FOLHA DE ROSTO

Integração de Conhecimento de Especialistas na Tipificação de Lojas de Retalho

Armando B. Mendes
Departamento de Matemática, Universidade dos Açores
R. da Mãe de Deus
9501-801 PONTA DELGADA
Telf.: 296 650 073
Fax.: 296 650 072
E-mail: amendes@notes.uac.pt

Maria Margarida G.M.S. Cardoso
Departamento de Métodos Quantitativos, ISCTE
Av. das Forças Armadas
1649-026 LISBOA
Telf.: 21 790 32 64
Fax.: 21 790 39 41
E-mail: margarida.cardoso@iscte.pt

INTEGRAÇÃO DE CONHECIMENTO DE ESPECIALISTAS NA TIPIFICAÇÃO DE LOJAS DE RETALHO

INTEGRATING MARKETING ANALYSTS KNOWLEDGE IN FOOD STORE CLUSTERING

Armando Brito Mendes[1]
DEP. DE MATEMÁTICA, UNIVERSIDADE DOS AÇORES
Margarida G.M.S. Cardoso[2]
DEP. DE MÉTODOS QUANTITATIVOS, ISCTE

PALAVRAS-CHAVE: MARKETING; ÁREAS DE INFLUÊNCIA; ANÁLISE CLASSIFICATÓRIA; *MULTIDIMENSIONAL SCALING*; MODELOS DE CLASSIFICAÇÃO SUPERVISIONADA, LOJAS DE RETALHO ALIMENTAR.

RESUMO: NESTE ARTIGO APRESENTAM-SE OS RESULTADOS OBTIDOS NA TIPIFICAÇÃO DE LOJAS DE RETALHO ALIMENTAR DE PEQUENA DIMENSÃO. O OBJECTIVO DA TIPIFICAÇÃO PRENDE-SE NÃO SÓ COM A REALIZAÇÃO FUTURA DE ACÇÕES DE MARKETING MAS TAMBÉM COM A PREVISÃO DE VENDAS POR MÉTODOS DE ANALOGIA. AINDA QUE SE TENHA RECOLHIDO MUITA INFORMAÇÃO RESULTANTE DE INQUÉRITOS A CLIENTES, BASES DE DADOS SOBRE CONCORRÊNCIA, DEMOGRAFIA DO INE E POR OBSERVAÇÃO DIRECTA DAS LOJAS E LOCALIZAÇÕES, O REDUZIDO NÚMERO DE LOJAS EXISTENTE DIFICULTA A AVALIAÇÃO INTERNA DAS TIPIFICAÇÕES OBTIDAS. ASSIM, SUGEREM-SE E APLICAM-SE DIFERENTES TÉCNICAS DE VALIDAÇÃO INTERNA E EXTERNA PARA CONSTRUÇÃO DOS AGRUPAMENTOS DE LOJAS COM UTILIZAÇÃO DE CONHECIMENTO DE ESPECIALISTAS. OS MÉTODOS APRESENTADOS DE INTEGRAÇÃO *A PRIORI* APRESENTAM BONS RESULTADOS. NO ENTANTO, CONCLUI-SE PELA MELHOR QUALIDADE DAS SOLUÇÕES RESULTANTES DA INTEGRAÇÃO DOS ESPECIALISTAS *A POSTERIORI,* VERIFICANDO-SE QUE A VARIAÇÃO INTRA-GRUPOS DAS VENDAS ANUAIS É MENOR SEGUNDO ESTA ABORDAGEM.

ABSTRACT: IN THIS ARTICLE WE USE SEVERAL CLUSTERING AND CLASSIFICATION TECNICHES TO DEFINE TYPES OF FOOD STORES OF SMALL DIMENSION. THE OBJECTIVE FOR THIS IS TO SUPPORT THE FUTURE ACCOMPLISHMENT OF MARKETING ACTIONS AND ALSO TO FORECAST SALES BY ANALOGY METHODS. ALTHOUGH WE HAD PICKED UP A LOT OF DATA BY MEANS OF INQUIRIES, EXISTING DATA BASES ON COMPETITION, DEMOGRAPHY, AND DIRECT OBSERVATION OF THE STORES AND ITS LOCATIONS, THE FEW EXISTING STORES MAKES INTERNAL VALIDATION OF OBTAINED CLUSTERS VERY DIFFICULT. THUS, WE SUGGEST HERE SEVERAL DIFFERENT TECHNIQUES OF EXTERNAL AND INTERNAL VALIDATION WITH THE EXPLICIT OR INPLICIT USE OF KNOWLEDGE OF SPECIALISTS. THE SUGGESTED *A PRIORI* METHODS OF SPECIALISTS KNOWLEDGE INTEGRATION PRESENT GOOD RESULTS. HOWEVER, WE FOUND BETTER QUALITY IN THE RESULTING SOLUTIONS FOR THE KNOWLEDGE INTEGRATION *A POSTERIORI*, AS THE CALCULATED INTRA-GRUPES VARIACTION OF THE ANNUAL SALES HAPPEN TO BE SMALLER.

[1] E-mail: amendes@notes.uac.pt
[2] E-mail: margarida.cardoso@iscte.pt

«A supplementary exercise in cluster description involves the investigation of the clusters in order to establish whether or not they can be given substantive interpretations (…). Such substantive descriptions do not make direct use of data, but require investigators to reflect on the results of classification studies.»

Gordon (1999)

1. INTRODUÇÃO

O sector do retalho em Portugal está a passar por uma reestruturação, tendo-se observado nos últimos anos a perda da liderança em termos de volume de vendas por parte das grandes superfícies em favor das pequenas e médias superfícies. Num mercado em fase de maturidade e cada vez mais competitivo o formato supermercado tem conseguido crescer continuamente tanto em número de unidades como em volume de vendas por unidade.

Já em 1996 os supermercados foram os únicos a registar um crescimento simultaneamente no número de lojas e no volume de vendas e consequentemente a aumentar a quota de mercado de 28 para 34% no universo Nielsen. Em 97 os supermercados atingiram a liderança e consolidaram a sua estratégia de expansão, principalmente no que se refere a cadeias como o *Pingo Doce*, *Intermarché*, *Dia* e *Lidl*. Segundo os dados mais recentes, em 2001, as vendas em supermercados eram já largamente superiores às vendas em hipermercados: 47% contra apenas 35% do total das vendas de lojas com produtos alimentares em Portugal (Figura 1).

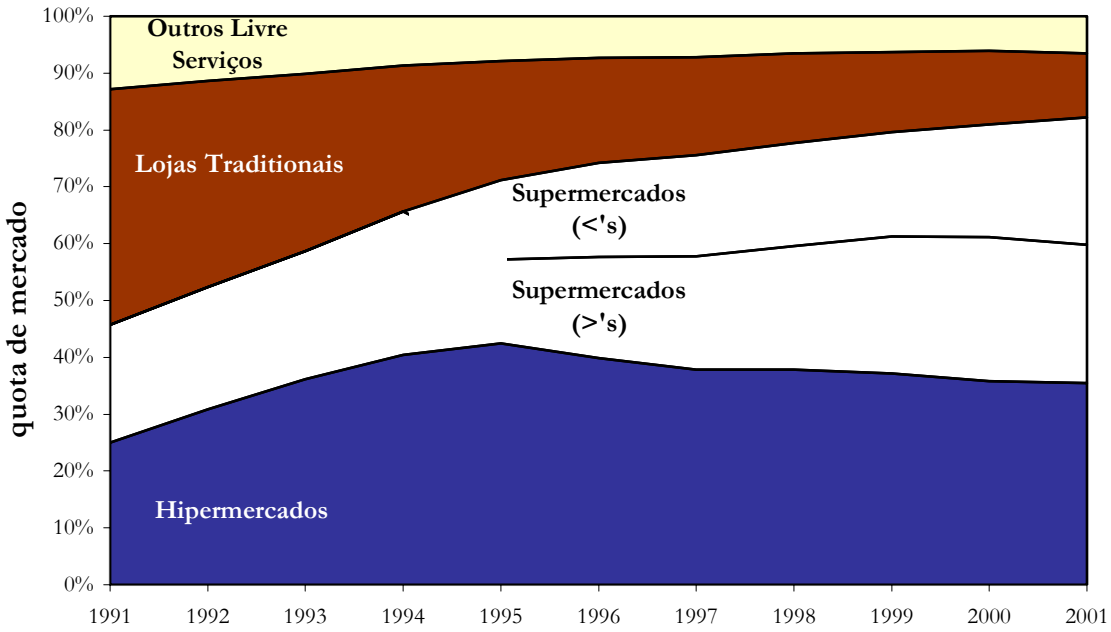

FIGURA 1 EVOLUÇÃO DA QUOTA DE MERCADO POR CONCEITO DE LOJA DE RETALHO ALIMENTAR
(Fonte: Distribuição Hoje, Novembro de 1997 e Dezembro de 2001[3]).

Perante este cenário alguns grupos de retalho Portugueses, à semelhança do que tem vindo a acontecer em vários países europeus (Mendes e Themido, 2003), estão a investir fortemente em lojas de menor dimensão, que ofereçam uma gama de produtos limitada, mas surjam como uma alternativa aos grandes espaços principalmente pela comodidade da proximidade e da rapidez de serviço.

No entanto, a localização de lojas de pequenas dimensões é especialmente crítica, uma vez que os investimentos apresentam poucas economias de escala e uma localização e \ ou gestão errada pode conduzir ao encerramento ou trespasse da loja (McGoldrick, 2000 e Salvaneschi, 1996)

[3] Note-se que as quotas de mercado foram calculadas considerando as vendas de lojas com produtos alimentares com exclusão das drogarias e agregando os puros alimentares e mercearias na denominação "Lojas Tradicionais".

Este trabalho situa-se no âmbito de um estudo de expansão de uma cadeia de lojas de retalho alimentar de pequena e média dimensão, pretendendo-se tipificar um conjunto de lojas existentes. Os referidos agrupamentos são úteis não apenas para se poder avaliar o desempenho relativo das localizações e gestão das lojas mas também, por utilização de métodos de previsão por analogia, para avaliação de novas localizações potenciais (Lilien e Rangaswamy, 2002 e Clarke, *et al.,* 2001).

