Estabilidade de uma Estrutura de Agrupamento – Segmentos de Clientes de uma Instituição Cultural

Margarida G. M. S. Cardoso
UNIDE, ISCTE-IUL
margarida.cardoso@iscte.pt

Resumo

Neste trabalho implementa-se, como meio de avaliação de estabilidade de um agrupamento, uma nova proposta de validação cruzada de agrupamentos que prescinde do uso de classificadores, recorrendo à utilização de amostras ponderadas de treino e teste (Cardoso, Faceli et al. 2009). Ilustra-se a metodologia proposta sobre um agrupamento de clientes do CCB - Centro Cultural de Belém. Este agrupamento é efetuado mediante estimação de um modelo de mistura finita. Na constituição dos grupos ou segmentos atende-se à natureza ordinal das variáveis base (medições em escala de tipo Likert), em alternativa à modelação habitual que consideraria as mesmas variáveis como métricas. Em complemento, são apontadas metodologias consideradas mais apropriadas para a interpretação e discriminação dos grupos obtidos.

Palavras-chave: avaliação de agrupamentos, índices de concordância, estabilidade

Title: Stability of a clustering structure: Client segments of a cultural institution

Abstract

This work implements, as a means of assessing the stability of a cluster, a new proposal for cross-validation of clusters that dispenses with the use of classifiers, resorting to the use of weighted samples of training and testing (Cardoso, Facel et al. 2009) We illustrate the proposed approach over a cluster of clients of CCB – Cultura Centre of Belem (Centro Cultural de Belém). The clustering is obtained by means of an estimation of a mixture finite model. In the constitution of the clusters or segments, it it taken in consideration the ordinal nature of the clustering base variables (measurements in Likert scale) in lieu of the usual modeling that would consider the same variables as metric. In addition, we point out to some methodologies that are considered more adequate to interpret and discriminate the segments obtained.

Keywords: clustering validation, indices of agreement, stability

1. Introdução

A análise de agrupamento (*clustering analysis*) tem como objetivo genérico constituir grupos cujos elementos sejam, de algum modo, homogéneos e distintos dos de outros grupos, em atributos que são considerados no processo de agrupamento.

A análise de agrupamento é, tipicamente, uma análise de interdependência (na terminologia habitual da Estatística Multivariada) ou não supervisionada (terminologia de *Data Mining*). Assim, um resultado obtido (agrupamento) não é suscetível de confrontação com qualquer estrutura alvo, conhecida ou observada, pelo que a avaliação de um agrupamento não pode passar por medidas de erro de estimativas associadas a quaisquer valores observados, como ocorre numa análise supervisionada. Há então que procurar outros meios para aferir a qualidade de um agrupamento.

A estabilidade é uma qualidade desejável de uma estrutura de agrupamento: ela traduz a capacidade de uma mesma estrutura de grupos ser suscetível de ser apreendida a partir de diferentes amostras que se associam a uma mesma população, usando um mesmo processo de agrupamento.

A questão que se coloca ao comparar diversos agrupamentos sobre diversas amostras para avaliar a estabilidade é a de medir o acordo entre partições constituídas em amostras diferentes, obrigando a exportar cada agrupamento de uma amostra de treino para uma amostra de teste, através de um classificador.

Neste trabalho implementa-se, como meio de avaliação de estabilidade de um agrupamento, uma nova proposta de validação cruzada de agrupamentos que prescinde do uso de classificadores, recorrendo à utilização de amostras ponderadas de treino e teste (Cardoso, Faceli et al. 2009). Ilustra-se a metodologia proposta sobre um agrupamento de clientes do CCB - Centro Cultural de Belém. Este agrupamento é efetuado mediante estimação de um modelo de mistura finita. Na constituição dos grupos ou segmentos atende-se à natureza ordinal das variáveis base, em alternativa à modelação habitual que consideraria as mesmas variáveis como métricas. Em complemento, são apontadas metodologias consideradas mais apropriadas para a interpretação e discriminação dos grupos obtidos.

2. Estabilidade

2.1. Estabilidade e Validação cruzada

A estabilidade é reconhecida como uma propriedade desejável de uma solução de agrupamento, e.g. (Mirkin 1996). Uma solução será estável se se mantiver razoavelmente inalterada quando o processo de agrupamento for sujeito a pequenas modificações, tais como parametrizações alternativas do algoritmo utilizado, introdução de ruído nos dados ou consideração de diferentes amostras.

A estabilidade pode ser aferida avaliando a concordância entre as diferentes partições resultantes das referidas modificações. Diversos índices de concordância (IC) podem ser usados com esse

fim. Se a modificação se referir às amostras base de agrupamento coloca-se, no entanto, a questão prévia de como medir concordância de partições constituídas em diferentes amostras, já que isso implica *transportar*, de alguma forma, uma das partições para a amostra base da outra. Esse *transporte* pode ser efetuado mediante um procedimento de validação cruzada desenhado para o efeito (McIntyre e Blashfield 1980), (Breckenridge 1989). Na validação cruzada de um agrupamento, o cerne da avaliação passa a ser não um conjunto de observações alvo conhecidas que se comparam com estimativas obtidas sobre uma amostra de teste, mas sim um agrupamento *transportado* da amostra de treino para a de teste (classes candidatas), que é confrontado com outro obtido diretamente sobre a amostra de teste. No final do procedimento de validação cruzada, valores de índices de concordância entre os grupos obtidos mediante agrupamento (sobre a amostra de teste) e classes resultantes de Análise Classificatória/Discriminante (sobre a mesma amostra), são usados como indicadores de estabilidade.

O procedimento de validação cruzada levanta algumas questões:

1) Selecionar um classificador adequado para exportar o agrupamento do treino para o teste já que, de acordo com (Lange, Roth et al. 2004): *by selecting an inappropriate classifier one can artificially increase the discrepancy between solutions* (p. 1304). *…the identification of optimal classifiers by analytical means seems unattainable. Therefore we have to resort to potentially suboptimal classifiers in practical applications* (p.1305).

2) Dispor de uma amostra original com dimensão suficiente para viabilizar a divisão em subamostras de treino e teste, divisão que pode ser replicada - e.g. (Tibshirani, Walther et al. 2001), (Levine e Domany 2001), (Dudoit e Fridlyand 2002), (Law e Jain 2003), (Lange, Roth et al. 2004).

