

# **Gestão da Cadeia de Suprimentos: proposição de uma metodologia para a associação de produtos a pontos de venda no canal indireto de distribuição**

*“Artigo a ser submetido ao periódico Gestão e Produção”*

Giordano Luiz Reis de Almeida (UFRGS) – *almeida.giordano@gmail.com*

Michel José Anzanello (UFRGS) – *anzanello@producao.ufrgs.br*

## **Resumo**

A gestão da distribuição de produtos no canal indireto de vendas é fundamental para as empresas que trabalham nesse contexto, demandando métodos estruturados para auxiliar na definição do *mix* de produtos a ser ofertado a cada ponto de venda (cliente). Nesse cenário, este estudo propõe, com base em variáveis quantitativas de venda, uma abordagem para selecionar os produtos mais promissores em termos de comercialização para cada grupo de clientes. Para tanto, realiza-se o agrupamento de produtos e pontos de venda com características similares e a identificação dos graus de relação entre eles. O método apresentado se baseia em ferramentas de Análise Multivariada de Dados (MVA), utilizando técnicas de Análise dos Componentes Principais (PCA) para a redução do número de variáveis e de clusterização para a segmentação dos produtos e clientes. Quando aplicado no canal indireto de distribuição de uma empresa multinacional do setor de alimentos e *commodities* agrícolas, o método proposto agrupou de forma satisfatória produtos e clientes com comportamentos similares e gerou associações entre os grupos que foram validadas pelo histórico de dados da empresa.

**Palavras-chave:** Supply Chain Management (SCM), canal indireto de distribuição, *mix* de produtos, PDVs (pontos de venda), Análise Multivariada de Dados (MVA), Análise dos Componentes Principais (PCA), *clusters*.

## **Abstract**

Products distribution management in the indirect sales channel is central for companies that work in this context. Thus, several companies require structured methods aimed at selecting the proper product mix to be offered to each point of sales (client). Hence, this study proposes a framework to select the most promising products in terms of commercialization to each group of clients. For that matter, products and points of sales with similar characteristics are grouped and the degrees

of relationship between them are identified. The proposed method relies on Multivariate Analysis (MVA) techniques, using Principal Component Analysis (PCA) to reduce the number of variables and clustering techniques to segment products and clients. When applied to the indirect distribution channel of a multinational company from the food and agricultural commodities sector, the proposed method satisfactory grouped products and clients with similar behaviors and generated associations between groups that were endorsed by previous trading.

**Keywords:** Supply Chain Management (SCM), indirect distribution channel, product mix, POS (point of sale), Multivariate Analysis (MVA), Principal Component Analysis (PCA), clusters.

## 1. Introdução

A busca das empresas por eficiência nos processos de distribuição e disponibilização de produtos aos consumidores realça a importância de tais aspectos para a obtenção de sucesso frente à concorrência. Nesse cenário, a gestão da cadeia de suprimentos tem apresentado crescente relevância, pois, a partir do gerenciamento das relações entre fornecedores e clientes, se objetiva proporcionar maior valor ao consumidor com base em menor custo para a rede de suprimentos (CRISTOPHER, 2011).

Um dos aspectos importantes abarcados no estudo dos canais de suprimentos é a escolha dos tipos de relação e das entidades envolvidas nas transações. Essa decisão está diretamente ligada à estratégia de cada empresa e terá forte impacto sobre o seu êxito. Conforme Ballou (2006), a eficiência e a eficácia da cadeia são decorrentes da formatação que para ela é escolhida.

Em grandes empresas dos setores de alimentos e bens de consumo, as quais produzem e comercializam uma gama de produtos, tais estratégias são fundamentais. Em geral, essas empresas escolhem o canal indireto de vendas, caracterizado pela existência de intermediários – nesse caso, distribuidores e pontos de venda (PDVs) – entre a empresa e o consumidor final. Entre os motivos da escolha estão agilidade, abrangência e capilaridade que esse modelo proporciona para atender aos consumidores.

Dado que os distribuidores atuam como o elo entre os fabricantes e o mercado, é primordial que as suas metas, regras e planos de ação estejam bem delineados, sobretudo aqueles relacionados à sua força de vendas. Todavia, comumente essas definições entre a indústria e os seus parceiros são feitas de forma não estruturada e com limitado embasamento. Definições

relevantes, como o *mix* de produtos a ser ofertado a cada PDV, não são precisamente realizadas por conta de falta de informação ou pelo grande volume de dados envolvidos sem que haja métodos estruturados para analisá-los. Como consequência, observa-se a falta de planos de ação explícitos para vendas dos distribuidores, o que acarreta perdas de oportunidades de venda. Além disso, lacunas de mercado não atendidas facilitam a entrada e a expansão da concorrência.

Tendo em vista essas circunstâncias, esse artigo propõe um método para associação de grupos de produtos a grupos de pontos de venda a partir de informações quantitativas. Objetiva-se gerar resultados que auxiliem no planejamento, na execução e no controle das ações de vendas do canal indireto, facilitando a definição do *mix* de produtos que deve ser destinado a cada ponto de venda. Para esse propósito, são utilizadas técnicas de Análise Multivariada de Dados (mais especificamente, Clusters e Análise de Componentes Principais) que permitem a análise de observações (RENCHER, 2002), como produtos e clientes (no contexto específico, representados pelos PDVs), por exemplo, e dados a respeito do comportamento de venda dos distribuidores e de compra dos PDVs, tipicamente descritos por múltiplas variáveis.

O artigo está estruturado em outras quatro seções, além da presente introdução: referencial teórico, procedimentos metodológicos, resultados e conclusão. Na seção seguinte, faz-se uma revisão bibliográfica abrangendo Análise Multivariada de Dados e estudos a respeito de agrupamentos de produtos e clientes, com o fim de embasar teoricamente o trabalho. Na seção de procedimentos metodológicos, descreve-se a sistemática e a sequência de ações aplicadas para o desenvolvimento do estudo. A seção subsequente apresenta os produtos gerados a partir da aplicação da metodologia e as análises resultantes. A seção de conclusão encerra o artigo.

## **2. Referencial Teórico**

A primeira parte desta seção apresenta fundamentos de Análise Multivariada de Dados (MVA), ao passo que a segunda revisa diferentes abordagens para a formação de grupos de produtos e de clientes com base em fatores quantitativos.

### **2.1. Análise Multivariada de Dados**

Metodologias destinadas a extrair informações de bases de dados descritas por diversas variáveis utilizando métodos estatísticos compõem o conjunto de conhecimentos de Análise Multivariada de Dados (JOHNSON e WICHERN, 2007). Nesse contexto, ferramentas de

clusterização e Análise de Componentes Principais (*Principal Component Analysis – PCA*) encontram ampla aplicabilidade, sendo seus fundamentos agora abordados.