2. MODELO DE INFORMAÇÃO E ANÁLISE

O problema da localização é complexo por ser necessário considerar inúmeras variáveis na avaliação de localizações de lojas (McGoldrick, 2000; Themido *et al.,* 1998 e Salvaneschi, 1996). Os retalhistas cedo se aperceberam da importância da localização, mas tentar perceber todos os aspectos do desempenho de lojas, potenciais localizações e comportamentos do consumidor obriga à recolha de enormes quantidades de informação de vários tipos como geográfica, demográfica, socioeconómica e referente a dinâmicas de competição.

Na Figura 2 apresenta-se uma classificação empírica dos factores explicativos das vendas de lojas de retalho alimentar de pequena dimensão. Os factores são divididos em três grandes grupos. Os factores endógenos pretendem avaliar aspectos apenas dependentes da loja e do local como as características da loja e da localização escolhida e a imagem da cadeia a que pertencem ou a gama e serviços associadas a essa imagem. De todas as características da loja, a área comercial é o factor de maior importância (Themido *et al.,* 1998 e Salvaneschi, 1996), o que é realçado no esquema por um ramo independente dos restantes. Os factores exógenos estão relacionados com a avaliação da área de influência da loja a nível do potencial de vendas (essencialmente variáveis demográficas) e da competição existente. É igualmente importante a caracterização socioeconómica dos clientes da loja, o conhecimento das suas preferências e da relação cliente \ loja.

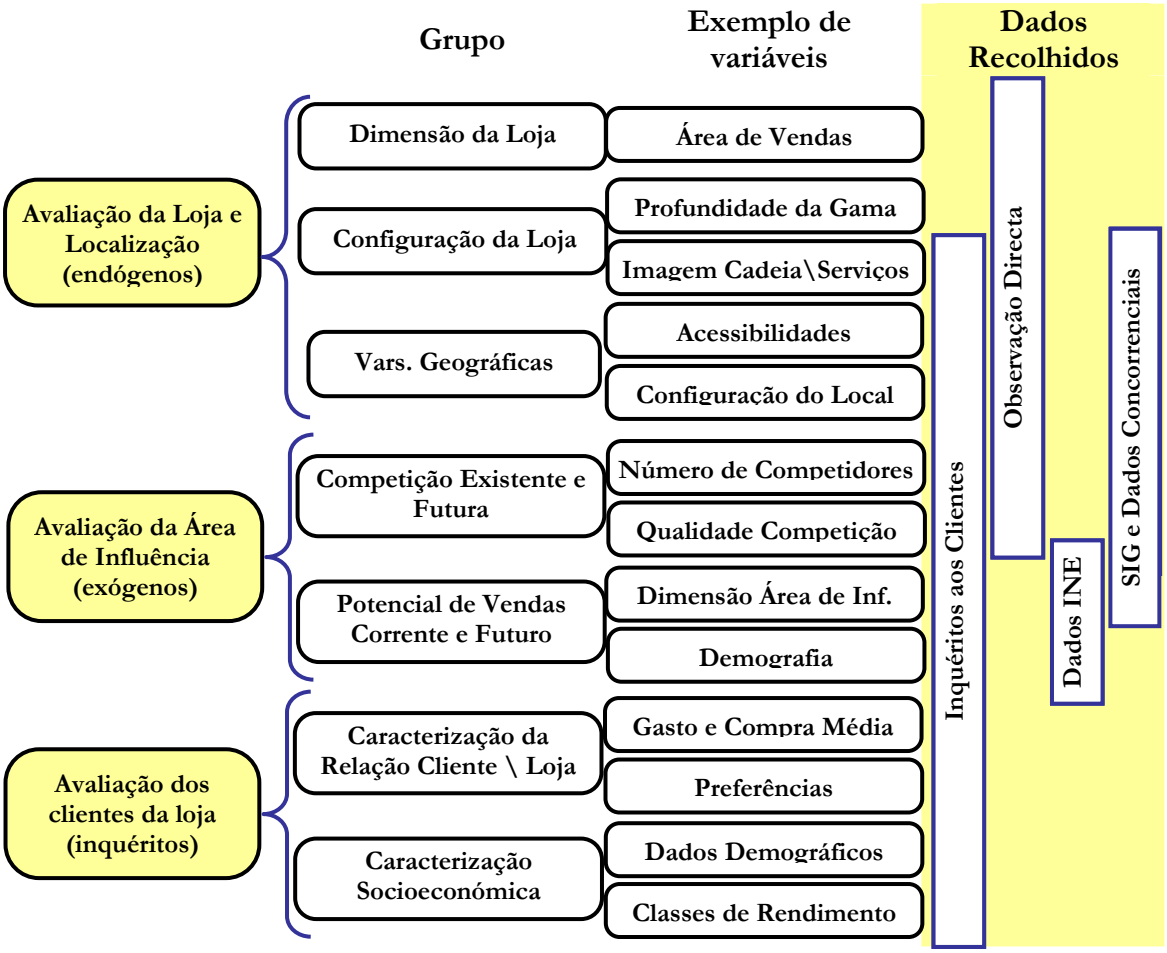

FIGURA 2 FACTORES QUE INFLUENCIAM AS VENDAS DE PEQUENAS LOJAS DE RETALHO E DADOS RECOLHIDOS.

A classificação de factores e variáveis sugerida é coerente com os resultados apresentados por Clarke, *et al.* (2001). Estes autores utilizaram mapas cognitivos, baseados em respostas a inquéritos por especialistas em localização das maiores cadeias retalhistas do Reino Unido, para identificar as principais variáveis utilizadas na prática neste tipo de decisões. Este processo resultou na identificação de 7 grupos de variáveis, sendo cinco deles directamente relacionados com as variáveis exploratórias definidas na Figura 2. As variáveis identificadas sobre o nome de *site / store configuration* e *retail composition* correspondem aos grupos "Configuração da Loja" e "Variáveis Geográficas". Clarke, *et al.* (2001) confirmam não só a classificação dos factores sugeridos como a complexidade das decisões relativas a localização.

3. METODOLOGIA DE RECOLHA DE DADOS

Na tentativa de obter um grande número de variáveis que cobrisse todos os aspectos da avaliação de localizações reuniram-se, neste estudo, dados de diferentes proveniências.

3.1. Recolha de informação por observação directa

Construiu-se uma *check list* com várias variáveis de localização e algumas relacionadas com competição e caracterização da área de influência, referidas a uma lista de itens preenchida por observação *in loco*. Esta lista de itens foi obtida para todas as lojas existentes pertencentes à cadeia em questão e para algumas lojas da competição mais importantes da área de influência. As variáveis resultantes são principalmente nominais mas algumas, resultantes de classificação de alguns aspectos da loja numa escala de nove pontos, apresentam-se numa escala ordinal.

3.2. Inquéritos aos clientes

Foram realizados inquéritos em todos os dias de duas semanas consecutivas, a todas as lojas existentes da cadeia, abrangendo um total de 3.766 inquiridos. Estes inquéritos permitiram avaliar a opinião dos clientes quanto a variáveis como a configuração da loja, acessibilidades e configuração da localização. Permitiram ainda a caracterização do cliente, caracterização da relação cliente - loja (motivação, meios de deslocação à loja, escolhas e preferências) e ainda identificação da concorrência. O inquérito foi realizado no ano de 2001 e repetido recentemente.

Dos inquéritos resultam essencialmente variáveis quantitativas como a percentagem de clientes que provém de casa, o volume médio de gastos na loja ou a compra média na loja. Em trabalhos anteriores apresentou-se já uma segmentação de clientes baseada nas respostas aos inquéritos com base em modelos de segmentos latentes. Desse trabalho, resultaram dois tipos de clientes que foram caracterizados como ***Clientes Preferenciais,*** os quais têm mais idade e níveis de escolaridade um pouco mais baixos, e os ***Clientes Eventuais*** (Cardoso e Mendes, 2002). No presente trabalho incluem-se como variável a percentagem de clientes do primeiro tipo em cada loja.

3.3. Dados demográficos do INE, BD concorrencial e análise espacial

Utiliza-se igualmente um grande número de variáveis quantitativas resultantes da base geográfica nacional do INE com informação demográfica do censo de 2001 e operacionalizada segundo um Sistemas de Informação Geográfico. Estas variáveis são utilizadas na avaliação de áreas de influência e na caracterização da concorrência. Na definição de áreas de influência utiliza-se igualmente uma base de dados com a localização de mais de 600 lojas de retalho alimentar em Portugal. Esta base de dados é mantida constantemente actualizada com a georeferenciação de novas lojas e recolha de alguns dados facilmente observáveis sobre as mesmas.

Tradicionalmente, as áreas de influência são delimitadas com *buffers* ou circunferências, com um raio adequado e calibrado utilizando resultados de inquéritos a clientes (Birkin *et al.,* 2002 e McMullin, 2000) ou delimitando coroas aproximadamente circulares baseadas em "tempos de viagem" na deslocação à loja e em algoritmos de caminho mais curto (Boots, 2002, Cowen *et al.,* 2000 e Salvaneschi, 1996). A informação sobre áreas de influência provém, no caso presente, da aplicação de um método que as define a partir de diagramas de Voronoi multiplicativos ponderados. Ao contrário dos restantes, este método permite, simultaneamente, incorporar a atractividade da loja e a presença de concorrência nas proximidades (Gonçalves e Mendes, 2002).

As variáveis utilizadas neste trabalho foram obtidas por cruzamento espacial entre as áreas de influência construídas por dois métodos distintos, os diagramas de Voronoi e as coroas obtidas por caminhos mais

curtos (Figura 3), os quais foram cruzados com duas técnicas de cálculo do valor das variáveis demográficas dentro da área de influência. No primeiro caso utiliza-se uma média ponderada das variáveis demográficas em que o peso é constituído pela fracção da área da secção estatística que se encontra dentro da área de influência delimitada para a loja.

