Segmentação de Clientes de Lojas de Pequena Dimensão

Margarida G.M.S. Cardoso, Dep. de Métodos Quantitativos. ISCTE
Av. das Forças Armadas 1649-026 LISBOA (email: margarida.cardoso@iscte.pt)
Armando B. Mendes, Dep. de Matemática, Universidade dos Açores.
R. da Mãe de Deus, 9501-801 PONTA DELGADA (email: amendes@notes.uac.pt)

Resumo

Neste trabalho apresentam-se resultados da segmentação de clientes de uma cadeia de retalho. A análise é baseada numa amostra de mais de 3500 respondentes a um inquérito. Para agrupar os indivíduos é estimado um modelo de segmentos latentes. A estimação tem como objectivo maximizar a função de probabilidade *a posteriori* e integra diversos algoritmos (implementados no software *Latent Gold*) usados em etapas sucessivas de um processo iterativo de estimação.

Os segmentos de clientes distribuem-se de modo diverso em quatro tipos de lojas que se determinam. A tipificação das lojas, usando um método hierárquico de agrupamento, atende a apreciações qualitativas de especialistas.

Palavras-chave: Segmentação, Modelos de Segmentos Latentes, Processos Hierárquicos de Agrupamento

Abstract

In this work we present results for a retail network client segmentation. The analysis is based on a sample of more than 3500 inquiry respondents. To cluster the individuals a model of latent segmentation is estimated. The objective of the estimate is to maximize the function of a posteriori probability and it integrates several algorithms (implemented in the software *Latent Gold*) used in successive stages of an iterative estimate process.

The client segments have different distributions in four clusters of stores that are determined. The store clustering uses a hierarchical method, and was assisted by specialists' qualitative appreciations.

Agradecimentos

Este texto é uma homenagem à Professora Isabel Hall Themido que tão entusiasticamente participou neste trabalho e a quem se devem parte dos resultados apresentados.

Introdução

Após um período em que se observou um grande crescimento do volume de vendas em grandes superfícies, assiste-se hoje em Portugal (e noutros países europeus) a uma fase de maior dinamismo nas pequenas e médias superfícies de retalho. Em resultado:

- os hipermercados e os supermercados são hoje, de acordo com dados Nielsen, as estruturas comerciais mais importantes (maior volume de vendas) em Portugal;
- os supermercados batem os hipermercados em diversas rubricas, tendo mesmo a sua quota de mercado global ultrapassado a dos hipermercados desde 1997.

A importância crescente de superfícies de menor dimensão induziu os grupos de retalho a investir fortemente em lojas desse tipo, apostando num retalho de proximidade. Este é um investimento elevado, com menores economias de escala e que obriga a um processo de decisão mais elaborado. A localização, dimensão e serviços oferecidos nas lojas (nomeadamente a sua adequação às necessidades dos segmentos ou nichos-alvo de mercado) estão a ter, em consequência, uma atenção crescente no processo de tomada de decisão.

Neste trabalho, apresentam-se algumas análises realizadas sobre dados relativos a uma cadeia de supermercados:

- Segmentação de clientes das lojas da cadeia, baseada em variáveis que caracterizam a relação dos clientes com o serviço prestado na loja. Nesta análise o agrupamento é apoiado na estimação de um modelo de segmentos latentes que integra as variáveis base com o pressuposto de que seguem uma distribuição multinomial.
- Tipificação das lojas a partir de um conjunto de variáveis cuja selecção é resultado de um longo processo de interacção com especialistas na apreciação de resultados alternativos de agrupamento. O processo de agrupamento utilizado é o método de Ward considerando distâncias euclidianas quadráticas.

Para além de constituírem um auxílio para a compreensão do mercado de clientes e da estrutura de oferta pretende-se que os resultados das análises realizadas possam vir a apoiar futuros modelos de previsão de vendas na cadeia de lojas.

