# Análise de Agrupamento Incremental – Segmentação de Pontos de Retalho

**Neuza Brito de Jesus**
EDP Comercial
neuza.jesus@edp.pt

**Margarida G.M.S Cardoso**
Dep. Métodos Quantitativos, ISCTE
margarida.cardoso@iscte.pt

**Resumo**

O presente artigo apresenta um estudo sobre a utilização do algoritmo incremental Two-Step para identificar grupos homogéneos de pontos de venda de um universo de pontos de retalho para produtos alimentares congelados. Com este trabalho pretende-se efectuar a segmentação desse universo, de forma a suportar a tomada de decisão por parte dos gestores de marketing e vendas. Os grupos são identificados utilizando informação proveniente de um *data warehouse* que agrega dados sobre as características de cada ponto de retalho e as vendas que origina. Os resultados obtidos permitiram a identificação de 4 grupos, cujo perfil foi traçado e avaliado mediante o recurso a alguns testes de hipóteses.

**Palavras-chave**: Data Mining, Agrupamento, Segmentação, Retalho.

**Title:** Incremental clustering analysis: Segmenting retail points

**Abstract**

The present study concerns the utilization of the Two-Step incremental clustering procedure to identify homogenous clusters of retail points that support the distribution network of frozen food products. The work is aimed to segment the retail points' universe, in order to support the marketing and sales' decision making. The segmentation is based on information stored in a data warehouse that includes stores characteristics and sales performances of each retail point. The results obtained allowed the identification of 4 clusters which profile was identified and evaluated using hypothesis tests.

**Keywords:** Data Mining, Clustering, Segmentation, Retail

## 1-Introdução

O Retalho tem sofrido diversas alterações nos últimos anos. É uma área bastante dinâmica e um elemento importante no sucesso das marcas num ambiente de consumo de massas. O papel do ponto de retalho na cadeia de distribuição Fabricante → Distribuidor/Concessionário → Retalho → Consumidor, ganha cada vez mais relevância. Num modelo de distribuição como este, em que toda a estratégia de marketing é delineada pelo fabricante, o retalho torna-se fundamental para operacionalizar e optimizar a sua implementação.

Constata-se também que a relação entre os retalhistas e os fabricantes se encontra mais próxima e transparente. O retalho tem crescido, tendo surgido entidades que apresentam um poder negocial mais forte (exemplo: grandes cadeias de retalho, hipermercados).

A abordagem do fabricante/marca em relação ao retalho passa pela criação de um laço mais forte, pelo aumento da eficiência da cadeia de distribuição, pela redução de custos e pela melhoria da comunicação com o retalho e consequentemente com o consumidor.

Tal como os consumidores, os pontos de retalho são diferentes, diferenciando-se (por exemplo) no ramo de actividade, na localização, ou no volume de negócios. Estas diferenças devem ser consideradas na definição de estratégias de marketing e políticas comerciais.

A crescente necessidade de que os fabricantes/marcas conheçam o seu universo de retalho tem levado a grandes investimentos na construção de sistemas que permitam armazenar e analisar informação sobre o retalho. Actualmente, as grandes marcas e fabricantes possuem sistemas desta natureza, que são frequentemente utilizados no apoio à tomada de decisão.

## 2-Objectivos

Neste trabalho pretendeu-se segmentar o universo de pontos de retalho que suportam a distribuição de produtos alimentares congelados para uma conhecida marca.

A identificação desses segmentos foi efectuada com base em características do ponto de retalho, nas gamas comercializadas e no volume de vendas.

Os objectivos do trabalho resumem-se a:
- Identificar grupos homogéneos de pontos de retalho utilizando informação sobre as características dos pontos de venda, bem como as suas vendas;
- Analisar o agrupamento de acordo com as regiões geográficas e os concessionários responsáveis pela distribuição;
- Identificar se existe canibalização entre as várias gamas da marca nos pontos de retalho;

Esta análise foi efectuada a partir de informação armazenada numa base de dados com cerca de 65.000 registos. A informação armazenada é proveniente de dados transaccionais enviados pelos concessionários que abastecem os pontos de retalho.

## 3-Metodologia

### 3.1-Selecção de variáveis para estudo

A primeira aplicação de análise de agrupamento na segmentação de mercados surgiu com Wendell Smith, em 1956 (v. Haley (1968), por exemplo). A ideia é organizar um mercado heterogéneo que possui um universo vasto de necessidades, motivações, valores e outras características, em subgrupos homogéneos.

De acordo com Kotler and Armstrong (1996) (por exemplo), não existe um método único para segmentar um mercado. Em particular, podem ser utilizadas várias variáveis para efectuar a segmentação e identificar a estrutura de um mercado.

Neste trabalho propõe-se que selecção das variáveis base para a segmentação seja baseada no modelo de Tankred (1999). Esta autora estuda a utilização de canais de venda alternativos para o mercado de alimentos ultracongelados na Suécia. Seguindo o objectivo proposto, a autora construiu um modelo - Modelo Simbiótico (MS) - que identificou a interdependência e variações do mercado relacionados com factores estruturais característicos da Oferta e da Procura. A Figura 1 mostra o modelo MS e ilustra esta relação de interdependência entre a Oferta e a Procura, bem como os factores que a definem.