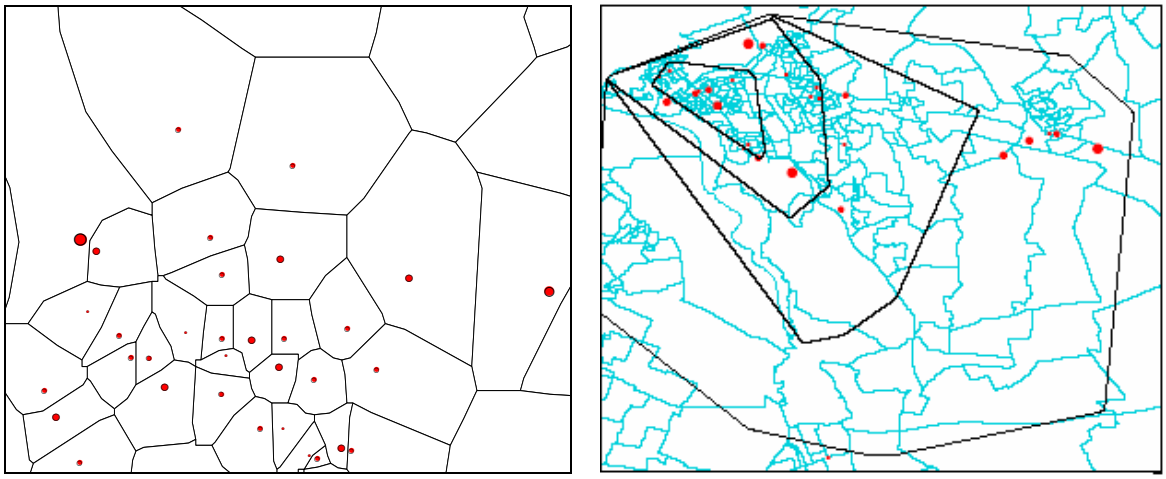

FIGURA 3 DIAGRAMAS DE VORONOI MULTIPLICATIVOS (ESQ.) E
COROAS DE CAMINHOS MAIS CURTOS A 2,5; 5; 10 E 15 MIN. DA LOJA (DIR.).
(Pontos representam lojas com raio proporcional à área de vendas.
À direita os polígonos mais claros delimitam as secções estatísticas do INE)

No segundo método utiliza-se uma regra de decisão para incluir ou não a variável demográfica associada a uma secção estatística na agregação referente a uma loja. A regra de decisão utilizada é igualmente baseada na fracção da área da secção estatística incluída na área de influência delimitada, sendo utilizado o parâmetro de 50%, *i.e.* se uma variável corresponder a uma secção estatística com mais de 50% da área dentro da área de influência da loja então é incluída na agregação. O segundo método tem a vantagem de ajustar as fronteiras da área de influência às fronteiras das secções estatísticas, o que pode ser mais adequado tendo em conta que as secções estatísticas delimitadas pelo INE têm em consideração barreiras geográficas.

Com base na agregação anterior foram calculadas variáveis relativas como percentagens de totais e densidades por hectare. Deste processo, e apesar de se ter feito uma selecção das variáveis disponibilizadas pelo INE, resultou um número incomportável de variáveis próximo do meio milhar. Para reduzir este número determinou-se a matriz de coeficientes de correlação de Pearson e foram retiradas todas as variáveis com correlações significativas muito elevadas (acima de 0,95) iniciando-se a eliminação de variáveis pelas que apresentavam maior número de correlações.

Em resumo, deste processo de recolha de dados resultaram um total de 250 variáveis em todas as escalas de medida e que cobrem todos os aspectos referidos na Figura 2, as quais foram utilizadas na tipificação das lojas de retalho.

4. ANÁLISE DE INFORMAÇÃO

Apesar da abundância de dados, o número de lojas com informação disponível é muito reduzido – apenas algumas dezenas - o que dificulta o processo de escolha de variáveis adequadas ao agrupamento ou tipificação das lojas e respectiva caracterização. Vários autores (ver por exemplo Jain e Dubes, 1988 e Wedel e Kamakura, 2000) distinguem validação externa, por utilização de conhecimento qualitativo não decorrente dos dados usados na tipificação e interna efectuada por utilização dos mesmos ou novos dados. No actual contexto, a utilização de validação externa revelou-se essencial já que a validação interna não é possível com rigor (ver por exemplo Naert e Leeflang, 1978). Note-se que a questão da validação é especialmente relevante, já que os métodos utilizados permitem sempre obter uma partição dos dados, a qual tem sempre de ser avaliada e comparada com outras tendo em conta os objectivos do estudo.

Embora poucos trabalhos tenham sido apresentados considerando a necessidade de integração de conhecimento de especialistas na validação de tipificações, Bay e Pazzani (2000), por exemplo, recorrem a um painel de especialistas para interpretar regras de classificação. Este trabalho conclui que muitas das referidas regras são inúteis ou redundantes e ainda que aponte para a subjectividade das interpretações dos especialistas, confirma a necessidade da utilização do seu conhecimento.

Neste trabalho consideram-se dois métodos básicos de integração do conhecimento de especialistas na tipificação das lojas. Uma primeira abordagem, mais formal, integrando o conhecimento de especialistas por meio da construção de uma matriz de dissemelhanças perceptuais entre as lojas, abordagem que denominaremos integração *a priori*. Uma segunda via em que a integração se dá na avaliação dos agrupamentos formados, denominada neste trabalho como integração *a posteriori*.

4.1. Integração do conhecimento de especialistas *a priori*

Nesta abordagem pretende-se integrar a opinião dos especialistas na selecção das variáveis base de agrupamento das lojas, procurando formalizar a sua contribuição.

Para tal solicitou-se a alguns especialistas, profundamente conhecedores das lojas, o preenchimento de um questionário onde se comparam pares de lojas segundo uma escala de dissemelhanças ordinal, com nove pontos. A comparação é genérica tendo, no entanto, sido realçado que tomassem em especial consideração os aspectos da localização, caracterização da loja e do desempenho da mesma.

A matriz simétrica de dissemelhanças utilizada neste trabalho foi obtida por consenso entre os vários especialistas. Os dois métodos a seguir descritos correspondem a duas abordagens diferentes para utilização desta informação:

- No método CLUST>MDL utiliza-se a matriz de dissemelhanças directamente, como base para obtenção de agrupamentos de lojas e, em seguida, utiliza-se um modelo discriminante lógico para selecção de variáveis utilizadas na caracterização e interpretação dos grupos.
- No método MDS>CLUST começa-se por realizar uma análise MDS - *Multidimensional Scaling* não métrica, com posterior extracção de variáveis relevantes para a quantificação das dissemelhanças (usando regressão) e por fim aplica-se uma análise de agrupamento sobre essas variáveis.

Método CLUST>MDL: agrupamento seguido de modelo discriminante lógico

Por utilização do método de Ward (Ward, 1963) sobre a matriz de dissemelhanças perceptuais foi possível agrupar as lojas em quatro grupos. Os resultados deste agrupamento foram consistentes com os obtidos por meio de outros métodos hierárquicos de agrupamento, como o método do vizinho mais afastado e o da mediana. Segundo o método das ligações médias e o método dos centróides, algumas lojas isolam-se primeiro, mas o essencial dos agrupamentos mantém-se.

Atendendo à presença de casos omissos e à não verificação dos pressupostos habituais em técnicas de análise discriminante paramétrica a caracterização dos grupos obtidos é realizada a partir de uma árvore de classificação (Cardoso, 2003). Esta abordagem tem a vantagem de permitir utilizar variáveis em todas as escalas de medida e de incorporar técnicas para lidar com valores omissos, o que evita a necessidade de se efectuarem extensos e demorados tratamentos prévios aos dados. Os seus resultados (regras proposicionais de classificação nos grupos de lojas) têm a vantagem de serem facialmente compreensíveis.

No caso particular em estudo, atendendo à reduzida dimensão dos dados, a metodologia utilizada considera o modelo resultante da aprendizagem sobre todas as lojas. A construção da árvore de classificação baseia-se no algoritmo CART- *Classification and Regression Trees* (Breiman *et al.,* 1984) segundo a implementação *AnswerTree v. 3.1*. Como dependente usa-se a variável nominal resultante do agrupamento das lojas pelo método de Ward. Como variáveis explicativas são considerados os inúmeros atributos disponíveis para caracterizar as lojas.

A escolha entre árvores alternativas obtidas (resultantes de diferentes parametrizações ou diferentes ramificações em caso de empate na escolha de um atributo explicativo) foi efectuada em função da precisão da classificação usando o erro de resubstituição e, ainda, o resultante do método *leave-one-out*. Esta segunda estimativa do erro permite avaliar, com algum realismo, a capacidade predictiva do modelo obtido.

Na Figura 4 apresenta-se a melhor árvore de classificação seleccionada segundo os critérios explicados anteriormente. Esta árvore classifica incorrectamente 2 lojas correspondendo a 79% de classificações correctas segundo a estimativa *leave-one-out*. Note-se ainda que estes resultados foram ainda validados com o conhecimento dos especialistas e que a sua intervenção foi também usada para seleccionar, em caso de empate, a variável mais adequada para ramificar, privilegiando a interpretação dos grupos de lojas.

