Sistemática de seleção de variáveis para classificação de produtos em categorias de modelos de reposição

"Artigo a ser submetido ao periódico Gestão e Produção"

César Augusto Mazzillo Júnior (UFRGS) - *mazzillo.junior@ufrgs.br*

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

Resumo

Empresas integradas em cadeias de suprimento buscam, cada vez mais, iniciativas para melhorar o desempenho global da cadeia, principalmente no que concerne o fornecimento de produtos e gerenciamento dos canais diretos e indiretos de distribuição. O VMI (*Vendor Managed Inventory*) auxilia as empresas a melhor gerenciar e balancear o estoque ao longo da cadeia. Para o seu correto funcionamento, é necessário que parâmetros e modelos de cálculo estejam bem definidos e aderentes às características dos produtos e locais de reposição, normalmente descritos por elevado número de variáveis. Esse artigo propõe uma sistemática de seleção de variáveis para classificação de produtos em modelos de reposição. Para tanto, utiliza a Análise dos Componentes Principais (ACP) em conjunto com as ferramentas de classificação algoritmo do vizinho mais próximo (kNN) e Análise Discriminante Linear (ADL). Ao ser aplicado em um estudo prático do setor de consultoria em *Supply Chain*, o método proposto alcançou uma acurácia de classificação de 90%, evidenciando que 55% das variáveis originais são fundamentais para a definição dos modelos.

Palavras-chave: SCM (*Supply Chain Management*), VMI (*Vendor Managed Inventory*), ACP (Análise dos componentes principais), kNN (algoritmo do vizinho mais próximo), ADL (Análise Discriminante Linear), MVA (Análise Multivariada de Dados)**.**

1. Introdução

Com o crescimento da competição mundial, a eficiência dos processos internos das organizações não é mais fator diferencial de mercado. Empresas integradas em uma cadeia de suprimentos devem atentar para todos os seus elos e gerenciar os processos entre eles com eficácia, flexibilidade e qualidade. Segundo Goldratt (1984), os esforços de melhoria nos processos devem focar no elo mais fraco da cadeia, visto que é ele que determina o desempenho global da cadeia – aquele percebido pelos clientes finais.

A Análise Multivariada de Dados (*Multivariate Analysis* - MVA) é uma ferramenta estatística que consiste na análise simultânea de múltiplas variáveis, as quais possuem a

capacidade de explicar um comportamento. A MVA pode auxiliar as empresas a melhor entenderem seus processos, permitindo respostas mais rápidas frente a mudanças de mercado, produtos e serviços mais alinhados com as necessidades dos clientes, e apoiando-se em menos recursos e em menor tempo. Além disso, a MVA permite transformar dados disponíveis em conhecimento para a tomada de decisão (HAIR JR *et al*., 2010; RENCHER; CHRISTENSEN, 2012).

A MVA possui aplicações em diversas áreas do mercado, como Engenharia, Medicina e Educação (RENCHER; CHRISTENSEN, 2012). Dentre as aplicações práticas, destaca-se o uso das técnicas para a operacionalização do *Data Mining*, ferramenta que permite extrair informações e conhecimento de grandes massas de dados para uso em análises de mercado, comportamento do consumidor e avaliação de produtos (HAN; KAMBER, 2006). As técnicas de MVA podem ser divididas entre as de relacionamento dependente e as de relacionamento interdependente, sendo dentre as interdependentes a Análise dos Componentes Principais (ACP) e a Análise Fatorial, as quais são tipicamente indicadas para estudos que envolvam elevado número de variáveis (HAIR JR *et al*., 2010). Neste contexto, Sahmer e Qannari (2008) indicam a importância de se trabalhar com um correto número de variáveis ou atributos relevantes e não redundantes, reduzindo tempo de análise e esforço dos analistas.

Por sua vez, o VMI (*Vendor Managed Inventory* – Estoque Gerenciado pelo Fornecedor) é uma iniciativa na qual o fornecedor passa a ser responsável pela reposição do estoque de produtos dos seus compradores, baseado em informações de vendas e nível de estoques dos elos a jusante da Cadeia de Suprimentos (KAZMIERCZAK NETO, 2009; SILVA, 2010). Empresas produtoras de bens de consumo, porém, possuem dificuldade na escolha do melhor modelo de reposição e dos melhores parâmetros para compô-lo, visto que trabalham com uma gama de mais de 1500 SKUs (*Stock keeping unit* – Unidade de Manutenção de Estoque) distintos com comportamento de venda e de mercado variáveis. Tendo em vista que o VMI busca proporcionar incremento de vendas, redução de falta de estoques, melhoria na disponibilização dos produtos e redução global dos estoques da cadeia, a definição correta do modelo de reposição e dos seus parâmetros possui caráter fundamental para a correta implantação da política (TAVARES, 2003).