Da parte da Procura foram identificados factores que influenciam o consumidor e a sua atitude na compra. Estes factores foram divididos em dois grupos. Os Factores Externos que se encontram fora do controlo do consumidor e que podem ser manipulados pelas marcas. Factores Internos que se baseiam em aspectos intrínsecos do consumidor, tais como as motivações de compra, necessidades, atitudes, personalidade, perfil cultural e percepção.

No que diz respeito à Oferta, foram também considerados Factores Internos e Externos. Os Factores Externos representam aspectos do ambiente envolvente que estão fora da esfera de actuação da Oferta no mercado. Factores Internos são aqueles sobre os quais a Oferta pode exercer influência directa no sentido de tornar a relação mais favorável.

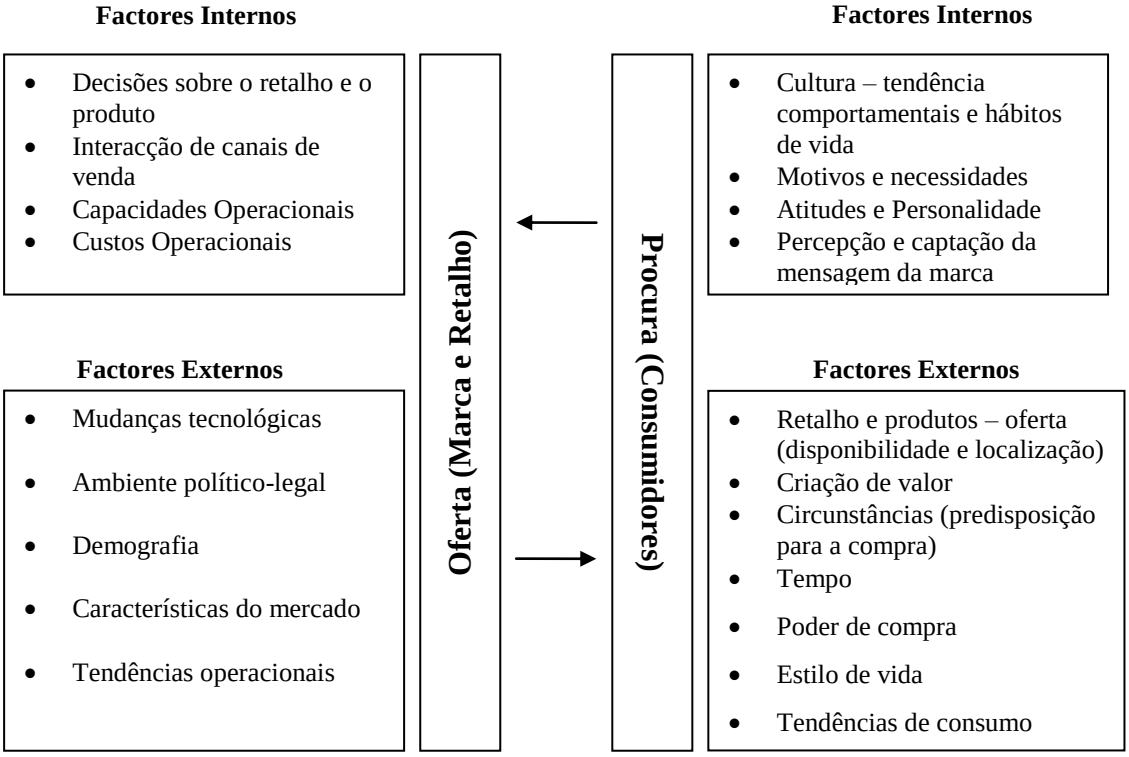


**Figura 1- Modelo Simbiótico**

Tendo em conta os objectivos do presente estudo adopta-se o referido MS nomeadamente no que se refere à investigação do modo como os factores ligados à Oferta poderão ajudar caracterizar o retalho. Pretende-se assim utilizar o MS como auxiliar na identificação das variáveis que irão servir de base para o trabalho de segmentação a efectuar.

### 3.2-Análise de Agrupamento

### 3.2.1-Vantagens do algoritmo Two-Step

Para a constituição dos segmentos pretende-se utilizar informação recolhida no data warehouse da empresa de produtos congelados, em estudo, organizada de acordo com as variáveis base identificadas tendo em conta os factores do Modelo MS relacionados com oferta - informação de aproximadamente 65.000 pontos de venda incluindo variáveis qualitativas e quantitativas.

Para efectuar a segmentação vai proceder-se a uma Análise de Agrupamento. Alguns dos procedimentos mais comuns em Análise de Agrupamento, como os algoritmos hierárquicos (algoritmo de Ward, por exemplo) e o K-Médias, mostraram-se inadequados para a presente aplicação devido às seguintes limitações:

– Algoritmos como o de Ward que têm um bom desempenho computacional e produzem bons resultados quando a base de dados é pequena, não são algoritmos adequados para bases de dados de maiores dimensões, como a que se trata aqui. Na verdade, os algoritmos mais utilizados são *in-memory*, isto é, exigem capacidade de memória da máquina para o seu processamento. Assim, quanto maior a base de dados, mais lento se torna o processo.