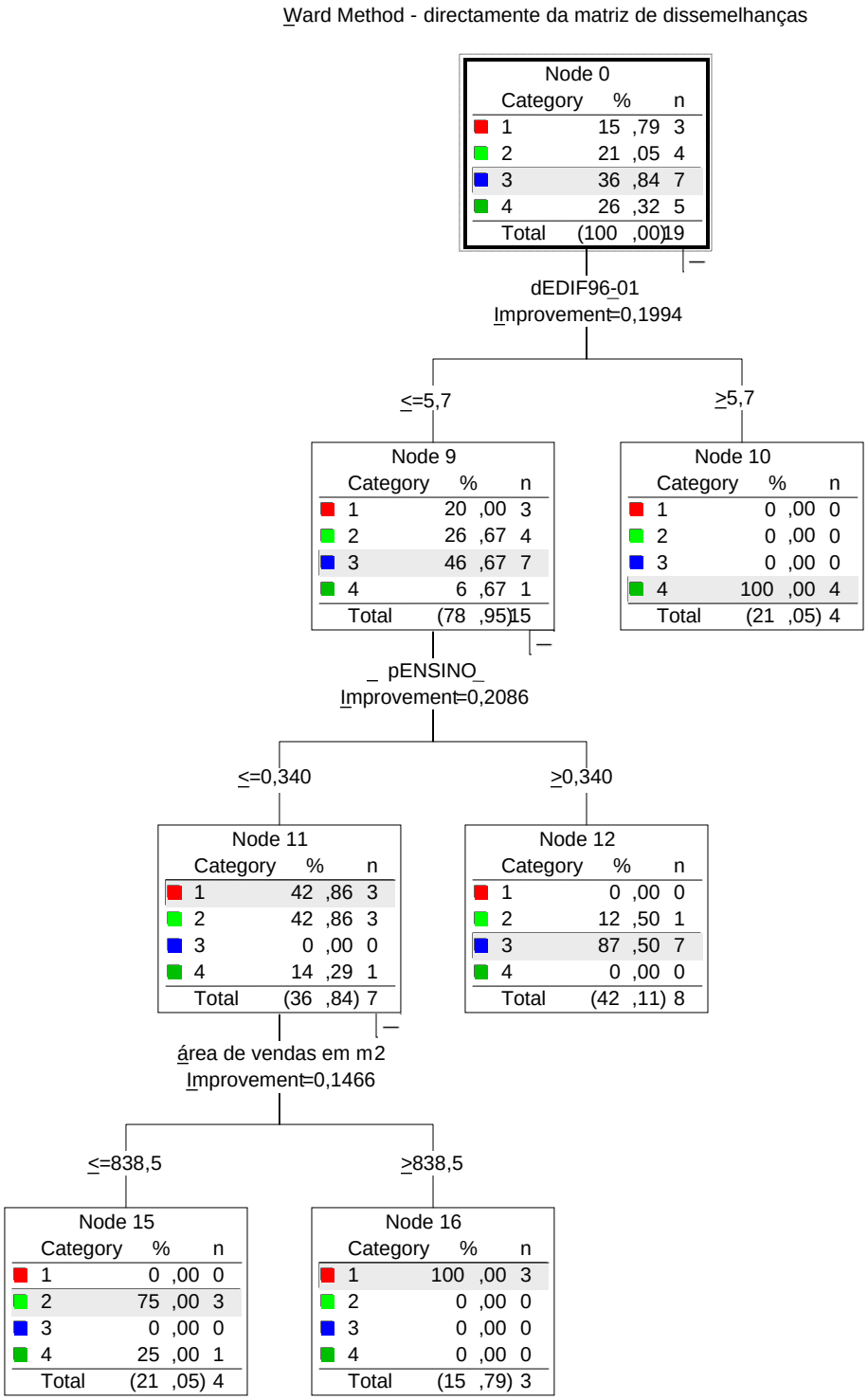

FIGURA 4 MELHOR ÁRVORE DISCRIMINANTE DAS CLASSIFICAÇÕES DAS LOJAS OBTIDAS DIRECTAMENTE DA MATRIZ DE DISSEMELHANÇAS.

- O primeiro nó folha (nó 10) separa o grupo 4 dos restantes segundo o valor da variável dEDIF96-01 correspondente à densidade de edifícios construídos entre 96 e 2001 por hectare com área de influência criada pelo método dos caminhos mais curtos a 2,5 minutos da loja. Logo, podemos considerar o grupo 4 como sendo constituído por lojas situadas em locais dinâmicos, com uma população em crescimento, designando-o por grupo de **lojas em crescimento**.
- O segundo nó folha (nó 12) separa o grupo 3 caracterizado por áreas de influência com uma população com elevadas qualificações académicas, considerado um indicador de elevados rendimentos. A variável pENSINO representa a percentagem de indivíduos residentes com o ensino secundário completo ou com um curso médio ou ainda com um curso superior, sendo a área de influência determinada como na variável anterior, pelo que denominaremos este grupo como **lojas nobres**.
- Os restantes nós terminais congregam lojas dos grupos 1 e 2, separadas segundo a dimensão da área de vendas. As lojas do grupo 1 caracterizam-se por apresentarem maiores dimensões relativamente ao grupo 4 pelo que o primeiro será chamado l**ojas maiores** e o segundo **lojas menores**. Note-se no entanto, que as denominações anteriores não se referem a todas as lojas mas apenas às que chegam ao nó 11 pelo que as restantes podem ou não ter áreas maiores.

Método MDS>CLUST: escolha de variáveis seguida de agrupamento

Segundo este outro caminho de integração do conhecimento de especialistas, realizou-se uma análise MDS não métrica, usando o algoritmo *ALSCAL*, um algoritmo de Takane, Young and Leeuw (descrito em Cox e Cox, 1994). Desta análise resultou uma solução com quatro dimensões, principais responsáveis pelas dissemelhanças entre as lojas, solução a que corresponde um valor de *Stress* de 7,8% e *RSQ* de 96%.
Posteriormente foram realizadas regressões lineares de diversas variáveis disponíveis, sobre as dimensões extraídas (*i.e.* considerando como explicativas as dimensões MDS extraídas), procedimento recomendado por vários autores (ver por exemplo Molinero e Cinca, 2000). Desta análise resultou a selecção de uma dezena de variáveis que eram muito bem explicadas pelas dimensões MDS com probabilidades de significância associadas ao teste F abaixo de 2%.
Para tentar reduzir este número e como se verificou que algumas dessas variáveis estavam altamente correlacionadas fizeram-se análises factoriais usando como método de extracção o das Componentes Principais. Dessa análise resulta de forma evidente, a partir do gráfico de valores próprios em função do número de factores, a decisão de extrair duas componentes principais, ainda que as percentagens de variância explicada não sejam muito elevadas (73%). Com o objectivo de caracterizar as dimensões extraídas apresenta-se, na Figura 5, o gráfico de dispersão das variáveis seleccionadas pelo processo anterior no espaço dos componentes principais extraídos após rotação pelo método Varimax com normalização de Kaiser.

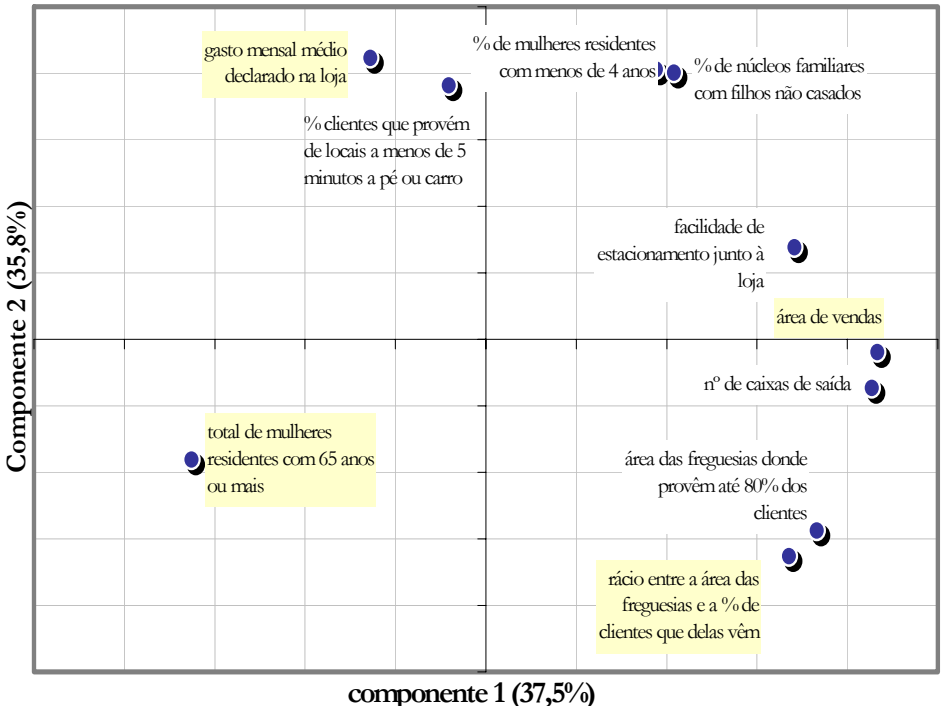

FIGURA 5 VARIÁVEIS ESTATISTICAMENTE SIGNIFICAS NA REGRESSÃO COM AS DIMENSÕES MDS NO ESPAÇO DOS COMPONENTES PRINCIPAIS

Estas variáveis apresentam correlações fortes como se verifica da proximidade dos pontos no gráfico anterior constituindo duas nuvens bem definidas. A componente 1 é caracterizada pela dimensão da loja e da sua área de influência enquanto a componente 2 é caracterizada pela proximidade dos clientes e pelos níveis de gastos.

Uma vez que muitas das variáveis estão altamente correlacionadas e a utilização de demasiadas variáveis pode, segundo Gnanadesikan (2001) mascarar a existência de grupos, a utilização dos componentes principais em substituição das variáveis originais poderia ser aconselhada. No entanto, vários autores desaconselham a utilização dos componentes principais directamente na análise de agrupamentos uma vez que estes podem não conseguir reproduzir o espaço multidimensional original e podem mascarar grupos existentes ou sugerir grupos inexistentes nos dados originais (Milligan, 1996).

Ainda que tal conclusão não seja consensual, adoptou-se uma solução de compromisso: um procedimento heurístico que consistiu em iniciar o agrupamento pelo Método de Ward por um número mínimo de variáveis (as duas com maiores pesos nos 2 componentes principais extraídos) e ir adicionando novas variáveis usando o critério de adicionar primeiro as variáveis menos correlacionadas com as já incluídas. Como critério de paragem utilizaram-se técnicas de validação interna como a variância explicada pelos agrupamentos (ver secção 4.3) e observação dos dendogramas formados.

Deste procedimento resultou um novo agrupamento das lojas, que se apresenta na Figura 6. Os grupos de lojas apresentados foram obtidos para as variáveis com fundo colorido na Figura 5. Estes grupos são bem definidos não tendo sido identificado qualquer loja atípica. Atendendo ao posicionamento das variáveis base de agrupamento no espaço definido pelas componentes principais extraídas denominaram-se os 3 grupos apresentados:

- **lojas de passagem:** um grupo de 4 lojas caracterizado por valores negativos no componente 2 relacionado com a proximidade dos clientes e com o gasto médio na loja;
- **lojas de bairro:** apresentando, pelo contrário, valores elevados no mesmo componente, sendo portanto caracterizadas por clientes com residência próxima e elevado gasto médio na loja;
- **lojas pequenas:** o grupo com maior número de lojas, apresentando valores negativos do componente 1 relacionado com a dimensão da loja e da área de influência.

Note-se que alguns destes agrupamentos aparentam alguma heterogeneidade nos valores das vendas anuais representada na Figura 6 pela área do círculo.