Segmentação dos Clientes

Variáveis Base

TABELA 1 – VARIÁVEIS BASE DE SEGMENTAÇÃO

VARIÁVEIS BASE	TIPO	NÍVEIS DAS VARIÁVEIS E FREQUÊNCIAS OBSERVADAS DAS RESPOSTAS					
FREQUÊNCIA DE COMPRAS (P01)	ordinal	**todos os dias**	**duas a três vezes por semana**	**uma vez por semana**	**uma vez por mês**	**ocasionalmente**	**primeira vez**
		31%	33%	15%	6%	12%	3%
HÁBITOS DE COMPRA (P02)	nominal	**durante a semana**	**ao fim de semana**	**em ambas as situações**			
		34%	11%	55%			
ORIGEM DA VIAGEM DE COMPRA (P03)	nominal	**de casa**	**do emprego**	**de outro local**			
		65%	26%	9%			
TEMPO DE VIAGEM À LOJA (P08)	ordinal	**até 2 min. a pé**	**2 a 5 min. a pé**	**5 a 10 min. a pé**	**Mais de 10 min. a pé**	**Até 5 min. de carro**	**5 a 15 min. de carro**	**mais de 15 min. de carro**
		13%	28%	15%	6%	10%	12%	15%
NÍVEL MÉDIO DE GASTO MENSAL NA LOJA (P19)	ordinal	**menos de 15 contos**	**15 a 30 contos**	**30 a 50 contos**	**50 a 75 contos**	**75 a 100 contos**	**mais de 100 contos**	
		38%	29%	12%	9%	6%	1%	
PERCENTAGEM DE GASTOS EM LOJAS DA CADEIA *(* GTP)	ordinal	**Menos de 20%**	**20% a 37%**	**37% a 60%**	**Mais de 60%**			
		25%	25%	25%	25%			

A selecção das variáveis base para segmentação dos clientes das lojas da cadeia atende, em primeiro lugar, ao interesse em considerar atributos da relação cliente-oferta. Tendo em conta que a intenção da constituição dos segmentos é, para além de proporcionar uma melhor compreensão do mercado, vir a estabelecer uma possível diferenciação na oferta, estes atributos são, naturalmente, relevantes. O valor dos segmentos constituídos terá, contudo, que ser posteriormente aferido também através do estudo da associação entre estes e outras variáveis (demográficas, por exemplo): apenas através de uma

caracterização mais completa se poderão avaliar questões como a acessibilidade dos segmentos, definição de políticas de *marketing* diferenciadas e previsão dos seus benefícios.

Após um estudo descritivo das variáveis atendendo, em particular, à sua variabilidade na amostra, foi estabelecido um conjunto de variáveis base de segmentação que se apresenta na Tabela 1.

Metodologia

Modelo de segmentos latentes

Para segmentar os clientes da cadeia de supermercados utiliza-se um modelo de segmentos latentes.

Em termos genéricos o modelo que se propõe considera que as variáveis base de segmentação ($\mathbf{Y_q}$, **q=1..6** coincidindo com os atributos apresentados na Tabela 1) são modeladas por uma mistura de distribuições multinomiais. A mistura atende à consideração da existência de **S** segmentos de clientes, cuja constituição é determinada no processo de ajustamento do modelo. Como resultado da estimação obtêm-se, então:

- Estimativas dos parâmetros das multinomiais para cada segmento.
- Probabilidades de pertença de cada cliente a cada segmento.

Segundo o modelo que se propõe

$$f(y_1 ... y_6 \mid \underline{\theta}) = \sum_{s=1}^{S} \lambda_s f(y_1 ... y_6 \mid \underline{\theta}_s)$$

em que:

S é o número de segmentos,

λ_s é o parâmetro representando o *peso* do segmento **s**,

f representa a f.p. conjunta das variáveis base $\mathbf{Y_q}$ (q=1…6),

$\underline{\theta}_s$ representa o vector de parâmetros modelando as característica $\mathbf{Y_q}$ dos indivíduos, intra-segmento **s** (incluindo, em particular, as probabilidades associadas aos níveis de $\mathbf{Y_q}$ no segmento **s**).

Os segmentos (variáveis latentes) podem ser representados por um vector $\underline{\mathbf{z}}=(\mathbf{Z_1 \ldots Z_S})$, para o qual se propõe uma distribuição multinomial em que os parâmetros $\underline{l}$ se identificam com os *pesos* já referidos

$$\underline{z} \sim M_{S-1}(\lambda_1, \ldots, \lambda_S)$$

$\mathbf{Z_s}$ (componente do vector $\underline{\mathbf{z}}$) representa, portanto, o número de observações da amostra que se integram no segmento **s** (**s=1…S**) e $\mathbf{z_{is}}$ é o indicador de pertença de cada entidade **i** (**i=1…I**) da amostra ao segmento **s** ($\mathbf{z_{is}}$ **=1** se **i** Î **s** e $\mathbf{z_{is}}$ **= 0** caso contrário) verificando-se que

$$Z_s = \sum_{i=1}^{I} z_{is}$$

No modelo proposto não se consideram interacções entre as variáveis latentes. No entanto, dependendo de resultados da estimação associada a modelos alternativos, admite-se a possibilidade de integrar interacções entre as variáveis base.