– Os algoritmos mais comuns são utilizados para variáveis base contínuas ou categorizadas e não mistas (incluindo variáveis contínuas e categorizadas). Embora existam medidas de distância para variáveis de naturezas mistas, baseadas na soma ponderada da distância das variáveis contínuas e distâncias de variáveis categorizadas, a escolha inadequada dos pesos facilmente se reflecte numa solução enviesada. De qualquer modo, não se trata de uma opção habitualmente implementada para os algoritmos referidos.

O algoritmo Two-Step, Chiu et al. (2001)[1], revela-se uma metodologia adequada para a presente aplicação de segmentação, pois possui características que permitem ultrapassar as limitações referidas para o algoritmo de Ward e o K-Médias. No caso do algoritmo de Ward, verifica-se a sua incapacidade de lidar com base de dados com a dimensão que se trata no presente trabalho.

No caso do K-Médias, estudos comparativos efectuados convergem no sentido de indicar o Two-Step com melhor desempenho computacional, Chiu et al. (2001). Por outro lado, o Two-Step permite lidar, directamente, com variáveis mistas (contínuas e categorizadas), o que não acontece com o K-Médias ou com o algoritmo de Ward.

O algoritmo Two-Step baseia-se no algoritmo de BIRCH- Balanced Iterative Reducing and Clustering using Hierarchies, Zhang et al. (1997) propondo um procedimento incremental executado em duas etapas que permite, com eficiência computacional, identificar grupos em base de dados de grande dimensão, integrando variáveis de natureza mista.

Para a sua descrição (sumária) usa-se a notação que se encontra na

Tabela **1**.

**Tabela 1 - Notação utilizada**

| |
|---|
| $Q^{cont}$ - número total de variáveis contínuas |
| $Q^{cat}$ - número total de variáveis categorizadas |
| N - número total de observações |
| d(k, k') – distância entre os grupos k e k' |

[1] . Encontra-se disponível no sotware SPSS versão 11.5 e seguintes

<k, k'> - índice que representa o grupo formado pela combinação dos grupos k e k'
$n_k$ - número de elementos do grupo k
$C_k$ – representação do grupo k

### 3.2.2-Medida de Distância

O algoritmo Two-Step utiliza uma medida de distância baseada na função de verosimilhança, que lida com variáveis contínuas e categorizadas. A medida de distância pressupõe o seguinte:

– As variáveis contínuas são modeladas por distribuições normais independentes intra-grupos com os parâmetros média e variância, a estimar.

– As variáveis categorizadas são modeladas por multinomiais independentes intra-grupos com probabilidades associadas às suas categorias a estimar.

A medida de distância é baseada no decréscimo da função de verosimilhança que resulta da união de dois grupos.

Sejam $C_k$ and $C_{k'}$ dois grupos que irão ser unidos resultando na constituição do grupo $C_{<k,k'>}$. Os dois grupos são substituídos pelo novo grupo $C_{<k,k'>}$ e uma nova estimativa do logaritmo da função de verosimilhança $l_{new}$ é obtida:

$$l_{new}^{\langle k,k'\rangle} = \sum_{s\neq k,k'}^{K} l(c_s) + l(c_{<k,k'>})$$

onde $l(C_s)$ é a contribuição do grupo $C_s$ para a função logaritmo da verosimilhança – l - dada por:

$$l = \sum_{s=1}^{K}\sum_{i\in I_s} \log f(\underline{y}_i \mid \underline{\theta}_k) = \sum_{s=1}^{K} l(C_s)$$

onde $I_s = \{i : \underline{y}_i \in C_s\}$ se refere aos índices das observações constantes no grupo $C_s$. Pode demonstrar-se que os $l(C_k)$ se decompõem em duas parcelas, isto é:

$$l(C_k) = l(C_k)^{met} + l(C_k)^{cat}$$

A primeira parcela refere-se às variáveis métricas e é baseada na variância[2]. A segunda parcela corresponde às variáveis categorizadas e é baseada na medida de entropia.

A distância log-verosimilhança define-se, então, em função do decréscimo $l$-$l_{lnew}$ na tentativa de minimizar a perda de informação (maximizar a função de verosimilhança). Considera-se, assim, que dois grupos são tão mais distantes quanto maior for o decréscimo da log-verosimilhança que resulta da sua agregação, i.e.

$$d(k,k') = l - l_{new}^{\langle k,k'\rangle}$$

O algoritmo Two-Step tem dois passos. No primeiro passo, os dados são analisados, um por um, e grupos homogéneos são formados tendo em conta a medida de distância. A cada grupo é associado um conjunto de estatísticas que sumarizam a informação sobre o grupo formado.

