendo a direção um fator importante não devemos escolher ligações muitos distantes do ponto  $P$  pelo que o comprimento é também um fator a ter em conta. Como podemos ver na Figura 4b, a função densidade de probabilidade diz-nos que os vetores de ligação mais pequenos são os mais prováveis.



**Figura 4** – O campo de vetores da vizinhança do ponto  $P$  (a) e o histograma da função densidade de probabilidade (b).

A questão que se coloca agora é a de escolher, das  $M$  possíveis ligações do ponto  $P$  com os seus vizinhos, a ligação “ideal”. Para isso vamos efetuar uma seriação comparando as  $M$  funções densidade de probabilidade que se obtêm se juntarmos cada uma das ligações  $p^i$  ao campo de ligações dos  $M$  vizinhos mais próximos de  $P$ . A entropia é uma das medidas que compara funções densidade de probabilidade pelo que iremos usá-la para seriar todas as possíveis ligações  $p^i$ . Basicamente, o que iremos fazer é seriar a ligação  $p^i$  de acordo com a variação que ela introduz na



função densidade de probabilidade. A ligação que introduz a menor desordem no sistema, a que menos aumenta a entropia do sistema, será a classificada como a ligação mais forte seguida das  $M-1$  restantes ligações por ordem decrescente.

Seja  $D = \{d_{ij}\}$ ,  $i, j = 1, 2, \dots, M$ ,  $i \neq j$  o conjunto de todos os vetores de ligação (campo de vetores dos  $M$  vizinhos mais próximos de  $P$ ). Seja  $H(D, p^i)$  a entropia associada com a ligação  $p^i$ , a entropia do conjunto de todas as ligações  $d_{ij}$  mais a ligação  $p^i$ , tal que

$$H(D, p^i) = H(D \cup \{p^i\}), i = 1, 2, \dots, M. \quad (9)$$

Esta medida entrópica será a nossa medida de dissemelhança<sup>2</sup>. Para cada ponto calculam-se as  $M$  entropias e constrói-se a matriz de dissemelhança e a respetiva matriz de vizinhança. Os elementos da primeira coluna da matriz de vizinhança serão aqueles com menor valor de dissemelhança entrópica. Esta será a coluna com a ligação entrópica mais forte, seguida das outras por ordem decrescente (crescente no valor de dissemelhança entrópica).

Regressemos ao exemplo da Figura 3. Na Tabela 1 estão representados os valores de dissemelhança entrópica calculados para o ponto  $P$  numa vizinhança  $M=9$ . Os pontos da figura estão referenciados de 1 a 14, da esquerda para a direita, de cima para baixo. Usou-se a entropia de Rényi como explicado na secção anterior.

**Tabela 1** – Os valores de dissemelhança entrópica (A) e de vizinhança (B) relativos ao ponto  $P$ .

A	Ponto	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	10	8,83	8,73	8,72	8,73	8,83			8,66	8,58		8,58	8,66		

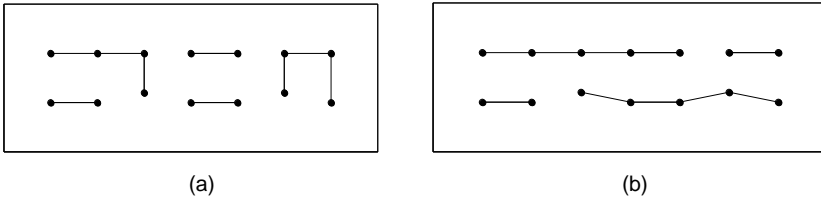
  

B		L1	L2	L3	L4	L5	L6	L6	L8	L9
	10	11	9	8	12	3	4	2	5	1

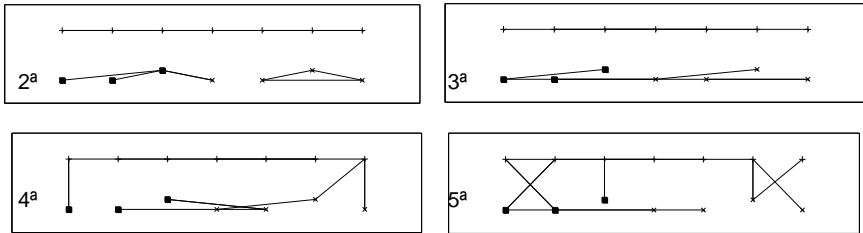
Na Figura 5 mostram-se as ligações referentes aos elementos da primeira coluna (L1) de uma matriz de vizinhança usando uma medida de dissemelhança Euclidiana (Figura 5a) e a medida entrópica proposta (Figura 5b). Aqui vê-se claramente que as ligações da primeira coluna seguem a estrutura horizontal dos dados e que, apesar do facto do ponto  $k_l$  ser o ponto mais próximo do ponto  $P$  em termos de distância Euclidiana, a ligação mais forte para este ponto é a ligação  $P-R$  como pretendido. Na Figura 6 mostram-se as ligações das colunas seguintes (até à

<sup>2</sup> De notar que esta medida de dissemelhança entrópica não obedece aos princípios de uma medida de distância já que não é simétrica, e pode não verificar a desigualdade triangular mas é, no entanto, suficiente para construir a matriz de vizinhança.

5ª) onde se pode ver ainda a tendência para ligações que seguem a estrutura horizontal dos dados.

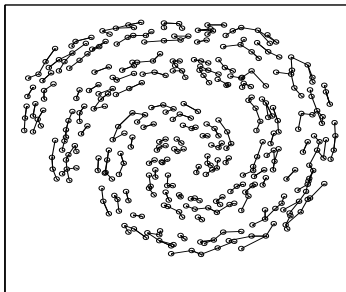


**Figura 5** – As ligações da primeira coluna usando a medida Euclidiana (a) e a medida entrópica proposta (b).



**Figura 6** – As ligações da 2ª, 3ª, 4ª e 5ª colunas. O elemento 10 liga-se aos elementos 9, 8, 12 e 3, de acordo com a matriz de vizinhança da Tabela 1.

Este comportamento pode também ser observado no conjunto de dados da Figura 1 e representado na Figura 7. As ligações relativas à primeira coluna seguem claramente a direção estrutural dos dados.



**Figura 7** – As ligações da primeira coluna seguem a estrutura dos dados quando se usa a medida entrópica.

Na Tabela 2 apresenta-se o pseudo-código para o cálculo da matriz de dissemelhança e a respetiva matriz de vizinhança.

Na Tabela 3 mostra-se um exemplo de uma matriz de dissemelhança entrópica (a) para um conjunto de dados de 16 elementos e a respetiva matriz de vizinhança (b). Neste exemplo usou-se o valor de  $M=5$  razão pela qual a matriz de vizinhança tem só 5 colunas.

A matriz de vizinhança baseada na medida entrópica pode ser utilizada por um algoritmo de agrupamento hierárquico.

**Tabela 2** – Pseudo-código para cálculo da matrix de dissemelhança entrópica e respetiva matriz de vizinhança.

---

For i=1 to $N$ (número de elementos a agrupar)
For j=1 to $M$ (número de vizinhos mais próximos)
Calcular $H(D, p^i) = H(D \cup \{p^i\})$ , $i = 1, 2, \dots, M$
end j
end i
Construir a matriz $N \times N$ de dissemelhança.
Construir a matriz $N \times M$ de vizinhança.

---

**Tabela 3** – Exemplo de uma matriz de dissemelhança entrópica (a) e respetiva matriz de vizinhança (b) para um conjunto de dados de 16 elementos ( $M=5$ ).

Points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	5.64	-	6.36	5.66	6.32	-	-	-	5.77	-	-	-	-	-	-
2	6.00	-	6.03	6.26	6.40	-	-	-	-	6.12	-	-	-	-	-	-
3	-	6.19	-	6.61	-	-	-	-	-	7.03	6.76	6.48	-	-	-	-
4	6.48	6.50	6.58	-	-	-	-	-	-	6.36	6.47	-	-	-	-	-
5	6.10	-	-	-	-	6.24	-	6.16	6.09	6.18	-	-	-	-	-	-
6	-	-	-	6.05	6.21	-	-	6.03	6.14	6.11	-	-	-	-	-	-
7	-	-	-	-	5.61	6.06	-	5.67	5.28	6.84	-	-	-	-	-	-
8	-	-	-	-	6.09	5.54	5.82	-	5.89	6.27	-	-	-	-	-	-
9	-	-	-	-	5.78	5.98	5.79	6.03	-	5.99	-	-	-	-	-	-
10	6.06	5.98	-	6.05	6.00	5.98	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-	3.74	4.41	4.73	3.93	3.86
12	-	-	-	-	-	-	-	-	-	-	3.86	-	3.88	4.36	4.60	3.95
13	-	-	-	-	-	-	-	-	-	-	4.39	3.78	-	3.79	4.75	3.87
14	-	-	-	-	-	-	-	-	-	-	4.71	4.36	3.80	-	3.93	3.86
15	-	-	-	-	-	-	-	-	-	-	3.85	4.81	5.01	3.82	-	3.71
16	-	-	-	-	-	-	-	-	-	-	4.19	4.18	4.18	4.18	4.18	-

(a)

Points	L1	L2	L3	L4	L5
<b>1</b>	2	5	10	6	4
<b>2</b>	1	3	10	4	5
<b>3</b>	2	4	11	1	10
<b>4</b>	10	3	6	2	11
<b>5</b>	9	1	8	10	6
<b>6</b>	8	4	10	9	5
<b>7</b>	9	5	8	6	10
<b>8</b>	6	7	9	5	10
<b>9</b>	5	7	10	6	8
<b>10</b>	6	2	5	4	1
<b>11</b>	12	16	15	13	14
<b>12</b>	11	13	16	14	15
<b>13</b>	12	14	16	11	15
<b>14</b>	13	16	15	12	11
<b>15</b>	16	14	11	12	13
<b>16</b>	14	11	13	12	15

b)

## 4 Conclusões

Neste trabalho apresenta-se uma nova matriz de dissemelhança construída com base numa medida entrópica. Esta medida entrópica de dissemelhança permite incorporar informação sobre a estrutura local dos dados, representada pela distribuição estatística das ligações na vizinhança de um ponto de referência, conseguindo assim um balanço entre a direcção estrutural dos dados e a sua proximidade (distância mínima). Deste modo é possível obter grupos com formas arbitrárias ao contrário por exemplo dos grupos globulares obtidos com o K-médias e outros algoritmos clássicos.

As experiências preliminares realizadas mostram o potencial desta matriz de dissemelhança e da sua possível aplicação em todo o tipo de problemas de agrupamento.

Um dos trabalhos futuros consistirá na utilização desta matriz de dissemelhança em algoritmos existentes e na procura de algoritmos novos que possam aproveitar as suas potencialidades. Algoritmos baseados na teoria de grafos serão objecto de estudo preferencial.

## Referências

- BERKHIN, P. (2002). Survey of clustering data mining techniques, *Accrue Software*, San Jose, CA, Tech. Rep.
- GUHA S., RASTOGI R. e SHIM K. (1998). CURE: an efficient clustering algorithm for large databases, in *ACM SIGMOD International Conference on Management of Data*, 73–84.
- GUHA S., RASTOGI R. e SHIM K. (2000). ROCK: A robust clustering algorithm for categorical attributes, *Information Systems*, 25, 5, 345–366.
- JAIN, A.K. e DUBES, R.C. (1988). *Algorithms for Clustering Data*. Prentice Hall International.
- JAIN, A.K., MURTY, M.N. e FLYNN, P.J. (1999). Data clustering: a review, *ACM Computing Surveys*, 31, 3, 264–323.
- JAIN, A.K., TOPCHY, A., LAW, M., e BUHMANN, J. (2004). Landscape of clustering algorithms, in *17th International Conference on Pattern Recognition*, 1, 260–263.
- Kamvar S. D., Klein D., e Manning C. D. (2002). Interpreting and extending classical agglomerative clustering algorithms using a model based approach, in *19th International Conference on Machine Learning*, 283–290.
- KARYPIS, G., HAN, S. e KUMAR, V. (1999). CHAMELEON: Hierarchical clustering using dynamic modeling, *IEEE Computer*, 32, 8, 68–75.
- KAUFMAN, L. e ROUSSEEuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: John Wiley and Sons.
- PARZEN, E. (1962). On the estimation of a probability density function and mode, *Annals of Mathematical Statistics*, 33, 1065–1076.
- RÉNYI, A. (1976). Some fundamental questions of information theory, *Selected Papers of Alfred Rényi*, 2, 526–552.
- SHANNON, C. (1948). A mathematical theory of communication, *Bell System Technical Journal*, 27, 379–423, 623–653.
- XU, D. e PRINCIPE, J. (1999). Training mlps layer-by-layer with the information potential, in *Intl. Joint Conf. on Neural Networks*, 1716–1720.
- ZHANG, T., RAMAKRISHNAN, R. e LIVNY, M. (1996). BIRCH: An efficient clustering method for very large databases, in *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada, 103–114.
- ZHANG, T., RAMAKRISHNAN, R., e LIVNY, M. (1997). BIRCH: A new data clustering algorithm and its applications, *Data Mining and Knowledge Discovery*, 1, 2, 141–182.

# Tipologia de Comportamentos de Crianças em Idade Pré-Escolar: Aplicação de Modelos de Análise Classificatória Hierárquica

António M. Caxaria<sup>1</sup> · Helena Bacelar-Nicolau<sup>2</sup> · Rita M. Leal<sup>3</sup>

© The Author(s) 2013

**Resumo** O inventário de comportamentos que se analisa foi elaborado para responder à necessidade de fornecer, aos educadores e às creches e jardins-de-infância, um instrumento de aplicação rápida e fácil para avaliação contínua do desenvolvimento de crianças e sinalização de possíveis perturbações. Efectuou-se a análise de dados sobre o conjunto de 172 crianças descritas por 123 itens, tendo-se utilizado modelos de classificação hierárquica ascendente sobre as variáveis e sobre os indivíduos. Compararam-se as partições / classificações e as hierarquias de classificações por forma a determinar o(s) modelo(s) de classificação que melhor se ajusta(m) aos dados. Encontraram-se as melhores partições para as variáveis e para os indivíduos. Identificaram-se os conjuntos de variáveis comportamentais mais importantes para os diferentes grupos de idade encontrados.

**Palavras-chave:** Análise Comportamental, Análise Classificatória Hierárquica Ascendente, Variáveis Binárias, Critérios de Agregação.

## 1 Introdução

O inventário de comportamentos que aqui analisamos contém 123 itens descritivos de respostas dicotómicas (sim/não) possíveis da criança a circunstâncias do ambiente. A criação dos itens baseou-se numa análise prévia da consistência das respostas obtidas, tendo-se passado de 157 questões iniciais para as 123 finais. A

---

<sup>1</sup>Instituto de Informática, Antonio.Caxaria@inst-informatica.pt

<sup>2</sup>Faculdade de Psicologia e LEAD, Universidade de Lisboa, hbacelar@fp.ul.pt

<sup>3</sup>Faculdade de Psicologia, Universidade de Lisboa e Escola Superior de Educação Maria Ulrich, mritaleal@hotmail.com

amostra de crianças utilizada tem idades compreendidas entre os 6 e os 63 meses de idade cronológica. Para cada criança, além dos itens relativos ao seu comportamento e idade em meses, registou-se o seu sexo. Estratificou-se previamente a amostra em 18 classes de acordo com a idade das crianças. Para cada classe escolheram-se aleatoriamente 10 crianças (excepto na primeira classe), ficando a amostra final com 172 crianças.