Ferramentas de clusterização (referida também por alguns autores como classificação, como em Everitt (1980)) têm o objetivo de analisar as relações dentro de uma base de dados a fim de determinar se é possível ou não descrever tais dados de forma resumida, por um pequeno número de classes de observações similares (GORDON, 1999). De acordo com Rencher (2002), os objetos dentro de um *cluster* devem ser similares aos demais inseridos no mesmo *cluster* (homogeneidade), e distintos de objetos inseridos em outros *clusters* (denotando separação).

Conforme Rencher (2002), as principais abordagens para a clusterização de observações descritas por variáveis quantitativas são os métodos hierárquicos e os não-hierárquicos. Os algoritmos hierárquicos – aglomerativos e divisivos – são um processo sequencial: a cada passo de um processo aglomerativo, uma observação ou um *cluster* é fundido a outro *cluster*; já nos processos divisivos, o processo inicia com um *cluster* contendo todas as observações e, a cada iteração, um grupo é dividido em dois. Ambos os tipos procuram encontrar a divisão ótima (de acordo com critérios pré-definidos) a cada estágio do agrupamento ou da separação dos *clusters* (EVERITT *et al.*, 2011).

Os métodos aglomerativos mais citados na literatura e mais comumente utilizados são os de ligação (no inglês, *linkage methods*). De acordo com Johnson e Wichern (2007), nesses métodos as observações ou *clusters* são fundidos de acordo com alguma medida de distância entre eles. No algoritmo de ligação simples, a cada passo se aglomeram os grupos com a menor distância entre os seus membros mais próximos. Já na ligação completa, grupos são unidos apoiando-se na distância entre as suas observações mais distantes. Na terceira sistemática mais comum, conhecida como *average linkage*, escolhem-se os *clusters* a serem fundidos de acordo com a distância média entre todos os pares de membros dos dois grupos.

Todavia, como ressaltado por Kaufman e Rousseeuw (1990), métodos hierárquicos apresentam como defeito o fato de as mudanças feitas em passos anteriores serem irreversíveis – depois que um agrupamento é realizado, ele não é mais desfeito. Ainda, são algoritmos que podem demandar um tempo longo de operacionalização, já que o número de soluções possíveis é igual ao número de observações contidas na base de dados em estudo (KHATTREE e NAIK, 2000).

Os métodos não-hierárquicos diferem dos métodos previamente apresentados por não utilizarem classificações hierárquicas das observações para formar os *clusters*. A partição em grupos é feita tendo em vista a minimização ou maximização de algum critério numérico (EVERITT *et al.*, 2011).

Uma das abordagens mais comuns entre os métodos não-hierárquicos é o *k-means*, que, ao contrário dos métodos hierárquicos, permite a realocação de itens de um *cluster* para outro. Primeiramente, são escolhidas *k* observações como sementes para dar o passo inicial. Tais pontos iniciais podem ser selecionados de diversas formas: separando *k* itens randomicamente, escolhendo os primeiros *k* pontos da base de dados, identificando os *k* pontos que estão mutualmente mais distantes, entre outras (RENCHEER, 2002). Jain e Dubes (1988) ressaltam que se deve ter atenção à forma de seleção dos pontos iniciais, pois os *clusters* finais podem ser alterados por diferentes escolhas de tais pontos.

Após a seleção das *k* observações iniciais, cada ponto é atribuído ao grupo mais próximo, de acordo com alguma medida de distância pré-determinada (RENCHEER, 2002). À medida que um item é adicionado a um *cluster*, seu centróide é recalculado e as distâncias para os outros pontos são calculadas com base nessa posição. Por fim, após todas as observações estarem alocadas a um *cluster*, é feita uma revisão completa: se um ponto está mais próximo do centróide de outro grupo, ele é movido para esse grupo e os centróides dos dois *clusters* alterados são recalculados. Esse processo continua até que nenhuma observação satisfaça o critério acima.

A segunda técnica multivariada abordada nesse estudo é a Análise dos Componentes Principais (ACP), a qual objetiva possibilitar a extração de informações de observações descritas por uma série de variáveis quantitativas e inter-relacionadas por meio da redução do número de dimensões que as caracterizam. Isso é feito por meio das combinações lineares, nas quais as variáveis originais são transformadas em um novo grupo de variáveis ortogonais – chamadas de componentes principais (CP); os CPs são não-correlacionados e retém o máximo possível da variabilidade presente nos dados originais (ABDI e WILLIAMS, 2010; JOLLIFFE, 2002). A combinação linear com a máxima variação será o primeiro CP; o segundo componente principal será a combinação linear ortogonal ao primeiro CP que apresenta a maior variação; e assim sucessivamente (RENCHEER, 2002).

Geometricamente, a ACP visa encontrar os eixos naturais das observações  $x_1, x_2, \dots, x_p$ , os quais terão origem no vetor médio  $\bar{x}$ . Após transladar a origem para  $\bar{x}$ , rotacionam-se os eixos

multiplicando cada observação  $x_i$  por uma matriz ortogonal  $A$  (Equação 1), que representa a participação de cada variável na variabilidade da base de dados (HE *et al.*, 2011; RENCHER, 2002). As novas observações ( $z_i$ ) que resultam dessa multiplicação são não-correlacionadas e mantêm as mesmas distâncias da origem com os novos eixos rotacionados (como mostra a Equação 2). Tais observações são os componentes principais do conjunto de dados e são definidos matematicamente pela Equação 3 (RENCHE, 2002).

$$z_i = Ax_i, \text{ onde } A = \begin{pmatrix} a'_1 \\ a'_2 \\ \vdots \\ a'_p \end{pmatrix} \quad (1)$$

$$z'_i z_i = (Ax_i)'(Ax_i) = x'_i A' A x_i = x'_i x_i \quad (2)$$

$$z_i = a'_i x = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ip}x_p \quad (3)$$

## 2.2. Métodos para a formação de grupos de produtos e de clientes

Diversos estudos têm abordado sistemáticas de classificação e agrupamento de produtos e de clientes, sobretudo em cenários em que há uma gama desses elementos. Na maior parte dos casos, o objetivo é simplificar o gerenciamento de tais entidades a partir da consideração de grupos em vez de itens individuais, reduzindo a complexidade do cenário. Esse é o intuito do estudo de Wallstrom e Hogan (2007), que desenvolveram um método de agrupamento não-supervisionado para produtos médicos de venda livre tendo em vista o seu monitoramento para a previsão de surtos de diversas doenças. A sistemática, baseada em modelos de agrupamento Bayesianos, aglomera os medicamentos de acordo com as suas séries temporais de demanda, formando classes de produtos que apresentam comportamentos similares ao longo do tempo.

Jie *et al.* (2011) destacam a importância de agrupar produtos com características similares em *clusters* para, por exemplo, enfrentar mercados competitivos. A fim de operacionalizar essa abordagem, propõem uma técnica de algoritmo genético baseado em *fuzzy clustering* para estabelecer o número ótimo de classes e a melhor classificação de um determinado conjunto de produtos, de acordo com variáveis pré-definidas e critérios traduzidos quantitativamente em uma função objetivo.