A utilização de um procedimento de estimação de um modelo de mistura finita para agrupamento tem a vantagem de evitar a questão **1),** já que permite não só constituir os grupos, mas também, obter um classificador que resulta da própria estimação do modelo e que pode ser utilizado sobre uma amostra de teste (Cardoso 2007). De facto, ao obter estimativas de máxima verosimilhança dos parâmetros de um modelo de mistura finita

$$f\left(\underline{y}|\,\underline{\lambda},\underline{\theta}\right) = \sum_{k=1}^{K} \lambda_k f\left(\underline{y}|\;\underline{\theta}_k\right)$$

(em que λ_k é o parâmetro representando o *peso* do grupo G_k, f representa a f.(d.)p. conjunta de $\underline{y}$ e $\underline{\theta}_k$ representa o vector de parâmetros distribucionais de f associado ao grupo G_k , pode determinar-se facilmente uma probabilidade *a posteriori* de pertença de cada observação $\underline{y}_i$ (i = 1 ... n) a cada grupo $G_{k'}$ dada por:

$$p\left(\underline{y} \in G_k|\underline{\lambda},\underline{\theta}\right) = \frac{\lambda_{k'} f\left(\underline{y}|\;\underline{\theta}_{k'}\right)}{\sum_{k=1}^{K} \lambda_k f\left(\underline{y}|\;\underline{\theta}_k\right)}$$

No entanto, a utilização desta metodologia de agrupamento não permite, por si só, contornar a questão (2). Assim, (Cardoso, Faceli et al. 2009) fazem uma proposta de metodologia de validação cruzada que radica no uso de uma amostra ponderada. Neste procedimento, a dimensão da amostra deixa de ser uma limitação relevante para a implementação da validação cruzada, já que os IC são baseados na amostra global ponderada e não numa amostra de teste propriamente dita. Para além disso, o uso de amostras ponderadas imita a constituição de subamostras aleatórias de treino e teste e deixa de haver a necessidade de construir um classificador, pelo que a questão (1) também não se coloca. Neste trabalho propõe-se fazer uso da metodologia de validação cruzada ponderada para obter um agrupamento de clientes de uma instituição cultural – um primeiro teste da metodologia em dados reais primários.

2.2. Índices de concordância

Na literatura encontram-se definidos múltiplos índices de concordância (IC) – e.g. (Faceli, Carvalho et al. 2005). Estes índices são frequentemente usados no contexto de avaliação externa de agrupamentos, em que é conhecido um agrupamento *a priori* e se pretende aferir a capacidade que diversas metodologias de agrupamento têm de recuperar essa estrutura. Neste trabalho os IC são utilizados numa análise conduzida no sentido de obter um agrupamento que se desconhece. Os valores dos IC deverão medir a concordância entre partições associadas a diferentes amostras e permitir aferir a estabilidade de uma solução candidata.

Em geral, os cálculos dos IC podem basear-se na tabela de contingência que associa duas partições consideradas, ou tabela de classificação cruzada (v. Tabela 1).

Tabela 1 – Tabela de classificação cruzada das partições P^K e P^Q

		$\mathbf{P^Q}$				
		C'_1	C'_2	...	C'_Q	*Total*
$\mathbf{P^K}$	C_1	n_{11}	n_{12}	...	n_{1Q}	$n_{1.}$
	C_2	n_{21}	n_{22}	...	n_{2Q}	$n_{2.}$
	...	...	...	...	...	...
	C_K	$n_{K,1}$	$n_{K,2}$	...	n_{KQ}	$n_{K.}$
	Total	$n_{.1}$	$n_{.2}$	...	$n_{.Q}$	n

Considerando uma tipologia dos IC que distingue IC simples e IC emparelhada (v. (Cardoso 2007), adota-se como IC simples a Estatística V de Crámer, v. (Agresti 2002) por exemplo, e o índice de Rand ajustado (Hubert e Arabie 1985) como um IC emparelhada.

A seleção destes índices decorre do facto de ambos considerarem limiares de concordância por acaso – sob a hipótese de independência das partições – nos próprios índices, o que não é comum nos inúmeros índices referidos na literatura.

A estatística V de Crámer – V – é uma medida do Qui-Quadrado (QQ) normalizada que varia convenientemente no intervalo [0,1]: 0 indica uma concordância por acaso e 1 concordância perfeita.

$$V(P^K, P^Q) = \sqrt{\frac{\frac{QQ(P^K, P^Q)}{n}}{min(K-1, Q-1)}}$$

em que

$$QQ(P^K, P^Q) = \sum_{k=1}^{K}\sum_{q=1}^{Q} \frac{\left(n_{kq} - \frac{n_{k.}n_{.q}}{n}\right)^2}{\frac{n_{k.}n_{.q}}{n}}$$

Note-se que no QQ são consideradas as frequências esperadas associadas a cada n_{kq} sob hipótese de independência de P^K e P^Q i.e. $\frac{n_{k.}n_{.q}}{n}$.

Os IC emparelhados consideram os números de pares de observações que duas partições concordam, ou não, em juntar ou separar, num dos correspondentes grupos. Concretamente, estes índices consideram:
a_{11} - O número de pares que P^K e P^Q concordam em juntar ou agrupar conjuntamente;
a_{10} - O número de pares que P^K junta mas P^Q separa;
a_{01} - O número de pares que P^K separa mas P^Q junta;
a_{00} - O número de pares que P^K e P^Q concordam em separar.

Tabela 2 – Tabela de concordância emparelhada entre as partições P^K e P^Q

		P^Q agrupa conjuntamente o par	
		1	**0**
P^K agrupa conjuntamente o par	**1**	a_{11}	a_{10}
	0	a_{01}	a_{00}

Os IC emparelhados podem ser vistos como medidas de dissemelhança entre duas variáveis binárias indicadoras de acordo no agrupar no mesmo grupo, ou separar em grupos diferentes, um par de observações (situações que tipicamente são codificadas com 1 e 0, respetivamente) - v. Tabela 2.