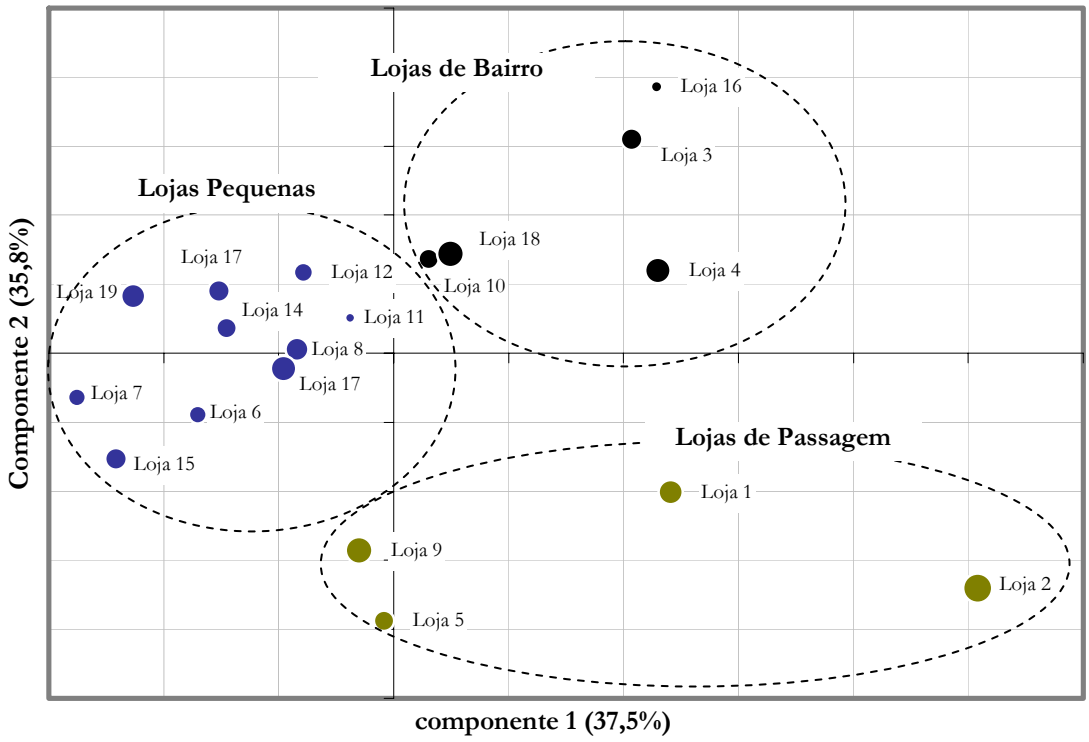

FIGURA 6 AGRUPAMENTOS DE WARD NO ESPAÇO DAS COMPONENTES PRINCIPAIS EXTRAÍDAS.
(A área do círculo é proporcional às vendas em 2002)

4.2. Integração do conhecimento de especialistas por validação *a posteriori*

Neste trabalho apresentam-se dois métodos para integração de conhecimento de especialistas por validação dos agrupamentos de lojas a *posteriori*:

- No primeiro - Método CLUST - utiliza-se um método de agrupamento baseado numa matriz de dissemelhança associada a um conjunto de variáveis seleccionadas pelos especialistas e cujo resultado é validado pelos mesmos especialistas.
- Na segunda abordagem – Método TREE - utiliza-se um método de aprendizagem supervisionada (árvore de regressão) usando como dependente uma variável métrica (traduzindo desempenho das lojas através das vendas) e como explicativas atributos associados às lojas.

Método CLUST: agrupamento sobre variáveis seleccionadas

O conhecimento de especialistas foi integrado na selecção das variáveis base para a tipificação das lojas e também na fase de apreciação de resultados de sucessivos agrupamentos hierárquicos (Cardoso e Mendes, 2002). A partir dos critérios de apreciação destes especialistas e de vários agrupamentos que foram sendo construídos concluiu-se pela importância de apenas 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 em termos de compras efectuadas;

O primeiro factor poderia ser traduzido pelas vendas realizadas 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.

A utilização de apenas duas variáveis pode levantar algumas questões uma vez que e se defendeu na secção 2 a necessidade de recolher um enorme volume de variáveis. Continuamos a advogar essa necessidade para a caracterização e definição dos *clusters*, numa perspectiva descritiva, mas não necessariamente na sua construção. Autores como Gnanadesikan (2001) defendem a utilização de um número mínimo de variáveis na construção dos agrupamentos e a utilização das restantes na sua interpretação e validação, já que afirmam ser a utilização de grande número de variáveis contraprudecente uma vez que pode mascarar a existência de grupos nos dados (ver Milligan, 1996, para uma análise mais completa).

Utilizando o método de Ward obteve-se uma tipificação que foi avaliada mediante a construção de inúmeros dendogramas com diversas combinações de métodos e medidas de distância sem se verificarem alterações de nota nos agrupamentos obtidos. Finalmente os resultados obtidos foram confirmados pelos especialistas de marketing ligados à cadeia de retalho em consideração, facto que concluiu o processo de validação.

Na Figura 9 apresentam-se os agrupamentos obtidos incluindo a denominação para os grupos de lojas baseada na caracterização destes. Assim têm-se:

- as **lojas de passagem** que se caracterizam por reduzidas percentagens de clientes inquiridos com deslocação exclusiva à loja, por elevadas percentagens de **clientes preferenciais** (segundo a segmentação efectuada em Cardoso e Mendes, 2002), valores médios de dimensão da loja, localizada em zonas de passagem ou centros comerciais movimentados, em áreas com elevado potencial e níveis de concorrência igualmente elevados;
- as **lojas grandes** que são as que mais vendem e são igualmente o grupo com maior dimensão, com menos de 50% dos clientes a deslocarem-se a pé e com o maior valor de compra média, de gasto na loja e de dimensão média do agregado familiar. No fundo são lojas que estão próximas de formatos maiores como as médias superfícies de comércio alimentar;
- as **lojas de bairro** que estão localizadas em zonas residenciais, aonde os clientes se deslocam maioritariamente a pé e muitas vezes fazem a viagem com a exclusiva intenção de fazer compras na loja. Estas lojas têm uma clientela maioritariamente de **clientes preferenciais**. Finalmente faz-se uma distinção entre as lojas de bairro maiores e menores que se baseia não só na dimensão da loja, como nas vendas, sendo as lojas maiores aquelas que apresentam maiores valores de vendas por m^2. Verifica-se igualmente que as lojas menores se situam em zonas mais densamente povoadas do que as maiores.

Por fim foram identificadas três lojas consideradas atípicas, já que duas delas não foram consideradas na análise de agrupamento por sugestão dos especialistas de marketing e a terceira foi identificada após a construção dos grupos. Esta última é, segundo opinião dos mesmos especialistas, provavelmente a semente para um novo agrupamento de lojas de rua caracterizadas por se localizarem em zonas com fluxos de passagem muito elevados.

Na Figura 9 apresentam-se vendas para os três anos disponíveis verificando-se que a variação dentro dos grupos tem apresentado alguma coerência. As maiores variações têm-se verificado nas lojas de bairro grandes e pequenas logo seguidas pelas lojas grandes. As lojas de passagem têm apresentado desempenhos piores apresentando mesmo variações negativas para o ano 2002. Os *outliers* apresentam desempenhos muito variáveis entre o mau (loja 16) e o muito bom (loja 2).

Método TREE: modelos de aprendizagem supervisionada

Nesta secção utiliza-se um método de aprendizagem supervisionada construindo uma árvore de regressão CART (Breiman *et al.,* 1984) capaz de simultaneamente prever o desenpenho das lojas (medida baseada nas vendas) e de constituir grupos (nós folha) aos quais se associam atributos específicos das lojas (ver por exemplo: Birn, 2002 e Chou *et al.,* 2000). Este modo de integração do conhecimento de especialistas pode classificar-se como um método *a posteriori* já que é utilizado na avaliação das árvores produzidas.

Na Figura 7 apresenta-se a melhor árvore segundo os critérios já descritos na secção 4.1. Note-se que foram efectuadas árvores utilizando diferentes variáveis dependentes. Nomeadamente as vendas nos diversos anos ou as vendas por área da loja uma vez que é um rácio muito comum na literatura (ver por exemplo Birkin *et al.,* 2002).

Foram rejeitadas árvores onde as variáveis não apresentavam o comportamento esperado face à variável dependente (por exemplo: se num nó folha uma variável que represente a dimensão da área de influência tiver valores superiores espera-se intuitivamente que o grupo de lojas que o constituem tenha um valor de vendas médio superior). Em caso de empates na selecção de uma variável responsável por ramificação a intervenção do especialista permitiu a selecção de variáveis produzindo resultados mais interpretáveis.

A partir dos histogramas apresentados na árvore da Figura 7 é possível identificar uma loja com vendas muito mais reduzidas do que as restantes agrupadas no primeiro nó folha. Retirar este ponto não altera a estrutura da árvore mas melhora muito as medidas de qualidade utilizadas na secção seguinte para comparar os diversos agrupamentos.

- Na árvore de regressão apresentada a primeira partição é efectuada segundo o valor de uma variável que pretende avaliar a dimensão da área de influência. Esta variável foi construída começando por ordenar as freguesias, para cada loja, segundo a ordem decrescente das percentagens de clientes que delas provinham. Depois somaram-se as áreas dessas freguesias até ao limite de 80% de inquiridos e calculou-se o rácio dessa área e a percentagem de clientes a que lhe correspondia. Obtém-se assim uma área de influência delimitada pelas freguesias. Esta variável permite separar as lojas com maiores vendas a que chamaremos **lojas grandes**.

- A segunda variável utilizada foi calculada dos dados do INE como o total de edifícios com um a dois pavimentos na área de influência calculada por caminhos mais curtos e usando a regra de decisão descrita na secção 2. Assim, esta variável mede o potencial da área de influência e permite separar as lojas que menos vendem a que chamaremos **lojas pequenas**.

- Por fim, a última variável de partição corresponde à percentagem de clientes preferenciais identificados em Cardoso e Mendes (2002) e permite separar as **lojas de passagem** (com baixas percentagens de clientes preferenciais) das **lojas de bairro**.