Na análise de dados efectuada sobre o conjunto das 172 crianças descritas pelos 123 itens, utilizámos modelos de classificação hierárquica ascendente sobre as variáveis binárias e sobre os indivíduos. Na secção seguinte apresenta-se uma breve descrição das variáveis, dos indivíduos e da metodologia utilizada. O inventário de comportamentos foi elaborado para responder à necessidade de fornecer, aos educadores e às creches e jardins-de-infância, um instrumento de aplicação rápida e fácil para avaliação contínua do desenvolvimento de crianças e sinalização de possíveis perturbações. Sendo assim, a procura de uma tipologia de comportamentos na amostra de crianças estudada poderá ser um ponto de partida para ajudar a definir estratégias com vista a essa avaliação.

Compararam-se as classificações e as hierarquias de classificações de forma a determinar modelos de classificação que melhor se ajustassem aos dados. Na Secção 3 apresentam-se resultados obtidos dessa análise exploratória sobre as variáveis e na Secção 4 resultados da análise exploratória sobre os indivíduos. As análises de um e de outro dos conjuntos, realizadas independentemente, foram depois postas em confronto. Na Secção 5 identificam-se os grupos de variáveis comportamentais mais importantes para os diferentes grupos de idade encontrados na análise classificatória dos indivíduos.

## **2 Descrição dos dados, objectivos e métodos**

O conjunto de dados tratados é resultado de um inquérito relativo ao comportamento de crianças com idade compreendida entre os 6 e os 63 meses (crianças de idade pré-escolar). O inquérito abrangeu a identificação de cada criança, através da idade e do sexo, e de itens relacionados com o seu comportamento. O sexo, que se revelou pouco discriminativo, foi retirado da análise dos dados a que se refere o presente trabalho. Os itens, as variáveis, são de natureza dicotómica, pelo que o estudo recai sobre dados binários, relacionando-se naturalmente o „sim” à presença/existência e o „não” à ausência/não existência de um determinado tipo de comportamento. O inquérito inicial, na sua primeira redacção, apresentava 157 questões, que mais tarde foram reduzidas para 123. São estas que aqui foram analisadas. Se representarmos os dados numa matriz, esta mostrará portanto 172 indivíduos descritos por 123 variáveis binárias que traduzem atitudes comportamentais. As 123 variáveis comportamentais são indicadas na secção seguinte, onde já estão reordenadas de acordo com os resultados obtidos pela análise classificatória efectuada. Quanto aos indivíduos, a amostra tinha sido

seleccionada em 18 grupos etários, o primeiro grupo com doze crianças e todos os restantes com dez crianças. Os limites de idade para cada grupo eram:

- entre os seis e os nove meses,
- entre os nove e os doze meses,
- entre os doze e os quinze meses,
- entre os quinze e os dezoito meses,
- entre os dezoito e os vinte e um meses,
- entre os vinte e um e os vinte e quatro meses,
- entre os vinte e quatro e os vinte e sete meses,
- entre os vinte e sete e os trinta meses,
- entre os trinta e os trinta e três meses,
- entre os trinta e três e os trinta e seis meses,
- entre os trinta e seis e os trinta e nove meses,
- entre os trinta e nove e os quarenta e dois meses,
- entre os quarenta e dois e os quarenta e cinco meses,
- entre os quarenta e cinco e os quarenta e oito meses,
- entre os quarenta e oito e os cinquenta e um meses,
- entre os cinquenta e um e os cinquenta e quatro meses,
- entre os cinquenta e quatro e os cinquenta e sete meses,
- e entre os cinquenta e sete e os sessenta e três meses.

A distribuição das crianças por grupo etário, idade e sexo vem apresentada nas tabelas da Secção 4, onde as crianças já estão reordenadas de acordo com os resultados obtidos pela análise classificatória efectuada. Pretendia-se, através de métodos de análise exploratória de dados multivariados, identificar os grupos de características/variáveis comportamentais mais importantes e reagrupar as crianças de acordo com essas características. A metodologia aqui apresentada refere-se à utilização de modelos de Análise Classificatória Hierárquica Ascendente. Foi utilizado o *software* SPSS.

Para efeito da escolha do método de classificação utilizaram-se:

- para as variáveis, os critérios de agregação “Average Linkage Between Groups”, “Average Linkage Within Groups”, “Complete Linkage / Furthest Neighbour” e “Centroid Method”;
- para os indivíduos / crianças, o “Ward’s Method”, além dos anteriores.

Como medidas de proximidade entre pares de variáveis ou entre pares de crianças utilizaram-se também vários dos coeficientes de semelhança e de dissemelhança propostos para dados binários.

### 3 Resultados para as variáveis

Relativamente ao agrupamento das variáveis, o resultado aqui apresentado foi obtido com o critério de agregação “Average Linkage Between Groups” e com o coeficiente de Ochiai, tendo os outros modelos classificatórios chegado a resultados bastante aproximados. As melhores partições (partições mais



“significativas”), são as que apresentam respectivamente 7, 5 e 3 classes de variáveis:

Grupo A: V003, V025, V026, V028, V029, V031, V032, V033, V034, V036, V039, V043; respectivamente: Diálogo pré-verbal; Objectos dentro fora; Reage a algumas palavras; Rotinas diárias; Apoia-se ergue-se de pé; Diz palavras com duas sílabas; Fasc. Aparece/desaparece; “Dá” um objecto; Entende o que é proibição; Cubos, bate-os com alegria; Agarra em pinça; Reconhece proibidos.

Grupo B: V027, V030, V035, V037, V040, V041, V042, V044, V045, V046, V047, V048, V049, V050, V051, V052, V053, V054, V055, V056, V057, V058, V059, V060, V061, V064, V065, V066, V067, V068, V069, V070, V071, V075, V077, V078, V092; respectivamente: Objectos rel. Corpo; Reconhece uma lenga-lenga; Faz “*tem-tem*”; Desloca-se apoiando-se em móveis; Aponta para objecto desejado; Reconhece rótulos verbais; Descalça as meias a pedido; Actos de “*imitação diferida*”; Enfia argolas em estacas; Aplica rótulos verbais holofrases; Sobe degraus gatinhando; Reconhece e evita obstáculos; Bebe por um copo sozinha; Serve-se de colher para comer a papa; Sobe e desce degraus; Colabora nas rotinas diárias; Brinca “*com panelinhas*”; Diz oito palavras (sem perfeição); Faz torre de dois cubos, a pedido; Pede para beber ou comer; Combina duas palavras; Repete um acto provocando o riso; A pedido, aponta para uma parte corpo; Desce degraus com passos alternados; Dá pontapé numa bola para frente; Faz torre de quatro cubos, a pedido; Segue instruções verbais; Ajuda a arrumar brinquedos, a pedido; A pedido, nomeia uma imagem; Brinca com outros “*ao lado*”; A pedido, aponta o lugar de objectos; Brinca mat mini função própria; Brincando desfaz e refaz formas; Brinca no formato destruir/ construir; Reconhece lugar coisas uso; Escolhe e leva adiante uma actividade; Distingue o nomes dos colegas.

Grupo C: V062, V063, V072, V073, V074, V076, V079, V080, V084, V090, V097, V098; respectivamente: Faz garatuja bordos papel; Formula “*frases*” de várias palavras; Distingue os termos “*grande*” e “*pequeno*”; Enfia contas de formato grande; Brinca com “*encaixes*”; Ensaia melodias; Discrimina verbalmente não presta; Brinca com cenários “*faz-de-conta*”; Recorda acontecimentos vividos; Nomeia as partes do corpo humano; Interessa-se por livros com imagens; Brinca com “*outros*”, cooperando.

Grupo D: V081, V082, V083, V085, V088, V091, V093, V094, V096, V099, V102, V103, V104, V108, V110; respectivamente: Verbalmente discriminar muito; Discrimina verbalmente o “*pesado*” e o “*leve*”; Discrimina verbalmente “*doce*”; Emprega os termos “*grande*” e “*pequeno*”; Discrimina verbalmente “*atrás*”; Emprega termos “*à frente/ atrás de si*”; Ordena material (nomeando); Calça meias e sapatos; Distingue verbalmente “*ao lado*”; Despe o casaco; Conta eventos, encadeando; Toma alguma iniciativa nas rotinas diárias; Veste roupa simples; Distingue verbalmente “*cheio*” e “*vazio*”; Imita gestos/ posições do corpo.

Grupo E: V086, V087, V089, V095, V112, V114, V127, V128; respectivamente: Emprega os termos “*alto*” e “*baixo*”; Distingue verbalmente “*alguns*” e “*todos*”; Discrimina verbalmente “*macio*”; Distingue verbalmente “*parecido*”; Mostra ter alguma noção de espaço; Observa as características dos animais; Encadeia frases – “*conversa*”; Verbaliza noções de parentesco.

Grupo F: V100, V101, V105, V106, V107, V109, V113, V115, V116, V121, V126, V129, V130, V132, V133, V134, V136; respectivamente: Desenha e representa graficamente; Reconhece diferença entre letras e números; Distingue/ nomeia materiais; Molda formas reconhecíveis; Distingue verbalmente “*manhã*”, etc.; Distingue verbalmente “*ontem*” e “*hoje*”; “*Lê*” a primeira letra do nome; Representa para outros; Fala nos tempos: “*hoje*”, “*ontem*” e “*amanhã*”; Descreve espontaneamente eventos; Joga dominó de imagens; Seria objectos por tamanhos; Copia grafismos redondos e quadrados; Jogos elementares com regras; Participa no planeamento de uma actividade; Quando arruma, faz montes; Persiste em tarefas deliberadas.

Grupo G: V111, V117, V118, V119, V120, V122, V123, V124, V125, V131, V135, V137, V138, V139, V140, V141, V142, V143, V144, V145, V146, V147; respectivamente: Observa as horas no relógio; Surge linha do céu/ linha da terra; Nomeia dia da semana e estações do ano; Escreve o nome (maiúsculas); Enuncia e escreve até 5; Identifica a mão direita e a mão esquerda; Observa o crescimento de sementes; Observa atenta fenómenos naturais; Faz colecções de materiais; Discrimina sinais de trânsito; Distingue verbalmente o “*parecer*” e o “*ser*”; Termina tarefas no tempo previsto; Enuncia e escreve até 10; Copia losango a pedido; Desenha a figura humana completa; Sérias simples (histórias); Classificações simples pedidas; Mapas de caminhos; Executa construções pequenas; Faz previsões de factos físicos; Dinheiro para pequenas compras; Desenha figura humana com vestuário.

Considerando a partição em 5 classes, as variáveis agrupam-se como segue:

- Classe 1 – coincide com o Grupo A;
- Classe 2 – coincide com o Grupo B;
- Classe 3 – coincide com o Grupo C;
- Classe 4 – agrega o Grupo D e o Grupo E;
- Classe 5 – junta o Grupo F e o Grupo G.

Finalmente, na partição menos fina das três, com 3 classes, as variáveis agrupam-se da seguinte forma:

- Classe 1 – agrega o Grupo A e o Grupo B;
- Classe 2 – constituída pelos Grupo C, Grupo D e Grupo E;
- Classe 3 – junta o Grupo F e o Grupo G.

## 4 Resultados para os indivíduos

Relativamente ao agrupamento dos indivíduos, referiremos o resultado obtido com o “Ward’s Method” e a medida de distância “Binary Squared Euclidian Distance”. As melhores partições são as que apresentam respectivamente 8, 5 e 3 classes de indivíduos. A partição das 172 crianças em oito classes de indivíduos, está indicada nos seguintes grupos:

Grupo Q:

1	7	9	2	3	4	5	6	8
10	11	12	15	13	14	16	17	18
19	20	21	22	24	26	29	27	28
30	31	32						

Grupo R:

25	23	35	38	44	52	70		
----	----	----	----	----	----	----	--	--

Grupo S:

33	34	36	37	39	41	42	45	43
48	49	47	50	53	55	60	57	59
69	63	68	66	67				

Grupo T:

40	51	54	56	62	58	61	64	71
72	77	78	74	79	84	85	83	88
89	96	99	103	107	108	122	124	136
138	140	149						

Grupo U:

46	73	82	76	75	80	91	92	97
95	98	106	111	115	117	146		

Grupo V:

81	87	102	104	105	110	121	132	145
----	----	-----	-----	-----	-----	-----	-----	-----

Grupo X:

65	86	90	93	94	100	101	112	113
114	116	118	119	120	123	125	127	128
129	130	133	134	135	141	142	156	161
169								

Grupo Z:

109	126	131	137	139	143	144	147	148
150	151	152	153	158	154	155	159	160
157	162	170	163	166	167	164	165	168
171	172							

Considerando a partição em 5 classes, os indivíduos agrupam-se da seguinte forma:

- Classe 1 – coincide com o Grupo Q;
- Classe 2 – agrega o Grupo R e o Grupo S;
- Classe 3 – coincide com o Grupo T;
- Classe 4 – agrega o Grupo U, o Grupo V e o Grupo X;
- Classe 5 – coincide com o Grupo Z.

Na partição em 3 classes, os indivíduos agrupam-se da seguinte forma:

- Classe 1 – coincide com o Grupo Q;
- Classe 2 – constituída pelos Grupo R, Grupo S e Grupo T;
- Classe 3 – engloba os Grupo U, Grupo V, Grupo X e Grupo Z.

A distribuição das crianças pelas oito classes, com indicação dos respectivos grupos etários, idade e sexo é a seguinte:

Criança n.º	Grupo etário	Idade em meses	Sexo
-------------	--------------	----------------	------

Grupo Q

Q 1	1	6	Masc.
Q 7	1	6	Fem.
Q 9	1	6	Fem.
Q 2	1	7	Masc.
Q 3	1	7	Fem.
Q 4	1	7	Fem.
Q 5	1	8	Fem.
Q 6	1	8	Masc.
Q 8	1	8	Fem.
Q 10	1	8	Fem.
Q 11	1	9	Masc.
Q 12	1	9	Masc.
Q 15	2	9	Masc.
Q 13	2	10	Fem.
Q 14	2	10	Masc.
Q 16	2	11	Fem.
Q 18	2	11	Masc.
Q 17	2	12	Fem.
Q 19	2	12	Masc.
Q 20	2	12	Masc.
Q 21	2	12	Fem.
Q 22	2	12	Fem.
Q 24	3	12	Fem.
Q 26	3	12	Fem.
Q 29	3	12	Fem.
Q 27	3	14	Masc.
Q 28	3	14	Masc.
Q 30	3	14	Fem.
Q 31	3	14	Fem.
Q 32	3	14	Masc.

Grupo R

R 25	3	12	Masc.
R 23	3	13	Masc.
R 35	4	15	Fem.
R 38	4	17	Fem.
R 44	5	18	Masc.
R 52	5	19	Masc.
R 70	7	25	Fem.

Grupo S

S 33	4	15	Masc.
S 34	4	15	Masc.
S 36	4	16	Masc.
S 37	4	16	Masc.
S 39	4	17	Fem.
S 41	4	17	Fem.
S 42	4	18	Fem.
S 45	5	18	Fem.
S 43	5	19	Fem.
S 48	5	19	Fem.
S 49	5	19	Masc.
S 47	5	20	Masc.
S 50	5	21	Fem.
S 53	6	21	Masc.
S 55	6	21	Fem.
S 60	6	21	Masc.
S 57	6	23	Masc.
S 59	6	23	Masc.
S 69	7	24	Masc.
S 63	7	25	Fem.
S 68	7	25	Fem.
S 66	7	26	Masc.
S 67	7	26	Masc.

Grupo T

T 40	4	16	Masc.
T 51	5	21	Masc.
T 54	6	21	Fem.
T 56	6	21	Fem.
T 62	6	21	Masc.
T 58	6	23	Fem.
T 61	6	23	Fem.
T 64	7	25	Masc.
T 71	7	25	Fem.
T 72	7	25	Masc.
T 77	8	28	Fem.
T 78	8	28	Fem.
T 74	8	29	Masc.
T 79	8	30	Fem.
T 84	9	30	Masc.
T 85	9	30	Fem.
T 83	9	32	Fem.
T 88	9	32	Masc.
T 89	9	32	Masc.
T 96	10	35	Fem.
T 99	10	35	Masc.
T 103	11	36	Masc.
T 107	11	36	Fem.
T 108	11	36	Fem.
T 122	12	42	Masc.
T 124	13	43	Fem.
T 138	14	45	Masc.
T 140	14	46	Fem.
T 136	14	47	Fem.
T 149	15	49	Masc.