O índice de Rand (Rand 1971) é, talvez, o mais popular dos IC. Ele quantifica a proporção de pares de observações que as duas partições concordam em juntar ou em separar em cada grupo:

$$Rand(P^K, P^Q) = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{00}}$$

(Hubert e Arabie 1985) estudam o índice de Rand sob a hipótese (H_0) de concordância por acaso e adotam, para exprimir esta hipótese, o modelo hipergeométrico generalizado, obtendo o índice de Rand ajustado:

$$Rand - a(P^K, P^Q) = \frac{Rand(P^K, P^Q) - E_{H0}[Rand(P^K, P^Q)]}{1 - E_{H0}[Rand(P^K, P^Q)]}$$

A partir da Tabela 1 este índice pode ser calculado por:

$$Rand - a(P^K, P^Q) = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}/\binom{n}{2}}{\frac{1}{2}\left(\sum_{k=1}^{K}\binom{n_{k.}}{2} + \sum_{q=1}^{Q}\binom{n_{.q}}{2}\right) - \sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}/\binom{n}{2}}$$

Comparando 5 IC emparelhada (Milligan e Cooper, 1986) concluem que o índice de Rand ajustado de (Hubert e Arabie 1985) pode ser usado com vantagem sobre os restantes, no âmbito da avaliação de agrupamentos. Esta conclusão contribui também para a seleção do referido índice.

3. Interpretar um agrupamento

Na sequência da constituição e avaliação de um agrupamento, torna-se imprescindível interpretar os grupos constituídos, já que a interpretabilidade de uma solução de agrupamento é inseparável da sua avaliação e, em última análise, condição da sua utilidade (Mirkin 1998).

Em geral, uma análise complementar de associação entre os grupos e cada variável de interesse X, disponível para os caracterizar facilita a interpretação do agrupamento considerado. Neste contexto, é aconselhável o uso de medidas de associação entre cada variável X qualitativa (nominal ou ordinal) e os grupos, ou cada variável X quantitativa (intervalar ou de razão) e os mesmos grupos. Para este efeito podem ser usadas medidas como o próprio coeficiente V de Crámer - $V(X, P^K)$ - anteriormente definido, ou o coeficiente Eta-quadrado - $\eta^2(X, P^K)$ - para X qualitativa ou métrica, respectivamente. Note-se que η^2 mede a variação entre-grupos, sobre a variação total correspondentes à variável métrica X de interesse, i.e.:

$$\eta^2(X, P^K) = \frac{\sum_{k=1}^{K} n_k(\bar{x}^k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

em que $n_K = \#G_K$ e $\sum_{k=1}^{K} n_k = n$.

Estas associações, sendo relevantes - pode medir-se o seu valor de prova e comparar com determinado nível de significância estatística pré-estabelecido - podem justificar um passo

adicional na análise para apoio à caraterização dos grupos. Assim, pode, eventualmente, realizar-se uma análise classificatória discriminante entre os grupos, exibindo a *outra face da moeda* da análise de agrupamento. Esta análise atende ao objetivo da própria análise de agrupamento – minimizar a variação intra-grupos e maximizar a variação entre-grupos - mas de modo supervisionado, construindo um classificador sobre os grupos. Este classificador (regras funcionais, proposicionais ou outras) permite alocar novos indivíduos aos grupos e, simultaneamente, proporcionar um apoio à interpretação dos mesmos grupos.

Na verdade, algumas metodologias de agrupamento podem facilitar esta análise que permite não só caracterizar os grupos de forma discriminante, mas imputar novas observações aos grupos. É o caso da estimação de modelos de mistura finita que proporciona um classificador nos grupos, função das variáveis base de agrupamento (v. também (Cardoso 2007)). Para a análise classificatória a realizar neste contexto, a escolha de classificadores específicos pode recair, com vantagem, sobre algoritmos de construção de árvores de classificação: CART- *Classification e Regression Trees*, (Breiman, Friedman et al. 1984), por exemplo, e de indução de regras proposicionais (Quinlan 1993), por exemplo. Estes tipos de classificadores proporcionam resultados muito fáceis de interpretar não só pelo analista, mas por especialistas do (qualquer) domínio de aplicação, v. (Cardoso e Moutinho 2003), por exemplo.

Finalmente, convém sublinhar a importância de se recorrer a conhecimento especializado (no domínio específico da aplicação) para avaliar um agrupamento.

4. Avaliação da estabilidade de um agrupamento de clientes de uma instituição cultural

4.1. Os dados base

Considere-se uma base de dados de respostas a um questionário dirigido a clientes do CCB - Centro Cultural de Belém, entre Fevereiro e Maio de 2007 (Duarte 2009). O inquérito inclui diversas questões relacionadas com o perfil de cliente, padrões de comportamento, perceção, avaliação de experiência e intenções relativas à instituição, num total de 31 questões.

A partir da referida amostra de clientes do CCB, procede-se ao seu agrupamento considerando como variáveis base $\underline{y} = (Y_1, \ldots, Y_5)$ as respostas às questões listadas na Tabela 3. Na análise são consideradas cerca de 700 respostas (completas) a estas questões. Na Tabela 3 incluem-se os itens que mais se correlacionam com as componentes principais extraídas das medidas multi-itens dos seguintes conceitos: Imagem percebida, Valor percebido, Satisfação do consumidor e, ainda, Lealdade. Na análise efetuada usa-se a técnica CATPCA- *Categorical Principal Components Analysis* atendendo à natureza ordinal das escalas e extraem-se as componentes com valores próprios associados superiores a 1, e.g. (Meulman, Van der Kooij et al. 2004).

Tabela 3 – Itens base de segmentação

Questão	**Escala: $l = 1, 2 \ldots L$**
Qual o seu grau de concordância com a afirmação: o CCB é uma instituição de confiança?	1= *discordo completamente, … 5= concordo completamente*
Considerando as suas expectativas, avalie o desempenho do CCB relativamente aos produtos e serviços que oferece?	1= *muito pior que o esperado,…5=muito melhor que o esperado*
Indique o seu grau de satisfação em relação à Localização do CCB	1=*nada satisfatório,… 5=totalmente satisfatório*
Indique o seu grau de satisfação em relação a Horários do CCB	1=*nada satisfatório,…5= totalmente satisfatório*
Recomendaria o CCB a um amigo?	1= *nunca,…5= com toda a certeza*

4.2. Agrupamento de clientes do CCB

Para o agrupamento ou segmentação dos clientes do CCB respondentes ao inquérito, atende-se à natureza ordinal das variáveis base de segmentação (v. Escala em Tabela 3) e às múltiplas vantagens do agrupamento via estimação de um modelo de mistura finita adotando-se o modelo seguinte:

$$f\left(\underline{y}|\underline{\Theta}\right) = \sum_{k=1}^{K} \lambda_k \prod_{j=1}^{J} f_j(\underline{y}_j|\underline{\pi}_j^k)$$