Este artigo apresenta uma abordagem para identificar as variáveis (critérios quantitativos de venda e mercado) mais relevantes para classificação dos produtos em classes de reposição.

Para tanto, a ACP é aplicada sobre os dados originais, e um índice de importância das variáveis é derivado com base nos parâmetros gerados pela ACP. Na sequência, inicia-se um processo iterativo de classificação de observações (modelos de produtos) e remoção de variáveis. As ferramentas de classificação testadas são o kNN e a ADL. Objetiva-se, com base em um conjunto reduzido de variáveis descritivas, aprimorar a acurácia de alocação dos produtos aos modelos de reposição, possibilitando uma melhor adequação ao processo do VMI e, consequentemente, um melhor desempenho nos indicadores associados.

Este artigo está estruturado como segue, além desta introdução. Na segunda seção, uma revisão teórica apresenta os fundamentos da Análise de Componentes Principais, Análise Discriminante, kNN (algoritmo do vizinho mais próximo) e Análise Multivariada, contextualizando sua aplicação nos conceitos de VMI. Na terceira seção, é apresentada a metodologia proposta para a definição das variáveis que melhor explicam a decisão pelos modelos de reposição e sua parametrização. A seção 4 apresenta um estudo de caso, ao passo que a quinta seção traz as conclusões.

2. Referencial Teórico

2.1. Ferramentas Multivariadas

Ferramentas multivariadas são métodos estatísticos e matemáticos que objetivam analisar a inter-relação entre um grande conjunto de variáveis ou sistemas complexos. Os diferentes métodos existem para serem aplicados em tipos e conjuntos específicos de dados, visando a elucidar diferentes análises (MALINOWSKI, 2002; BARTHOLOMEW, 2010). Fundamentos de ACP, algoritmo do vizinho mais próximo (kNN) e Análise Discriminante Linear (ADL) são agora apresentados.

2.1.1. Análise dos Componentes Principais (ACP)

A ACP é uma técnica de ordenação multivariada que objetiva encontrar padrões de comportamento em dados interdependentes e não agrupados, reduzindo o número de dimensões e, então, exibindo a posição dos dados em variáveis latentes (componentes principais - CPs), os quais são não correlacionados. Dessa forma, a ACP busca reter a maior quantidade de informação e variabilidade existente nos dados, com a menor quantidade de componentes principais (JOLLIFFE, 2006; SYMS, 2008; HE *et al.*, 2011; RENCHER; CHRISTENSEN, 2012).

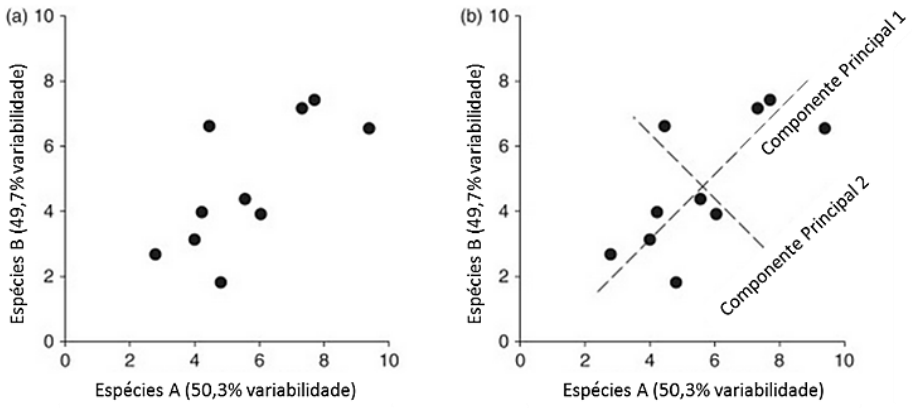
Matematicamente, a ACP consiste em uma transformação linear ortogonal, cuja aplicação transforma os dados em um novo sistema de coordenadas (JOLLIFFE, 2002; MOROPOULOU; POLIKRETI, 2009). O primeiro CP formado é a combinação linear que representa a maior variação do conjunto de dados, enquanto o segundo, ortogonal ao primeiro, possui a segunda maior variação, e assim por diante. Os valores dessas novas variáveis são chamados de valores fatoriais e podem ser interpretados geometricamente (ABDI; WILLIAMS, 2010; RENCHER; CHRISTENSEN, 2012). O número de componentes principais formado é menor ou igual ao número de variáveis do conjunto de dados (MOROPOULOU; POLIKRETI, 2009).