Grupo U

U 46	5	20	Fem.
U 73	8	27	Masc.
U 82	8	27	Masc.
U 76	8	28	Masc.
U 75	8	29	Fem.
U 80	8	30	Fem.
U 91	9	31	Fem.
U 92	9	31	Masc.
U 97	10	33	Fem.
U 95	10	34	Masc.
U 98	10	35	Masc.
U 106	11	36	Masc.
U 111	11	37	Fem.
U 115	12	40	Masc.
U 117	12	41	Masc.
U 146	15	50	Masc.

Grupo V

V 81	8	30	Masc.
V 87	9	31	Masc.
V 102	10	34	Masc.
V 104	11	37	Masc.
V 105	11	37	Masc.
V 110	11	39	Fem.
V 121	12	41	Fem.
V 132	13	42	Fem.
V 145	15	50	Masc.

Grupo X

X 65	7	25	Fem.
X 86	9	31	Fem.
X 90	9	32	Fem.
X 93	10	34	Fem.
X 94	10	34	Masc.
X 100	10	34	Masc.
X 101	10	34	Fem.
X 112	11	37	Fem.
X 113	12	40	Masc.
X 114	12	40	Masc.
X 116	12	41	Fem.
X 118	12	41	Fem.
X 119	12	42	Fem.
X 120	12	42	Fem.
X 123	13	42	Fem.
X 130	13	42	Masc.
X 128	13	43	Masc.
X 125	13	44	Masc.
X 127	13	44	Masc.
X 129	13	44	Masc.
X 133	14	45	Masc.
X 134	14	46	Fem.
X 135	14	47	Fem.
X 141	14	47	Masc.
X 142	14	48	Masc.
X 156	16	55	Masc.
X 161	16	57	Masc.
X 169	17	58	Masc.

Grupo Z

Z 109	11	37	Masc.
Z 126	13	43	Masc.
Z 131	13	44	Masc.
Z 137	14	47	Masc.
Z 139	14	48	Fem.
Z 144	15	49	Masc.
Z 147	15	50	Masc.
Z 148	15	50	Fem.
Z 150	15	52	Masc.
Z 151	15	52	Fem.
Z 152	15	53	Fem.
Z 143	15	54	Fem.
Z 153	16	54	Masc.
Z 158	16	54	Fem.
Z 154	16	55	Masc.
Z 155	16	55	Fem.
Z 159	16	55	Fem.
Z 160	16	55	Masc.
Z 157	16	56	Fem.
Z 162	16	57	Fem.
Z 170	17	58	Fem.
Z 163	17	60	Masc.
Z 166	17	60	Fem.
Z 167	17	61	Fem.
Z 164	17	62	Fem.
Z 165	17	62	Fem.
Z 168	17	63	Fem.
Z 171	17	63	Masc.
Z 172	17	63	Masc.

## 5 Conclusões

De acordo com os resultados obtidos e os objectivos do estudo, vamos considerar a partição dos indivíduos e a correspondente partição das variáveis, em 5 classes ou em 3 classes. Assim, em vez dos 18 grupos iniciais de crianças, teremos 5 (ou 3) grupos, cada um descrito pela classe de variáveis mais adequada.

As idades das crianças que limitam a partição em 5 classes (excluindo os extremos) são:

1ª Classe	6 a 14 meses			
2ª Classe		13 a 26 meses		
3ª Classe			21 a 47 meses	
4ª Classe				25 a 57 meses
5ª Classe				43 a 63 meses

Ou seja, em classes mutuamente exclusivas:

1ª Classe	6 a 14 meses			
2ª Classe		15 a 26 meses		
3ª Classe			27 a 47 meses	
4ª Classe				48 a 57 meses
5ª Classe				58 a 63 meses

As idades das crianças que limitam a partição em 3 classes (excluindo os extremos) são:

1ª Classe	6 a 14 meses		
2ª Classe		13 a 47 meses	
3ª Classe			25 a 63 meses

Em classes mutuamente exclusivas, vem, respectivamente:

1ª Classe	6 a 14 meses		
2ª Classe		15 a 47 meses	
3ª Classe			48 a 63 meses

A primeira está em correspondência com a classificação das variáveis em cinco classes e a segunda está em correspondência com a classificação das variáveis em três classes. De acordo com o objectivo do trabalho, verificou-se que a partição em três classes não traria, porém, qualquer utilidade do ponto de vista da acção educativa. Propõe-se portanto a partição dos indivíduos e a correspondente partição das variáveis, em 5 classes.

Assim, na partição em cinco classes:

**A primeira classe, Grupo Q, das crianças** com idades compreendidas entre os **6 e os 14 meses**, é principalmente **caracterizada pelo Grupo A de variáveis**, isto é, V003-Diálogo pré-verbal, V025-Objectos dentro fora, V026-Reage a algumas palavras, V028-Rotinas diárias, V029-Apoio-se ergue-se de pé, V031-Diz palavras com duas sílabas, V031-Diz palavras com duas sílabas, V032-Fasc. Aparece/desaparece, V033-“*Dá*” um objecto, V034-Entende o que é proibição, V036-Cubos, bate-os alegria, V039-Agarra em pinça, V043-Reconhece proibidos.

**A segunda classe**, que agrega o **Grupo R** e o **Grupo S das crianças** com idades entre **13 e 26 meses** (15 a 26, numa classificação com classes não sobrepostas), é principalmente **caracterizada pelo Grupo B de variáveis**, isto é, V027-Objectos rel. Corpo, V030-Reconhece uma lenga-lenga, V035-Faz “*tem-tem*”, V037-Desloca-se apoiando-se em móveis, V040-Aponta para objecto desejado, V041-Reconhecer rótulos verbais, V042-Descalça as meias a pedido, V044-Actos de “*imitação diferida*”, V045-Enfia argolas em estacas, V046-Aplica rótulos verbais holofrases, V047-Sobe degraus gatinhando, V048-Reconhece e evita obstáculos, V049-Bebe por um copo sozinha, V050-Serve-se de colher para comer a papa, V051-Sobe e desce degraus, V052-Colabora nas rotinas diárias, V053-Brinca “*com panelinhas*”, V054-Diz oito palavras (sem perfeição), V055-Faz torre de dois cubos, a pedido, V056-Pede para beber ou comer, V057-Combina duas palavras, V058-Repete um acto provocando o riso, V059-A pedido, aponta para uma parte corpo, V060-Desce degraus com passos alternados, V061-Dá pontapé numa bola para frente, V064-Faz torre de quatro cubos, a pedido, V065-Segue instruções verbais, V066-Ajuda a arrumar brinquedos, a pedido, V067-A pedido, nomeia uma imagem, V068-Brinca com outros “*ao lado*”, V069-A pedido, aponta o lugar de objectos, V070-Brinca mat mini função própria, V071-Brincando desfaz e refaz formas, V075-Brinca no formato destruir/ construir, V077-Reconhece lugar coisas uso, V078-Escolhe e leva adiante uma actividade, V092-Distingue o nomes dos colegas.

**A terceira classe, Grupo T**, entre **21 e 47 meses** (27 a 47, numa classificação com classes não sobrepostas), é principalmente **caracterizada pelo Grupo C de variáveis**, isto é, V062-Faz garatuja bordos papel, V063-Formula “*frases*” de várias palavras, V072-Distingue os termos “*grande*” e “*pequeno*”, V073-Enfia contas de formato grande, V074-Brinca com “*encaixes*”, V076-Ensaia melodias, V079-Discrimina verbalmente não presta, V080-Brinca com cenários “*faz-de-conta*”, V084-Recorda acontecimentos vividos, V090-Nomeia as partes do corpo humano, V097-Interessa-se por livros com imagens, V098-Brinca com “*outros*”, cooperando.

**A quarta classe**, que agrega o **Grupo U**, o **Grupo V** e o **Grupo X** de crianças, entre **25 e 57 meses** (48 a 57, numa classificação com classes não sobrepostas), é principalmente **caracterizada pela Classe 4 de variáveis**, que agrega o **Grupo D e o Grupo E**, isto é, V081-Verbalmente discriminar muito, V082-Discrimina verbalmente o “*pesado*” e o “*leve*”, V083-Discrimina verbalmente “*doce*”, V085-

Emprega os termos “grande” e “pequeno”, V088-Discrimina verbalmente “atrás”, V091-Emprega termos “frente/atrás si”, V093-Ordena material (nomeando), V094-Calça meias e sapatos, V096-Distingue verbalmente “ao lado”, V099-Despe o casaco, V102-Conta eventos, encadeando, V103-Toma alguma iniciativa nas rotinas diárias, V104-Veste roupa simples, V108-Distingue verbalmente “cheio” e “vazio”, V110-Imita gestos/ posições do corpo, V086-Emprega os termos “alto” e “baixo”, V087-Distingue verbalmente “alguns” e “todos”, V089-Discrimina verbalmente “macio”, V095-Distingue verbalmente “parecido”, V112-Mostra ter alguma noção de espaço, V114-Observa as características dos animais, V127-Encadeia frases – “conversa”, V128-Verbaliza noções de parentesco.

**A quinta classe, que coincide com o Grupo Z** de crianças, entre **43 a 63 meses** (58 a 63, numa classificação com classes não sobrepostas), é principalmente **caracterizada pela Classe 5 de variáveis, que junta o Grupo F e o Grupo G**, isto é, V100-Desenha e representa graficamente, V101-Reconhece diferença entre letras e números, V105-Distinguir/ nomear materiais, V106-Molda formas reconhecíveis, V107-Distingue verbalmente “manhã”, etc., V109-Distingue verbalmente “ontem” e “hoje”, V113-“Lê” a primeira letra do nome, V115-Representa para outros, V116-Fala nos tempos: “hoje”, “ontem” e “amanhã”, V121-Descreve espontaneamente eventos, V126-Joga dominó de imagens, V129-Seria objectos por tamanhos, V130-Copia grafismos redondos e quadrados, V132-Jogos elementares com regras, V133-Participa no planeamento de uma actividade, V134-Quando arruma, faz montes, V136-Persiste em tarefas deliberadas, V111-Observa as horas no relógio, V117-Surge linha céu ou linha da terra, V118-Nomeia dia da semana e estações do ano, V119-Escreve o nome (maiúsculas), V120-Enuncia e escreve até 5, V122-Identifica a mão direita e a mão esquerda, V123-Observa o crescimento de sementes, V124-Observa atenta fenómenos naturais, V125-Faz colecções de materiais, V131-Discrimina sinais de trânsito, V135-Distingue verbalmente o “parecer” e o “ser”, V137-Termina tarefas no tempo previsto, V138-Enuncia e escreve até 10, V139-Copia losango a pedido, V140-Desenha a figura humana completa, V141-Seriações simples (histórias), V142-Classificações simples pedidas, V143-Mapas de caminhos, V144-Executa construções pequenas, V145-Faz previsões factos físicos, V146-Dinheiro para pequenas compras, V147-Desenha figura humana com vestuário.

A correspondência entre as novas classes etárias e as classes de variáveis que as caracterizam, bem como a respectiva interpretação dos resultados, tendo em conta a informação a priori sobre a caracterização dos grupos inicialmente propostos, permitiu identificar os grupos de variáveis comportamentais mais importantes e reagrupar as crianças de acordo com esses novos grupos.

Os resultados obtidos são bem complementados com os que foram fornecidos pelo método da análise factorial das correspondências efectuada sobre os mesmos dados.



## Referências

- BACELAR-NICOLAU, H.(1981). *Contribuições ao Estudo da Comparação de Coeficientes em Análise Classificatória*. Tese Doutoramento, Universidade de Lisboa.
- BACELAR-NICOLAU, H. (2005). Introdução à Análise Classificatória Hierárquica (Modelos de ACH Ascendente). *Notas e Comunicações do LEAD/ FPUL*.
- CAXARIA, A.M. (1998). *Análise Multivariada de Dados num Modelo de Avaliação das Tecnologias de Informação na Administração Pública*, Tese de Mestrado, ISEGI.
- GORDON, A.D. (1999). *Classification*, 2<sup>nd</sup> Edition, Chapman & Hall, London.
- LEAL, R.M. (1971). *A Avaliação do Nível de Funcionamento Psico-Social no Jardim Infantil*. Escola A.E.I de S. Tomé e S.C.A.I.D., Lisboa.
- LEAL, R. M. (1985). *Introdução ao Estudo dos Processos de Socialização Precoce da Criança*. (Tradução do inglês, pela autora, Associação de Estudantes da Faculdade de Psicologia da Universidade de Lisboa. Re-editado pelo IPAF-S.Paulo, 2004.
- LERMAN, I.C. (1981). *Classification et Analyse Ordinale des Données*, Dunod, Paris.
- NICOLAU, F., BACELAR-NICOLAU, H. (1981). Novos aspectos da Análise Classificatória. *Actas do II Colóquio de Estatística e Investigação Operacional*, 322-341, Fundação / Covilhã.
- NICOLAU, F., BACELAR-NICOLAU, H. (1998). Some Trends in the Classification of Variables, in *Data Science, Classification, and Related Methods* (eds. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, Y. Baba), Springer-Verlag, 89-98.
- RODRIGUES, P. (2004). Aplicações de Estatística e Análise de Dados. *Instituto de Informática*.

# Estimação da Abundância e das Taxas de Incidência e Prevalência em Populações Elusivas

Anabela Afonso<sup>1</sup> · João F. Monteiro<sup>2</sup> · Sónia Batista<sup>3</sup> · Russell Alpizar-Jara<sup>4</sup>

© The Author(s) 2013

**Resumo** Na literatura sobre a estimação de parâmetros populacionais é possível encontrar vários métodos para estimar a abundância e as taxas de incidência e prevalência. A escolha do método depende não só do tipo e características da população em estudo como também do conjunto de pressupostos em que os métodos de amostragem assentam. Neste trabalho efectua-se uma breve descrição da amostragem por distâncias, do método de captura-recaptura e da combinação dos dois métodos anteriores, e são analisados dois casos de estudo onde se utilizaram estes métodos de amostragem.

**Palavras-chave:** abundância, amostragem por distâncias, modelos de captura-recaptura, taxa de incidência, taxa de prevalência.

## 1 Introdução

Cada vez é mais comum surgirem no dia-a-dia notícias, nos meios de comunicação social, onde se apresentam estimativas para o número total de indivíduos que possuem uma certa característica. Eis alguns exemplos:

- Governo adquire 400 mil vacinas contra o vírus do papiloma humano (HPV) ... um vírus que mata 300 mulheres por ano em Portugal. (Diário de Notícias, 31/12/2008).

---

<sup>1</sup>Departamento de Matemática e CIMA, Universidade de Évora, aafonso@uevora.pt

<sup>2</sup>Departamento de Matemática e CIMA, Universidade de Évora, jfgm@uevora.pt

<sup>3</sup>Departamento de Matemática, Universidade de Évora, sraquelpc@hotmail.com

<sup>4</sup>Departamento de Matemática e CIMA, Universidade de Évora, alpizar@uevora.pt

- O mais recente estudo de avaliação da vida animal mostra que pelo menos 1141 das 5487 espécies de mamíferos estão ameaçadas de extinção. (Diário de Notícias, 7/10/2008).
- Uma média de 15 portugueses são todos os anos infectados pela leishmaniose, uma doença canina transmissível ao homem ainda pouco conhecida mas que pode ser fatal para animais e humanos quando não tratada. (Expresso, edição online, 12/1/2009).