Neste modelo os Y_j referem-se às variáveis base de agrupamento (no caso, $J = 5$) e supõe-se a sua independência condicional ou intra-grupos. Atendendo à natureza qualitativa das variáveis base, considera-se

$$f\left(\underline{y}_j\middle|\underline{\pi}_j^k\right) = \prod_{l=1}^{L} (\pi_{jl}^k)^{y_{jl}}$$

sendo $\pi_{jl}^k = P\left(Y_j = l\middle|\underline{y} \in G_k\right)$ e $\underline{y}_j = (y_{j1}, \ldots, y_{jL})$ com y_{jl} indicador binário (1 ou 0) da categoria l da resposta (no caso, $L = 5$). Tem-se, ainda,

$$log\left(\frac{\pi_{jl}^k}{\pi_{jL}^k}\right) = \eta_{jl}^k$$

ou,

$$\pi_{jl}^k = \frac{e^{\eta_{jl}^k}}{\sum_{l=1}^{L} e^{\eta_{jl}^k}}$$

E, de modo a incorporar a natureza ordinal das respostas, utiliza-se a proposta de modelo ordinal de categorias adjacentes, considerando $\eta_{jl}^k = \beta_{jl}^0 + \beta_l^k$ (Vermunt e Magidson 2005). Nesta proposta tem-se em conta o logaritmo do *odd* de categorias adjacentes, v. (Agresti 2002):

$$log\left(\frac{\pi_{jl}^k}{\pi_{j,l+1}^k}\right) = \alpha_{jl} + \alpha^k$$

pelo que

$$log\left(\frac{\pi_{jl}^{k}}{\pi_{jL}^{k}}\right) = log\left(\frac{\pi_{jl}^{k}}{\pi_{j,l+1}^{k}}\right) + log\left(\frac{\pi_{j,l+1}^{k}}{\pi_{j,l+2}^{k}}\right) + \cdots log\left(\frac{\pi_{j,L-1}^{k}}{\pi_{jL}^{k}}\right) = \sum_{m=l}^{L-1} \alpha_{jm} + (L-l)\alpha^{k}.$$

As estimativas dos parâmetros obtêm-se usando o algoritmo *Latent Gold* (Vermunt e Magidson 2005), na tentativa de maximizar a função de probabilidade *a posteriori* (estimativas MAP-*Maximum a posteriori*). Considera-se, assim, o objetivo de maximizar $logL + logp(\underline{\Theta})$ em que L indica a função de verosimilhança associada ao referido modelo de mistura e $p(\underline{\Theta})$ a função de probabilidade *a priori* dos parâmetros do modelo.

Na determinação do número de grupos (K) têm-se em conta os critérios BIC - Bayesian Information Criterion (Schwarz 1978) e AIC- Akaike Information Criterion (Akaike 1974). Os resultados associados a estimações do modelo proposto (k=1…10) apresentam-se na Tabela 4 e figuram na Ilustração 1 e na Ilustração 2 onde se observa um mínimo local de BIC para 3 grupos e de AIC para 5 grupos.

Tabela 4 – Critérios para a determinação do número de grupos

Nº de grupos	LL	BIC(LL)	AIC(LL)
1	-3831.5	7787.7	7701.1
2	-3648.5	7461.1	7347.1
3	-3620.5	**7444.3**	7302.9
4	-3610.2	7463.0	7294.3
5	-3593.7	7469.5	**7273.4**
6	-3587.9	7497.3	7273.8
7	-3580.3	7521.3	7270.5
8	-3576.5	7553.1	7275.0
9	-3565.0	7569.5	7264.0
10	-3562.0	7602.8	7270.0

Ilustração 1 – Critério BIC

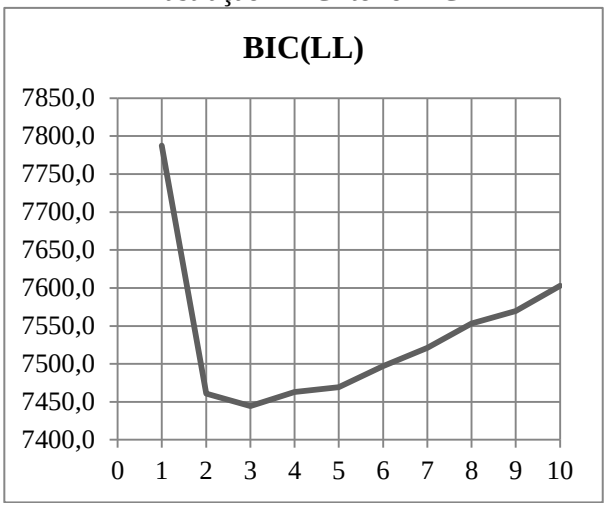
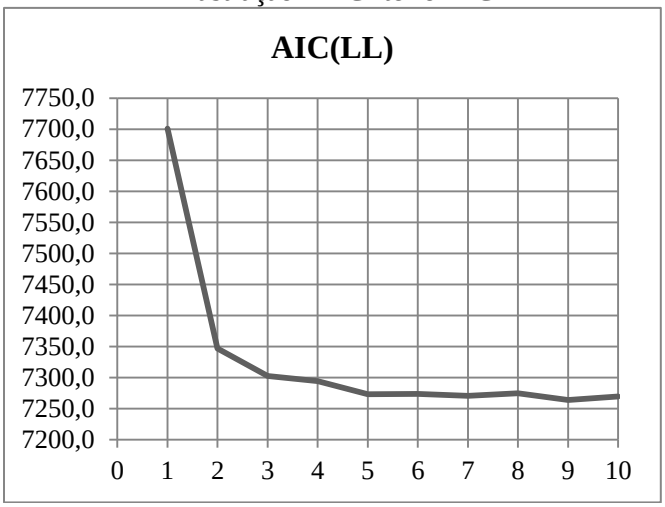


Ilustração 2 – Critério AIC



4.3. Estabilidade de agrupamentos de clientes do CCB

A distribuição das soluções com 3 e 5 grupos apresenta-se na Tabela 5. À partida, a expressão dos grupos G4 e G5 da solução com 5 grupos é insignificante pelo que, do ponto de vista prático, não parece justificar-se a constituição de mais que 3 grupos. Segue-se uma análise da estabilidade das duas soluções alternativas.