No contexto da categorização de clientes, Ho *et al.* (2012) combinam a técnica de clusterização *k-means* com algoritmos genéticos para alocar consumidores a grupos. O algoritmo genético é utilizado para encontrar uma partição inicial das observações próxima à ótima, resolvendo o problema da divisão de partida do método *k-means*. Por sua vez, Mo *et al.* (2010) utilizam métodos de clusterização *k-means* e *self-organizing map* (metodologia baseada em redes neurais) para propor um algoritmo em dois estágios: no primeiro, cidades com características similares são agrupadas com base em variáveis de comportamento dos consumidores; no segundo, os clientes inseridos em cada *cluster* de cidades são categorizados. Segundo os autores, benefícios da sistemática incluem a redução do tempo necessário para segmentar clientes e a melhora no resultado final do algoritmo (sendo o segundo decorrente da disponibilidade de maior número de observações nos grupos de cidades do que em cada cidade de forma isolada).

Wang *et al.* (2014) abordam a questão de agrupamento de clientes com características similares no contexto de redes logísticas, a fim de possibilitar redução de custos operacionais e melhorar o nível de serviço. A abordagem utilizada é um algoritmo de *fuzzy clustering* aplicado em variáveis com diferentes níveis de importância para a determinação dos perfis dos clientes, os quais são definidos por uma análise hierárquica. Comparando tal método com outros três algoritmos de *fuzzy clustering*, percebe-se que essa metodologia captura com maior acurácia a similaridade entre clientes, de acordo com o índice de validação escolhido.

Para associar o resultado de agrupamentos de dois conjuntos de itens diferentes, Lingras *et al.* (2014) desenvolveram uma metodologia de clusterização iterativa na qual os resultados do agrupamento de um dos conjuntos de observações influenciam a descrição do outro conjunto. O método é testado em uma loja de varejo: clientes e produtos são representados por variáveis estáticas – dados históricos das transações – e por uma parte dinâmica – número de produtos contidos em cada *cluster* que são comprados por cada cliente e número de clientes de cada *cluster* que adquirem cada um dos produtos. As iterações param quando a parte dinâmica dos conjuntos se estabiliza. Os resultados mostram que é possível obter conclusões significativas quanto à associação de perfis de clientes e produtos.

Observa-se que os métodos de clusterização mais utilizados nos estudos que analisam produtos e clientes têm sido os não-hierárquicos, sobretudo devido aos melhores resultados obtidos em relação aos métodos hierárquicos. Uma análise nesse sentido é feita em Wajrock *et al.* (2008), os quais comparam diversos métodos de clusterização para uma base de dados de

clientes. A qualidade de cada método é avaliada por quatro índices, os quais levam em conta a compactação (o quanto os pontos de um mesmo *cluster* estão próximos entre si) e a separação (o quanto os *clusters* estão distantes uns dos outros) dos *clusters*. Conclui-se que, para três dos quatro índices utilizados, os métodos não-hierárquicos oferecem melhores resultados que os hierárquicos.

### 3. Procedimentos Metodológicos

Esta seção caracteriza o método de pesquisa, detalhando sua natureza, abordagem, objetivos e procedimento e, por fim, apresenta os passos do método proposto.

O método de pesquisa se enquadra como de natureza aplicada, devido à utilização de ferramental estatístico para um fim prático. A abordagem do estudo é majoritariamente quantitativa, tendo em vista que a metodologia se baseia em métricas relacionadas a transações de venda e compra de produtos e utiliza técnicas de análise multivariada de dados. Aspectos qualitativos são utilizados na interpretação e avaliação dos resultados obtidos com a aplicação das ferramentas de MVA. Já os objetivos são exploratórios, uma vez que o propósito da pesquisa é gerar um resultado final a partir do tratamento quantitativo e interpretação dos dados utilizados na investigação de um determinado contexto. Quanto ao procedimento, caracteriza-se como estudo de caso, dado que é utilizada uma situação real e contemporânea para aplicar e testar a metodologia proposta a fim de validá-la (YIN, 2009).

O método proposto, por sua vez, é operacionalizado em quatro etapas: (i) Coleta de dados; (ii) Aplicação de PCA nas matrizes de dados; (iii) Clusterização das observações a partir dos componentes principais; e (iv) Definição dos índices de proximidade entre *clusters*. Tais etapas são detalhadas na sequência.

Na etapa de coleta de dados são geradas duas matrizes. A primeira matriz,  $Z$ , recebe os dados referentes aos produtos e seus comportamentos de venda. Cada produto consiste em uma observação  $p_i$  ( $i = 1, 2, \dots, I$ ; onde  $I$  é o número de produtos na matriz), a qual é descrita por métricas  $x_a$  ( $a = 1, 2, \dots, A$ ; onde  $A$  é o número de métricas que representam as observações da matriz  $Z$ ). A segunda matriz,  $W$ , recebe os dados relacionados aos clientes e seus padrões de compra. Reforça-se que os clientes são, no sistema em análise, os PDVs, os quais são os clientes diretos dos distribuidores. Os clientes constituem as observações  $c_j$  ( $j = 1, 2, \dots, J$ ; onde  $J$  é o

número de clientes na matriz) e são descritos por métricas  $y_b$  ( $b = 1, 2, \dots, B$ ; onde  $B$  é o número de métricas que descrevem as observações da matriz  $W$ ).

Para o passo subsequente do método, aplica-se PCA nas matrizes  $Z$  e  $W$ . O objetivo desta etapa está na redução das dimensões para descrever as observações de cada matriz, visando simplificar a execução da etapa posterior de clusterização. A fim de determinar o número de componentes principais a serem retidos, utiliza-se o método *scree graph*, o qual auxilia graficamente na identificação dos CPs que retêm a maior variabilidade dos dados originais (RENCHER, 2002). Com a aplicação do PCA, as observações da matriz  $Z$  passam a ser descritas por métricas  $t_r$  ( $r = 1, 2, \dots, R$ ; onde  $R$  é o número de CPs retidos para os itens da matriz  $Z$ ), enquanto as observações da matriz  $W$  passam a ser caracterizadas por métricas  $t_s$  ( $s = 1, 2, \dots, S$ ; onde  $S$  é o número de CPs retidos para os itens da matriz  $W$ ).