Contudo uma má estimação dos parâmetros de interesse tem vários riscos associados. Do ponto de vista epidemiológico, se por um lado a subestimação do número de portadores de uma certa doença pode originar um surto ou uma epidemia, com a sobrestimação verifica-se um desperdício de recursos hospitalares. Um exemplo bem recente, e muito divulgado, de uma má estimação foi o iminente surto de gripe A (vírus H1N1), que originou a produção de um número elevado de vacinas e a alocação excessiva de recursos hospitalares.

Do ponto de vista biológico, as medidas de gestão dos recursos naturais baseiam-se, entre outros factores, na biodiversidade. Para tal é fundamental conhecer quais as espécies em vias de extinção e quais as espécies em sobreabundância que podem interferir no equilíbrio ecológico.

Na literatura é possível encontrar diversos métodos de estimação e nem sempre é fácil escolher o método adequado. Outras vezes nem sempre é viável a sua aplicação, porque os pressupostos inerentes ao método não se verificam (Alpizar-Jara *et al.*, 2008). As populações elusivas caracterizam-se por serem difíceis de capturar ou detectar. Neste trabalho pretendemos divulgar dois métodos de amostragem, deste tipo de populações, que têm tido um grande desenvolvimento nos últimos anos: o método de captura-recaptura e a amostragem por distâncias. Deste modo, na secção que se segue introduzem-se estes métodos e posteriormente apresentam-se dois exemplos de aplicação. Na última secção discutem-se as limitações destes métodos.

## **2 Métodos de amostragem por distâncias e captura-recaptura**

### **2.1 Amostragem por distâncias**

A amostragem por distâncias é uma técnica bastante utilizada para estimar a abundância ou densidade de indivíduos numa determinada área de estudo, bem como a probabilidade de detecção. Esta técnica tem sido bastante utilizada em populações animais e vegetais.

Nos métodos mais usuais de amostragem por distâncias, o observador percorre uma série de linhas (ou pontos), designados por transectos lineares (ou pontuais), e regista a distância perpendicular (ou radial) dos indivíduos detectados a essas linhas (ou pontos). Uma vez que nem toda a área de estudo é observada nem todos

os indivíduos na área coberta são detectados. O estimador da abundância de indivíduos,  $\hat{N}$ , na área em estudo,  $A$ , é dado pelo número de indivíduos detectados,  $n$ , corrigido pela probabilidade de cobertura,  $P_c$ , e pela probabilidade de detecção,  $P_a$ , i. e.

$$\hat{N} = \frac{n}{P_c \hat{P}_a}, \text{ com } 0 < P_c \leq 1 \text{ e } 0 < \hat{P}_a \leq 1.$$

A probabilidade de cobertura,  $P_c$ , assume-se conhecida e dada por  $a/A$ , onde  $a$  representa a área amostrada. No entanto a probabilidade de detecção,  $P_a$ , é desconhecida e portanto tem de ser estimada. O ponto chave da amostragem por distâncias assenta na forma como esta probabilidade é estimada. Com base nas distâncias observadas, constrói-se um histograma ao qual é ajustada uma função, que se designa por função de detecção,  $\hat{g}(x)$ , e representa a  $P(\text{detectar}|\text{distância } x)$  (Buckland *et al.*, 2001). Como a presença de distâncias atípicas dificultam o processo de estimação da função  $g(x)$ , é usual truncar parte das maiores distâncias observadas, definindo-se  $w$  como a distância de truncatura. A probabilidade de detecção é estimada por

$$\hat{P}_a = \frac{1}{w} \int_0^w \hat{g}(x) dx.$$

Os pressupostos base dos modelos clássicos da amostragem por distâncias são: i) todos os objectos sobre a linha ou ponto são detectados, i.e.,  $g(0) = 1$ ; ii) os objectos são detectados antes de efectuarem qualquer movimento em reacção ao observador; e iii) as distâncias são medidas correctamente. Outros pressupostos que também são considerados são: iv) as detecções são independentes; e v) as distâncias dos objectos aos transectos são uniformes, sendo este último pressuposto garantido com a colocação aleatória ou sistemática dos transectos (Buckland *et al.*, 2001). Para desenvolvimentos deste método de amostragem ver, por exemplo, Buckland *et al.* (2004).

## 2.2 Amostragem por captura-recaptura

A amostragem por captura-recaptura tem sido usada em diversas áreas, como por exemplo biologia, epidemiologia e ciências sociais. Os modelos de captura-recaptura permitem estimar as taxas de incidência ou prevalência de uma certa doença, as taxas de natalidade, mortalidade ou morbilidade e ainda a dimensão (abundância) de uma população possuidora de uma determinada característica. O método consiste na captura, marcação e libertação dos indivíduos da população em vários instantes de amostragem. Com base na informação das recapturas é possível estimar a abundância.

O modelo de captura-recaptura mais simples é composto por dois instantes de amostragem. Na primeira ocasião recolhe-se uma amostra de  $n_1$  indivíduos, que são marcados e devolvidos à população com  $N$  animais, para que se misturem

aleatoriamente com a população com  $N - n_1$  animais não marcados. Após um certo período de tempo suficientemente curto para que não se verifiquem entradas nem saídas da população (assume-se uma população fechada), recolhe-se uma nova amostra de  $n_1$  indivíduos, verificando-se que destes  $n_{11}$  indivíduos já estavam marcados (recapturas). Admitindo que todos os indivíduos têm igual probabilidade de captura (homogeneidade) e que as marcas não se perdem nem são contadas mais do que uma vez pelo observador, então é de esperar que a proporção de animais marcados na segunda amostra seja semelhante à de animais marcados na população total. Desta simples relação resulta o estimador de Lincoln-Petersen para a dimensão da população (Lincoln, 1930; Petersen, 1896):

$$\hat{N} = \frac{n_1 \cdot n_1}{n_{11}}, \text{ com } n_{11} > 0.$$

Para muitas populações, o pressuposto de homogeneidade das probabilidades de captura é pouco realista, quer devido a factores externos entre diferentes momentos de amostragem ( $t$  = temporal), quer a características intrínsecas dos indivíduos ( $h$  = heterogeneidade) ou à reacção dos indivíduos ao método de marcação ( $b$  = comportamento). Deste modo, Otis *et al.* (1978) propuseram 8 modelos que combinam estas fontes de heterogeneidade e têm como parâmetros  $N$ , o tamanho da população, e  $p_{ij}$ , a probabilidade do indivíduo  $i$  ser capturado no instante  $j$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, k$ . Devido à sua complexidade, no modelo que considera as três fontes de variabilidade, designado por  $M_{tbh}$ , não é possível estimar todos os seus  $N \times k + 1$  parâmetros, sem que sejam impostas algumas restrições sobre os mesmos.

Mais tarde foram propostos modelos para populações abertas, i.e. cujo tamanho da população pode variar ao longo do tempo devido a entradas (nascimentos ou imigrações) e saídas (mortes ou emigrações). Nestes modelos para além da estimação das probabilidades de captura, são estimadas as dimensões da população nos vários instantes de amostragem bem como das taxas de sobrevivência e de mortalidade (ver por exemplo Williams *et al.*, 2002).

Pollock (1982) propôs a combinação de modelos para populações fechadas e abertas numa análise integrada que permite obter estimadores dos parâmetros populacionais que sejam robustos à eventual violação dos pressupostos. Este delineamento é conhecido por delineamento robusto de Pollock.

## 2.3 Combinação da amostragem por distâncias com os métodos de captura-recaptura

O pressuposto de detecção certa sobre a linha ou ponto,  $g(0) = 1$ , é muitas vezes violado, podendo o estimador convencional ser bastante enviesado. Para estimar a probabilidade de detecção sobre a linha, Alpizar-Jara e Pollock (1996, 1999) e Borchers *et al.* (1998a,b) propuseram modelos que combinam a amostragem por transectos lineares com os métodos de captura-recaptura.

Um possível caso de combinação apresenta-se quando na população existem  $n_1$  indivíduos previamente marcados como resultado de uma possível monitorização. Num segundo momento de amostragem é realizado uma amostragem por transectos lineares registando-se a distância aos  $n_1$  indivíduos detectados e o número  $n_{11}$  de indivíduos marcados. Com a informação recolhida são estimadas as probabilidades de captura e de detecção sobre a linha do transecto,  $g_0$ , que permitirão corrigir o enviesamento do estimador da abundância causado pela violação do pressuposto  $g(0) = 1$ .

O estimador da abundância para o modelo combinado mostrou ser mais eficiente e menos enviesado que o estimador baseado em apenas um dos métodos. No modelo combinado, para além dos pressupostos subjacentes aos modelos de amostragem por distâncias, assume-se ainda que: i) os indivíduos marcados são identificados univocamente, ii) os indivíduos têm igual probabilidade de serem marcados ou avistados dentro de cada amostra, e iii) existe independência entre os animais marcados e avistados.

Tradicionalmente, a estimação dos parâmetros é efectuada pelo método da máxima verosimilhança. Monteiro e Alpizar-Jara (2006) estendem a metodologia utilizando uma abordagem de estimação bayesiana. Esta metodologia apresenta como principais vantagens a facilidade de interpretação dos intervalos de confiança para os parâmetros de interesse e a possibilidade de modelar heterogeneidade não observável nas probabilidades de detecção. Os métodos bayesianos permitem ajustar modelos complexos sem ser necessário recorrer a aproximações assintóticas (Brooks *et al.*, 2000; 2002).

### 3 Aplicações

#### 3.1 Estimação da taxa de prevalência da diabetes

Apresenta-se nesta subsecção um exemplo fictício para ilustrar esta abordagem. Numa certa cidade e num determinado ano, verificou-se que dos 123 pacientes com diagnóstico de diabetes, que estiveram nos centros de saúde/médicos de família, 27 também recorreram às urgências do Hospital dessa cidade, e no total estiveram 64 pacientes nas urgências do referido Hospital (Tabela 1). Desconhece-se o número de pacientes com diabetes que não recorreram a estes serviços.

A taxa de prevalência representa a proporção de pessoas com um determinado resultado de saúde relativamente à população de interesse. Admitindo que nessa cidade existiam 50.000 habitantes e considerando apenas 160 casos de diabetes diagnosticados, estimou-se uma taxa de prevalência de 3,2 por 1000 habitantes. No entanto este valor subestima a verdadeira taxa de prevalência, uma vez que não são considerados os pacientes que não recorreram aos centros de saúde e ao Hospital. Com base no estimador da abundância de Lincoln-Petersen o número estimado de pacientes com diabetes é 292, sendo a taxa de prevalência estimada de 5,8 por

1000 habitantes (erro-padrão estimado = 0,754), ou seja 80% superior à que não toma em consideração a ausência de registos de pacientes com diabetes.

**Tabela 1** – Número de pacientes com diagnóstico de diabetes que estiveram presentes nos centros de saúde/médicos de família e/ou nas urgências do Hospital, numa certa cidade.

		Pacientes diagnosticados nos centros de saúde		Total
		Presente	Ausente	
Pacientes diagnosticados nas urgências	Presente	27 ( $n_{11}$ )	37 ( $n_{10}$ )	64 ( $n_{1.}$ )
	Ausente	96 ( $n_{01}$ )	? ( $n_{00}$ )	? ( $n_{0.}$ )
Total		123 ( $n_{.1}$ )	? ( $n_{.0}$ )	? ( $N$ )

A diabetes é uma doença crónica prolongada em expansão nos países desenvolvidos e atinge cada vez mais pessoas. Causa complicações graves como sejam problemas cardiovasculares, hipertensão, insuficiência renal, cegueira, amputação e no extremo a morte do paciente. Actualmente, segundo a Organização Mundial de Saúde (OMS), esta doença é a quarta causa de morte na maioria dos países desenvolvidos, e em 2009 em Portugal os custos associados são de cerca 0,6% a 0,8% do PIB (PNPCD, 2011). Uma correcta estimação da taxa de prevalência desta doença revela-se assim de vital importância.

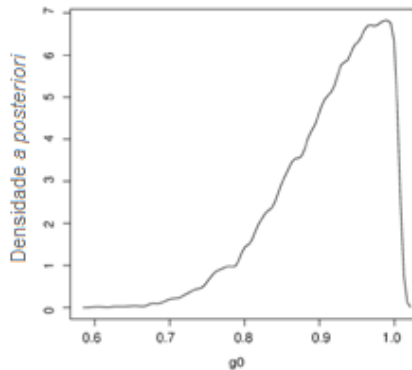
### 3.2 Estimação de ungulados de montanha (*Rupicapra p. pyrenaica*)

No Parque Nacional de Pirenéus no sudoeste de França, existe uma população de rebecos (*Rupicapra p. pyrenaica*) que é monitorizada há já vários anos sendo conhecida a sua dimensão. Esta característica permite testar a validade e aplicabilidade, a esta população real, de alguns métodos de estimação.

A amostra em análise foi recolhida em Outubro de 2002, na zona de Clot-Cayan onde se sabia que albergava aproximadamente 230 rebecos, dos quais 53 foram marcados num estudo realizado anteriormente. Realizou-se uma amostragem de um transecto linear com 2 km de comprimento. O transecto foi percorrido a pé por uma equipa com 3 observadores que registaram o sexo, a idade e a distância perpendicular dos animais detectados ao transecto linear. No total foram detectados 100 rebecos, dos quais 23 estavam marcados.

Para estimar os parâmetros de interesse utilizou-se a abordagem bayesiana proposta por Monteiro e Alpizar-Jara (2006). A distribuição *a posteriori* de  $g_0$  apresenta uma grande assimetria negativa (Figura 1), sendo por isso usada a mediana da distribuição para estimar o parâmetro pretendido ( $\hat{g}_0 = 0,992$ ). Esta

assimetria é também tida em conta no intervalo de maior densidade *a posteriori* (abreviadamente, intervalo HPD) a 95%, onde com 95% de credibilidade o parâmetro  $g_0$  está entre 0,78 e 1,00. A diferença entre a abundância estimada ( $\hat{N} = 242$ ) e a verdadeira abundância ( $N = 230$ ) é reduzida e o intervalo HPD a 95% obtido [210; 281] contém o verdadeiro valor do parâmetro. Foram ainda estimadas a probabilidade de um indivíduo ter sido marcado ( $\hat{p}_1 = 0,22$ ;  $\widehat{CV} = 14,13\%$ ) e a probabilidade de detecção no transecto linear ( $\hat{p}_2 = 0,40$ ;  $\widehat{CV} = 7,87\%$ ).



**Figura 1** – Densidade *a posteriori* de  $g_0$ , i.e. probabilidade de detecção de um indivíduo sobre a linha transecto.

Esta metodologia tem vantagens relativamente à proposta por Alpizar-Jara e Pollock (1999) no sentido de que restringe o espaço paramétrico de  $g_0$ , impedindo que este assumia valores superiores a 1, que carecem de sentido biológico.

#### 4 Limitações e soluções

Nos últimos anos tem-se assistido a um aumento significativo no número de artigos científicos publicados sobre os métodos de amostragem descritos neste trabalho. Para além de aplicações, nestes trabalhos apresentam-se também avanços teóricos que permitem lidar com a violação dos pressupostos base.

A abordagem bayesiana foi proposta não só para estimar a probabilidade de detecção dos indivíduos sobre a linha ou ponto como também modelar a heterogeneidade nas probabilidades de captura ou detecção (Monteiro *et al.*, 2008).

A dependência entre as localizações dos indivíduos pode ser tida em conta recorrendo a métodos de amostragem adaptativos (Thompson e Seber, 1996) que podem ser combinados com a amostragem por distâncias (Pollard e Buckland,



1997; Afonso e Alpizar-Jara, 2006) ou modelada através dos processos pontuais (Hedley e Buckland, 2004; Borchers e Efford, 2008).

Em terrenos montanhosos com declive irregular, os estimadores convencionais da amostragem por transectos lineares podem apresentar enviesamentos significativos. Afonso (2010) propõe um estimador para a probabilidade de detecção menos enviesado que o convencional em terrenos com declive muito irregular.