Tabela 5 – Distribuição de soluções com 3 e 5 grupos (Ag3 e Ag5, respetivamente)

Ag3	Frequência	Percentagem	Ag5	Frequência	Percentagem
G1	390	55.2	G1	360	51.0
G2	173	24.5	G2	209	29.6
G3	143	20.3	G3	118	16.7
			G4	11	1.6
			G5	8	1.1
Total	706	100	Total	706	100

Para a análise da estabilidade de cada solução são constituídas 2 amostras ponderadas (v. Tabela 6) em que se atribui valor 1 a observações ditas de treino e 10^{-10} a observações ditas de teste. Seguidamente, efetuam-se agrupamentos em cada uma das 2 amostras.

Tabela 6 – Distribuição de amostra de treino ponderada

Pesos	Frequência	Percentagem
10^{-10}	353	50
1	353	50
Total	706	100

Os resultados de validação cruzada para soluções com 3 e 5 grupos apresentam-se nas

Tabela 7 e Tabela 8. Nestas tabelas de contingência, as colunas referem-se aos grupos constituídos via amostra de treino ponderada e as linhas aos obtidos com amostra de teste ponderada (em que os pesos indicados na Tabela 6 são trocados).

A estabilidade observada para as 2 soluções consideradas é estimada pelos indicadores presentes na Tabela 9.

Tabela 7 – Tabela de classificação cruzada (soluções com 3 grupos)

		grupos via a. treino			Total
		G1	G2	G3	
grupos via a. teste	G'1	234	176	1	411
	G'2	108	0	48	156
	G'3	8	131	0	139
Total		350	307	49	706

Tabela 8 – Tabela de classificação cruzada (soluções com 5 grupos)

		grupos via a. treino					
		G1	G2	G3	G4	G5	Total
grupos via a. teste	G'1	179	87	39	0	1	306
	G'2	129	0	0	75	17	221
	G'3	9	23	102	0	2	136
	G'4	2	0	0	27	7	36
	G'5	3	0	0	0	4	7
Total		322	110	141	102	31	706

De acordo com os resultados obtidos, os índices de associação simples e emparelhada empatam as duas soluções no que se refere à propriedade da estabilidade: o V de Cramer seleciona a solução com 3 e o Rand ajustado com 5 grupos. Decide-se optar pela constituição de 3 grupos, privilegiando a parcimónia e um número razoável de indivíduos por segmento.

Tabela 9 – Indicadores de estabilidade

	3 grupos	5 grupos
V de Crámer	0,522	0,503
Rand ajustado	0,190	0,256

4.4. Segmentos de clientes do CCB

A solução que se considera para avaliação é a que resulta de afetação modal dos clientes do CCB i.e. cada consumidor é classificado no grupo ao qual se associa maior probabilidade de pertença (estimativa *a posteriori* obtida pelo modelo).

Uma caracterização breve da solução com 3 grupos associa-se, em primeiro lugar, às variáveis base de segmentação e, seguidamente, a outras variáveis que descrevem o perfil dos clientes do CCB.

Em geral, os 3 grupos constituídos associam-se significativamente (nível de significância 0.01) com as variáveis base de segmentação, como esperado. Na Ilustração 5 e na Ilustração 3 são visíveis as diferenças entre os grupos nas variáveis base de segmentação, evidenciando-se o G2 como o dos clientes do CCB mais confiantes no CCB, mais satisfeitos e que mais tencionam recomendar a instituição a um amigo.

A árvore de classificação CART representada na Ilustração 6 (3 primeiros níveis) ajuda a completar a descrição dos grupos a partir das variáveis base de agrupamento. A parametrização da árvore completa inclui 5 níveis e números mínimos de observações em nós pais e filhos de 11 e 10, respetivamente. A sua precisão medida em amostra de treino vs teste (50% de amostra global) apresenta-se na Tabela 10. Trata-se, como seria de esperar, de uma árvore com elevada precisão já que se baseia em variáveis preditivas dos grupos que são as próprias variáveis de agrupamento.

Quanto às variáveis demográficas, apenas se regista uma associação significativa entre os grupos e o nível etário. O G2 e o G3 destacam-se com uma percentagem relativamente mais elevada de indivíduos mais velhos e mais novos, respetivamente. As restantes variáveis demográficas não discriminam entre os grupos de consumidores do CCB.

Ilustração 3 – Perfis dos grupos segundo as variáveis base (I)