Dado um conjunto de vetores de dados $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$ com dimensões iguais a $\mathbf{p}$, pode-se formar uma matriz $\mathbf{M_{pxn}}$ que represente o conjunto inicial dos dados para estudo e calcular-se o vetor médio $\bar{\mathbf{y}}$ através da Equação (1) (HE *et al.*, 2011). Objetiva-se encontrar os eixos naturais do conjunto de dados, os quais possuem origem em $\bar{\mathbf{y}}$. Isso é realizado ao se transladar a origem para $\bar{\mathbf{y}}$ e então rotacionar os eixos (Figura 1). Após a rotação, as novas variáveis geradas (componentes principais) serão independentes (RENCHER; CHRISTENSEN, 2012).

$$\bar{y} = \frac{1}{n} * \sum_{i=1}^{n} y_i \tag{1}$$

Figura 1 – (a) Conjunto original de dados exemplo. (b) Novos eixos criados com centro em $\bar{y}$, obtendo os dois componentes principais não correlacionados.

Fonte: SYMS (2008).



Conforme He *et al.* (2011) e Rencher e Christensen (2012), para rotacionar os eixos, cada vetor observação $\boldsymbol{y_i}$ deve ser multiplicado por **A,** vetor que traz os pesos (influência na variabilidade) de cada variável no conjunto de dados**.** Como **A** é ortogonal, pode-se afirmar

que a distância da origem não foi alterada (conforme a Equação (2)), transformando $\mathbf{y_i}$ em um novo ponto $\mathbf{t_i}$, o qual dista o mesmo valor da origem com os eixos rotacionados. Ao encontrar os novos eixos, e realizar operações algébricas de transposição e ortogonalização, encontra-se o vetor **A** (Equação (3)) para que os componentes principais $\mathbf{t_1}, \mathbf{t_2}, \dots, \mathbf{t_n}$ em $\mathbf{t = Ay}$ sejam não relacionados. Os CPs (variáveis latentes) consistem nas variáveis $\mathbf{t_1} = \boldsymbol{a_{1'}}\mathbf{y}, \dots, \mathbf{t_p} = \boldsymbol{a_{p'}}\mathbf{y}$ em $\mathbf{z = Ay}$. Análises baseadas nos coeficientes de importância do ACP são estruturadas nas variáveis latentes formadas, as quais seguem o exemplo da Equação (4), na qual $\mathbf{a_{np}}$ representa o peso da variável $\mathbf{y_p}$ na variabilidade do conjunto de dados.

$$\mathbf{t'_i t_i} = (\mathbf{Ay_i})'(\mathbf{Ay_i}) = \mathbf{y'_i A' Ay_i} = \mathbf{t'_1 t_i} \tag{2}$$

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{a_{1'}} \\ \boldsymbol{a_{2'}} \\ \vdots \\ \boldsymbol{a_{p'}} \end{pmatrix} \tag{3}$$

$$\mathbf{t_1} = \mathbf{a_{11} y_1} + \mathbf{a_{12} y_2} + \cdots + \mathbf{a_{1p} y_p} \tag{4}$$

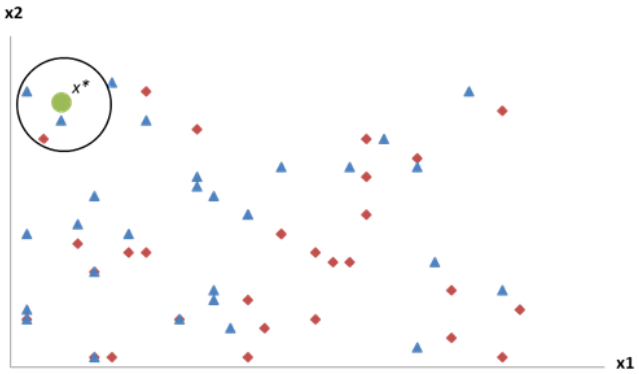
2.1.2. Algoritmo do vizinho mais próximo (kNN)

O kNN é um método supervisionado de classificação de dados baseado na proximidade de seus vizinhos em um espaço amostral (DAKHLAOUI, *et al.*, 2012). Seu objetivo é formar uma generalização com base em um conjunto de treinamento, maximizando a acurácia da classificação de novos dados (GAO; GAO, 2010). O algoritmo pressupõe que o conjunto de treinamento é composto pelas variáveis descritivas e pela sua classificação; o KNN então utiliza tais variáveis para classificar um novo item (SU, 2011).