Nas ciências médicas as amostras são geralmente conjuntos de dados obtidos junto de duas ou mais fontes (por exemplo: hospitais e centros de saúde) podendo existir uma grande dependência entre elas. A modelação desta dependência pode ser realizada através dos modelos log-lineares (IWGDMF, 1995a,b).

## 5 Agradecimentos

Os autores AA, JFM e RAJ são membros do CIMA-UE, centro de investigação financiado no âmbito do FEDER pelo Programa de Financiamento Plurianual da FCT.

Os autores agradecem a Jesús Pérez, Jean-Paul Campre e Emanuel Serrano que recolheram e facultaram os dados dos rebecos para análise.

## Referências

- AFONSO, A. (2010). *Amostragem por distâncias: efeito da distribuição espacial e adaptação para terrenos montanhosos*, tese de doutoramento, Universidade de Évora.
- AFONSO, A. e ALPIZAR-JARA, R. (2006). Amostragem por distâncias adaptativa: combinação de transectos lineares com pontuais, em *Ciência Estatística*, Actas do XIII Congresso Anual da Sociedade Portuguesa de Estatística (Canto, L.C., Martins, E.G., Rocha, C., Oliveira, M.F., Leal, M.M. e Rosado, F., eds.), Edições SPE, 163-172.
- ALPIZAR-JARA, R.; AFONSO, A. e MONTEIRO, J.F. (2008). Estimação da abundância animal e de populações humanas móveis. *Boletim da Sociedade Portuguesa de Estatística*, 2, 18-25.
- ALPIZAR-JARA, R. e POLLOCK, K.H. (1996). A combination line transect and capture-recapture sampling model for multiple observers in aerial surveys. *Journal Environmental and Ecological Statistics*, 3, 311-327.
- ALPIZAR-JARA, R. e POLLOCK, K.H. (1999). Combining line transect capture-recapture for mark-resighting studies. Em *Marine Mammal Survey and Assessment Methods*, (Garner, G. W, Amstrup, S. C., Laake, J. L., Manly, B. F. J., McDonald, L. L. e Robertson, D. G. eds). A. A. Balkema, Rotterdam, 99-114.

- BORCHERS, D.L.; BUCKLAND, S.T.; GOEDHART, P.; CLARKE, E. e HEDLEY, S. (1998a). Horvitz-Thompson estimators for double-platform line transect surveys, *Biometrics*, 54, 1221-1237.
- BORCHERS, D.L. e EFFORD, M.G. (2008). Spatially Explicit Maximum Likelihood Methods for Capture-Recapture Studies, *Biometrics*, 64, 377-385.
- BORCHERS, D.L.; ZUCCHINI, W. e FEWSTER, R. (1998b). Mark-recapture models for line transect surveys, *Biometrics*, 54, 1207-1220.
- BROOKS, S.P.; CATCHPOLE, E.A. e MORGAN, B.J.T. (2000). Bayesian animal survival estimation, *Statistical Science*, 15, 357-376.
- BROOKS, S.P.; CATCHPOLE, E.A.; MORGAN, B.J.T. e HARRIS, M.P. (2002). Bayesian methods for analysing ringing data, *Journal of Applied Statistics*, 29, 187-206.
- BUCKLAND, S.T.; ANDERSON, D.R.; BURNHAM, K.P.; LAAKE, J. L.; BORCHERS, D. e THOMAS, L. (2001). *Introduction to distance sampling - estimating abundance of biological populations*. Oxford University Press, Oxford.
- BUCKLAND, S.T.; ANDERSON, D.R.; BURNHAM, K.P.; LAAKE, J.L.; BORCHERS, D. e THOMAS, L. (2004). *Advanced distance sampling - estimating abundance of biological populations*. Oxford University Press, Oxford.
- HEDLEY, S.L. e BUCKLAND, S.T. (2004). Spatial models for line transect sampling, *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 181-199.
- International Working Group for Disease Monitoring and Forecasting (1995a). Capture-recapture and multiple-record system estimation, I: History and theoretical development, *American Journal of Epidemiology*, 141, 1047-1058.
- International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple-record system estimation, II: Applications in human diseases, *American Journal of Epidemiology*, 141, 1059-1088.
- LINCOLN, F.C. (1930). *Calculating waterfowl abundance on the basis of banding returns*, U.S. G.P.O., Washington D.C., 4 pp.
- MONTEIRO, J.F.G. e ALPIZAR-JARA, R. (2006). Estimación Bayesiana de  $g_0$  em amostragem por distâncias, em *Ciência Estatística*, Actas do XIII Congresso Anual da Sociedade Portuguesa de Estatística (Canto, L.C., Martins, E.G., Rocha, C., Oliveira, M.F., Leal, M.M. e Rosado, F., eds.), Edições SPE, 501-510.
- MONTEIRO, J.F.G.; ALPIZAR-JARA, R.; SERRANO, E., CRAMPE, J.P. e PÉREZ, J.M. (2008). Estimación Bayesiana de  $g_0$  usando el muestreo por distancias y su aplicación en las estimas de densidad de ungulados de montaña, em *Tendencias Actuales en el Estudio y Conservación de los Caprinos Europeos* (Torres, J.E.G., León, J.C-M., Paris, P.F. e de Llano Aguilar, R.C. eds). II Congreso Internacional del género CAPRA en Europa, Jaén, 207-217.

- OTIS, D.L.; BURNHAM, K.P.; WHITE, G.C. e ANDERSON, D.R. (1978). Statistical inference from capture data on closed animal populations, *Wildlife Monographs*, 62, 3-135.
- PETERSEN, G.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea, *Report of Danish Biology Statistics*, 6, 1-48.
- POLLARD, J.H. e BUCKLAND, S.T. (1997). A strategy for adaptive sampling in shipboard line transect surveys, *Report of the International Whaling Commission*, 47, 921-31.
- POLLOCK, K.H. (1982). A capture–recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46, 752-757.
- Programa Nacional de Prevenção e Controlo da Diabetes (2011). *Diabetes: Factos e Números 2010*, Relatório Anual do Observatório Nacional da Diabetes. Direcção – Geral da Saúde. Disponível em <http://www.dgs.pt/upload/membro.id/ficheiros/i013881.pdf>. Consultado a 3 de Fevereiro de 2011.
- THOMPSON, S.K. e SEBER, G.A.F. (1996). *Adaptive sampling*, Wiley, New York.
- WILLIAMS, B.K., NICHOLS, J.D. e CONROY, M.J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego.

## **Caracterização, Classificação e Discriminação de Doentes Atendidos no Serviço de Urgência Devido a Resultados Clínicos Negativos da Farmacoterapia**

**Margarida Cavaco<sup>1</sup> · Luís S. Dias<sup>2\*</sup> · Fernando Fernández-Llimós<sup>3</sup>**

© The Author(s) 2013

**Resumo** Resultados Clínicos Negativos da Farmacoterapia (RCNF) como motivo de atendimento em Serviços de Urgência foram investigados através de um questionário semi-fechado. As respostas foram analisadas por análise das correspondências e classificação hierárquica. Identificaram-se três grupos de doentes com RCNF, um essencialmente de sobre-representação de necessidade de medicamentos, outro essencialmente com sobre-representação de ineffectividade de medicamentos, o terceiro com representação equitativa de necessidade e ineffectividade. Considerando o conjunto dos doentes com e sem RCNF elaborou-se uma árvore de decisão binária para identificação de doentes com RCNF usando só informação passível de ser conhecida antes da ida à Urgência, minimizando a probabilidade de não identificar doentes com RCNF. Os descritores da árvore obtida são total de medicamentos tomado há menos de sete dias, total tomado há um ano ou mais, idade, reconhecimento de melhoras, toma de medicamentos em SOS e toma de omeprazol (antiulceroso, inibidor da bomba de prótons).

**Palavras-chave:** Análise das correspondências; Classificação hierárquica; Resultados Clínicos Negativos da Farmacoterapia; Segmentação binária, Serviço de Urgência.

---

<sup>1</sup>Grupo de Acompanhamento Farmacoterapêutico de Évora, margcavaco@gmail.com

<sup>2</sup>Departamento de Biologia, Universidade de Évora, lsdias@uevora.pt

<sup>3</sup>Subgrupo de Sócio-Farmácia, Universidade de Lisboa, s-llimos@ff.ul.pt

## 1 Introdução

Resultados Clínicos Negativos da Farmacoterapia (RCNF), entendidos como resultados na saúde do doente não adequados ao objectivo da farmacoterapia e associados ao uso ou ao erro de uso de medicamentos (Comité de Consenso, 2007) são uma causa importante, ainda que muito frequentemente evitável, de atendimento em urgências hospitalares (Baena *et al.*, 2002; Parejo, 2003; Cavaco, 2009).

A redução da incidência de RCNF proporcionará uma efectiva melhoria da qualidade de vida dos doentes. Poderá também permitir uma utilização mais racional dos recursos evitando atendimentos, tratamentos e internamentos hospitalares.

Pretendeu-se com este estudo caracterizar os doentes com RCNF, os factores que lhes estão associados e contribuir para o planeamento de programas de acompanhamento farmacoterapêutico usando instrumentos simples e eficazes de identificação precoce de doentes passíveis de vir a ter RCNF.

## 2 Métodos

Questionários semi-fechados baseados no questionário de Baena *et al.* (2001) foram aplicados com consentimento informado durante 20 dias entre 19 de Setembro e 18 de Novembro de 2006 a doentes admitidos no Serviço de Urgência Geral do Hospital de Faro entre as 16 h e as 20 h, após triagem pelo sistema de Manchester e antes da avaliação médica. Foram excluídos os doentes menores de 12 anos, grávidas, acidentados, tentativas de suicídio e doentes reincidentes durante o período de estudo. Em cada dia inquiriu-se o primeiro doente física e psiquicamente apto a responder e após o fim da entrevista, o primeiro doente nas mesmas condições.

Os doentes foram inquiridos quanto ao género, idade, profissão/actividade, local de residência, peso e altura, alergias, tabagismo, fidelidade a uma farmácia, medicamentos tomados, quem os prescreveu ou indicou, há quanto tempo e de que forma os toma, conhecimento sobre para que os toma e apreciação do resultado do seu uso, motivo do atendimento no Serviço de Urgência e outros problemas de saúde existentes, nos dois casos incluindo período de instalação. Posteriormente registou-se o destino do doente após atendimento (internamento, encaminhamento para consulta externa ou alta clínica).

A profissão foi codificada de acordo com o primeiro nível do Instituto Nacional de Estatística (INE, 1994) acrescido das categorias estudante, desempregado, doméstica e reformado/aposentado. Os medicamentos foram codificados de acordo com a Anatomical Therapeutic Classification até ao nível mais baixo (ATC/DDD, 2003). Os motivos de atendimento e outros problemas de saúde foram codificados

de acordo com a Classificação Internacional de Cuidados Primários até ao nível mais baixo (ICPC-2, 2006). A caracterização dos RCNF, efectuada com apoio de médico do serviço de urgência, seguiu a adoptada no Segundo Consenso de Granada (Comité de Consenso, 2002).

Todas as respostas foram expressas como descritores primários binários (1898 descritores primários) a partir dos quais se obtiveram 23 descritores secundários (e.g. total de medicamentos tomados ou total de medicamentos prescritos por médico) tendo sido eliminados todos os descritores que apresentavam o mesmo valor em todos os doentes.

A primeira etapa da análise dos 112 doentes com RCNF (descritos por 1416 descritores) consistiu numa análise das correspondências retendo-se  $f+1$  factores, sendo  $f$  o número de factores cujos valores próprios tinham valores-teste com  $P < 0.05$  (Lebart *et al.* 2000). A caracterização dos factores retidos foi feita considerando as contribuições absolutas e relativas dos descritores. A segunda etapa consistiu na classificação hierárquica usando as coordenadas dos  $f+1$  factores retidos com pesquisa da melhor partição entre 2 e 50 classes pelo critério de Ward seguindo os procedimentos descritos em Lebart *et al.* (1984, 2000). As classes formadas na melhor partição foram caracterizadas com base nos valores-teste dos descritores em cada classe tomando como referência um nível de significância de experiência de  $P = 0.001$  calculado pelo método de Dunn-Šidák (Ury, 1976; Sokal e Rohlf, 1995).

Finalmente, usando só os descritores que podem ser conhecidos em ambulatório investigou-se quais os que permitiriam identificar a ocorrência de RCNF usando árvores de segmentação binária. A análise de segmentação binária envolveu a criação de 10 árvores óptimas obtidas a partir de corridas independentes, com afectação aleatória de 66% dos doentes ao grupo base e 33% ao grupo teste e custos a priori de identificações erradas iguais e unitários.

Análise das correspondências, classificação hierárquica e árvores de segmentação binária foram realizadas usando SPAD (2007). Descrição e algoritmos dos procedimentos podem ser encontrados em Lebart *et al.* (1984, 2000) incluindo o cálculo e utilização de valores-teste e, para as árvores de segmentação binária, também em Breiman *et al.* (1984) e em Gueguen e Nakache (1988).

### 3 Resultados e discussão

Foram inquiridos 213 doentes constituindo 25% e 5% respectivamente de todos os admitidos na urgência no intervalo horário e nos dias em que a amostragem foi realizada. Nenhum dos doentes se recusou a responder.

Dos 213 inquiridos, 112 (cerca de 53%) recorreram à urgência devido a RCNF. Nestes predominava a ineffectividade (46%) e a necessidade (37%) entendendo-se

por inefectividade o doente estar a tomar medicamentos que não atingem de modo suficiente os objectivos terapêuticos esperados e por necessidade o doente com um problema de saúde concreto não realizar a terapêutica apropriada por falta de prescrição ou por não adesão (Comité de Consenso, 2007).

Considerando os inquiridos com RCNF o doente tipo é uma mulher, com cerca de 61 anos, residente no distrito de Faro, profissionalmente activa, com peso (avaliado pelo Índice de Massa Corporal) normal ou excessivo, fiel a uma farmácia, tomando em média 3.1 medicamentos, quase sempre há vários anos, quase sempre prescritos por médico e quase sempre sabendo para que os toma, nunca tendo ido à Urgência ou tendo ido há um ano ou mais.

Na análise das correspondências dos doentes com RCNF só o primeiro factor tinha um valor-teste com  $P < 0.05$ . O exame das coordenadas dos doentes nos dois primeiros factores sugeria a existência de um doente potencialmente atípico. A classificação hierárquica nas coordenadas dos dois primeiros factores reforçou essa sugestão já que a melhor partição resultou em três classes, uma constituída por 45 doentes, outra por 66 doentes e a terceira pelo doente referenciado anteriormente. Consequentemente, realizou-se nova análise das correspondências com esse doente como suplementar. Tal como na análise anterior, só o primeiro factor tinha um valor-teste com  $P < 0.05$  e surgia, ainda que de forma menos evidente que na primeira análise, um novo doente potencialmente anómalo. Tendo em conta que não surgia individualizado na melhor partição, foi mantido como doente activo. Os dois primeiros factores representam 3.2% da inércia total, um valor baixo mas compreensível em situações em que quase todas as variáveis são binárias (Lebart *et al.* 2000) ainda mais quando mais de 98% da matriz de dados é constituída por valores zero.