[2] Na verdade, ao operacionalizar esta medida, adiciona-se à estimativa da variância condicional a estimativa da variância incondicional, procurando evitar a ocorrência de problemas de cáculo do logaritmo no caso da variância condicional (intra-grupo) ser nula.

No segundo passo, os grupos obtidos na etapa um são tratados como observações individuais e um algortimo hierárquico é utilizado para criar o agrupamento.

### 3.2.3-Primeira etapa: sumarização

O objectivo da primeira etapa do algoritmo é reduzir a dimensão do problema através da identificação de regiões densas (nuvens de observações que estatisticamente se encontram próximas), mediante a construção de uma árvore de objectos simbólicos.

A etapa de sumarização usa uma aproximação de agrupamento incremental. Explora as observações uma a uma e decide se a observação deve ser integrada nos grupos formados previamente, ou se, por outro lado, se deve começar um novo grupo, tendo como base um critério que limita a distância máxima entre duas observações do mesmo grupo.

Cada observação é processada uma única vez sendo encaminhada para o nó folha e correspondente subgrupo mais próximo, que se designa por *Cluster Feature Entry* (CFE). Cada subconjunto $C_k$ é identificado por um conjunto de estatísticas que sumariza informação sobre ele, a que se dá o nome de *Cluster Feature* (CF). A árvore construída chama-se *Cluster Feature-Tree (CF-Tree).*

Cada *CF* tem informação sobre o número de elementos do subgrupo, a média e variância para cada variável contínua e a frequência associada a cada categoria de cada variável categorizada. Assim, cada nó folha da *CF-Tree* é caracterizado pelo $CF_k$ em que se inclui informação sobre:

- o número de observações no grupo $C_k$;
- a soma dos valores de cada atributo contínuo associado às observações do grupo $C_k$;
- a soma dos quadrados dos valores de cada atributo contínuo para as observações do grupo $C_k$;
- um vector onde se reúnem os números de observações do grupo $C_k$ por cada categoria de cada variável categorizada.

A constituição de *CF* é um factor determinante na eficiência computacional alcançada pelo algoritmo, pois armazena menos informação que se mostra ser a necessária para o cálculo de todas as medidas para efectuar o agrupamento.

Quando dois grupos são reunidos a *CF* do novo grupo corresponde à "soma" das duas *Cluster Features* dos dois grupos iniciais.

Da análise de pré-agrupamento que é realizada na primeira etapa resulta, assim, uma *CF-Tree*. A construção desta árvore encontra-se dependente de três parâmetros: *B - Branching Factor* que limita o número de nós descendentes numa ramificação; *T – Threshold Value* que limita a distância de fusão de dois subgrupos; *D - Maximum Tree Depth* que é o número máximo de níveis que a árvore pode ter.

Deste modo cada nó tem, no máximo, B entradas da forma $[CF_k, Child_k]$ onde $k=1, \ldots, B$. $Child_i$ representa o sub-grupo que é uma das entradas para o nó-folha e $CF_k$ representa o *CF* de cada folha. Em cada nó deverá verificar-se que todos os registos estão a uma distância máxima T uns dos outros. Caso contrário, se o número de nós descendentes de uma ramificação for inferior ao limite B inicia-se um novo subgrupo. Caso seja excedido o número de ramificações B, o nó-folha é dividido em dois, as *Cluster Features Entries* são

redistribuídas pelos dois novos nós. A partição inicializa-se pelas duas mais afastadas, sendo as restantes afectadas segundo a proximidade, aos novos nós.

Se a árvore crescer para além do tamanho máximo permitido, ela será reconstruída utilizando um valor maior para T.

O valor de T condiciona o número de regiões densas. Por um lado, valores elevados de T originam menos sub-grupos, potencialmente com mais heterogeneidade intra-grupos, mas computacionalmente o processo torna-se mais eficiente. Por outro lado, um valor pequeno de T pode originar um número elevado de sub-grupos, mais homogéneos, mas simultaneamente pode tornar o processo de agrupamento mais lento. Assim sendo, a escolha do valor T deverá ser efectuada de forma cautelosa, de forma a não diminuir a eficiência computacional e por outro lado, não produzir resultados desajustados.

### 3.2.4-Segunda Etapa: Agrupamento

Após a construção da árvore, os sub-grupos identificados constituem as novas entidades a agrupar. Estas novas entidades são, naturalmente, em número menor do que o número de observações na base de dados. As estatísticas entretanto armazenadas nos *Cluster Features* são suficientes para calcular a distância e podem ser aplicadas posteriormente.

### 3.2.5-Outliers

O algoritmo possui um procedimento que permite detectar a existência de *outliers* e o tratamento dos mesmos.

Na primeira etapa, as observações que não são integradas em nenhum sub-grupo e os sub-grupos que têm menos observações do que uma determinada proporção do número de observações do nó com o maior número de registos (normalmente 25%), são identificados como *outliers*, constituindo um grupo separado.