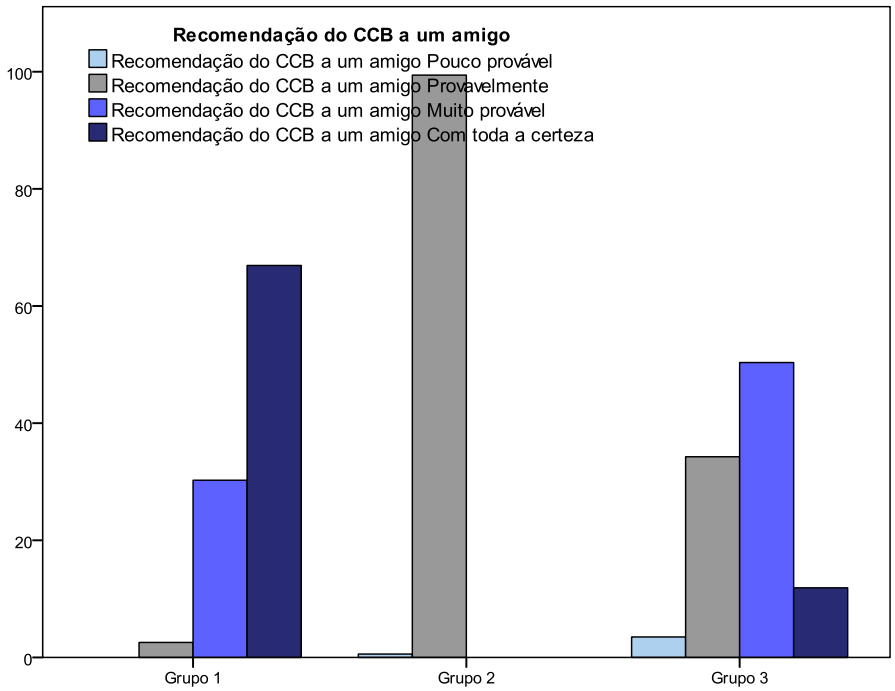
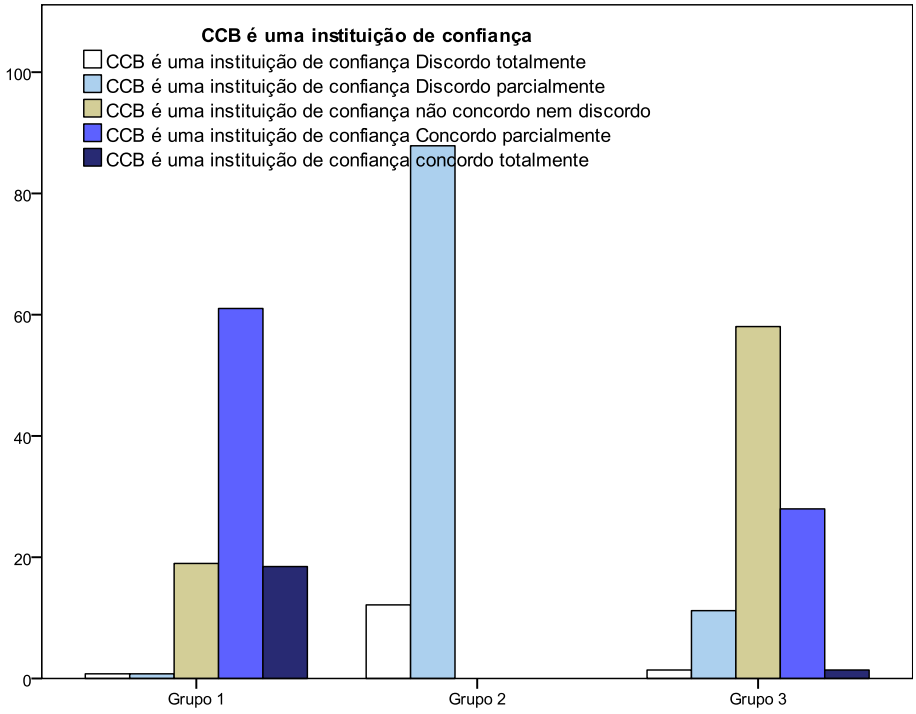


Ilustração 4 – Perfis dos grupos segundo as variáveis base (II)



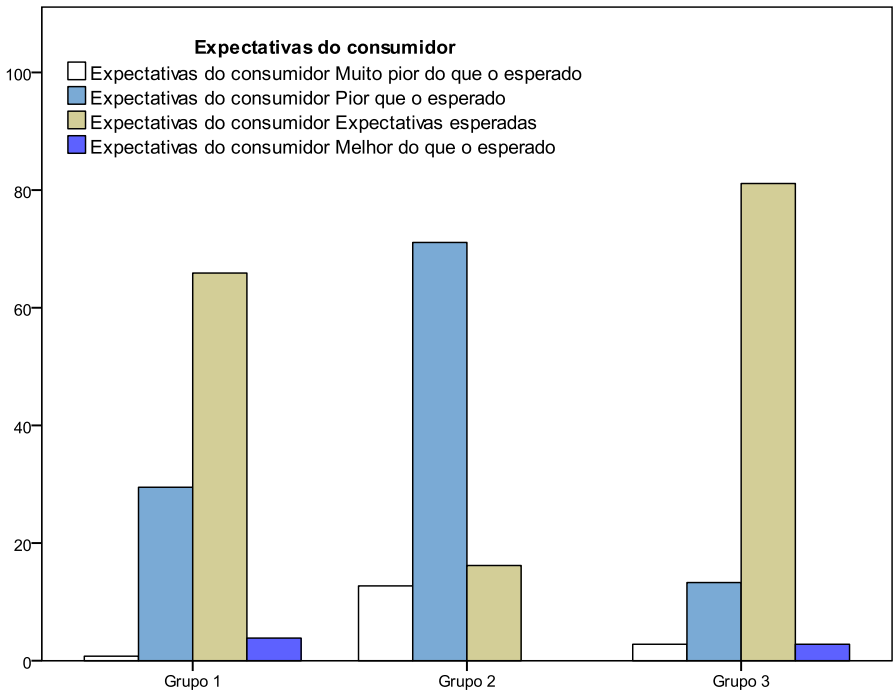
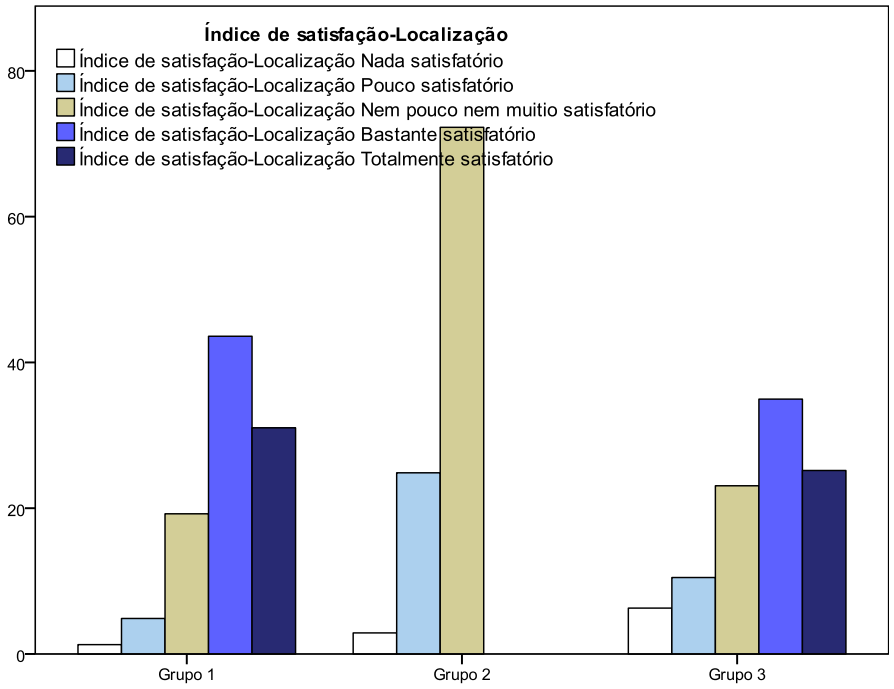