A etapa seguinte da metodologia realiza a clusterização das observações de cada matriz utilizando as variáveis  $t_r$  e  $t_s$ . Inicialmente, com vistas à definição do número de *clusters* a serem formados ( $l$  é o número de *clusters* da matriz  $Z$  e  $m$  é o número de *clusters* da matriz  $W$ ), é realizada uma análise qualitativa a respeito das soluções válidas, as quais são identificadas com o uso de duas ferramentas quantitativas – dendogramas e *silhouette index* (calculado de acordo com a Equação 4). Para cada matriz, constrói-se um dendograma, ferramenta que apresenta em formato de diagrama a sequência de formação de *clusters* e as distâncias as quais eles são unidos quando um método hierárquico é usado (RENCHER, 2002). Ainda, analisa-se o *silhouette index* médio de cada *cluster* e os *silhouette index* médios das matrizes ( $\bar{S}$ ) – os quais medem a qualidade dos *clusters* formados a partir de sua homogeneidade e separação (ROUSSEEUW, 1987) – para diversas opções de números de *clusters*. Na sequência, é feita uma análise qualitativa das opções direcionadas pelas ferramentas quantitativas a fim de definir o melhor número de *clusters* em cada matriz (a Figura 1 exemplifica a geração de *clusters* genéricos formados a partir de duas matrizes diferentes em um mesmo sistema de coordenadas). Ao longo dessa etapa, o método de clusterização sugerido para gerar os diferentes agrupamentos é o *k-means*.

$$SI_i = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (4)$$

\*  $a(i)$  é a distância média da  $i$ -ésima observação para todas as outras em seu cluster

\*  $b(i)$  é a distância média da  $i$ -ésima observação para todas as outras contidas no cluster vizinho mais próximo

Por fim, calculam-se as distâncias Euclidianas entre os centróides de cada *cluster* da matriz  $Z$  e os centróides de cada *cluster* da matriz  $W$  (resultados armazenados na estrutura da Tabela 1) com o objetivo de determinar índices de proximidade entre estes grupos. Os índices de proximidade (IP) são calculados dividindo o inverso da distância entre um *cluster* da matriz  $W$  (denotado por  $f$ ) e um *cluster* da matriz  $Z$  (denotado por  $g$ ) pela soma dos inversos das distâncias do respectivo *cluster* da matriz  $W$  para todos os *clusters* da matriz  $Z$  (Equação 5). Este critério gera índices de proximidade entre 0 e 1, sendo os *clusters* mais relacionados à medida que esse grau é maior. Em termos práticos, esse índice pode ser utilizado para direcionar grupos de produtos a grupos de clientes mais promissores em termos de comercialização, dentre outras aplicações.

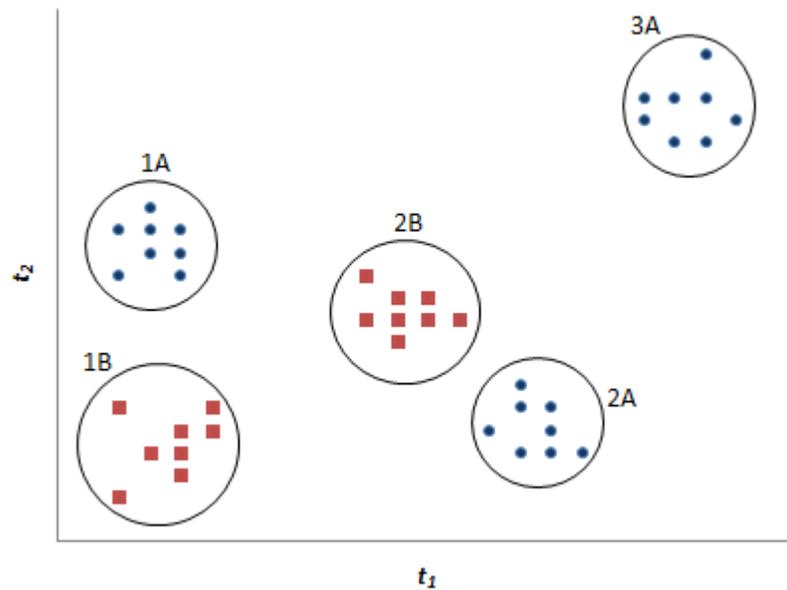


Figura 1 – Exemplo de plotagem de *clusters*

Fonte: Elaborado pelos autores.

Tabela 1 – Distâncias entre *clusters*

Fonte: Elaborado pelos autores.

<i>Clusters</i>	<b>W1</b>	...	<b>Wm</b>
<b>Z1</b>	distância Z1-W1	-	distância Z1-Wm
...	-	-	-
<b>Zl</b>	distância Zl-W1	-	distância Zl-Wm

$$IP_{f-g} = \left( \frac{1}{d_{f-g}} \right) / \left( \sum_{i=1}^l \frac{1}{d_{f-i}} \right) \quad (5)$$

#### 4. Resultados

O método proposto foi aplicado em uma companhia multinacional que produz, processa e comercializa alimentos e *commodities* agrícolas. No Brasil, está presente em todas as 27 unidades federativas. A divisão selecionada para a aplicação do método é a de mercadorias voltadas ao consumidor final, em razão da adequação ao ambiente do estudo (canal indireto de distribuição) e disponibilidade de dados.

Esta divisão possui, aproximadamente, 130 produtos e conta com 37 distribuidores no país. As companhias de distribuição têm o papel de descentralizar as vendas, possibilitando que a empresa de alimentos esteja presente não só nas grandes redes de varejo, mas também nos diversos PDVs de pequeno e médio porte presentes na maioria das cidades do território nacional.

Para acompanhar o desempenho dos distribuidores, esta multinacional utiliza uma solução de troca de dados – iniciativa conhecida no mercado como *data sharing intelligence* (DSI) – de uma empresa de tecnologia da informação e consultoria de negócios situada na região da Grande Porto Alegre, no estado do Rio Grande do Sul. Com essa solução, os distribuidores informam diariamente, por meio de troca eletrônica e automatizada de dados, as transações relativas aos produtos da empresa fornecedora. Essa empresa, por sua vez, utiliza a troca eletrônica de dados para informar aos distribuidores as suas metas. Todas as partes envolvidas têm acesso aos dados que devem analisar em uma mesma plataforma *online*, desenvolvida e gerenciada pela empresa de tecnologia e consultoria contratada, a qual também é responsável

pelas integrações e banco de dados. Entretanto, mesmo com dados disponíveis em grande quantidade, a definição do *mix* de produtos a ser ofertado aos PDVs é uma dificuldade encontrada pela empresa e seus distribuidores, sobretudo em razão da falta de métodos estruturados para analisar essas informações.

Tendo em vista esse cenário e a facilidade de acesso aos profissionais e ao banco de dados da empresa de tecnologia e consultoria, a primeira etapa do método proposto – coleta de dados – foi realizada de forma presencial durante um dia, iniciando com a exposição do método planejado de trabalho aos consultores. Na sequência, foram definidas em conjunto as métricas  $x_a$  que melhor caracterizam os produtos e as métricas  $y_b$  que melhor descrevem os clientes dos distribuidores (PDVs), valendo-se do conhecimento dos profissionais. O resultado foi a seleção de oito métricas  $x_a$  e de cinco métricas  $y_b$ , cujas descrições são apresentadas nas Tabelas 2 e 3, respectivamente.