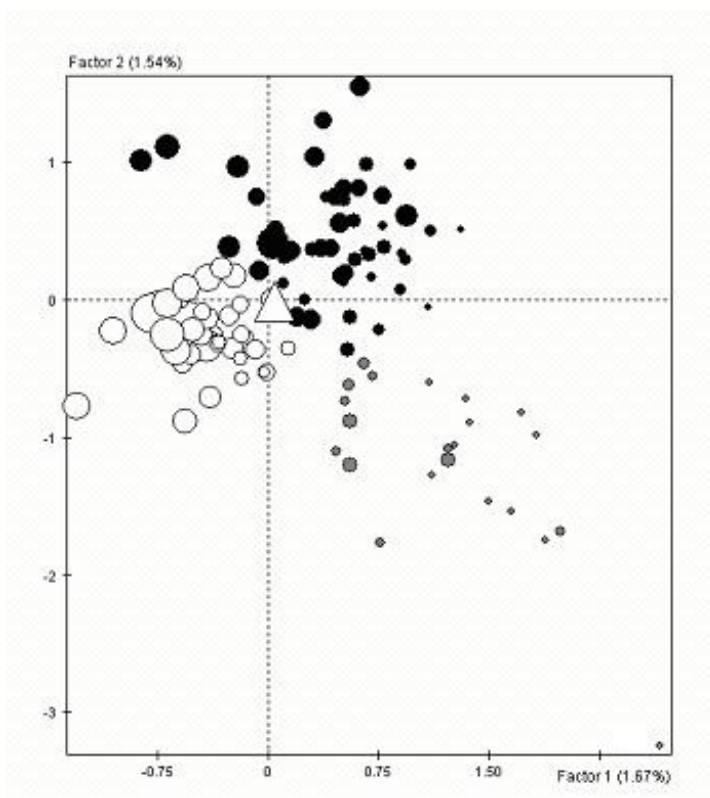
O primeiro factor (1.7% da inércia total) é caracterizado pelos descritores idade entre 25 e 44 anos, total de medicamentos prescritos por médico e total de motivos de atendimento na urgência, com o segundo descritor oposto aos restantes. O segundo factor (1.5% da inércia total) é caracterizado pelos descritores RCNF por necessidade, total de medicamentos relativamente aos quais não está melhor e total de medicamentos que toma há sete dias ou menos, com o primeiro descritor oposto aos restantes.

A melhor partição originou três classes (Figura 1). A classe 1 (38 doentes) tem uma representação equitativa de RCNF por inefectividade e por necessidade, subrepresentação de doentes com 25 a 44 anos e sobrerepresentação de doentes com 65 a 74 anos, de reformados/aposentados, do total de medicamentos que tomam, que sabem para que tomam e relativamente aos quais se sentem melhor, do total de tomas diárias, do total de medicamentos prescritos por médico, do total de medicamentos que tomam há um ano ou mais e de diabetes não insulino-dependente como problema de saúde não responsável pela ida à urgência.

A classe 2 (22 doentes) tem uma sobrerepresentação de doentes com RCNF por necessidade e uma subrepresentação de doentes com RCNF por inefectividade bem

como do total de medicamentos que tomam, que sabem para que tomam e relativamente aos quais se sentem melhor, do total de tomas diárias e do de medicamentos prescritos por médico.

A classe 3 (51 doentes) tem uma subrepresentação de doentes com RCNF por necessidade, de reformados, do total de medicamentos que tomam há um ano ou mais e do total de problemas de saúde que não foram a causa da ida à urgência e uma sobrerepresentação de doentes com RCNF por inefectividade, do consumo de ibuprofeno (anti-inflamatório não esteróide e antirreumatismal) e de medicamentos indicados por outro que não médico ou farmacêutico.



**Figura 1** – Representação dos doentes com resultados clínicos negativos da farmacoterapia nos dois primeiros factores da análise das correspondências. Dimensão dos pontos proporcional ao peso. Círculos brancos para os doentes da classe 1, cinzentos para os da classe 2, pretos para os da classe 3. Doente suplementar representado por um triângulo.



No que diz respeito ao tipo de RCNF a classe 2 e 3 são claramente opostas. A classe 2 é constituída por doentes com RCNF por necessidade, tomando um reduzido número de medicamentos, poucas tomas diárias, fraco conhecimento dos objectivos terapêuticos desses medicamentos e um reduzido reconhecimento de melhoras associadas aos medicamentos. Em contrapartida, a classe 3 é constituída por doentes com RCNF por inefectividade com elevado consumo de ibuprofeno, geralmente não reformados, tomando poucos medicamentos há mais de um ano e com poucos problemas de saúde além daquele ou daqueles que os levaram a recorrer à urgência.

A classe 1 (à qual é afectado o doente analisado como suplementar) é uma classe de doentes mais velhos, reformados, naturalmente tomando mais medicamentos, mais vezes por dia e há mais tempo mas com uma representação equitativa de RCNF por inefectividade e necessidade. Talvez surpreendentemente, mas só à primeira vista, é a classe em que os doentes melhor sabem para que tomam os medicamentos.

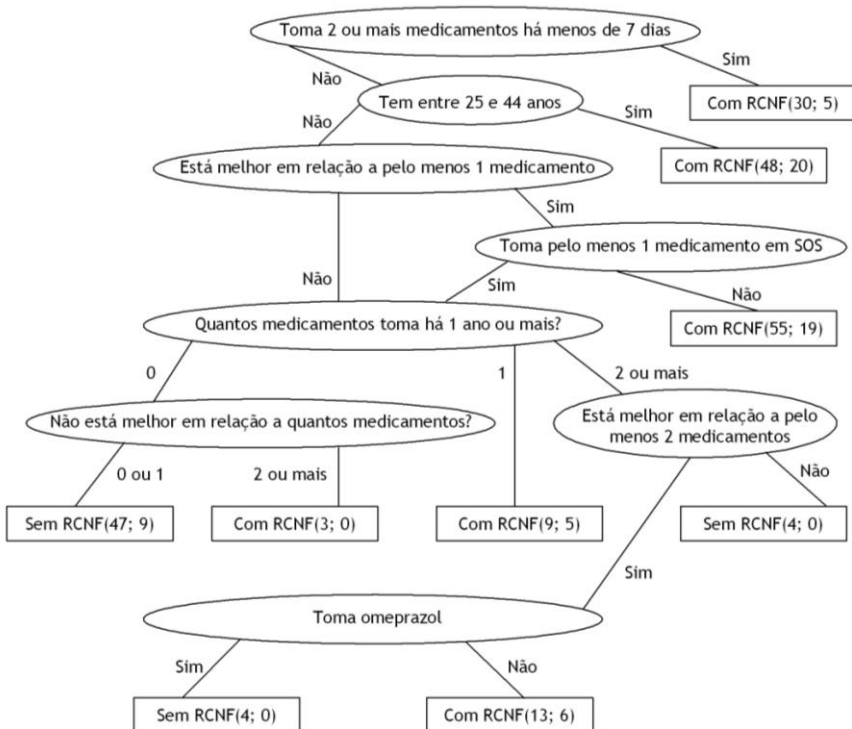
De acordo com os critérios de Baena *et al.* (2002) 82% dos casos em que a ocorrência de RCNF foi identificada eram potencialmente evitáveis, um valor extremamente elevado, ainda mais tendo em conta que na literatura são referidos valores de evitabilidade entre 19% e 70% (Parejo, 2003). Consequentemente, era potencialmente evitável a ida à urgência de 43% dos doentes inquiridos e o internamento posterior de 5% deles.

Na perspectiva de contribuir para uma detecção precoce de risco de RCNF e usando só os descritores que para serem conhecidos não exigem a admissão no Serviço de Urgência seleccionou-se a árvore de decisão binária que quando aplicada a todos os doentes minimizava a quantidade de doentes com RCNF erradamente identificados como não tendo RCNF. Apesar de ser a melhor árvore de acordo com o critério de não errar a identificação de doentes com RCNF a percentagem de identificações erradas ainda era apreciável (21% dos doentes com RCNF identificados como não tendo RCNF). O aumento do custo a priori deste erro de identificação não permitiu obter árvores de segmentação binária óptimas pelo que se aplicaram todas as restantes árvores previamente obtidas, só aos doentes identificados como não tendo RCNF pela melhor árvore optando-se pela árvore que neste subgrupo minimizava a quantidade de doentes com RCNF erradamente identificados como não tendo RCNF. Finalmente, combinaram-se numa só as duas árvores de segmentação binária conseguindo assim obter uma árvore de decisão binária que identifica correctamente 70% dos doentes, identifica erradamente 26% dos doentes como tendo RCNF e 4% dos doentes como não tendo RCNF, reduzindo de 21% para 8% os doentes que tendo RCNF são erradamente identificados como não tendo.

Os descritores envolvidos são: total de medicamentos tomado há menos de sete dias ou há um ano ou mais, a idade, o reconhecimento de melhoras desde que toma

os medicamentos, a toma de medicamentos em SOS e a toma de omeprazol, um antiulceroso inibidor da bomba de protões (Figura 2).

Tendo em conta que se privilegiou a redução de identificações erradas de doentes que de facto apresentavam RCNF, da árvore obtida resulta uma relativamente elevada percentagem de identificações erradas de RCNF em doentes que não os apresentavam. Em todo o caso, a probabilidade de declarar erradamente que um doente pode vir a ter RCNF (26%) representa um custo que consideramos aceitável dada a reduzida probabilidade (4%) de falhar a identificação de potenciais RCNF.



**Figura 2** – Árvore de segmentação binária dos doentes com e sem Resultados Clínicos Negativos da Farmacoterapia (RCNF). Entre parêntesis o número de pacientes e o número de pacientes identificados erradamente.

Apesar disso e apesar dos resultados obtidos poderem não ser facilmente extrapoláveis para contextos não hospitalares devido à especificidade própria da necessidade de deslocação ao Serviço de Urgência, o exame da árvore de segmentação binária obtida permite lançar alguma luz sobre os factores associados ao surgimento de RCNF. Quase metade dos doentes identificados como tendo RCNF (e quase metade dos doentes com RCNF) são-no por estarem a tomar dois ou mais medicamentos nos 7 dias anteriores à ida ao Serviço de Urgência ou por terem entre 25 e 44 anos. Adicionalmente, em doentes que não tomam ou só estão a tomar um medicamento nos 7 dias anteriores à ida ao Serviço de Urgência e com idade compreendidas fora daquele intervalo (quase sempre mais velhos), não tomar medicamentos em SOS (mesmo estando melhor em relação a pelo menos um medicamento) praticamente completa a identificação de doentes como tendo RCNF.

Em contrapartida, a identificação de doentes como não tendo RCNF requer a verificação de um conjunto maior de condições, tendo em comum tomar menos de 2 medicamentos há menos de 7 dias, não ter entre 25 e 44 anos e não estar melhor em relação a pelo menos um medicamento e acaba por estar, algo paradoxalmente quase sempre associado a não estar melhor em relação a zero ou um medicamento. No entanto, o exame das respostas revela que 44 desses 47 doentes identificados como não tendo RCNF de facto responderam que não havia medicamentos para os não sentiam melhoras, incluindo os 9 doentes erradamente identificados como não tendo RCNF (8 dos quais por necessidade).

## Agradecimentos

Os autores agradecem aos revisores científicos as sugestões e comentários feitos às versões anteriores deste trabalho.

## Referências

- ATC/DDD (2003). *The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses* (ACT/DDD), World Health Organization.
- BAENA, M.I., CALLEJA, M.A., ROMERO, J.M., VARGAS, J., ZARZUELO, A., JIMÉNEZ-MARTÍN, J. e FAUS, M.J. (2001). Validación de un Cuestionario para la Identificación de Problemas Relacionados con los Medicamentos en Usuarios de un Servicio de Urgencias Hospitalario. *Ars Pharmaceutica*, 42, 3-4, 147-171.
- BAENA, M.I., MARÍN, R., MARTÍNEZ OLMOS, J., FAJARDO, P., VARGAS, J. e FAUS, M.J. (2002). Nuevos Criterios para Determinar la Evitabilidad de los

- Problemas Relacionados con los Medicamentos. Una Revisión Actualizada a Partir de la Experiencia con 2.558 Personas. *Pharmaceutical Care España*, 4, 393-396.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. e STONE, C.J. (1984) *Classification and Regression Trees*, Boca Raton, USA, Chapman & Hall/CRC.
- CAVACO, M.I.M.G. (2009). *Resultados Clínicos Negativos da Farmacoterapia como Motivo de Atendimento no Serviço de Urgência*, Dissertação de Mestrado, Universidade de Évora, Évora, Portugal.
- COMITÉ DE CONSENSO (2002). Segundo Consenso de Granada sobre Problemas Relacionados con Medicamentos. *Ars Pharmaceutica*, 43, 179-187.
- COMITÉ DE CONSENSO (2007). Tercer Consenso de Granada sobre Problemas Relacionados con Medicamentos (PRM) y Resultados Negativos Asociados a la Medicación (RNM). *Ars Pharmaceutica*, 48, 5-17.
- GUEGUEN, A. e NAKACHE, J.P. (1988) Méthode de Discrimination Basée sur la Construction d'un Arbre de Décision Binaire. *Revue de Statistique Appliquée*, XXXVI, 19-38.
- ICPC-2 (2006). *Classificação Internacional de Cuidados Primários*, 2.<sup>a</sup> edição, Comissão Internacional de Classificações de Wonca.
- INE (1994). *Classificação Nacional de Profissões - 1994*, Instituto Nacional de Estatística, <http://www.ine.pt/prodserv/nomenclaturas/cnp1994.asp> (acedido a 25/6/2007).
- LEBART, L., MORINEAU, A. e PIRON, M. (2000). *Statistique Exploratoire Multidimensionnelle*, 3e éd., Paris, France, Dunod.
- LEBART, L., MORINEAU, A. e WARWICK, K.M. (1984). *Multivariate Descriptive Statistical Analysis*, New York, USA, Wiley.
- PAREJO, M.I.B. (2003). *Problemas Relacionados con los Medicamentos como causa de Consulta en el Servicio de Urgencias del Hospital Universitario Virgen de las Nieves de Granada*, Tesis Doctoral, Universidad de Granada, Granada, España.
- SOKAL, R.R. e ROHLF, F.J. (1995) *Biometry. The Principles and Practice of Statistics in Biological Research*, 3rd ed., New York, USA, Freeman.
- SPAD (2007) SPAD. *Data Mining & Text Mining*, v. 6.5.0, Paris, France.
- URY, H.K. (1976). A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary simple sizes. *Technometrics*, 18, 89-97.



# Contributos para Previsão do Consumo de Energia Eléctrica na Ilha de São Miguel

Armando B. Mendes<sup>1</sup>

© The Author(s) 2013

**Resumo** Na sequência de trabalhos anteriores, identificam-se algumas das principais causas das variações no consumo horário de energia eléctrica com base na identificação de padrões e relações entre os dados de consumo e variáveis climáticas e identificadoras de sazonalidades. Para tal, testam-se vários algoritmos de *data mining*, nomeadamente os disponíveis nos Analysis Services do *software* MS SQL Server.

Assim, foi possível verificar que as variáveis climáticas têm influência muito significativa na produção de energia eléctrica quando combinadas com variáveis dicotómicas que assinalem padrões horários. Utilizando árvores de modelos de regressão, foi possível prever os consumos de 2007 com um erro absoluto médio de 1,4 MW. O melhor modelo foi validado usando não apenas as estatísticas de desempenho obtidas para uma amostra de teste, mas também o conhecimento de domínio, que permitiu confirmar os padrões no comportamento do consumo.

**Palavras-chave:** consumo de energia, factores climáticos, *data mining*, árvores de modelos de regressão, previsão.

## 1 Previsão do consumo de energia eléctrica

Vários estudos se têm debruçado sobre o difícil problema de prever consumos de energia eléctrica. Como se sabe, não existe nenhuma forma eficiente de armazenar energia eléctrica, pelo que há uma necessidade constante de ajustar a produção ao consumo. Se for possível estimar o consumo com alguma antecedência a produção poderá ser ajustada, tornando todo o processo mais simples e eficiente.

Na verdade, a produção ultrapassa este problema produzindo em excesso, ou mais exactamente consumindo combustíveis fósseis (ou energia eólica) em excesso

---

<sup>1</sup> Departamento de Matemática e CEEApIA, Universidade dos Açores, amendes@uac.pt

sem produzir a energia eléctrica correspondente. Qualquer ajuste da produção que permita reduzir estes excessos mantendo o nível de serviço ao consumidor tem óbvias vantagens económicas e operacionais.