Por outro lado, no modelo, atende-se, do modo seguinte, ao nível de medição das variáveis considerado na representação das probabilidades de observar cada nível l de uma variável $\mathbf{Y_q}$ num segmento **s**:

- se $\mathbf{Y_q}$ é variável nominal com $\mathbf{l=1..L_q}$ níveis

$$\theta_{q_l s} = \beta^s + \beta^0_{q_l} + \beta^s_{q_l}$$

- se $\mathbf{Y_q}$ é variável ordinal com $\mathbf{l=1..L_q}$ níveis

$$\theta_{q_l s} = \beta^s + \beta^0_{q_l} + \beta^s_q \times l$$

A eventual consideração de dependências locais entre duas variáveis base ($\mathbf{Y_q}$ e $\mathbf{Y_{q^*}}$) levará à introdução de parcelas adicionais (na expressão que define $q_{\mathbf{q,s}}$ e $q_{\mathbf{q^*,s}}$) modelando a interacção:

$$\theta_{(q_l, q^*_k)s} = \beta^s + \beta^0_{q_l} + \beta^0_{q^*_k} + \beta^0_{(q_l, q^*_k)} + \beta^s_{q_l} + \beta^s_{q^*_k}$$

Integrada na modelação está, ainda, a proposta de distribuições conjugadas para os parâmetros a estimar. No caso do modelo proposto considera-se uma distribuição Dirichelet conjugada de cada distribuição multinomial associada quer às variáveis base, quer à variável que modela os segmentos latentes.

Processo de Estimação

Sendo **L(q|y)** a função de verosimilhança e **h(q)** a f.p. (conjunta) *a priori*, o processo de estimação de q procura maximizar a probabilidade *a posteriori* **h(q|y)** - dadas as observações e o conhecimento *a priori* - no sentido de obter um modelo mais credível. A integração da informação *a priori* e da informação contida nas observações de $\mathbf{Y_1 \ldots Y_6}$ na função objectivo faz-se usando a função de verosimilhança, através do Teorema de Bayes:

$$h(\underline{\theta} \mid \underline{y}) = \frac{L(\underline{\theta} \mid \underline{y})h(\underline{\theta})}{\int_{\Theta} L(\underline{\theta} \mid \underline{y})h(\underline{\theta})d\underline{\theta}}$$

Tendo em conta que

$$h(\underline{\theta}|\underline{y}) \propto L(\underline{\theta}|\underline{y})h(\underline{\theta})$$

e usando a transformação logarítmica, o processo de estimação (integrado no software *Latent Gold*) tem como objectivo maximizar

$$\log L(\underline{\theta}|\underline{y}) + \log h(\underline{\theta}).$$

O processo usado na estimação integra, basicamente, dois algoritmos que são usados em fases sucessivas:

- uma variante do algoritmo EM-*Expectation Maximization* de Dempster, Laird e Rubin em 1977 (usado numa fase inicial do processo);
- o algoritmo de Newton-Raphson (usado numa fase mais próxima da convergência para uma solução).

Análise e Resultados

A análise realizada permitiu, por meio de sucessivas modificações do modelo, obter uma solução de agrupamento considerada satisfatória, atendendo aos indicadores quantitativos e à sua interpretabilidade.

Na avaliação dos modelos alternativos usaram-se indicadores quantitativos habituais no domínio da Teoria da Informação. Estes critérios procuram simultaneamente maximizar a verosimilhança e minimizar a complexidade do modelo proposto (função do número de parâmetros a estimar). Eles têm, em geral, como objectivo, minimizar

$$-2LnL + \alpha \times d$$

em que:

α = constante de penalização;

d = número de parâmetros a estimar.

Para avaliar resultados alternativos de estimação para os diversos modelos propostos usaram-se:

- Critério AIC-*Akaike Information Criterion,* de Akaike, em 1974, obtido com α**=2**;
- Critério *BIC-Bayesian Information Criterion* de Schwartz, em 1978, obtido com α**=ln(I)**;
- CAIC-*Consistent Akaike Information Criterion* de Bozdogan, em 1987, obtido com α**=ln(I)+1**.

Numa das alternativas de modelação consideradas foi ensaiada (por exemplo) a integração da variável *Percentagem de gastos em lojas da cadeia* (GTP) como contínua, pressupondo a sua normalidade. Os modelos construídos foram posteriormente descartados em função de:

- Não se verificar a normalidade
  (aplicou-se Teste de ajustamento de Kolmogorov-Smirnov sobre a variável referida em cada um dos grupos constituídos);
- O ajustamento dos modelos ser de qualidade inferior ao ajustamento obtido mediante a consideração da variável como ordinal;
  De facto, para qualquer número de classes considerado, obtém-se um valor de BIC da ordem de 60.000 no caso desta variável ser integrada no modelo como contínua e no caso de ela ser considerada ordinal o BIC é da ordem de 40.000. Em consequência optou-se por perder informação mediante a categorização da variável GTP (obtida através dos quartis), tal como já se tinha apresentado na Tabela 1.