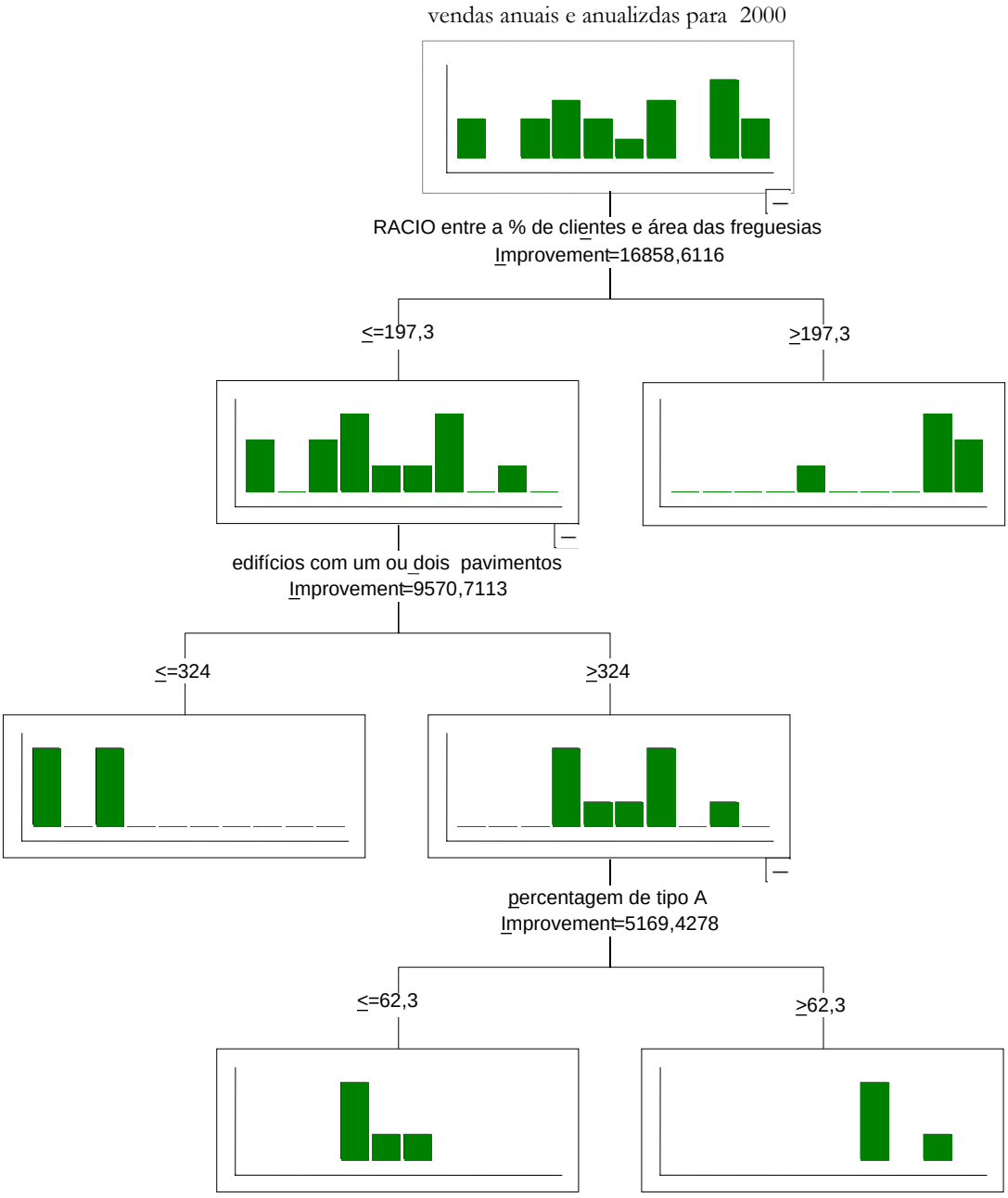

FIGURA 7 ÁRVORE DE CLASSIFICAÇÃO OBTIDA PELO MÉTODO CART.
(Os gráficos de barras representam os histogramas da variável dependente em cada nó)

4.3. Análise dos resultados e comparações entre as tipificações obtidas

No Quadro 1 resumem-se as principais características das diferentes metodologias utilizadas para tipificar as lojas de retalho alimentar de pequena dimensão. Do quadro é evidente a variedade de vias seguidas e de variáveis de base utilizadas no agrupamento das lojas e respectiva caracterização.

Para comparar as tipificações das lojas de retalho obtidas através do uso das diferentes metodologias de integração do conhecimento de especialistas podem observar-se os gráficos de extremos e quartis da Figura 8. Nestes gráficos pode avaliar-se o grau de coesão dos diferentes agrupamentos obtidos e identificar *outliers* (assim considerados em função da variável vendas anuais) no caso do método utilizado não os ter já identificado previamente.

QUADRO 1 VARIÁVEIS BASE DE AGRUPAMENTO E RESUMO DAS DIFERENTES METODOLOGIAS

Metodologia de integração de conhecimento de especialistas	***a priori***		***a posteriori***	
Metodologia de tipificação	*CLUST>MDL*	*MDS>CLUST*	*CLUST*	*TREE*
Variáveis base de agrupamento	Directamente da matriz de dissemelhanças.	10 Variáveis com maior correlação com as dimensões MDS.	Vendas anuais; Percentagem de clientes com origem e destino casa.	*Dependente: vendas em 2000* Rácio entre a percentagem de clientes das freguesias com maior proveniência e a área dessas freguesias; Número de edifícios com 1 e 2 pavimentos; Percentagem de clientes preferenciais.
Variáveis usada na caracterização	Densidade de edifícios construídos nos últimos 5 anos; Residentes com ensino secundário ou +; Área de vendas.	Área de vendas; Gasto mensal declarado na loja, Rácio entre a percentagem de clientes das freguesias com maior proveniência e a área e Total de mulheres residentes com mais de 65 anos.		
Agrupamentos obtidos	Lojas em Crescimento; Lojas Nobres; Lojas Maiores; Lojas Menores.	Lojas de Bairro; Lojas Pequenas; Lojas de Passagem.	Lojas Grandes; Lojas Passagem; Lojas de Bairro Grandes e Pequenas.	Lojas grandes; Lojas Pequenas; Lojas de Bairro; Lojas de Passagem.

Vários dos métodos não conseguiram identificar convenientemente os *outliers* presentes nos dados. É o caso da integração de conhecimento de especialistas *a priori* pelo método CLUST>MDL e da árvore de regressão construída onde a loja 5 constitui *outlier* para as vendas nos anos 2001 e 2002. Quanto à homogeneidade dentro dos agrupamentos formados parece ser evidente uma superioridade dos agrupamentos formados pelos métodos *a posteriori*.

Para comparação adicional das diferentes tipificações obtidas utiliza-se o valor da variância de vendas explicada pelos agrupamentos formados (variância inter-grupos dividida pela variância total) calculado excluindo as lojas atípicas identificadas nos gráficos da Figura 8. Teve-se o cuidado de excluir sempre o mesmo número de lojas em todos os anos para que os valores fossem comparáveis. No Quadro 2 apresentam-se os valores referentes às vendas anuais nos três anos disponíveis e a algumas outras variáveis igualmente relevantes na avaliação do desempenho das lojas.

Os métodos com integração formal de conhecimento de especialistas *a priori*, *i.e.* que utilizaram a matriz de dissemelhanças obtida por inquérito directo aos especialistas, apresentam resultados fracos para o caso particular em estudo. Tal resultado pode ser atribuído ao facto de, apesar de terem sido integradas as variáveis relacionadas com vendas no grupo de variáveis que poderiam ser escolhidas, os métodos utilizados para construir os grupos e discriminar as variáveis não as seleccionaram e logo o desempenho avaliado por utilização deste tipo de variáveis apresenta valores muito mais reduzidos. Esta observação sugere a utilização por parte dos especialistas de outras medidas de desempenho não explicitadas.

Assim, ou os especialistas não conseguiram traduzir nas comparações entre pares de lojas as verdadeiras avaliações de desempenho que se pretendiam ou os métodos utilizados para escolha de variáveis, para o caso particular em estudo e nomeadamente para o facto de apenas se disporem de muito poucas lojas, não foram os mais adequados. Pode-se ainda concluir pela maior facilidade em aceitar \ rejeitar os agrupamentos após a sua constituição do que a utilização de grande número de comparações binárias. Estas foram consideradas difíceis pelos especialistas por ser necessário manter uma visão global das restantes comparações (os especialistas consideraram ser frequentemente necessário rever classificações já atribuídas por comparação com novas classificações).