A grande maioria dos trabalhos publicados centra-se na obtenção de previsões o mais precisas possível e com o máximo de antecedência (ver por exemplo: Troutt *et. al.*, 1991; Engle *et al.* 1992; Liu e Harris, 1993). Todas estas publicações reconhecem a importância de perceber a relação entre os factores climáticos e sazonais e o consumo de energia. A previsão dos factores climáticos é a principal fonte de incerteza na previsão do consumo de energia (Troutt *et. al.*, 1991).

Este problema é mais complexo em ambientes insulares especialmente sujeitos a alterações climáticas. Assim, para quem tem de dar resposta a essa solicitação de consumo, é necessário caracterizar com cuidado o meio envolvente. As empresas que produzem a energia eléctrica têm de tomar decisões em tempo real de quanto produzir por meios flexíveis, principalmente centrais termoeléctricas, de modo a satisfazer a procura em cada segundo.

Neste contexto, qualquer conhecimento adicional sobre a forma como se comporta o consumo de energia eléctrica numa dada região é muito valorizado.

Na ilha de São Miguel a energia eléctrica é obtida principalmente através de centrais termoeléctricas, centrais geotérmicas, mini-hídricas e de sistemas particulares de produção de energia baseados em biogás estando igualmente a começar a ligar-se à rede os primeiros clientes com microprodução. Nos últimos anos tem-se verificado um forte aumento da produção de energia proveniente de fontes hidrotermais, energia geotérmica, tendo atingido para os dados mais recentes (informação de Novembro de 2010, [www.eda.pt](http://www.eda.pt)) uma produção acumulada de 37,9% do total da ilha. Estes dois tipos de energia passam a representar, neste período, 7,6% da produção total. Segundo a mesma fonte a produção de energia eléctrica tem aumentado continuamente, verificando-se um crescimento da produção de 2,4% entre Janeiro a Novembro de 2010, comparativamente a igual período do ano transacto. Neste período a produção de energia geotérmica teve um crescimento de 5,2% em comparação com igual período do ano anterior. Verificou-se ainda um crescimento de 43,2% de produção hídrica e um crescimento de 8,6% de produção eólica.

O grupo EDA – Electricidade dos Açores, mantém grandes volumes de dados relativos a consumos ou produções horárias de energia eléctrica. Note-se que, neste contexto, o consumo coincide com a produção uma vez que não se considera a possibilidade de armazenamento. Estes dados podem ser trabalhados de forma a caracterizar o consumo de energia e as necessidades energéticas, a obter informação e conhecimento e, assim, tomar decisões devidamente fundamentadas.

Este trabalho pretende ser um contributo para identificar as causas que levam a variações no consumo de energia com base na identificação de padrões e relações entre os dados de consumo e várias variáveis climatéricas e representantes da sazonalidade. Como meio para alcançar esse objectivo utilizam-se técnicas de *data*

*mining*, para a descoberta de conhecimento, construção de modelos e previsões segundo a descrição de um caso de estudo.

Para a concretização desse objectivo, utiliza-se neste projecto a metodologia CRISP-DM *CRoss Industry Standard Process for Data Mining*, já utilizada e descrita pelo autor em trabalhos anteriores (Mendes *et al.*, 2008; Mendes, 2010). A utilização desta metodologia em problemas abordados por técnicas de *data mining* tem-se revelado muito útil por, no essencial, permitir estruturar e disciplinar o processo, evitando a aplicação indiscriminada de algoritmos como reacção natural à disponibilização dos mesmos em *software* de simples utilização.

Note-se, no entanto, que as seis fases do modelo processual nem sempre surgem de forma sequencial, verificando-se frequentemente a necessidade de voltar um passo atrás. Estes retornos entre as fases descritas no modelo processual estão previstos na metodologia e constituem a espiral de modelação e extracção de conhecimento (Lavrač *et al.*, 2004). Para uma descrição completa desta metodologia ver o documento original de Chapman *et al.* (2000) e o sítio Web criado pelo projecto: [www.crisp-dm.org](http://www.crisp-dm.org).

## 2 Obter dados e pré-processamento

As primeiras dificuldades prenderam-se com a obtenção dos dados e o seu indispensável tratamento para que seja possível a sua utilização com os algoritmos pretendidos.

A base de dados sobre a qual se trabalhou, refere-se aos dados meteorológicos recolhidos no Aeroporto de Ponta Delgada, no período compreendido entre os anos de 1998 e 2007. Fez-se o *download* do sítio [wunderground.com](http://wunderground.com), recorrendo a um programa em Java, construído propositadamente para este fim. Salienta-se que, antes de construir esta ferramenta de trabalho, o *download* dos dados era uma tarefa impensável e quase impraticável, tendo em conta o tempo para a elaboração do projecto, uma vez que era necessário descarregar a informação de 3.300 dias aproximadamente, sendo possível obter apenas um dia de cada vez.

Quanto aos dados dos registos instantâneos das potências de cada um dos sistemas electroprodutores da Ilha de São Miguel, para os anos 2005 e 2006, existiam já em meio digital de fácil acesso, tornando fácil a importação dos mesmos. No entanto, em relação ao ano de 2007 foi necessário descarregá-lo via intranet do sistema de informação da empresa produtora. Devido a limitações no sistema de informação e a restrições de segurança, o *download* destes dados só foi possível em períodos de dez dias. Além disso, estes estavam separados por áreas de produção, dificultando ainda mais a sua recolha.

Além destas dificuldades de acesso aos dados, verificaram-se igualmente alguns problemas durante a exploração dos dados que tiveram por consequência a necessidade de implementar delicadas tarefas de limpeza. Recorreu-se ao *software*



SQL Server 2005, usando ferramentas OLAP (*Online Analytical Processing*) para construir um cubo como ferramenta para resumo e exploração dos dados. A imagem do cubo é usada para descrever tabelas de contingência (ou resumo uma vez que são usadas outras medidas descritivas além de contagens) multidimensionais, sobre as quais o *software* permite várias operações como rotação (*roll*) e agregar (*drill-up*) ou desagregar (*drill-down*) em hierarquias de dimensões.

O processo de construção do cubo e de implementação de fluxos de processamento (*process flows*) foi semelhante ao descrito em trabalho anterior (Mendes *et al.*, 2008). Importa referir que este *software* permite criar tabelas de dados *on-the-fly* por indicação de um atributo chave, efectuando-se a agregação dos dados (por soma para variáveis métricas e contagens para as não métricas) imediatamente antes da sua utilização no ajuste de modelos.

Dos problemas mais frequentes, também descritos noutros trabalhos, de que o excelente livro de Chen (2001) é um bom exemplo, verificaram-se potências com valores inexistentes ou com valores abaixo do normal e que foram posteriormente considerados como valores omissos.

As condições climáticas são uma descrição textual simplificada das condições que caracterizam o comportamento dos elementos climáticos num dado momento. Verificou-se a existência de algum ruído nas variáveis métricas, conduzindo a valores atípicos facilmente identificados no caso de *outliers* univariados por representações gráficas como os gráficos de extremos e quartis. Estes valores devem-se certamente a erros na transmissão ou registo de dados e foram removidos. O critério usado na eliminação foi o de 3 distâncias interquartílicas, tendo-se verificado, no entanto, que este tipo de erros nos dados era fácil de identificar por apresentarem valores claramente fora do intervalo razoável admitido para a variável. Os *outliers* multivariados foram identificados durante a construção de modelos como descrito na secção seguinte.

Finalmente, após ter todos os dados em formato relacional e tendo efectuado as correcções necessárias, fundiram-se numa só tabela todos os atributos referentes ao clima e à potência, sendo assim possível executar um novo cubo com a integração das variáveis climatológicas.

Uma dificuldade relevante nesta fase prende-se com o facto de as chaves nem sempre terem correspondência nos dois conjuntos de dados. Este tipo de problema é conhecido como um problema de fusão de dados (*data fusion*) no sentido em que surge da compatibilização de dados provenientes de fontes diversificadas, ver por exemplo Chen (2001) e Saporta (2002). Desta forma, foi necessário confrontar as duas fontes de dados e extrair somente as instâncias cujas horas coincidiam. Tal foi efectuado igualmente implementando fluxos processuais adequados.

No final da fase de exploração e pré-processamento, dispunha-se de uma base de dados bem estruturada e com dados no nível de qualidade adequado ao estudo pretendido.

### 3 Exploração de dados e construção de modelos

Dado o objectivo de prever consumos de energia eléctrica, dos algoritmos disponíveis no MS SQL Server, os mais adequados para análise previsional de uma variável quantitativa são os Microsoft Decision Trees, Microsoft Linear Regression e Microsoft Neural Network. Uma breve descrição dos algoritmos disponíveis no *software* pode ser encontrada em Larson (2006). No entanto, antes de construir modelos de previsão, faz-se, segundo a metodologia indicada, exploração de dados na tentativa de identificar relações entre o clima e o consumo de energia, as quais são essenciais para validação dos modelos e conhecimento obtido da análise.

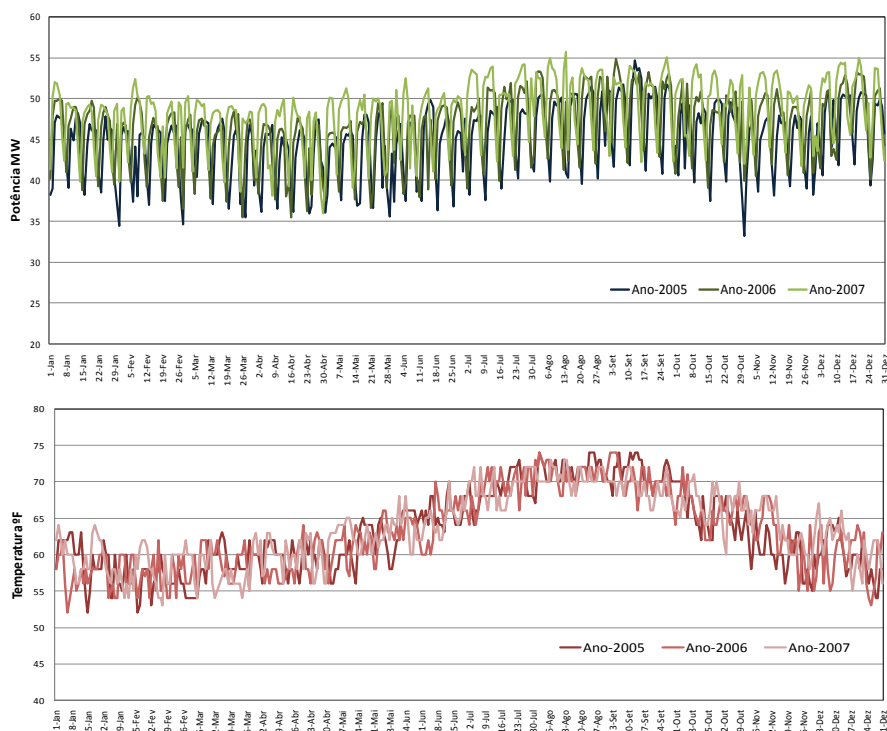
Como se considera a potência consumida como a variável dependente em todos os modelos, a compreensão do comportamento desta variável ao longo do tempo é particularmente relevante e pode observar-se nas três séries de valores médios diários para os três anos da Figura 1. Neste cronograma é possível observar a existência de sazonalidades tanto semanais, apresentando valores menores durante o fim-de-semana, como durante o ano observando-se um aumento do consumo durante os meses de verão e no final do ano. Na mesma figura é igualmente possível observar um aumento anual das potências utilizadas.

No que se refere à temperatura ambiente, o aumento durante os meses de verão é igualmente visível, observando-se uma notável coincidência entre as duas curvas.

Para caracterizar as variáveis utilizadas, começou-se por usar o algoritmo Naïve Bayes. A divisão em classes ou discretização das variáveis (também conhecido por *binning* não supervisionado) foi realizada pelo mesmo algoritmo e resultou em cinco classes. Posteriormente o algoritmo determina probabilidades condicionadas, as quais são usadas para medir o grau de associação entre diferentes classes de variáveis a prever e variáveis explicativas. O facto de considerar as variáveis como independentes, mesmo sabendo-se que é muito frequente a existência de dependências, dá o nome Naïve ao método.

Por aplicação deste algoritmo, foi possível verificar, de forma clara, que para temperaturas acima dos 21 °C as potências situam-se entre os 50 MW e 56 MW, e para valores de temperatura abaixo dos 13 °C, as potências são menores do que 35,6 MW.

A rede de dependências (*dependency network*) é uma forma de visualização onde as variáveis são representadas por nós, com os arcos a representar a existência de relações de causalidade (arcos orientados) entre os nós (Larson, 2006).



**Figura 1** – Três séries de potências médias diárias e as séries correspondentes de temperaturas médias diárias.

A intensidade da relação é dada pela capacidade da variável explicativa prever a variável dependente. Concluiu-se que as melhores variáveis para prever o consumo de energia eléctrica são, por ordem: a humidade, ponto de orvalho e temperatura. As piores variáveis são a velocidade do vento e as condições climáticas.

O ponto de orvalho ou de condensação ( $T_{PO}$ ), é a temperatura a que o ar tem de ser arrefecido, mantendo a pressão constante, para que fique saturado de vapor de água. A esta temperatura começa a condensação. Prova-se que este valor é uma função não linear da temperatura ambiente ( $T$ ) e da humidade relativa ( $H$ ):

$$T_{PO} = H \times (112 + 0,9T) - 112 + 0,1T. \quad (1)$$

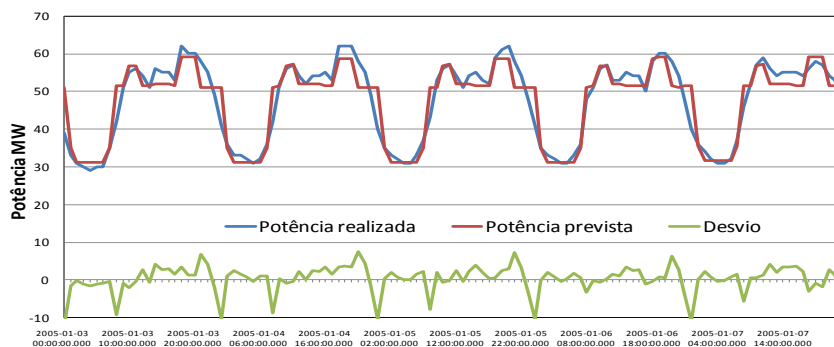
Por utilização do algoritmo de *Clustering*, utilizando o atributo “Dia” como chave, verifica-se que este algoritmo não supervisionado criou um conjunto de dez *clusters*. Alguns desses *clusters* confirmam os resultados anteriores obtidos pelo algoritmo Naïve Bayes. Por exemplo, num desses grupos a frequência de horas

com a potência superior a 47 MW é de 82%, e a frequência da temperatura entre 19,7 e os 27°C é de 98,3%.

Para a construção de modelos começou-se pela abordagem mais simples estimando um modelo de regressão linear múltipla com a Potência como variável dependente e como possíveis variáveis explicativas todas as restantes. O modelo obtido revelou-se de baixa qualidade, não permitindo, nomeadamente, uma boa adaptação a picos de consumo de energia ocorridos durante o dia e que se repetem de segunda a sexta-feira.

Para melhorar a qualidade do ajustamento do modelo aos dados, foram introduzidas variáveis binárias em períodos horários correspondentes a picos de consumo. Adicionou-se ainda uma variável binária que indicasse a existência de fins-de-semana e feriados, uma vez que o comportamento nestes dias é claramente diferente dos restantes dias da semana. Por fim, criou-se uma nova variável Horas retirada do campo Data, que identifica a hora de cada registo. Deste modo, consegue-se modelar com algum rigor os efeitos sazonais do consumo energético.