Tabela 2 – Caracterização das métricas  $x_a$  (produtos)

Fonte: Elaborado pelos autores.

<b>Índice</b>	<b>Métrica</b>	<b>Descrição</b>
$x_1$	<i>Sell out</i> médio mensal (R\$)	Média mensal do balanço de saída do produto do distribuidor para os PDVs em reais ( <i>Sell out</i> = venda + bonificação – cancelamento – devolução).
$x_2$	<i>Sell out</i> médio mensal (unidades)	Média mensal do balanço de saída do produto do distribuidor para os PDVs em unidades ( <i>Sell out</i> = venda + bonificação – cancelamento – devolução).
$x_3$	Preço médio	Preço médio de venda de uma unidade do produto.
$x_4$	Dias com venda	Número de dias do ano nos quais o produto teve alguma unidade vendida.
$x_5$	Número de pedidos	Número de pedidos feitos pelos PDVs nos quais o produto esteve contido.
$x_6$	Clientes positivados	Número de PDVs que compraram o produto.
$x_7$	Clientes conquistados	Número de PDVs que compraram pela primeira vez um item do fornecedor em análise quando adquiriram o produto.
$x_8$	Vendedores com positividade	Número de vendedores que efetivaram venda do produto.

Tabela 3 – Caracterização das métricas  $y_b$  (clientes)

Fonte: Elaborado pelos autores.

Índice	Métrica	Descrição
$y_1$	<i>Sell out</i> médio mensal (R\$)	Médio mensal do balanço de entrada dos produtos do distribuidor para o PDV em reais ( <i>Sell out</i> = venda + bonificação – cancelamento – devolução)
$y_2$	<i>Sell out</i> médio mensal (unidades)	Média mensal do balanço de entrada dos produtos do distribuidor para o PDV em unidades ( <i>Sell out</i> = venda + bonificação – cancelamento – devolução)
$y_3$	Número de pedidos	Número de pedidos realizados pelo PDV ao seu distribuidor.
$y_4$	Produtos positivados	Número de produtos distintos comprados pelo PDV.
$y_5$	<i>Ticket</i> médio (R\$)	Valor médio (em reais) dos pedidos realizados pelo PDV ao seu distribuidor.

Com as métricas selecionadas, partiu-se para a definição do conjunto de dados a ser analisado. A partir do conhecimento dos consultores, foram estabelecidos dois critérios para proporcionar robustez e consistência aos dados: eles deveriam abranger o máximo possível do território nacional e estar validados pela empresa de alimentos. Assim sendo, optou-se por utilizar dados de cinco distribuidores, cada um localizado em uma das regiões do Brasil (sul, sudeste, centro-oeste, norte e nordeste). Em termos do tamanho da amostra, utilizaram-se dados que contemplam todo o ano de 2013, visto que possui informações acuradas e que abrangem um ciclo fechado de planejamento da empresa fornecedora de alimentos.

Por fim, com o auxílio da mesma equipe de consultores, foi feita a coleta das observações diretamente no banco de dados da empresa de tecnologia e consultoria. Como resultado, a matriz  $Z$  (referente aos produtos) foi composta por  $I = 135$  produtos e por  $A = 8$  métricas (resultando em uma matriz  $135 \times 8$ ); já os dados referentes aos clientes formaram a matriz  $W$ , a qual foi composta por  $J = 11.467$  clientes e  $B = 5$  métricas (resultando em uma matriz  $11.467 \times 5$ ).

De posse das matrizes  $Z$  e  $W$  provenientes da aplicação da primeira etapa do método, foi realizado o segundo passo: a aplicação de PCA nas matrizes. A decomposição das observações em componentes principais gerou os *scree graphs* apresentados nas Figuras 2 e 3. Analisando tais gráficos, percebe-se que, a partir do terceiro componente, níveis pouco expressivos de variabilidade são acrescentados aos dados com o uso de componentes principais adicionais; de tal forma, infere-se que os fatores localizados antes desse ponto são os mais importantes, pois carregam a maior parte da variabilidade dos dados. Com o cálculo do percentual da variância dos dados explicado por cada CP, foi verificado que os dois primeiros CPs da matriz  $Z$  explicam

99,01% da variância desse conjunto, ao passo que os dois primeiros CPs da matriz  $W$  explicam 99,46% da variância dessa matriz. Portanto, dados esses resultados e também para possibilitar a análise gráfica nas etapas subsequentes, optou-se pela seleção de dois componentes principais para descrever as observações de cada matriz, resultando em  $R = 2$  e  $S = 2$ .

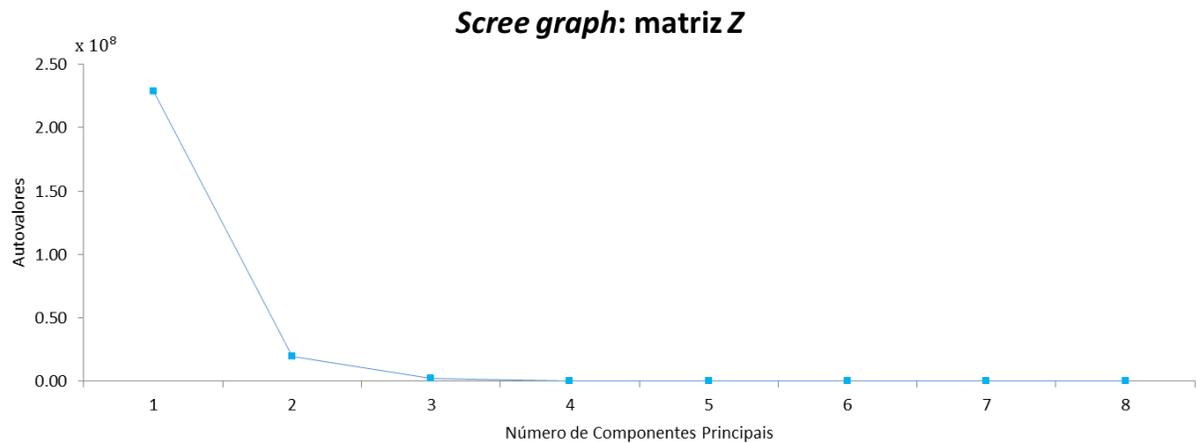


Figura 2 – *Scree graph* da matriz  $Z$

Fonte: Elaborado pelos autores.