Na Tabela 2 apresentam-se os melhores dos muitos modelos testados e algumas das medidas de desempenho obtidas. As medidas de desempenho utilizadas para comparar os diferentes modelos são baseadas na função de verosimilhança: no logaritmo desta função (LL) ou seu quadrado (L^2) sendo as medidas BIC, AIC e CAIC definidas de forma idêntica nos dois casos (diferindo apenas na primeira parcela).

TABELA 2 – MEDIDAS DE DESEMPENHO ASSOCIADAS A MODELOS ALTERNATIVOS

Modelo		Desempenho							
		L^2				LL			
Número de Segmentos	Resíduos bivariados	L^2	BIC_L^2	AIC_L^2	CAIC_L^2	LL	BIC_LL	AIC_LL	CAIC_LL
2	_	4793	-67749	-13285	-76788	-22391	45039	44846	45071
2	P03 x P02	4586	-67924	-13484	-76959	-22288	44864	44647	44900
2	P03 x P02 P02 x P01	**4304**	-68191	-13762	-77224	-22146	**44598**	**44369**	**44636**
3	_	**4368**	-68102	-13692	-77132	-22178	**44686**	**44439**	**44727**

De acordo com os valores das medidas de desempenho representadas na Tabela 2 conclui-se que o modelo com dois segmentos e dois resíduos bivariados (considerando a interacção entre *Origem da Viagem de Compra* (P03) e *Hábitos de Compra* (P02) e esta variável e *Frequência de Compras* (P01) é o que apresenta melhores resultados (tendo associado um mínimo local para o BIC_LL, por exemplo), tendo sido o adoptado.

Após a afectação determinística dos clientes aos segmentos (cada indivíduo é afecto ao segmento modal) constitui-se um segmento A a que pertencem 57% dos clientes e um segmento B com os restantes 43%.

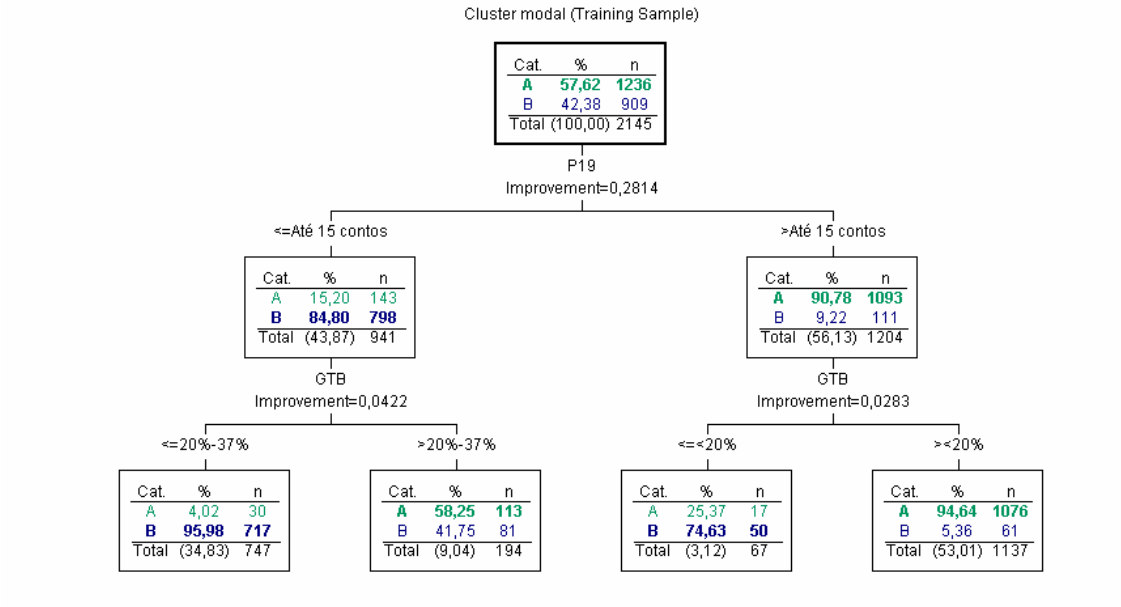
Descrição dos segmentos constituídos

Os resultados da segmentação foram avaliados por meio de uma árvore de decisão constituída segundo o algoritmo CART-*Classification and Regression Trees* de Breiman *et al.* (1984). Na modelação considerou-se uma amostra de treino (70% das observações) e uma amostra de teste (restantes 30%). A Figura 1 mostra uma árvore de classificação CART (um modelo discriminante lógico) apoiada nas variáveis base de segmentação.