Com base em um conjunto de treinamento, formado por *n* observações previamente classificadas, como $\mathbf{X} = \{x_1, x_2, \dots, x_t\}$, e uma observação a ser classificada, formada por $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ em um espaço *t*-dimensional, tem-se que a distância Euclidiana, na qual o algoritmo baseia-se, pode ser calculada para cada ponto do conjunto de treinamento utilizando-se os valores geométricos de cada um dos pontos. Com base nesta distância, os *k* vizinhos mais próximos são identificados e, com base em um *k* escolhido, esse relativo ao número de grupos possíveis de classificação, uma nova observação é atribuída à classe com maior número de observações em *k*, conforme ilustrado na Figura 2 (SU, 2011; DAKHLAOUI; DUDA *et al.*, 2012).

Figura 2 – Uma observação x* é testada em um conjunto treinamento com *k = 3*, formado pelas observações com menor distância euclidiana do ponto x*. No caso apresentado, a observação x* é alocada à categoria dos pontos representados pelos triângulos.

Fonte: Adaptado de DUDA *et al.* (2012).



2.1.3. Análise Discriminante Linear (ADL)

A Análise Discriminante Linear (*Linear Discriminant Analysis*) é um dos métodos estatísticos supervisionados mais conhecidos para classificação de dados (XU *et al.*, 2008). Seu objetivo é, com base em um conjunto de treinamento de dados, buscar uma combinação linear de variáveis que melhor explique aqueles dados (FISHER, 1937; MARTINEZ; KAK, 2001). Proposto por Fisher (1937), o método baseia-se em um conjunto de aprendizado com *n* observações e $\boldsymbol{K}$ classes, o qual formará duas matrizes, a de dispersão intraclasses $\boldsymbol{S_w}$ (que computa a variância dos padrões em relação à classe que pertence) e a de dispersão entre as classes $\boldsymbol{S_b}$ (computa a variância entre as classes), as quais são definidas pelas equações (5) e (6), respectivamente:

$$\boldsymbol{S_w} = \sum_{k=1}^{K} \boldsymbol{n_k} * (\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \bar{\boldsymbol{x}}_{\boldsymbol{..}}) * (\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \bar{\boldsymbol{x}}_{\boldsymbol{..}})^{\boldsymbol{T}} \tag{5}$$

$$\boldsymbol{S_b} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \boldsymbol{n_k} * (\bar{\boldsymbol{x}}_{\boldsymbol{ki}} - \bar{\boldsymbol{x}}_{\boldsymbol{k.}}) * (\bar{\boldsymbol{x}}_{\boldsymbol{ki}} - \bar{\boldsymbol{x}}_{\boldsymbol{k.}})^{\boldsymbol{T}} \tag{6}$$

nas quais $\boldsymbol{x_{ki}}$ é a observação $\boldsymbol{i}$ em um espaço *p*-dimensional da classe *k*, $\boldsymbol{n_k}$ é o número de amostras do conjunto de aprendizado da classe *k* e $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ $\bar{\boldsymbol{x}}_{\boldsymbol{..}}$, representam o vetor médio da classe *k* e a média geral, respectivamente (PARK; LEE, 2007; KITANI; THOMAZ, 2007; XU *et al.*, 2008). Segundo Hardle e Simar (2003) e Kitani e Thomaz (2007), o objetivo é encontrar a razão entre o determinante da matriz $\boldsymbol{S_b}$ e a matriz $\boldsymbol{S_w}$, conhecido como critério de Fischer (Equação (7)).

$$\textbf{Critério de Fischer} = \max \frac{|S_b|}{|S_w|} \tag{7}$$

O vetor de projeção responsável pelo máximo critério de Fischer denota a maior separação entre as classes.

2.2. Métodos para seleção de variáveis para classificação de dados

Sistemáticas para seleção de variáveis em aplicações industriais e corporativas têm recebido crescente atenção do meio acadêmico por possibilitarem classificações e predições mais acuradas. Wold *et al*. (2001) foram pioneiros na proposição de técnicas de PLSR (*Partial least squares projection regression* – Regressão dos mínimos quadrados parciais) aliadas a outras técnicas multivariadas (ACP, kNN e ADL) com vistas à criação de índices de importância de variáveis para predição (VIP – *Variable importance for the projection*). Choi *et al.* (2011) ressaltam que tais métodos buscam descartar variáveis irrelevantes para aperfeiçoar a classificação de dados e poupar esforço computacional no processamento de dados.