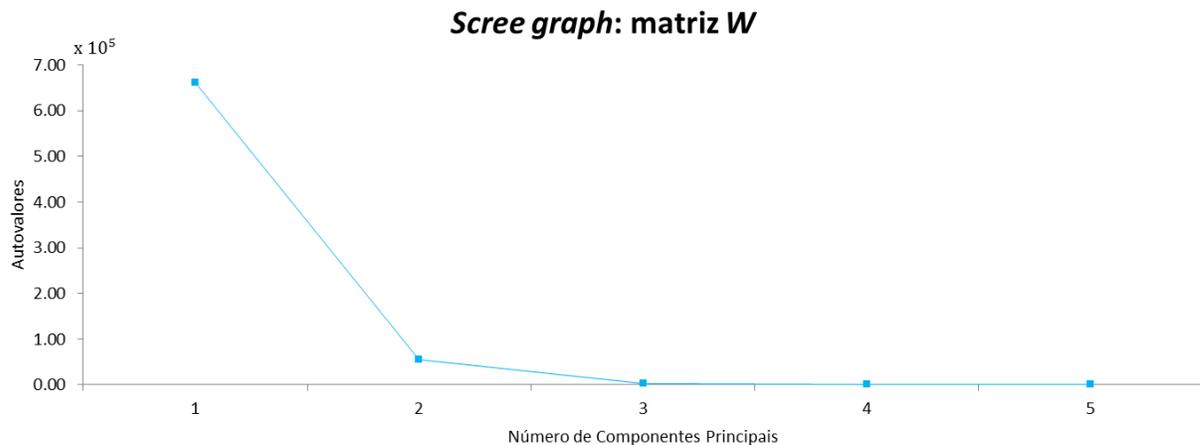


Figura 3 – *Scree graph* da matriz  $W$

Fonte: Elaborado pelos autores.

No terceiro passo do método, as análises são realizadas sobre as duas matrizes de observações caracterizadas pelos CPs definidos no passo anterior e não mais pelas métricas originais definidas na etapa 1. Para a definição do número de *clusters* a serem formados em cada

14

conjunto de dados, o uso dos dendogramas como ferramenta direcionadora foi descartado, já que a sua análise visual não trouxe conclusões importantes devido ao elevado número de observações em cada matriz (o que tornou a representação muito difícil de ser interpretada). De tal forma, recomenda-se que o dendograma seja utilizado no método apenas quando o número de itens do conjunto em estudo for reduzido o suficiente para gerar um diagrama que possa ser interpretado de forma visual.

As matrizes  $Z$  e  $W$  foram clusterizadas utilizando o método  $k$ -means, variando-se o número de  $clusters$  de dois a quinze; o limite máximo de quinze  $clusters$  por matriz foi estabelecido pelos consultores especialistas, que julgaram em análise prévia que os dados não estariam divididos em um número maior de categorias. Para cada número de  $clusters$  construídos, a qualidade do agrupamento gerado foi avaliada pelo  $\bar{S}$ ; o agrupamento com maior  $\bar{S}$  indica o número recomendado de  $clusters$  a serem gerados (ROUSSEEUW, 1987) para posterior análise qualitativa. O perfil de variação do  $\bar{S}$  para  $Z$  é apresentado na Figura 4 e para  $W$  na Figura 5.

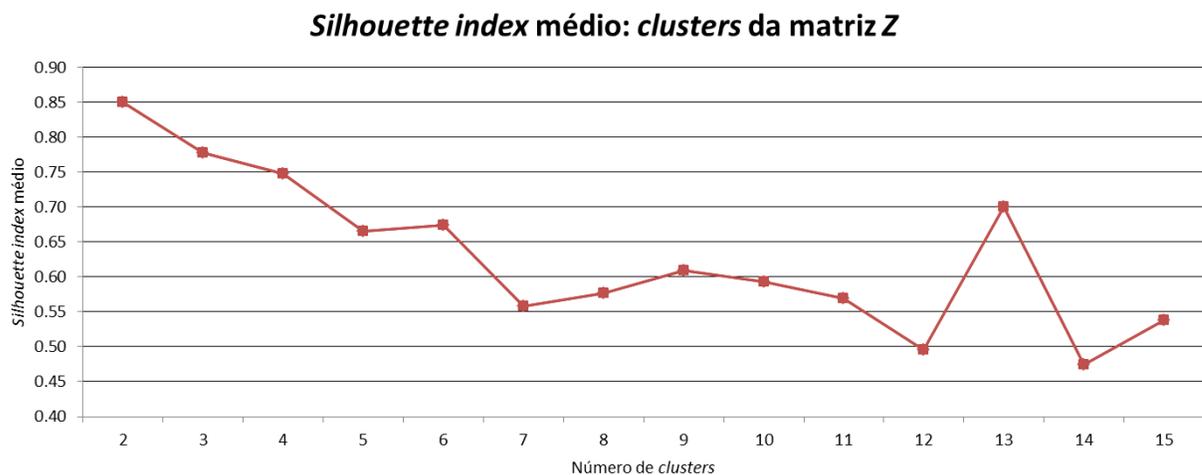


Figura 4 -  $\bar{S}$  dos  $clusters$  da matriz  $Z$

Fonte: Elaborado pelos autores.

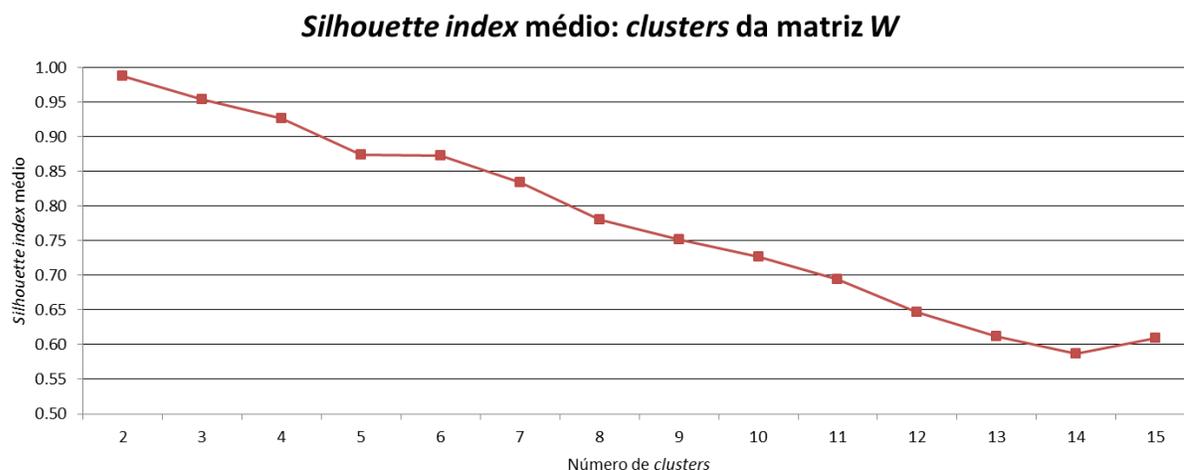


Figura 5 –  $\bar{S}$  dos *clusters* da matriz *W*

Fonte: Elaborado pelos autores.