A partir da árvore na Figura 1 é fácil concluir que 85% dos clientes que gastam menos de 15 contos por mês numa (qualquer) loja da cadeia são do segmento B e 91% dos que gastam mais são do segmento A. Apenas esta variável seria suficiente para obter um erro de classificação (percentagem de indivíduos incorrectamente classificados) inferior a 12%. Se conjugarmos esta variável com a *Percentagem de gastos em lojas da cadeia* (relativos ao total de gastos para a casa dispendidos em diversos tipos de estabelecimentos) obtêm-se erros de classificação inferiores (cerca de 9%, quer na

amostra de treino, quer na amostra de teste). De entre as variáveis base de segmentação, *Nível Médio de Gasto Mensal na Loja* e *Percentagem de gastos em lojas da cadeia* são as variáveis que principalmente distinguem os segmentos, A e B. A terceira variável com maior poder discriminante é a *Frequência de compras em lojas da cadeia* cuja integração no modelo em árvore resulta num erro de classificação de 8%, calculado sobre a amostra de teste. De acordo com este modelo passam a designar-se os segmentos A e B por ***Clientes Preferenciais*** e ***Clientes Eventuais***, respectivamente.

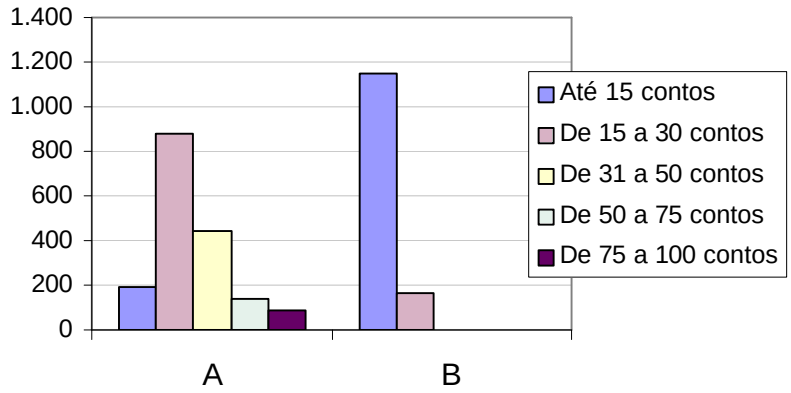
FIGURA 1 – ÁRVORE DE CLASSIFICAÇÃO SOBRE AS VARIÁVEIS BASE DE SEGMENTAÇÃO



Uma caracterização sumária dos dois segmentos de clientes é suportada por vários testes de Qui-quadrado de independência permitindo concluir que se verificam associações significativas entre estes e todas as variáveis base de segmentação (a um nível de significância de 0,01). Na Figura 2 ilustra-se a associação entre os segmentos e a variável que melhor os discrimina.

De acordo com análises semelhantes sobre atributos caracterizando o perfil dos segmentos pode, ainda, acrescentar-se que os ***Clientes Preferenciais*** têm mais idade e níveis de escolaridade um pouco mais baixos do que os ***Clientes Eventuais***.

FIGURA 2 – ASSOCIAÇÃO ENTRE OS SEGMENTOS E O *NÍVEL DE GASTOS EM LOJAS DA CADEIA*



Tipificação das Lojas

Os retalhistas sempre entenderam a localização como um factor crítico do sucesso de uma nova loja (Moore e Attenwell, 1991). No entanto, entender todos os factores relacionados com o potencial da localização e comportamento do consumidor exige grande quantidade de informação relevante de natureza geográfica, demográfica, socioeconómica e concorrencial (Themido e Mendes, 2001). A tipificação das 19 lojas objecto do presente estudo deverá proporcionar uma estrutura que facilite a compreensão das relações entre estes factores, assim como a futura previsão de vendas nas lojas.

Na selecção das variáveis base para a tipificação das lojas atendeu-se, em particular, aos critérios dos responsáveis das lojas. A partir dos seus critérios de apreciação, concluiu-se que estes consideravam, essencialmente, dois factores no agrupamento das lojas:

- uma medida da dimensão da loja e das vendas;
- uma medida da proporção de clientes residenciais *versus* clientes de passagem já que estes dois tipos de clientela eram, *a priori*, percebidos como distintos;

O primeiro factor poderia ser traduzido pelas *Vendas realizadas em 2000* (VEND2000) ou pela *área da loja*. Optou-se pela primeira variável tendo em conta a sua maior dispersão relativa.