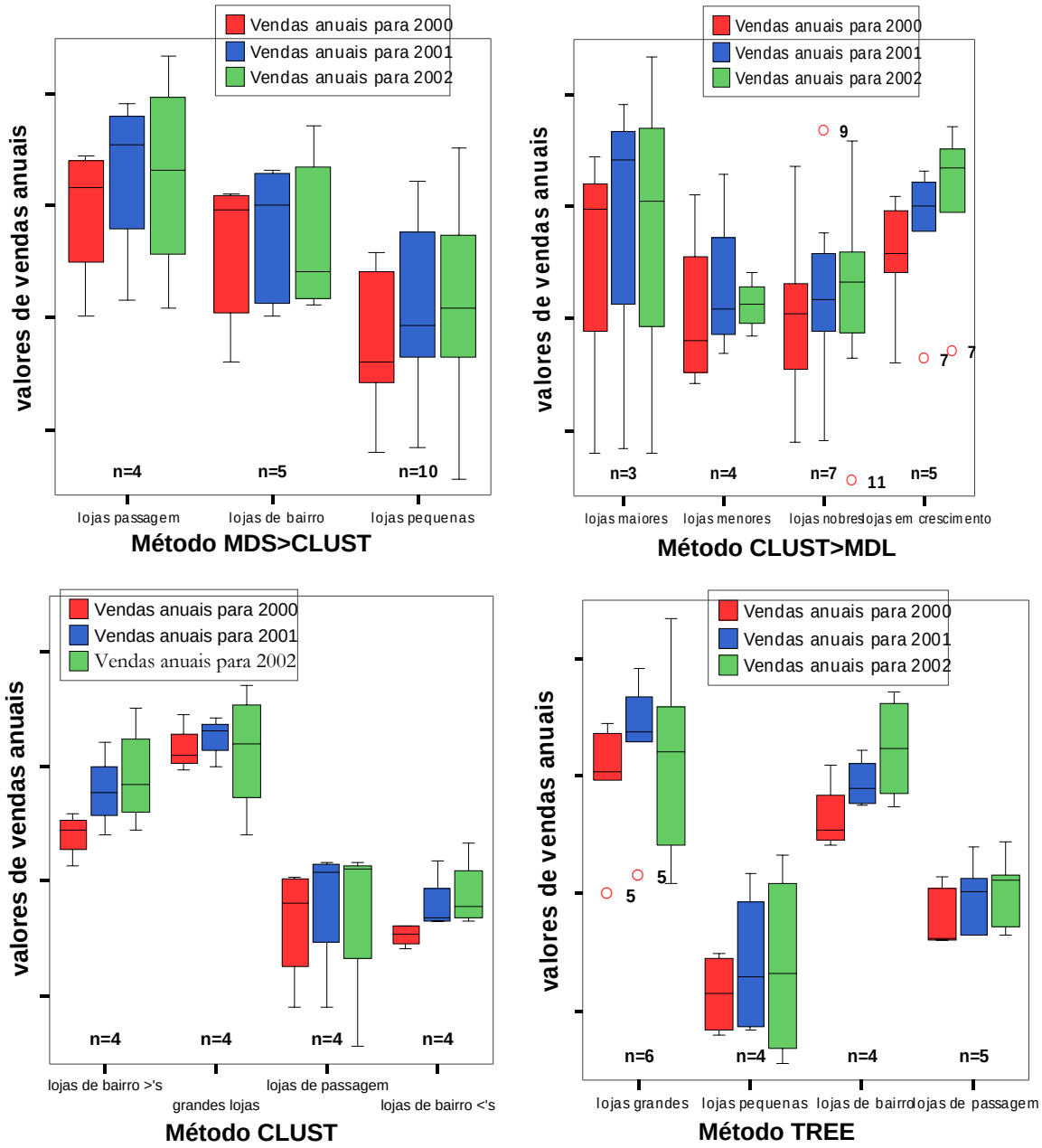

FIGURA 8 GRÁFICOS DE EXTREMOS E QUARTIS PARA AS VENDAS EM CADA UM DOS GRUPAMENTOS OBTIDOS PELAS VÁRIAS METODOLOGIAS

Note-se, no entanto, a relativa superioridade do método MDS>CLUST sobre o método CLUST>MDL. Poder-se-ia pensar que o método CLUST>MDL ao efectuar agregações directamente a partir da matriz de dissemelhanças conseguiria traduzir mais fielmente os sentimentos dos especialistas. Tal não se verificou talvez por ser difícil a avaliação do desempenho das lojas sem observação de dados, o que pode ser ultrapassado em parte pela escolha de variáveis antes de se efectuar a tipificação das lojas, como se procedeu para o método MDS>CLUST.

Do Quadro 2 verifica-se igualmente uma redução da percentagem de variância das vendas explicada pelo agrupamento à medida que mais dados de vendas vão sendo disponibilizados, o que não é de estranhar já que os agrupamentos foram construídos com vendas de 2000. Utilizaram-se os valores de vendas mais antigos com intenção de avaliar a degradação do desempenho ao longo do tempo. Apesar de essa degradação ser evidente os valores para 2002 são bastante bons, bem acima dos 50%, o que parece indicar que estes agrupamentos são, até certo ponto, resistentes à passagem do tempo. No entanto, as variações de vendas são muito pouco explicadas por estes agrupamentos o que indica que dentro de cada agrupamento as variações não são muito coerentes existindo variações bastante distintas. Note-se, no entanto, que se se tiver o cuidado de efectuar novas classificações sempre que se disponham de novos dados os resultado melhoram significativamente.

QUADRO 2 FRACÇÃO DA VARIÂNCIA EXPLICADA PELOS AGRUPAMENTOS PARA VÁRIAS VARIÁVEIS RELACIONADAS COM VENDAS E PARA OS VÁRIOS MÉTODOS EM COMPARAÇÃO

Metodologia de integração de conhecimento de especialistas	***a priori***		***a posteriori***		
Metodologia de tipificação	*CLUST>MDL*	*MDS>CLUST*	*CLUST*	*TREE*	
Exclusão de outliers	*2 lojas*	*0 lojas*	*3 lojas*	*0 lojas*	*1 loja*
vendas anuais para 2000	16,6%	51,2%	86,1%	81,0%	90,4%
2001	19,8%	47,8%	78,9%	74,8%	85,4%
2002	35,1%	35,3%	62,5%	62,2%	69,6%
vendas de 2002 por m²	27,8%	15,9%	37,8%	33,3%	33,3%
variação de vendas 02 - 01	6,9%	12,7%	10,3%	25,9%	25,9%
01 - 00	2,4%	9,6%	23,5%	5,3%	6,9%

Usando como indicador da qualidade das soluções a proporção de variância explicada pelos agrupamentos era previsível (e verifica-se por observação do Quadro 2) a superioridade dos métodos de aprendizagem supervisionada (TREE). Sublinha-se, contudo, a necessidade de uma validação *a posteriori* de qualquer árvore produzida por este método atendendo ao reduzido número de lojas. Note-se, a este propósito, que a exclusão de apenas uma loja corresponde a uma melhoria na percentagem de variância explicada pelos agrupamentos de aproximadamente 10% nos vários anos para os quais se dispõe de valores de vendas.

O método CLUST com integração de conhecimento de especialistas *a posteriori* obteve igualmente bons resultados quanto a variâncias explicadas (*vide* novamente o Quadro 2). Este método apresenta os melhores valores para as vendas por unidade de área e para a variação das vendas entre 2002 e 2001 ainda que sejam valores muito baixos.

Na verdade os especialistas de marketing envolvidos neste estudo preferem a tipificação segundo este método. Este facto pode revelar que a forma como foram envolvidos nesta tipificação lhes transmitiu mais confiança nos resultados e \ ou a medida de desempenho utilizada neste estudo não é a mais valorizada pelos especialistas em localização de lojas.

5. CONCLUSÕES:

Uma dificuldade que resulta directamente da complexidade de caracterização de lojas de retalho alimentar é o envolvimento de um grande número e variedade de atributos, facto que acarreta uma grande operação de recolha e tratamento de dados a qual, no caso presente, se prolonga há mais de dois anos. Em consequência desta complexidade e atendendo a que a cadeia de retalho alimentar em estudo dispõe apenas de algumas dezenas de lojas, colocam-se também dificuldades particulares na tipificação das lojas, que segundo se advoga neste trabalho, apenas poderão ser ultrapassadas com a integração de conhecimento de especialistas.

Neste artigo apresentam-se duas abordagens distintas para integrar conhecimento de especialistas na tipificação de um conjunto de lojas de retalho alimentar de dimensão pequena a média, pertencentes a uma cadeia nacional: integração *a priori* e *a posteriori*.

Nos casos de integração *a priori* os especialistas proporcionaram medidas de dissemelhança entre as lojas, tendo o agrupamento sido efectuado ou sobre essas mesmas medidas, ou sobre variáveis consideradas explicativas dessas dissemelhanças. Na integração *a posteriori* a intervenção dos especialistas ocorreu sobretudo na avaliação de diversos agrupamentos propostos e, por vezes, na identificação de *outliers* e na selecção de variáveis base de agrupamento.

As análises de agrupamento propostas realizaram-se ou pelo método hierárquico e não supervisionado de Ward ou por meio de uma árvore de regressão (abordagem supervisionada). Utilizando a variação do valor de vendas anuais das lojas intra-grupos como medida de erro, conclui-se que o melhor agrupamento foi obtido integrando a opinião dos especialistas *a posteriori*, apenas na identificação de lojas *outliers* e com base na observação de árvores de regressão alternativas. Mesmo para amostras de pequena dimensão, e desde que se

disponha de um número de variáveis explicativas suficiente, as árvores de regressão construídas demonstraram ser um método adequado para construção e caracterização de agrupamentos de lojas.

Neste caso particular, uma cuidadosa avaliação *a posteriori* dos resultados obtidos, realizada por especialistas, conduziu à selecção de uma tipificação constituída pelo método de integração *a posteriori* não supervisionado. Esta análise considera apenas duas variáveis: as vendas anuais na loja e a percentagem de clientes que se deslocam à loja propositadamente, *i.e.* aqueles que afirmam nos inquéritos provir de casa e voltar para casa após as compras, e em que se usa o método de Ward (ver Figura 9). Assim, e apesar dos bons resultados obtidos pelos modelos supervisionados os especialistas em localização de lojas preferem a tipificação obtida pelo método CLUST.

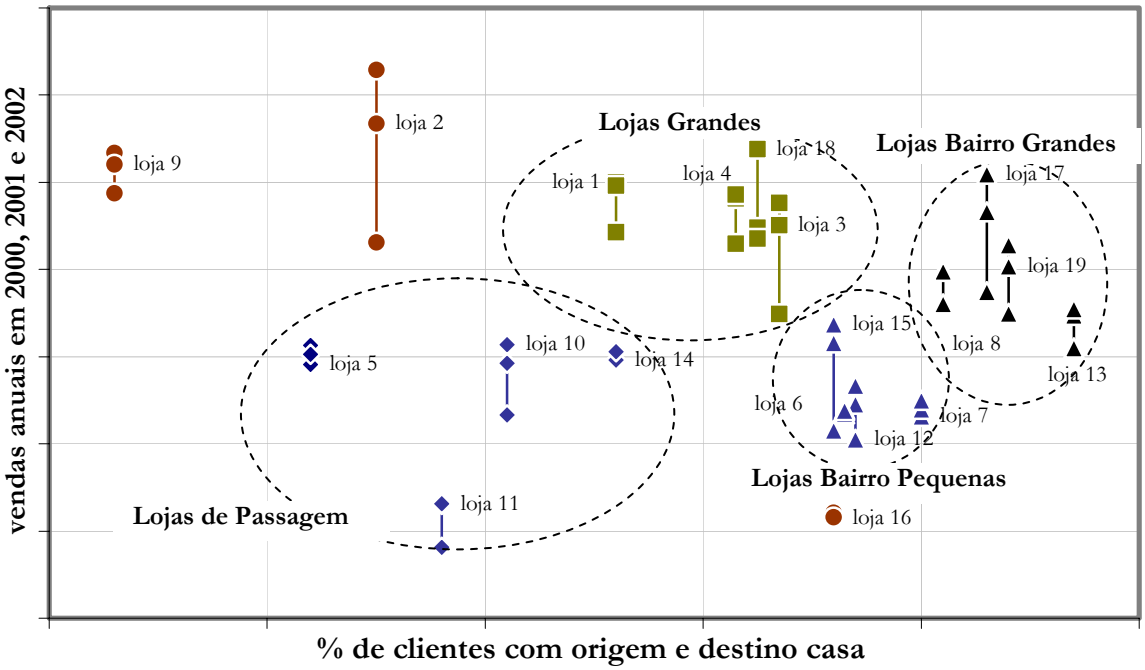

FIGURA 9 TIPIFICAÇÃO DAS LOJAS OBTIDA POR VALIDAÇÃO *A POSTERIORI* PELO MÉTODO CLUST.