Quando a árvore é reconstruída (aumentando o valor de T) é verificado se os registos identificados como *outliers* assim se mantêm, ou seja, se estão a uma distância superior ao valor limite T dos sub-grupos; caso contrário, são integrados nessas regiões e deixam de ser visto como *outliers*.

Este processo prossegue iterativamente em cada reconstrução da árvore. Quando todas as observações são analisadas e integradas no grupo, chega-se ao final da primeira etapa e obtém-se o grupo de *outliers*, resistentes ao incremento de T.

### 3.2.6-Determinação do número de grupos

O número de grupos resultante de um processo de agrupamento pode ser indicado *a priori*, mediante critérios indicados por especialistas do domínio de aplicação. Em alternativa, procura-se que a estrutura de segmentos se ajuste automaticamente aos dados.

No Two-Step, os critérios mais comummente utilizados na determinação do número de segmentos são o BIC -*Bayesian Information Criterion,* Schwarz (1978), e AIC-*Aikaike's Information Criterion,* Akaike (1973).

Na primeira etapa do procedimento Two-Step, a cada união de dois sub-grupos, o valor de um critério de informação é calculado e uma primeira estimativa do número de grupos é obtida que corresponde ao abrandamento no decréscimo do valor do BIC e/ou AIC. Segundo os

autores do Two-Step, Chiu et al. (2001), obtém-se, assim, uma boa aproximação do número máximo de grupos.

Na segunda etapa do procedimento, procura refinar-se esta estimativa inicial do número de grupos. Para tal é utilizado como critério de determinação do número de grupos o rácio que mede a alteração na distância entre os dois grupos mais próximos, como resultado da fusão de dois grupos pelo Agrupamento Hierárquico. Escolhe-se então o número de grupos para o valor no qual o rácio da distância atinge o seu máximo.

## 4-Análise de dados

### 4.1-Variáveis base de segmentação

No presente artigo trata-se, como já referido, da segmentação de um universo de pontos de retalho que suportam a distribuição de produtos alimentares congelados. Com o objectivo de identificar e traçar o perfil de grupos homogéneos de pontos de venda utiliza-se a informação sobre as características (por exemplo: região geográfica, localização, concessionário responsável pelo abastecimento, volume de vendas, entre outras) de aproximadamente 65.000 dos pontos de venda que constituem o universo dos pontos de retalho.

Tendo adoptado para a selecção das variáveis base de segmentação na base de dados composta por informação do retalho o Modelo Simbiótico (MS), incluíram-se, para análise, as variáveis base de segmentação que se apresentam na

**Figura 2**.


**Figura 2 – Variáveis base de segmentação**

Como a questão de segmentação se coloca do ponto de vista da Oferta foram identificadas, mediante recurso ao modelo MS, variáveis que caracterizam os factores internos e externos e que irão servir de base ao agrupamento. A Figura 2 relaciona cada factor com o conjunto de variáveis base identificadas. Estas variáveis são de natureza contínua e categorizada.

As variáveis categorizadas incluem: Canal de Vendas, Localização, Ramo de Actividade, Grupo Económico, Região Geográfica, Distribuidor/Concessionário, Política de Merchandising . As variáveis contínuas são: valor e volume global de vendas, vendas em valor por categoria (foram consideradas 6 categorias de produto às quais foi atribuída a classificação A,B, C,D,E, F e G), descontos concedidos e número de encomendas anuais

**4.2-Principais resultados do algoritmo Two-Step**

Antes de proceder ao agrupamento com o Two-Step foi indicada uma parametrização para este algoritmo. Os parâmetros iniciais definidos para a construção da *CF-Tree* foram: T= 0; B= 8; D= 3. Para o caso em estudo, a definição dos parâmetros iniciais seguiu as recomendações dadas pelos autores (v. Chiu et al. (2001) e Zhang et al. (1997)) para a utilização do algoritmo.

Para o parâmetro T (Threshold Value), a escolha do valor inicial é efectuada de forma a minimizar o número de vezes que a CF-Tree é reconstruída. O valor inicial é incrementado iterativamente de acordo com uma heurística definida pelos autores do BIRCH, Zhang et al. (1997), pelo que é adequado fixar o valor inicial de T de uma forma conservadora. Os autores atribuem zero ao valor inicial de T, pois este permite obter melhores resultados na execução computacional (tempo de execução e memória necessária).

Os parâmetros B (Branching factor) e D (Maximum Depth) são fixos utilizando as recomendações do SPSS para a utilização do algoritmo. A utilização dos valores B=8 e D=3 é recomendada tendo em conta o número de pré-clusters que esta paramentrização permite $8^3$= 512. Vários ensaios realizados sobre múltiplas bases de dados permitiram concluir que esta parametrização conduz a pré-clusters em número suficiente para produzir bons resultados e, em simultâneo, não compromete o tempo da etapa de agrupamento (para mais detalhes ver referência documentação de SPSS (2001)).