A escolha da variável para traduzir o segundo factor atendeu, também, a critérios de dispersão. Optou-se, neste caso, pelo cruzamento de duas perguntas efectuadas no

inquérito, definindo, assim, a *Percentagem de clientes que declararam provir de casa e voltar para casa após as compras* (ORG_DST_CASA).

O agrupamento das lojas foi obtido usando as duas variáveis base referidas (VEND2000 e ORG_DST_CASA padronizadas), utilizando a distância Euclidiana quadrada e o método de Ward. *A priori* foram excluídas da análise duas lojas consideradas atípicas pelos responsáveis da cadeia: lojas 18 e 19. O dendograma correspondente apresenta-se na Figura 3.

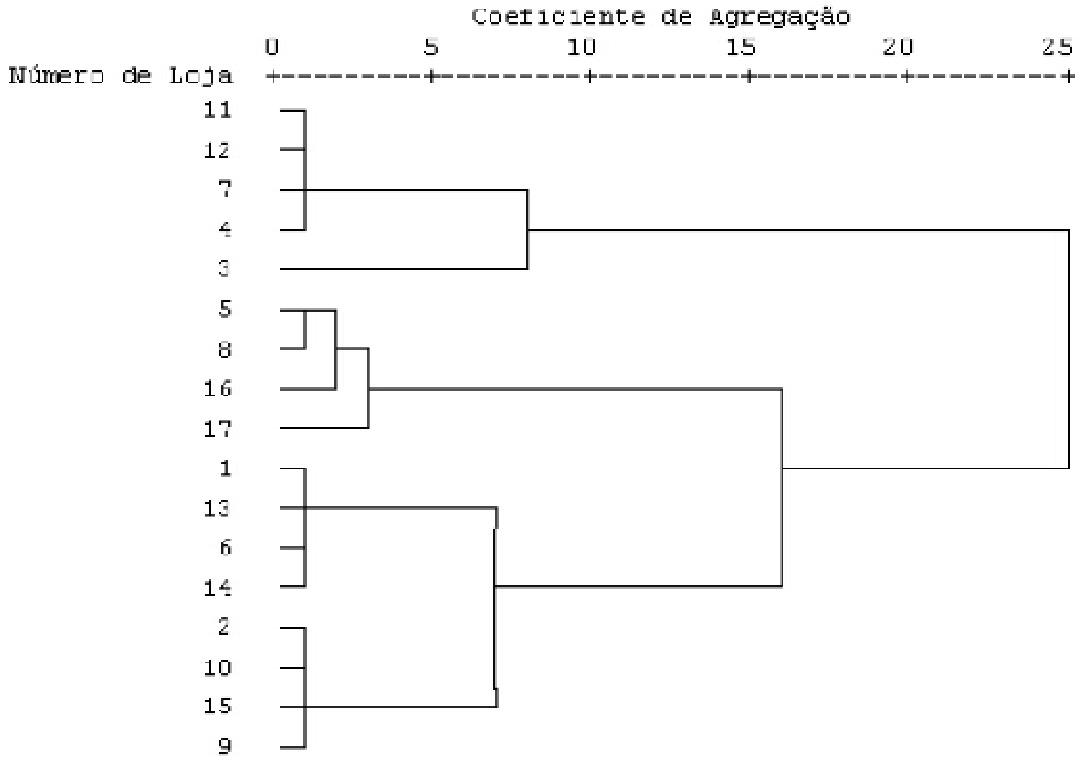
No sentido de determinar o ponto de corte do dendograma analisou-se a variação dos coeficientes de agregação com o número de grupos, tendo-se optado por uma partição em quatro grupos mais uma loja atípica constituindo um grupo singular: a loja 3.

Várias análises complementares foram realizados para validar os grupos formados (Halkidi *et al.,* 2001):

1) Utilizaram-se outros métodos hierárquicos de agrupamento como o método do Vizinho Mais Afastado e o da Mediana tendo-se obtido resultados semelhantes aos do método de Ward. Segundo o método das Ligações Médias e o método dos Centroides, a loja 3 isola-se primeiro, mas o essencial dos agrupamentos mantém-se.
2) Utilizaram-se medidas alternativas de (dis)semelhança entre pontos como a distância Euclideana, a distância de Chebychev, a distância Absoluta (*Block*) e a distância de Mahalanobis, tendo-se confirmado os resultados obtidos pelo recurso à distância Euclideana quadrada (usando o método de Ward).
3) Observou-se que os resultados eram pouco sensíveis a alterações nos métodos de padronização das variáveis.
4) Verificou-se, ainda, que os resultados eram pouco sensíveis à introdução ou remoção das lojas consideradas atípicas: lojas 3, 18 e 19.

Finalmente os resultados obtidos foram de encontro às expectativas dos responsáveis pelas lojas, facto que concluiu o processo de validação.