Por se tratar de uma análise exploratória, a avaliação do perfil de  $\bar{S}$  para a matriz *Z* sugere que dois, três, quatro e treze *clusters* possam constituir soluções possíveis, já que os seus  $\bar{S}$  são consistentes (acima de 0,7) e muito próximos (0,8503; 0,7775; 0,7479 e 0,7007 respectivamente), com apenas 0,1496 de diferença absoluta entre o maior e o menor  $\bar{S}$ . De forma análoga, a análise do gráfico de  $\bar{S}$  para a matriz *W* sugere as opções com dois, três, quatro, cinco e seis *clusters*, todos com o índice acima de 0,85, denotando boa separação entre os grupos, e com 0,1154 de diferença absoluta entre o maior e o menor  $\bar{S}$ . Enfatiza-se, nesse instante, que por se tratar de uma análise exploratória, deve-se conciliar valores elevados de  $\bar{S}$  com a possibilidade de análise e interpretação qualitativa dos *clusters* gerados. Em outras palavras, optar diretamente pelo máximo  $\bar{S}$  em detrimento a agrupamentos com  $\bar{S}$ 's próximos pode dificultar a interpretação dos *clusters* gerados (o que não é recomendado em cenários exploratórios).

Os resultados acima foram então enviados aos consultores que haviam auxiliado na etapa 1 com vistas à definição do número de agrupamentos mais coerentes para os produtos e clientes. Em análise detalhada feita em reunião com os especialistas, a opção de treze *clusters* para a matriz *Z* foi descartada em razão de que produtos com comportamento tipicamente similar estavam em grupos distintos e porque cinco grupos possuíam apenas um ou dois produtos, o que não corresponde ao perfil de produtos da empresa. Ao avaliar as outras alternativas, os especialistas optaram pela geração de quatro agrupamentos tanto para a matriz *Z* (produtos) quanto para a matriz *W* (clientes). Após a efetiva inserção dos produtos e clientes nos *clusters* de

destino definidos pelo método *k-means*, os especialistas julgaram que os agrupamentos gerados representam de forma adequada as diferenças de características dos produtos e dos clientes, agrupando de forma satisfatória produtos com reconhecido comportamento semelhante pelos profissionais da área. Além disso, os especialistas corroboraram que os números de *clusters* gerados facilitam a definição do *mix* de produtos e o gerenciamento das ações de venda.

Por fim, estimaram-se os centróides de cada *cluster* das matrizes *Z* e *W*, as distâncias entre os centróides dos *clusters* das duas matrizes e, após, calcularam-se os IP segundo a Equação 5 (todas as etapas foram realizadas utilizando os CPs definidos no passo 2). Os IP resultantes são apresentados na Tabela 4 e os centróides dos *clusters* de *Z* e de *W* são apresentados na Figura 6.

Tabela 4 – Índices de proximidade (IP) entre *clusters*

Fonte: Elaborado pelos autores.

<b>Cluster</b>	<b>Z1</b>	<b>Z2</b>	<b>Z3</b>	<b>Z4</b>
<b>W1</b>	0,0668	0,1681	0,1950	0,5700
<b>W2</b>	0,0065	0,0269	0,0158	0,9507
<b>W3</b>	0,0297	0,1467	0,0686	0,7551
<b>W4</b>	0,4028	0,1119	0,3355	0,1497

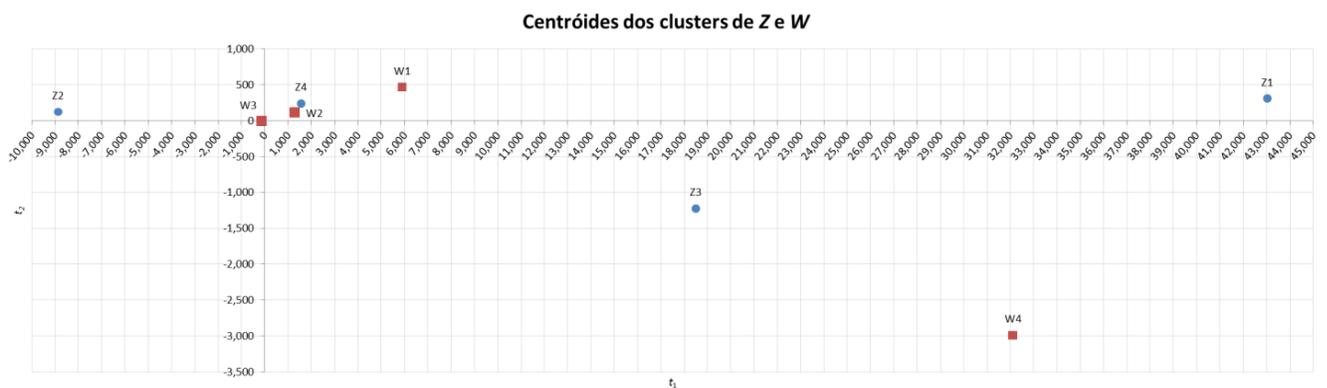


Figura 6 – Centróides dos *clusters* de *Z* e *W*

Fonte: Elaborado pelos autores.

Os índices de proximidade calculados apresentam o grau de relacionamento entre os *clusters* de produtos e clientes (perceba que o propósito de análise deste estudo está na relação

entre *clusters* de produtos e clientes, e não entre *clusters* de clientes). Quanto mais próximo de 1, mais relacionados dois *clusters* são. Ainda, observando-se que a soma dos IP de um grupo de clientes sempre resulta em 1, é possível perceber que esse índice indica a porcentagem de composição do *mix* de produtos para cada cliente. Por exemplo, os índices da Tabela 4 indicam que o *mix* de produtos ofertado aos clientes do *cluster* W2 deve ser composto por, aproximadamente, 95% de produtos do *cluster* Z4. Em análise análoga, deve-se compor o *mix* de produtos dos clientes do grupo W4 priorizando os grupos de produtos Z1 e Z3 com um *share* em torno de 40% e 34%, respectivamente, porém também com alguns produtos dos *clusters* Z2 e Z4, que devem ter uma participação em torno de 11% e 15%, respectivamente. Esse direcionamento na definição do *mix* de produtos reduz a complexidade do cenário e serve como recurso para a decisão dos analistas.

Em uma perspectiva complementar de análise, os resultados também possibilitam a identificação dos *clusters* de produtos que devem ser preferencialmente oferecidos aos clientes. A Tabela 4 indica que o *cluster* Z4 apresenta a maior soma dos IP: 2,4255. Dividindo esse resultado pela soma de todos os índices, percebe-se que esse grupo de produtos representa em torno de 60,64% do *mix* geral de produtos da empresa. Analogamente, percebe-se que Z1 representa 12,65%, Z2 contempla 11,34% e Z3 detém 15,37% de participação. Essa informação pode ser útil para o gerenciamento das entregas dos produtos aos distribuidores e dos seus estoques, por exemplo, e também para decisões em níveis anteriores ao canal de distribuição, como o planejamento da produção.