De acordo com o procedimento Two-Step, assim parametrizado, foram constituídos 4 grupos e um grupo adicional de *outliers*. Os grupos foram identificados como:

- *Grupo 1* – Pontos de retalho com baixo poder económico e que não se encontram associados a grandes cadeias de retalho. Situam-se nas regiões do interior e centro de Portugal;
- *Grupo 2* - Pontos de retalho situados nas zonas litorais Norte e Sul de Portugal e que se encontram associados a grandes cadeias de retalho como a GCT-Gestão de Comércio Total;
- *Grupo 3* – Pontos de Retalho com o melhor desempenho, maior poder económico e localizações priveligiadas para o consumo das gamas de produto em estudo;
- *Grupo 4* – Pontos de Retalho localizados em áreas urbanas;

Os números de observações pertencentes a cada grupo encontram-se na

.

**Tabela 2 – Distribuição dos grupos de pontos de venda**

| | | Nº de pontos | % de pontos |
|---|---|---|---|
| Grupos | 1 | 12128 | 18.7% |
| | 2 | 14158 | 21.8% |

| | | | |
|---|---|---|---|
| | 3 | 14511 | 22.3% |
| | 4 | 23743 | 36.5% |
| | *Outliers* | 426 | .7% |
| | Total | 64966 | 100.0% |

O número de grupos foi determinado do modo seguinte: numa primeira etapa foi utilizado o critério Bayesian Information Criterion (BIC), tendo-se verificado que o decréscimo no BIC atenua com 4 grupos; na segunda etapa, verificou-se que a alteração do rácio da distância tinha um valor máximo para um agrupamento com 4 grupos, tendo sido esta a solução seleccionada.

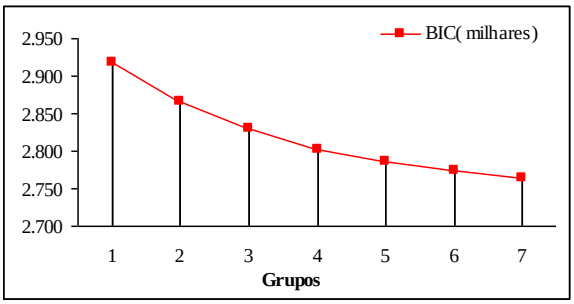


**Figura 3 - Valores de BIC**

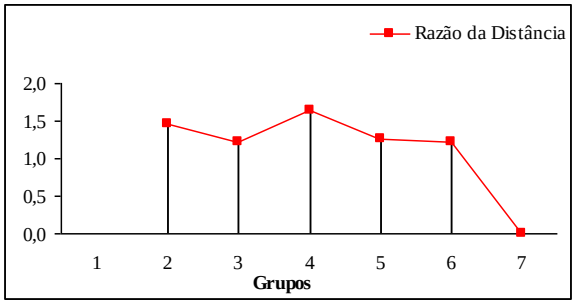


**Figura 4 – Razão de distâncias**

Esta solução encontrada foi, ainda, sujeita a uma análise de sensibilidade. Ensaiou-se então uma solução com 3 grupos, tendo-se concluído que esta não era suficientemente ilustrativa do universo de pontos de retalho, não isolando os pontos de retalho identificados no Grupo 3, os pontos de venda com melhor desempenho.

Em alternativa ensaiou-se, também, a constituição de 5 grupos. Neste caso existiriam pelo menos dois grupos com perfis semelhantes, comprometendo a heterogeneidade entre-grupos.

**4.3-Descrição e identificação do perfil dos grupos**

Os grupos que resultaram do Two-Step foram analisados para determinar o seu perfil. Foi assim analisado o comportamento de cada variável base de segmentação, em cada grupo. Procurou-se, ainda, identificar a correlação entre as gamas de produtos em cada grupo, recorrendo-se para tal à Análise em Componentes Principais.

Com o objectivo de determinar se o algoritmo de agrupamento efectivamente produzira grupos homogéneos e heterogéneos entre si, nomeadamente no que se refere às variáveis base de segmentação, foram conduzidos testes de hipóteses.

Na análise das variáveis contínuas os pré-requesitos para a utilização do teste *ANOVA* não foram verificados (homocedasticidade em particular), tendo sido utilizado em alternativa, o teste não paramétrico Kruskall-Wallis. Os testes foram efectuados considerando um nível de significância de 5%. Foram observadas probabilidades de significância (p-values) para cada variável muito inferiores a 5%, o que levou à rejeição da hipótese nula formulada: valores iguais para a mediana entre os grupos. Concluiu-se, portanto, pela existência de diferenças significativas entre os grupos para todas as variáveis base de segmentação contínuas.

A análise das variáveis categorizadas foi feita com o Teste do Qui-Quadrado de independência, considerando a hipótese nula de não existir associação entre as variáveis em estudo e o agrupamento. Para todas as variáveis se obtiveram valores elevados da estatística do qui-quadrado, indicadores da rejeição da hipótese nula e de associação entre o agrupamento e as variáveis categorizadas. No entanto, na aplicação deste teste, o pré-requisito de menos de 5% de células com frequências nulas não foi verificado. Optou-se então por cingir a conclusão de existência de associação aos dados analisados (em alternativa a proceder ao agrupamento de algumas categorias das variáveis em análise, viabilizando a inferência mas comprometendo a interpretabilidade).