FIGURA 3 – DENDOGRAMA ASSOCIADO À ANÁLISE DE AGRUPAMENTO SOBRE 17 LOJAS



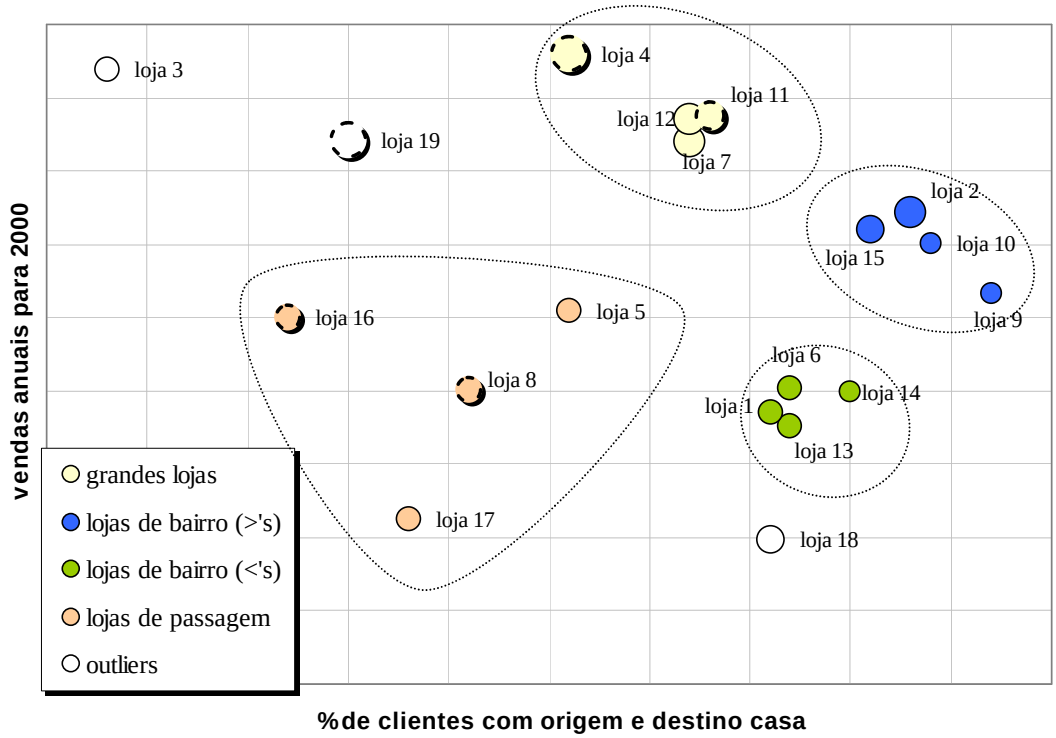
Na Figura 4 apresenta-se um gráfico de dispersão onde se podem visualizar os grupos formados no espaço das variáveis base de segmentação. Nesta figura complementa-se a caracterização das lojas com a sua dimensão (representada pelo diâmetro do círculo que a assinala) e a indicação da sua localização em centros comerciais (assinalada a tracejado). Esta representação suporta as seguintes denominações dos grupos constituídos: ***Grandes Lojas, Lojas de Bairro e Lojas de Passagem***.

Os agrupamentos surgem com aspecto bastante homogéneo com excepção do grupo denominado ***Lojas de Passagem*** onde a dispersão é mais elevada em ambos os eixos. Em geral é, ainda, visível uma associação entre a área da loja e as vendas. Algumas excepções, como a loja 3, poderão (segundo interpretação dos especialistas) corresponder a uma localização especialmente adequada, originando vendas superiores às previstas para lojas de dimensão semelhante. Assim, pensa-se que esta loja provavelmente constituirá uma semente para um novo agrupamento mais do que um simples *outlier*.

As lojas inseridas em centros comerciais apresentam valores baixos na variável que indica a percentagem de clientes residenciais, resultado que era expectável já que os

centros comerciais apresentam múltiplos pontos de atracção. A excepção da loja 11 refere-se a uma galeria comercial particularmente pequena onde a referida loja é a principal âncora.