A fim de validar os resultados obtidos com o método, foi feita a verificação do comportamento real de compra de clientes inseridos em alguns dos *clusters* quanto ao seu *mix* de produtos médio. Estes dados foram coletados diretamente no banco de dados da empresa de consultoria, com auxílio dos consultores e da equipe técnica. Os resultados indicam uma boa aderência do método: clientes do *cluster* W2 compram, em média, 83% do seu *mix* de produtos de itens que pertencem ao *cluster* Z4, enquanto o método indica um IP de 0,9507. Outra relação importante é que, para o grupo de clientes W3, os produtos do *cluster* Z4 compõem em torno de 68% do *mix* e os produtos do *cluster* Z2 representam em torno de 18%, enquanto os IP resultantes destas relações foram de 0,7551 e 0,1467, respectivamente. Ainda assim, nota-se que o *cluster* de produtos Z4 foi supervalorizado com o método, assumindo um papel mais importante que o real. Tal situação pode ser compreendida pela Figura 6, que mostra este grupo muito

próximo de três grupos de clientes. Esse aspecto reforça que as saídas do método proposto não são definitivas, mas representam um apoio valioso para a tomada de decisão, devendo ser refinadas com conhecimentos de mercado, clientes, produtos e de outros aspectos relevantes para a obtenção de bons resultados.

## 5. Conclusão

A distribuição e disponibilização de produtos aos consumidores é fundamental para o sucesso das empresas no mercado. O *mix* de produtos ofertado a cada cliente é uma decisão importante, sobretudo no contexto do canal indireto de distribuição. Entretanto, percebe-se a carência de métodos estruturados para embasar a tomada desta decisão, o que ocasiona perda de oportunidades de negócio. De tal forma, este artigo propôs um método para a definição dos itens a serem oferecidos a cada cliente no canal indireto de vendas.

Informações quantitativas sobre a venda de produtos e compra de clientes – tipicamente descritas por múltiplas variáveis de naturezas distintas – e ferramentas de análise multivariada de dados compõem a estrutura principal do método apresentado, apoiado em quatro etapas. Para testar e validar a sistemática proposta, ela foi aplicada a dados reais da linha de produtos voltada ao consumidor final de uma empresa que processa e comercializa alimentos e *commodities agrícolas*, a qual trabalha com canal indireto de distribuição e, portanto, se adequa ao ambiente de estudo.

Os resultados obtidos apontaram que o método é capaz de agrupar produtos e clientes com características similares e definir o grau de relacionamento entre eles – representado pelos Índices de Proximidade (IP). Estes índices, que variam de 0 a 1, indicam como deve ser a composição do *mix* de produtos para cada grupo de cliente, relacionando-os aos grupos de produto. Para embasar o estudo, a validação dos índices obtidos com o *mix* de produtos médio real comprado pelos clientes demonstrou que o método proporciona resultados válidos e que podem ser úteis para o gerenciamento das ações de venda do canal indireto.

Em análise dos resultados, destaca-se que os produtos do método proposto devem ser usados num caráter de direcionadores da tomada de decisão dos analistas, sendo refinados com conhecimento quantitativo e qualitativo do mercado específico, de aspectos culturais e demográficos dos clientes, de questões econômicas, de ações de *marketing*, entre outros pertinentes. Como oportunidades para estudos futuros, sugerem-se (i) a inclusão de variáveis

demográficas e econômicas para a descrição do perfil dos clientes com vistas a aumentar a precisão de sua clusterização, melhorando a acurácia dos resultados do método; (ii) a adaptação do método para o canal direto de vendas, com a identificação das métricas que mais se adequam e melhor descrevem o comportamento deste canal; e (iii) a sistematização da incorporação da análise qualitativa dos especialistas no método, organizando as suas contribuições e os seus relacionamentos com os resultados quantitativos.

## Referências

ABDI, H.; WILLIAMS, L. J. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 433-459, 2010.

BALLOU, R. H. Gerenciamento da cadeia de suprimentos/logística empresarial. Porto Alegre: Bookman, 2006.

CHRISTOPHER, M. Logística e gerenciamento na cadeia de suprimentos. São Paulo: Cengage Learning, 2011.

EVERITT, B. *Cluster Analysis* (2nd Edition). New York: Halsted Press, 1980.

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. *Cluster Analysis* (5th Edition). Chichester: John Wiley & Sons, 2011.

GORDON, A. *Classification* (2nd Edition). London: Chapman and Hall-CRC, 1999.

HE, S.; WANG, G. A.; COOK, D. F. Multivariate measurement system analysis in multisite testing: An online technique using principal component analysis. *Expert Systems with Applications*, 38, 14602-14608, 2011.

HO, G. T. S.; IP, W. H.; LEE, C. K. M.; MOU, W. L. Customer grouping for better resources allocatin using GA based *clustering* technique. *Expert Systems with Applications*, 39, 1979-1987, 2012.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. New Jersey: Prentice Hall, 1988.

- JIE, L.; AILING, G.; LIANYONG, Y. Application of Product *Cluster* Method based on Fuzzy *Cluster* and Genetic Algorithm. *Procedia Engineering*, 15, 4930-4935, 2011.
- JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall, 2007.
- JOLLIFFE, I. T. *Principal Component Analysis*. New York: Springer, 2002.
- KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.
- KHATTREE, R.; NAIK, D. N. *Multivariate Data Reduction and Discrimination with SAS<sup>®</sup> Software*. Cary, NC: SAS Institute Inc., 2000.
- LINGRAS, P.; ELAGAMY, A.; AMMAR, A.; ELOUEDI, Z. Iterative meta-*clustering* through granular hierarchy of supermarket customers and products. *Information Sciences*, 257, 14-31, 2014.
- MO, J.; KIANG, M. Y.; ZOU, P.; LI, Y. A two-stage *clustering* approach for multi-region segmentation. *Expert Systems with Applications*, 37, 7120-7131, 2010.
- RENCHER, A. C. *Methods of multivariate analysis*. New Jersey: Wiley-Interscience, 2002.
- ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65, 1987.
- WAJROCK, S.; ANTILLE, N.; RYTZ, A.; PINEAU, N.; HAGER, C. Partitioning methods outperform hierarchical methods for *clustering* consumers in preference mapping. *Food Quality and Preference*, 19, 662-669, 2008.
- WALLSTROM, G. L.; HOGAN, W. R. Unsupervised *clustering* of over-the-counter healthcare products into product categories. *Journal of Biomedical Informatics*, 40, 642-648, 2007.

WANG, Y.; MA, X.; LAO, Y.; WANG, Y. A fuzzy-based customer *clustering* approach with hierarchical structure for logistics network optimization. *Expert Systems with Applications*, 41, 521-534, 2014.

YIN, R. K. *Case study research: design and methods*. Thousand Oaks, CA: Sage, 2009.