Na sequência da análise efectuada segue-se uma descrição mais detalhada de cada grupo:

- O grupo 1 reúne pontos de venda de pequena dimensão. Este grupo representa 18,7% do total do universo estudado. Geograficamente os pontos de venda deste grupo situam-se, na sua maioria, na região centro do país. Têm, ainda, um baixo poder económico e negocial com os seus fornecedores, não sendo alvos preferenciais dos investimentos feitos pelos fabricantes no retalho. Têm localizações diversificadas e não se encontram associados a grandes grupos económicos. Estes são pontos de retalho que efectuam abastecimentos poucos frequentes e que se centram na principal gama de produtos comercializados pela marca fabricante.
- O grupo 2 reúne 21,8% dos pontos de retalho, que são na sua maioria mini-mercados. Estes pontos de venda associam-se, frequentemente, a grandes cadeias grossistas, que por terem maior poder negocial conseguem melhores preços junto dos fabricantes. Geograficamente, estes pontos de venda situam-se em zonas litorais e fora dos grandes centros urbanos. Estão localizados junto a parques de campismo, hóteis, estações de autocarro e universidades.
- O grupo 3 contém 22,3% dos pontos de retalho, reunindo os que têm maiores vendas em valor e volum. Neste grupo concentram-se pontos de retalho com poder económico e negocial. Localizam-se em zonas litorais com grande afluência de consumidores. Os ramos de actividade que prevalecem são: bares e restaurantes de praia, pastelarias, lojas de gelados, lojas de conveniência e estações de serviço, localizando-se junto a praias, piscinas, parques de entretenimento e cinemas. Uma vez que estes pontos de retalho têm um bom desempenho e elevado potencial, a marca realiza neles muitos investimentos.
- No grupo 4 encontram-se 36% dos pontos de retalho, que se situam em áreas urbanas densamente povoadas. Estes pontos de venda dedicam-se ao retalho especializado como: lojas de café, lojas de doces, cantinas, restaurantes, tabacaria e quiosques. As localizações mais frequentes são: zoo, centros comerciais, escolas, universidades, estações de comboio, cinemas. Estes pontos de venda têm maiores valores de consumo na gama de restauração.

O grupo de observações identificadas como *outliers* foi de igual modo estudado. Trata-se de um pequeno grupo que reúne 0.7% do total de pontos de retalho, sendo responsável por 9% de vendas ao retalho. Estes pontos de venda têm elevados valores de vendas dos produtos, mais elevados do que no grupo 3, efectuando abastecimentos com frequência. No entanto, não é possível traçar um perfil deste grupo precisamente porque é constituído por observações *outliers*. É, no entanto, possível destacar que estes pontos de venda estão associados a grandes grupos económicos tais como hotéis, cadeias de retalho, cadeias de restauração. Por outro lado, a variável Canal de Venda é determinante na identificação destes clientes outliers, pois tratam-se de pontos de venda que são geridos pelo canal que trata de clientes especiais que possuem características distintas do retalho em geral: dimensão, associação a grupos económicos, ramo de actividade e localização.

No sentido de complementar o traçado do perfil dos grupos analisaram-se as correlações existentes entre os valores de vendas das gamas de produtos – A, B, C, D, E, F e G. Assim, em cada grupo constituído foi realizada uma Análise em Componentes Principais, seguida de rotação VARIMAX. Os principais resultados das análises ilustram-se na Figura 5.

A gama A é a mais consumida nos 4 grupos, variando de um grupo para outro a proporção de consumo face às restantes gamas.

Foi observado que existem fortes correlações entre as gamas e que se mantêm semelhantes nos vários grupos. As gamas A e B estão correlacionadas, bem como as gamas C e D.

No que diz respeito ao grupo de *outliers*, verifica-se que a principal categoria consumida nos grupos perde peso para as restantes, verificando-se ainda que existe uma correlação negativa entre a gama C e as gamas A e B.