Ilustração 5 – Perfis dos grupos segundo as variáveis base (III)



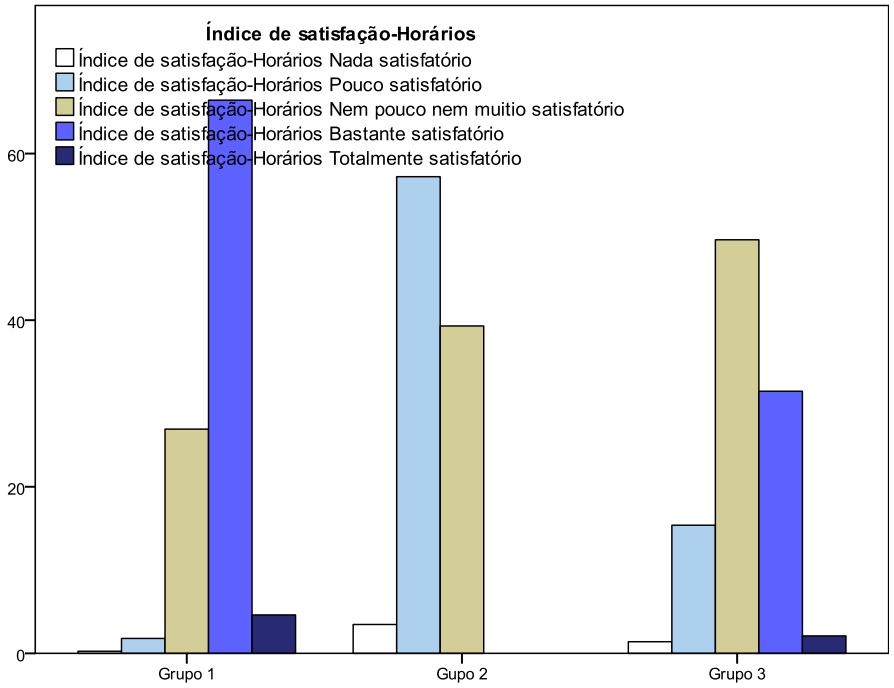


Tabela 10 – Precisão de CART (5 níveis) avaliada em amostra de treino vs teste

		Grupo Previsto			
Amostra	Grupo efetivo	G1	G2	G3	Percent. Class. correcta/
Treino	G1	184	15	4	90,6%
	G2	12	76	0	86,4%
	G3	15	0	47	75,8%
					87,0%
Teste	G1	173	14	0	92,5%
	G2	5	80	0	94,1%
	G3	18	0	63	77,8%
					89,5%

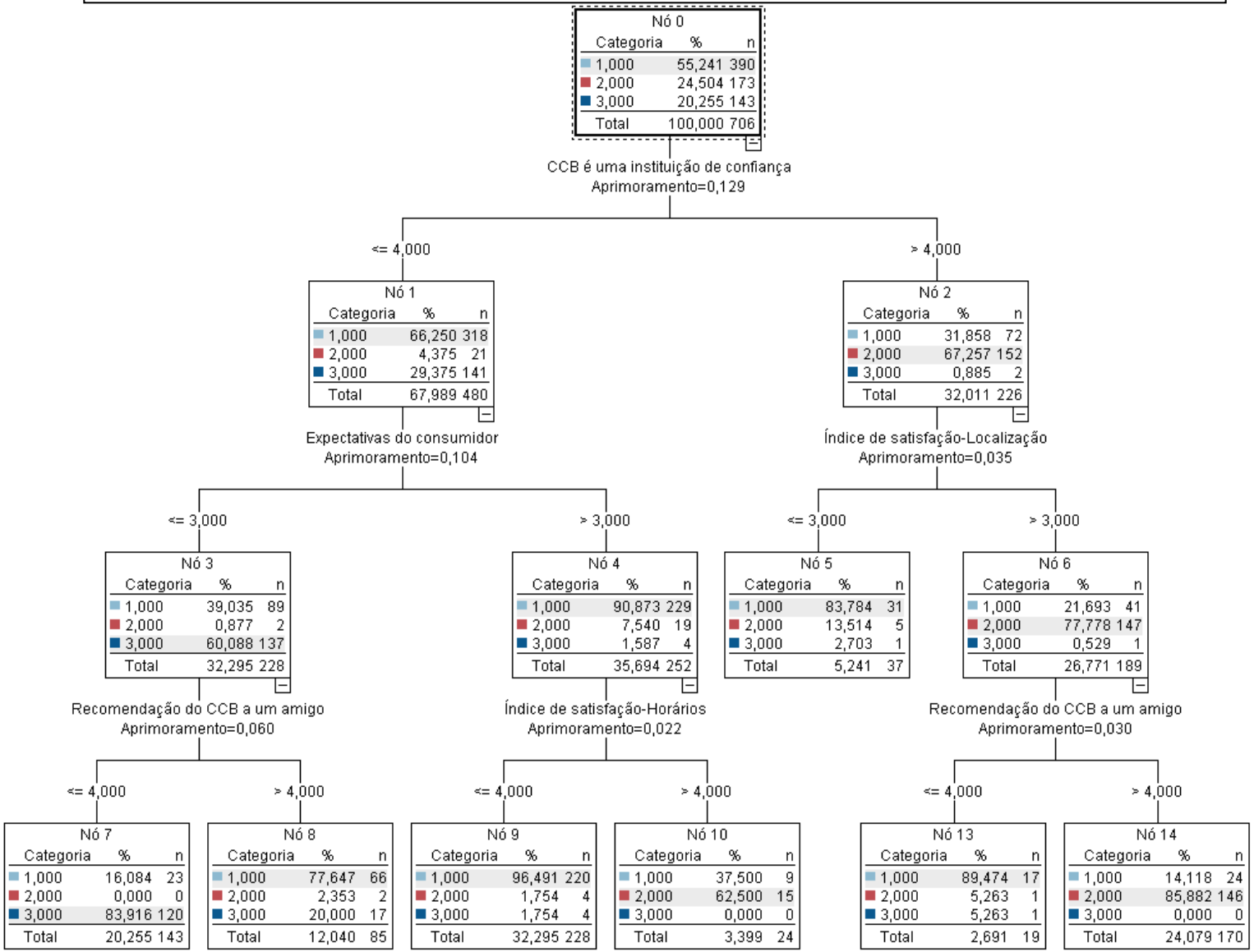
5. Discussão e perspetivas

Neste trabalho trata-se a avaliação de agrupamentos, em particular, a avaliação de estabilidade de partições resultantes de análises de agrupamento.