Na Figura 2 pode-se observar que em parte este objectivo foi conseguido. Contudo, ainda existiam diferenças entre a potência prevista e a realizada com erros significativos e um padrão nos resíduos que não tinha sido modelado. Nesse sentido foram adicionadas novas variáveis binárias que vieram a diminuir os erros nesses pontos.



**Figura 2** – Valores reais e previstos por um modelo de Regressão linear, apenas para Janeiro de 2005.

Todo o processo anterior foi repetido, para um espaço temporal mais alargado, utilizando-se registos dos anos 2005 e 2006 num total de 15.666. Esta tabela serviu como dados de treino. Usando todas as variáveis disponíveis construíram-se vários modelos com os algoritmos MS Decision Trees, MS Linear Regression e MS Neural Network e verificou-se que aqueles que apresentam melhor ajuste aos dados são os que utilizam o algoritmo MS Decision Trees.

Na tentativa de melhorar os modelos obtidos introduziram-se igualmente outras variáveis que traduzissem o mês e os dias da semana (variáveis ordinais). Como se

verificou na Figura 1, existe forte sazonalidade nos dados ao longo do ano o que justifica ainda a introdução de variáveis binárias identificativas das estações do ano. Para modelar o aumento médio no consumo de ano para ano, criou-se uma nova variável que identificasse o ano.

A opção de introduzir todas estas variáveis binárias, revelou-se acertada uma vez que estas novas variáveis foram escolhidas pelo algoritmo tanto na parte lógica como na parte funcional para a construção dos melhores modelos.

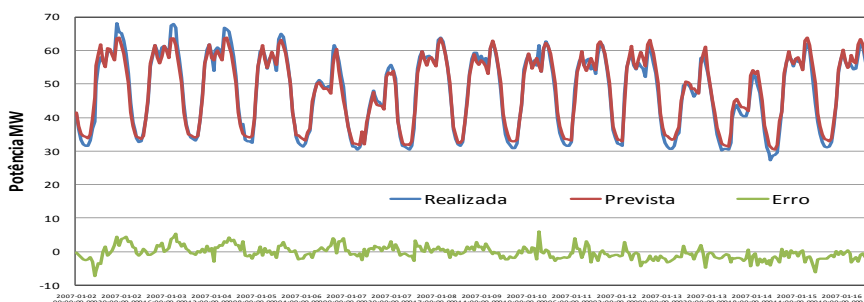
Nos novos modelos estimados, os resultados foram muito melhores, obtendo-se os resultados com melhor score novamente para o modelo criado pelo algoritmo MS Decision Trees. Note-se que este algoritmo, quando são usadas variáveis dependentes numéricas, induz árvores de modelos. Ou seja, é construída uma árvore de decisão de pequena dimensão, de tal forma que em cada nó folha restem um elevado número de observações as quais são utilizadas para ajustar um modelo de regressão linear. Obtêm-se, assim, modelos de regressão para cada segmento de observações ou registos que obedecem às condições lógicas correspondentes ao caminho na árvore que conduz a esse nó folha. As regras classificatórias são induzidas dos dados por partição recursiva, construindo uma árvore da raiz para as folhas, por divisão das instâncias segundo condições simples sobre as variáveis explicativas. As condições são escolhidas de modo a minimizar a variância dos grupos obtidos.

Durante a modelação, verificou-se com alguma frequência o aparecimento de valores de desvio anómalos, os quais, após algum esforço de pesquisa foram identificados como sendo devidos a valores atípicos das potências em situações de ocorrência de “disparo geral” ou outras semelhantes. Para corrigir estes valores foi necessário recuar à fase de tratamento dos dados. Foi ainda possível concluir que variáveis como a Velocidade do Vento, Direcção do Vento e Condições Climáticas, não são adequadas para explicar o comportamento do consumo de energia eléctrica na ilha de São Miguel.

A validação do melhor modelo obtido com o algoritmo MS Decision Trees, foi efectuada com a base de dados de teste correspondente a valores de consumo horário para o ano de 2007. Na Tabela 1 apresentam-se algumas estatísticas de qualidade das previsões obtidas pelo modelo. Da leitura da tabela e por observação do gráfico na Figura 3 é possível observar que a qualidade do modelo é muito boa.

**Tabela 1** – Estatísticas de qualidade de ajuste para o melhor modelo obtido.

Coefficiente de determinação múltipla R2	0,94
Erro quadrado médio (EQM)	3,52
Desvio Padrão (DP)	1,87
Erro Médio (EM)	0,21
Média do erro percentual absoluto (MEPA)	2,92 %
Erro relativo absoluto (ERA)	1,4 %

**Figura 3** – Potência prevista e realizada para o melhor modelo obtido, ajustado com dados de 2005 e 2006. Apresenta-se apenas os dados para o mês de Janeiro de 2007.

## 4 Resultados e conclusões

A estrutura da árvore de modelos que apresentou os melhores resultados, esquematizada na Figura 4 representa-se sob a forma de uma tabela de decisão com variáveis e intervalos de valores nas colunas, lendo-se a sequência de condições lógicas nas linhas. É possível observar que as variáveis usadas nas ramificações são principalmente relacionadas com períodos temporais como é o caso da hora do dia, dia da semana e mês, traduzindo períodos sazonais.

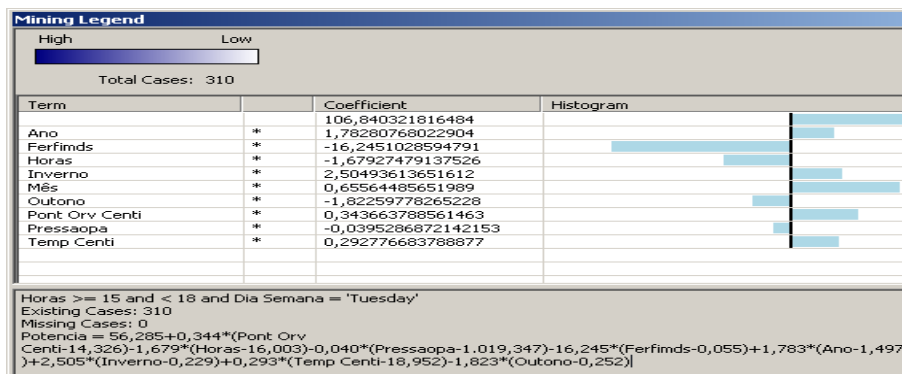
A árvore obtida é constituída por 59 ramos com os consequentes 59 modelos de regressão linear em cada nó folha. Na Figura 5 apresenta-se a expressão do modelo de regressão do nó terminal da ramificação cuja hora está entre as 15 e as 17 às terças-feiras.

O modelo obtido pode ser facilmente interpretado. Para todas as terças-feiras entre as 15 e as 18 horas para o aumento de 1 °C acima da média da temperatura ambiente a potência consumida aumenta 0,293 MW (considerando todas as outras variáveis independentes constantes). Para o mesmo aumento no valor do ponto de

orvalho a potência aumenta 0,344 MW. Existe ainda um aumento anual médio de 1,783 MW, um aumento mensal de 0,656 MW, um aumento no Inverno de  $2,505 \cdot (1 - 0,229)$  cujo valor é 1,93 MW, e uma diminuição no Outono de  $1,823 \cdot (1 - 0,252)$  o que perfaz 1,36 MW.

horas do dia	dia da semana	mês	dia da semana	horas do dia	mês	temperatura	nº ramo	
0-2	segunda			0			1	
	terça a domingo	1-4		1-2			2	
		5-8					3	
		9-12					4	
3-5	segunda						5	
	terça						6	
	quarta						7	
	quinta						8	
	sexta						9	
	sábado						10	
	domingo						11	
6-7	domingo						12	
	segunda						13	
	terça a sábado						14	
8			1-6				15	
			7-12				16	
9	domingo						17	
	segunda						18	
	terça a sexta						19	
	sábado						20	
9-11	segunda						21	
	terça						22	
	quarta						23	
	quinta						24	
	sexta						25	
	sábado						26	
	domingo						27	
12-14	domingo e sábado						28	
	segunda a sexta			12-13	1-3	< 18,857	29	
					4-5		30	
					6-9		31	
				10-12	32			
			14				33	
					≥ 18,857		34	
							35	
15-17	segunda							36
	terça							37
	quarta							38
	quinta							39
	sexta							40
	sábado							41
	domingo							42
18-20	1-4	domingo e sábado					43	
		segunda a sexta	18			44		
			19			45		
			20			46		
	5-8	domingo					47	
		segunda					48	
		terça a sexta	5-6			49		
			7-8			50		
			sábado			51		
	9-10						52	
11-12						53		
21-23	1-2					54		
	3-4					55		
	5-8	21-22				56		
		22-23				57		
	9-10					58		
	11-12					59		

Figura 4 – Estrutura lógica da árvore de modelos de regressão completa.



**Figura 5** – Um exemplo de um nó folha da árvore de modelos.

De modo semelhante nas restantes ramificações da árvore de modelos, verifica-se que a temperatura, o ponto de orvalho e a humidade contribuem para a variação do consumo de energia, isto é, quando aumentam os valores destas variáveis há um aumento da potência.

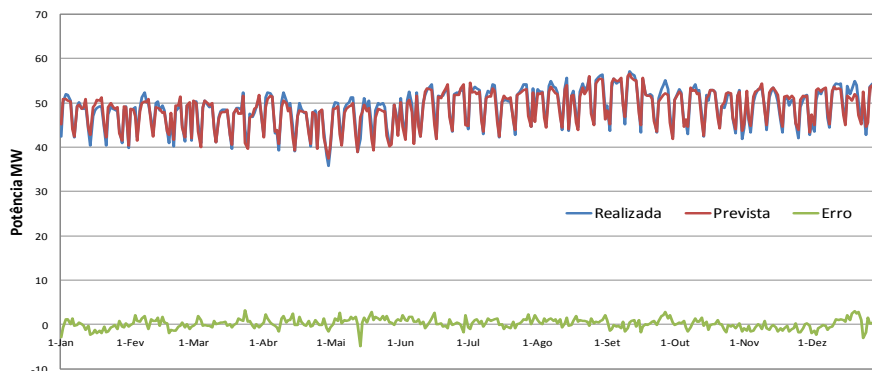
Como se sabe de conhecimento de domínio ou simples senso comum, a existência de elevadas temperaturas obriga-nos a recorrer a sistemas de refrigeração. Por outro lado, quando a humidade relativa do ar é elevada, sabemos que nos sistemas de transporte de energia eléctrica existem perdas devido à corrente de fuga através dos isoladores, verificando-se ainda um aumento da utilização dos desumidificadores por parte dos consumidores.

Ainda assim, a influência de factores climáticos como o ponto de orvalho ou a pressão atmosférica no consumo horário de energia eléctrica não é totalmente evidente do conhecimento de domínio. Considera-se igualmente a possibilidade destas variáveis estarem a funcionar como *proxies* ou substitutas de factores com influência mais directamente reconhecível no consumo de energia como a intensidade luminosa do sol ao longo do dia, ou o aumento da população verificado em períodos mais quentes.

Note-se que se considera o modelo de árvores de decisão adequado, não apenas pelos resultados obtidos pelas estatísticas de qualidade de ajuste, mas porque, como se pode observar na Figura 1, as relações com algumas variáveis climáticas não se mantêm ao longo de todo o ano. Por exemplo, a temperatura pode explicar o comportamento no verão, mas não durante o Natal. Tentar ajustar a mesma regressão a todo o ano resultaria em ajustes de pouca qualidade. Assim, o ajuste por segmentos ou períodos sazonais faz, certamente, mais sentido. Aliás este tipo de comportamento das séries de consumo eléctrico é reconhecido em publicações anteriores sendo recomendada a modelação de períodos específicos como as horas de pico (ver, por exemplo: Engle *et al.*, 1992 e Liu e Harris, 1993).



A fim de obtermos uma visão global para o ano de 2007, foram calculadas as médias diárias da potência realizada e da potência prevista como se pode observar na Figura 2. Os valores previstos foram obtidos com os modelos estimados para os anos de 2005 e 2006.



**Figura 9** – Valores médios diários de potência prevista e realizada para o ano de 2007 com a série dos resíduos.

O muito bom ajuste observado confirma a qualidade das previsões. O desvio absoluto médio não ultrapassa 1,4 MW o que significa que esse seria o erro médio da previsão se fosse possível obter valores de previsão das variáveis climáticas sem erro.

Ao considerar os resultados apresentados pelos modelos é inevitável concluir que o consumo de energia é maior nos períodos de maior temperatura. No Verão, a média das potências diárias é mais elevada, embora no Inverno também se verifique um consumo elevado por altura da época natalícia.

Este estudo veio mostrar que é possível explicar a influência do clima na produção horária de energia eléctrica.

## Agradecimentos

O autor agradece ao Dr. José M.S. Ferreira por colaboração em grande parte dos resultados apresentados e a todos os dirigentes da Electricidade dos Açores pela colaboração e autorizações prestadas, com uma menção muito especial ao Eng.º Aires Ferreira. O autor agradece ainda os comentários dos revisores do artigo.

---

## Referências

- CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. e WIRTH, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*. SPSS Inc., USA.
- CHEN, Z. (2001). *Intelligent Data Warehousing: From data preparation to data mining*. CRC Press: Boca Raton, USA.
- ENGLE, R.F.; MUSTAFA, C. e RICE, J. (1992). Modelling e peak electricity demand. *Journal of Forecasting*, 11, 241-251.
- LARSON, B. (2006). *Delivering Business Intelligence with Microsoft SQL Server 2005*. McGraw-Hill: Emeryville, USA.
- LAVRAČ, N.; MOTODA, H.; FAWCETT, T.; HOLTE, R.; LANGLEY, P. e ADRIAANS, P. (2004). Introduction: Lessons learned from data mining applications and collaborative problem solving. *Machine Learning*, 57, 13-34.
- LIU, LON-MU e HARRIS, J.L. (1993). Dynamic structural analysis and forecasting of residential electricity consumption. *International Journal of Forecasting*, 9, 437-455.
- MENDES, A.B., FERREIRA, A. e ALFARO, P.J. (2008). Suporte à decisão em tecnologias de comunicação: Utilização de OLAP e data mining, *In Actas CISTI2008*, Cota, M.P., Editor. @LibroTex, 973-984.
- MENDES, A.B. (2010). BI and data warehouse solutions for energy production industry: Application of the CRISP-DM methodology, *In Bridging the Socio-technical Gap in Decision Support Systems: Challenges for the next decade, Frontiers in Artificial Intelligence and Applications*, Respício, A.; Adam, F.; Phillips-Wren, G.; Teixeira, C. e Telhada, J. (Eds.), IOS Press, 211-222. ISBN: 978-1-60750-576-1.
- SAPORTA, G. (2002). Data fusion and data grafting. *Computational Statistics & Data Analysis*, 38, 465-473.
- SMITH, D.G.C. (1989). Combination of forecasts in electricity demand prediction. *Journal of Forecasting*, 8, 349-356.
- TROUTT, M.D.; MUMFORD, L.G. e SCHULTZ, D.E. (1991). Using spreadsheet simulation to generate a distribution of forecasts for electric power demand. *Journal of the Operational Research Society*, 42, 931-939.

## CLASSIFICAÇÃO E ANÁLISE DE DADOS: MÉTODOS E APLICAÇÕES

vem responder à vontade expressa por muitos dos participantes nas sucessivas reuniões anuais da CLAD, de que se desse visibilidade aos trabalhos apresentados nas JOCLAD – Jornadas de Classificação e Análise de Dados, assim os valorizando e divulgando. O conjunto dos artigos aqui incluídos deve ser entendido como uma amostra não aleatória, dos trabalhos apresentados entre as JOCLAD2004 e as JOCLAD2010 e entretanto submetidos a esta publicação.