Este facto pode ser devido à forma como foram envolvidos no método de tipificação e por compreenderem melhor os métodos utilizados o que lhes transmitiu mais confiança nos resultados. Na verdade esta tipificação foi já adoptada encontrando-se já a ser utilizada como base para acções de promoção diferenciadas (acções distintas em horários igualmente distintos em cada tipo de lojas definido).

Embora a intervenção dos especialistas possa observar-se em diversas fases da análise de tipificação, realçam-se aqui as seguintes:

- escolha das variáveis sobre as quais se recolhem dados e construção de variáveis compostas a partir de outras obtidas directamente;
- na selecção das variáveis base de agrupamento;
- na fase de selecção de agrupamentos alternativos e de validação dos agrupamentos formados;
- na possibilidade de seleccionar lojas atípicas ou *outliers* que podem prejudicar uma tipificação;
- na selecção de variáveis usadas na ramificação de uma árvore de decisão (explicativa de uma medida de desempenho das lojas) quando o algoritmo usado identifica casos de variáveis empatadas.

Em conclusão, advoga-se que é indispensável, especialmente no caso de amostras pequenas, a integração, ainda que apenas qualitativa, do conhecimento de especialistas de marketing na tipificação de lojas de retalho. Finalmente é de sublinhar a importância da interpretabilidade da solução promovendo a sua utilidade prática.

6. NOTA FINAL

Ainda que a presente tipificação das lojas de retalho se encontre validada pelos especialistas, na verdade, a avaliação que estes especialistas utilizam não se baseia apenas nas vendas das lojas mas num potencial de vendas a médio e longo prazo que terá de ser avaliado por outro tipo de medidas de desempenho.

Assim, uma evolução futura deste trabalho poderá consistir na utilização de métodos de teoria de decisão multiobjectivo para construir uma medida de desempenho mais adequada à avaliação das tipificações de lojas em prazos mais alargados. Ou, em alternativa, na construção de várias tipificações, usando árvores de regressão com diferentes variáveis dependentes, cujos resultados poderão se combinados numa tipificação de consenso (ver por exemplo Gordon, 1999). Estas variáveis dependentes deverão ser construídas numa tentativa de traduzir, de forma mais fiel, a medida de avaliação empírica utilizada, intuitivamente, pelos especialistas.

7. BIBLIOGRAFIA

BAY, S.D. e PAZZANI, M.J. - *Discovering and describing category differences: What makes a discovered difference insightful?* Em: *Proceedings*. Annual Meeting of the Cognitive Science Society 22ª ed., 2000. Texto completo em www.ics.uci.edu/~pazzani/Publications/Publications.html.

BIRKIN, Mark; CLARKE, Graham e CLARKE, Martin - *Retail Geography and Intelligent Network Planning*. Chischester, U.K.: John Wiley & Sons, 2002. ISBN: 0-471-49803-3.

BIRN, Robin J. (Ed.) - *The International Handbook of Market Research Techniques*. 2ª ed., London, U.K.: Kogan Page, 2002. ISBN: 0-749-43865-7.

BOOTS, Barry - *Using local statistics for boundary characterization*. Em: BOOTS, Barry; OKABE, Atsuyuki e THOMAS, Richard (Eds.) *Modelling Geographical Systems: Statistical and computational applications*. Dordrecht, Netherlands: Kluwer Academic Publishers, 2002. ISBN: 140200821X.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., STONE, C. - *Classification and Regression Trees*. California: Wadsworth, Inc., 1984. ISBN: 0-534-98053-8.

CARDOSO, Margarida G.M.S. - *Modelos discriminantes lógicos na caracterização de uma estrutura de segmentos*. Em: REIS, Elizabeth e HILL, Manuela Magalhães (Eds.) *Temas em Métodos Quantitativos*. Lisboa: Edições Sílabo, 2003. ISBN: 972-618-291-1. p. 181-192.

CARDOSO, Margarida G.M.S. e MENDES, Armando B. - *Segmentação de clientes de lojas de pequena dimensão*. Em: CARVALHO, Lucília; BRILHANTE, Fátima e ROSADO, Fernando (Eds.) *Novos Rumos em Estatística*. Actas do Congresso Anual da Sociedade Portuguesa de Estatística 9ª ed.. Ponta Delgada: Sociedade Portuguesa de Estatística, 2002. ISBN: 972-98619-4-3. p. 157-170. Resumo em www.uac.pt/~amendes.

CHOU, Paul B.; GROSSMAN, Edna; GUNOPULOS, Dimitrios e KAMESAM, Pasumarti - *Identifying prospective customers*. Em: *Proceedings ACM SIGKDD*. International Conference on Knowledge Discovery and Data Mining 6ª ed., New York, USA: ACM press, 2000. ISBN: 1-58113-233-6. p. 447-456.

CLARKE, Ian; MACKANESS, William; BALL, Barbara e HORITA, Masahide - *The devil is in the detail: Visualising analogical thought in retail location decision-making*. Em: *Recent Advances in Retailing & Services Science*. European Institute of Retailing & Services Science Conference 8ª ed. Vancouver : EIRASS, 2001.

COWEN, David J.; JENSEN, John R.; SHIRLEY, W. Lynn; ZHOU, Yingming e REMINGTON, Kevin - *Commercial real estate GIS site evaluation models: Interfaces to ArcView GIS*. Em: *Proceedings*. Annual ESRI International User Conference 20ª ed. ESRI online Library, 2000. p. 140-145. Texto completo em www.esri.com/library/userconf/proc00/professional/papers/

COX, Trevor F. e COX, Michael A.A. - *Multidimensional Scaling*. Chapman & Hall, 1994. ISBN: 1584-88094-5.

GNANADESIKAN, Ramanathan - *Cluster analysis: An overview of aims, aids, & challanges*. Em: NEVES, Manuela; COELHO, Carlos Agra e CADIMA, Jorge *Actas*. Congresso Anual da Sociedade Portuguesa de Estatística 8ª ed., Peniche, Portugal: Sociedade Portuguesa de Estatística, 2001. p. 39-57.

GONÇALVES, Alexandre B. e MENDES, Armando B. - *Caracterização de áreas de influência de lojas de retalho alimentar de pequena dimensão com base em diagramas de Voronoi ponderados*. Em: ESIG'2002. Encontro de

Utilizadores de Informação Geográfica 7ª ed.. Lisboa: USIG, 2002. p. 1-7 (publicado em CD-ROM). Texto completo em www.uac.pt/~amendes.

GORDON, A.D. - *Classification.* Monographs on Statistics and Applied Probability, 2ª ed., Boca Raton: CRC Press, 1999. ISBN: 1-58488-013-9.

JAIN, Anil K. e DUBES, Richard C. - *Algorithms for Clustering Data.* Advanced Reference Series: Computer Science, Englewood Cliffs, USA: Prentice Hall, 1988. ISBN: 013022278X.

LILIEN, Gary L. e RANGASWAMY, Arvind - *Marketing Engineering: Computer-assisted marketing analysis and planning.* 2ª ed., Prentice Hall, 2002. ISBN: 0-130-35549-6.

McGOLDRICK, P. - *Retail Marketing.* 2ª ed., Maidenhead: McGraw-Hill, 2000. ISBN: 0-07709-250-3.

McMULLIN, Shaun K. - *Where are your customers raster based modeling for customer prospecting.* Em: *Proceedings.* Annual ESRI International User Conference 20ª ed. ESRI online Library. 2000. p. 795-823. Texto completo em www.esri.com/library/userconf/proc00/professional/papers/

MENDES, Armando B. e THEMIDO, Isabel Hall - *Multi outlet retail site location assessment: A state of the art.* International Transactions in Operations Research. U.K.: Pergamon. 11 : 1, 2004. p. 1-18.

MILLIGAN, Glenn W. - *Clustering validation: Results and implications for applied analyses.* Em: ARABIE, P.; HUBERT, L.J. e DE SOETE, G. *Clustering and Classification*., Singapore: World Scientific, 1996. ISBN: 981-02-1287-9 p. 341-375.

MOLINERO, Cecilio Mar e CINCA, Carlos Serrano - *Finding treasure buried in data - Maps in management: multidimensional scaling.* OR Insight. U.K.: Operational Research Society. 13 : 1, Jan/Mar, 2000. p. 18-26.

NAERT, Philippe A. e LEEFLANG, Peter S.H. - *Building Implementable Marketing Models.* Boston: Kluwer Academic Press, 1978. ISBN: 90-207-674-8.

SALVANESCHI, Luigi - *Location, Location, Location: How to select the best site for your business.* Psi Successful Business Library, Psi Research - Oasis Press, 1996. ISBN: 1-55571-376-9.

THEMIDO, Isabel Hall; QUINTINO, António e LEITÃO, José - *Modelling the retail sales of gasoline in a Portuguese metropolitan area.* International Transactions in Operations Research. U.K.: Pergamon. 5 : 2, 1998. p. 89-102.

WARD, J.H., Jr. - *Hierarchical grouping to optimize an objective function.* Journal of the American Statistical Association. U.S.: ASA. 58, 1963. p. 236-244.

WEDEL, Michel e KAMAKURA, Wagner A. - *Market Segmentaion: Conceptual and methodological foundations.* International Series in Quantitative Marketing, 2ª ed., Massachusetts : Kluwer Academic Publishers, 2000. ISBN: 0-7923-8635-3.