O aspeto inovador do caso prático apresentado refere-se, em primeiro lugar, à constituição de segmentos de clientes do CCB - Centro Cultural de Belém. Nesta aplicação o desafio é lidar com variáveis ordinais, tipicamente não contempladas na análise de agrupamento que sistematicamente considera as medidas tipo Likert (como as de Tabela 3) como variáveis métricas. Considera-se que embora as referidas medidas pudessem ser tidas como intervalares - admitindo igual espaçamento entre as categorias, o que se considera razoável – a abordagem proposta prescinde de pressupostos sendo, por isso, mais útil em aplicações.

Quanto à avaliação da estabilidade do agrupamento obtido, a metodologia utilizada – (Cardoso et al. 2009) - permite efetuar a validação cruzada da solução de agrupamento a partir de uma amostra de dimensão moderada, recorrendo à introdução de pesos associados às observações. Neste trabalho, a consideração específica dos pesos 10^{-10} procura imitar a constituição tradicional de amostras de treino e teste. No entanto, e embora não se tenham testado pesos alternativos, esta metodologia é flexível, permitindo a um analista análises comparativas resultantes da consideração de pesos diversos.

Ilustração 6- Árvore de classificação CART para solução de agrupamento sobre amostra global (3 níveis)



Os resultados obtidos na aplicação ilustram a dificuldade prática da avaliação de agrupamentos, não sendo clara a indicação de uma partição candidata (3 ou 5 grupos) quando se atende à estabilidade. A este propósito será reconfortante a leitura de (Jain e Dubes 1988): *The validation of clustering structures is the most difficult and frustrating part of cluster analysis …* (p. 222)

Seria naturalmente útil considerar mais índices de concordância (IC) entre partições. No entanto, a escolha dos índices V de Cramer e Rand ajustado atende (como foi dito) à incorporação de limiares de concordância por acaso, sob a hipótese de independência das partições o que não é considerado na maioria de propostas de IC.

Sendo assim, a investigação futura deverá considerar uma metodologia adequada para incorporar limiares de concordância em IC comummente utilizados na avaliação de resultados em análise de agrupamento. Uma primeira proposta neste sentido é a que se apresenta, sobre dados simulados, em (Amorim e Cardoso 2010). Nesta proposta, a simulação de tabelas de classificação cruzada sob hipótese de independência restrita é utilizada como meio para determinar limiares úteis para quaisquer valores de IC entre partições. Este será futuramente um meio para melhor aferir a propriedade da estabilidade em soluções de agrupamento.

Referências

Agresti, A. (2002). Categorical Data Analysis, Wiley.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716–723.

Amorim, M. J. e M. G. M. S. Cardoso (2010). Limiares de concordância entre duas partições. XVIII Congresso Anual da Sociedade Portuguesa de Estatística. S. Pedro do Sul**:** 47-49.

Breckenridge, J. (1989). "Replicating cluster analysis: method, consistency and validity." Multivariate Behavioral Research **24**: 147-161.

Breiman, L., J. H. Friedman, et al. (1984). Classification and Regression Trees. California, Wadsworth, Inc. .

Cardoso, M. G. e L. Moutinho (2003). "A Logical Type Discriminant Model for Profiling a Segment Structure." Journal of Targeting, Measurement e Analysis for Marketing **12**(1): 27-41.

Cardoso, M. G. M. S. (2007). Clustering e cross-validation. IASC 07 - Statistics for Data Mining, Learning e Knowledge Extraction, Aveiro, Portugal.

Cardoso, M. G. M. S., K. Faceli, et al. (2009). Evaluation of clustering results: the trade-off bias-variability. Classification as a Tool for Research.11th IFCS Biennial Conference, Dresden, Springer. Berlin-Heidelberg-New York. Hermann Locarek-Junge, Claus Weihs (editors). P. 201-208

Duarte, A. A. (2009). A satisfação do consumidor nas instituições culturais. O caso do Centro Cultural de Belém Tese de Mestrado em Marketing ISCTE-IUL.

Dudoit, S. e J. Fridlyand (2002). "A prediction-based resampling method for estimating the number of clusters in a data set." Genome Biology **3**(7): 0036.0031-0036.0021.

Faceli, K., A. Carvalho, et al. (2005). Validação de algoritmos de agrupamento. São Carlos, ICMC - Universidade de São Paulo.

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

Jain, A. K. e R. C. Dubes (1988). Algorithms for clustering data, Englewood Cliffs, N.J.: Prentice Hall.

Lange, T., V. Roth, et al. (2004). "Stability based validation of clustering solutions." Neural Computation **16**: 1299-1323.

Law, M. H. e A. K. Jain (2003). Cluster validity by bootstrapping partitions, Department of Computer Science e Engineering. Michigan State University.

Levine, E. e E. Domany (2001). "Resampling method for unsupervised estimation of cluster validity." Neural Computation **13**: 2573-2593.

McIntyre, R. M. e R. K. Blashfield (1980). "A nearest-centroid technique for evaluating the minimum-variance clustering procedure." Multivariate Behavioral Research **2**: 225-238.

Meulman, J. J., A. J. Van der Kooij, et al. (2004). Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal e Nominal Data., Sage Publications.

Milligan, G.W. and Cooper, M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21 441-458.
Mirkin, B. (1996). Mathematical Classication and Clustering, Kluwer Academic Publisher

Mirkin, B. (1996). Mathematical Classification e Clustering. Dordrecht/ Boston/ London, Kluwer Academic Publishers.

Mirkin, B. (1998). Data Analysis e Classification. International Summer School on Knowledge Discovery in Databases e Data Mining: Methods e Applications, Caminha, Portugal.

Quinlan, J. (1993). C4.5: Programs for Machine Learning. California, Morgan Kaufmann Publishers.

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical Association **66**: 846-850.
Schwarz, G. (1978). "Estimating the Dimenson of a Model." The Annals of Statistics **6**: 461-464.

Tibshirani, R., G. Walther, et al. (2001). Cluster validation by prediction strength, Department of Statistics, Stanford University.

Vermunt, J. K. e J. Magidson (2005). Technical Guide for Latent GOLD 4.0: Basic e Advanced. Belmont, Statistical Innovations Inc.