FIGURA 4 – POSICIONAMENTO DOS TIPOS DE LOJA NAS DIMENSÕES BASE DE AGRUPAMENTO



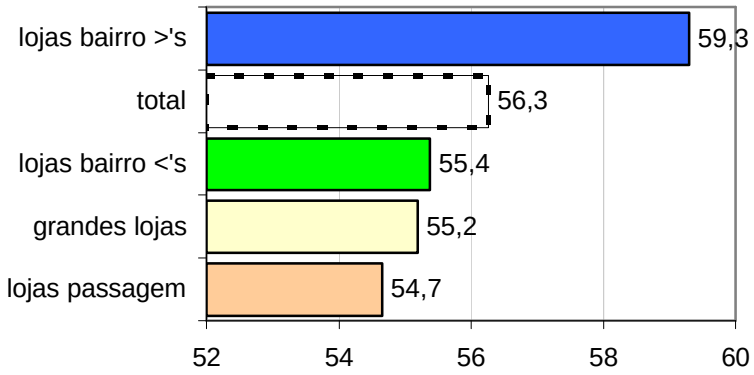
Nota: O raio do círculo representa a área comercial da loja e o tracejado assinala lojas em centros comerciais.

Utilizando as médias dentro de cada grupo para prever vendas para as lojas em estudo, conseguem-se obter erros relativos médios de 12%, valor bastante aceitável de acordo com a literatura (ver, por exemplo, Blankenship *et al.*, 1998).

Em complemento das análises efectuadas realiza-se, a seguir, o cruzamento dos segmentos de clientes e dos tipos de lojas no intuito de apoiar futuras decisões no sentido de adequar a oferta à procura.

Na Figura 5 pode notar-se que as ***Lojas de Passagem*** são as que têm menos ***Clientes Preferenciais***. As ***Lojas de Bairro*** pequenas atraem menos clientes deste segmento que as grandes. Em média as ***Grandes Lojas*** atraem tantos ***Clientes Preferenciais*** como as pequenas ***Lojas de Bairro.***

FIGURA 5 – PERCENTAGEM DE *CLIENTES PREFERENCIAIS NOS* DIVERSOS TIPOS DE LOJAS



Notas Finais e Perspectivas

Para a determinação dos segmentos de clientes utilizou-se um modelo de segmentos latentes. No processo de estimação Bayesiana deste modelo notou-se que o andamento da função de probabilidade *a posteriori* (associado ao crescimento do número de segmentos) se mostrou idêntico ao andamento da função de verosimilhança. Assim, o processo de estimação implementado no software *Latent Gold* acaba por não penalizar *suficientemente* o aumento da complexidade do modelo através da consideração de um conhecimento *a priori* que resulta vago. Este é um papel que acaba por ser desempenhado por outros indicadores (BIC, por exemplo). Será interessante, no futuro, comparar o desempenho deste algoritmo com o de outros processos alternativos de estimação Bayesiana, propondo aproximações diversas.

No que diz respeito à presente aplicação será, ainda, necessário complementar a caracterização dos segmentos para viabilizar uma eventual política de diferenciação da oferta.

A análise de agrupamento das 19 lojas estudadas apresenta bons resultados de acordo com o processo de validação adoptado. A opção pela utilização de apenas duas variáveis base para a tipificação das lojas atendeu ao critério de especialistas.

De futuro valerá a pena aprofundar a questão da integração de informação proveniente da opinião de especialistas quer no processo de selecção das variáveis base, quer no processo de avaliação de agrupamentos.

Referências

Blankenship, A.B.; Breen, G.E. e Dutka, A.F. (1998). State of the art marketing research. (2ª ed.). NTC Business Books: Chicago.

Breiman, L; Friedman, J.H.; Olshen, R.A.; Stone, C.J. (1984). Classification and Regression Trees. Wadsworth, Inc.: California.

Dempster, A. P., Laird, N. M., e Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society*. Series B39, pp.1-38.

Halkidi, M.; Batistakis, Y. e Vazirgiannis, M. (2001). On clustering validation techniques. Technical report, Dept of Informatics, Athens Univ. of Economics & Business, Athens, Greece (Hellas).

Magidson, J., e Vermunt, J. (2000). Latent class cluster analysis. Em J. A. Agenaars e A. L. McCutcheon (Eds.). Applied latent class analysis. Cambridge: Cambridge University Press.

Moore, S. e Attenwell, G. (1991). To be and where not to be - The Tesco approach to locational analysis. *OR Insight*, 4, pp. 21-24.

Themido, I.H. e Mendes, A.B. (2001). Multi outlet retail site location assessment: A state of the art. Relatório do CESUR. IST: Lisboa.

Vermunt, J. e Magidson, J. (2000). Latent Gold's user's guide. Statistical Innovations Inc.

Wedel, M. e De Sarbo, W. S. (1994). A review of recent developments in latent class regression. Em R. P. Bagozzi (Ed.). Advanced methods of marketing research. Blackwell Publishers Ltd.: UK, pp. 352-388.

Wedel, M. e Kamakura, W. (2001). Market segmentation - Conceptual and methodological foundations. (2ª ed.). Kluwer Academic Publishers.