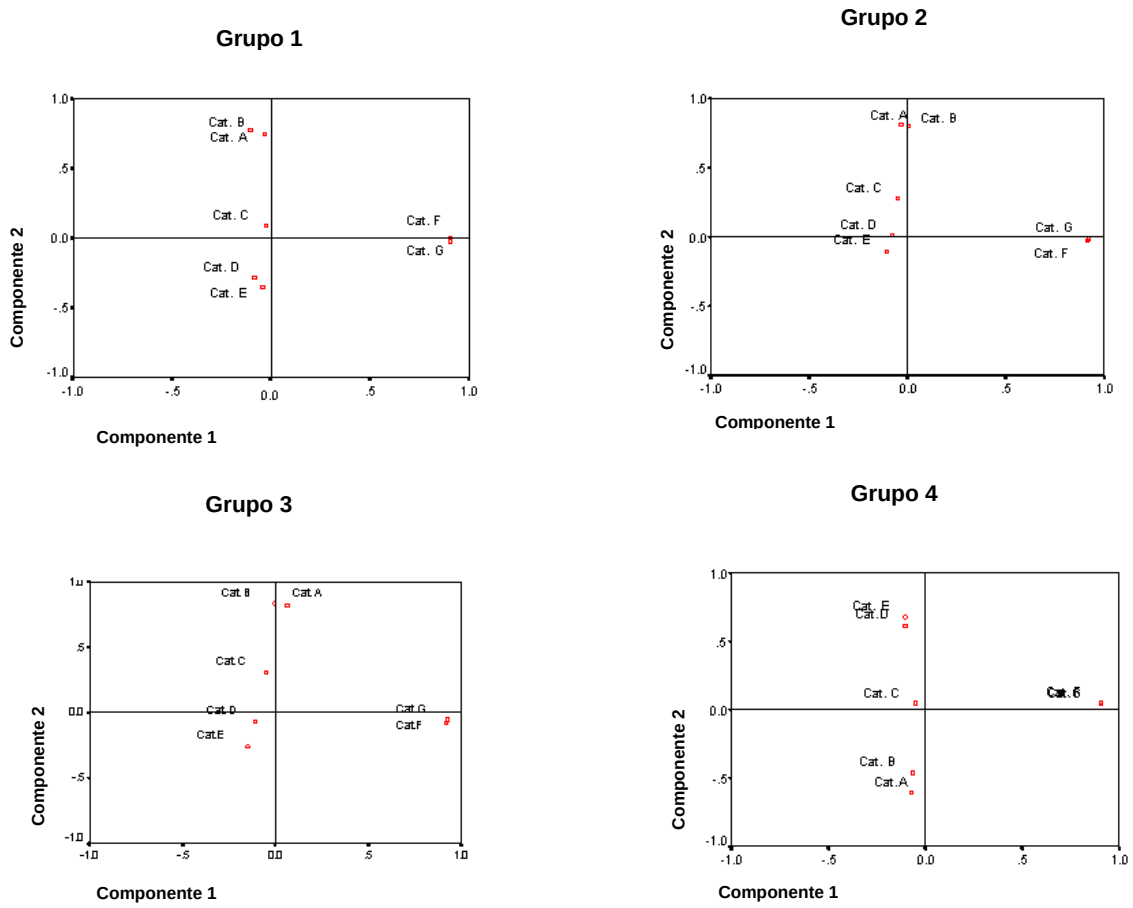


**Figura 5 – Correlações entre as vendas das gamas de produtos: resultados de Análises em Componentes Principais nos 4 grupos**

## 5-Conclusões

O estudo desenvolvido procurou responder a um desafio lançado por uma marca fabricante de produtos alimentares que pretendia conhecer, de forma mais aprofundada, o universo de pontos de retalho que dão suporte à sua cadeia de distribuição.

Neste trabalho pretendeu-se desenvolver uma metodologia que permitisse analisar a informação que a marca possui sobre o retalho e que se encontra armazenada num data warehouse da empresa.
Com o intuito de efectuar a segmentação do universo dos cerca de 65000 pontos de retalho, recorreu-se ao algoritmo de agrupamento Two-Step. A utilização deste algoritmo veio possibilitar a análise de uma base de dados de grande dimensão, com variáveis de natureza mista. Para tal contribui o facto de se tratar um procedimento incremental divido em duas etapas (uma primeira etapa de sumarização dos dados e uma segunda de agrupamento) e de utilizar um medida de distância (distância de verosimilhança) que lida com variáveis contínuas e categorizadas. A utilização do Two-Step permitiu que a determinação do número de grupos fosse baseada em critérios de informação adequados para o efeito.

A utilização do Two-Step permitiu, ainda, a detecção de *outliers*, excluindo-os da análise através da criação de um grupo separado para análise posterior.

A aplicação do algoritmo conduziu à identificação de 4 grupos, cujo perfil e principais diferenças existentes foram identificadas.

Os resultados deste agrupamento permitem ao fabricante/marca ter um maior e melhor conhecimento do universo de pontos de retalho através dos quais faz os seus produtos chegar ao consumidor. Este maior conhecimento traduz-se numa vantagem competitiva para o crescimento da marca, pois permite-lhe ajustar esforços e investimentos aos vários perfis de pontos de retalho identificados.

**Bibliografia**

(2001). Two-Step Cluster Component: A scalable component enabling more efficient customer segmentation. SPSS White Paper.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autorregressive moving average models. *Biometrika*, 60(2) 255-265.

Chiu, T., D. Fang, et al. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Haley, R. I. (1968). Benefit Segmentation - A Decision-Oriented Research Tool. *Journal of Marketing*, 32 226-233.

Kotler, P. and G. Armstrong (1996). *Principles of Marketing*. London, Prentice-Hall.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) 461-464.

Tankred, S. (1999). Evaluation of Alternative Retail Channels in Food Industry: A Consumer Approach. School of Economics and Management, Lund University, Sweden.

Zhang, T., R. Ramakrishnan, et al. (1997). BIRCH-A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2).