Anzanello *et al.* (2009) propõem a criação de uma série de diferentes índices de importância de variáveis (IIV), um deles baseado nos pesos das variáveis das combinações lineares independentes originadas pela regressão PLS, e posterior classificação pelo algoritmo de kNN para remover as variáveis menos relevantes e aumentar a acurácia do conjunto de treinamento (Figura 3). Choi *et al.* (2011) também apresentam um método baseado em pesos das variáveis oriundos da ACP e posterior aplicação de ADL para classificação dos dados. Por sua vez, Gertheiss e Tutz (2009) propõem a aplicação de um conjunto de métodos baseados no algoritmo de KNN para estimar a probabilidade de um determinado dado pertencer a uma classe e depois classificá-lo de acordo com tal análise. Os métodos avaliados consideram os algoritmos clássicos de kNN e ADL e um algoritmo cujas variáveis possuem pesos de acordo com sua influência.

Ao abranger outros métodos de MVA, Ballabio *et al.* (2010) propõem a construção da Medida Canônica de Correlação (Índice CMC), calculado com base em dois conjuntos de dados, o das variáveis independentes e o da matriz das classes existentes. Os índices são calculados por variável e um *ranking* é gerado com base na contribuição das variáveis em discriminar as classes. Rao e Lakshminarayanan (2006) apresentam um método para classificação baseado na matriz dos coeficientes de correlação parciais e no coeficiente de Pearson, medida que define a associação entre variáveis contínuas, comparando-o com uma aplicação de ADL na mesma base de dados. Por fim, Shreve *et al.* (2011) propõem uma sistemática baseada em simulação de Monte Carlo para possibilitar a comparação de métodos

de classificação com base na estabilidade dos modelos e validade da seleção de variáveis; os autores concluem que um conjunto de dados amostral suficientemente grande permite a comparação dos métodos.

3. Procedimentos Metodológicos

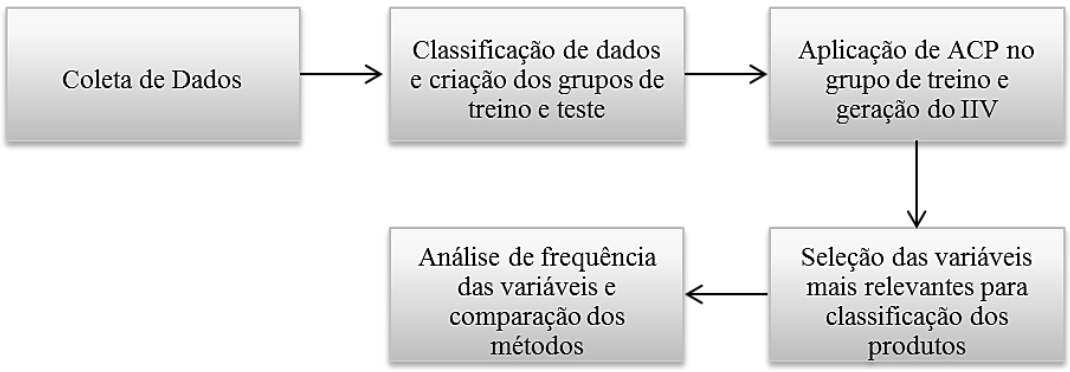
Esta seção descreve a classificação da pesquisa quanto ao seu método, no qual se define a natureza do estudo, sua abordagem, objetivos e procedimentos. Após, descreve-se o método aplicado no trabalho, detalhando os passos para o desenvolvimento da sistemática de seleção de variáveis para inserção de produtos em categorias de reposição automática de estoques.

O presente estudo é definido como de natureza aplicada, sendo a sistemática desenvolvida com base na solução de VMI da empresa. Quanto à sua abordagem, a pesquisa é de caráter quantitativo, pois busca o desenvolvimento de uma sistemática com base em dados históricos de vendas e de demanda para classificar os produtos em métodos de reposição automática de estoques. Os objetivos do estudo são definidos como exploratórios e explicativos, visto que parte de uma massa informações, formulando hipóteses para alcançar os seus resultados. Por fim, quanto ao procedimento, o estudo é caracterizado como experimental, visto que se baseará na manipulação de fatores e variáveis para observar um resultado (YIN, 2003).

As etapas propostas, ilustradas na Figura 3, apoiam-se nas técnicas ACP, ADL e kNN, e são detalhadas na sequência.

Figura 3 – Fluxograma das etapas para o desenvolvimento do método

Fonte: Elaborado pelos autores.



3.1. Coleta de dados

Para a montagem da base de dados foi estruturada uma matriz nos moldes da Tabela 1, contendo as variáveis pertinentes ao processo de distribuição disponíveis no sistema VMI e

que poderiam influenciar na escolha do modelo de reposição para uma relação produto/local de reposição (no caso estudado, centros de distribuição e lojas). Cada observação $(\boldsymbol{y_i}, \boldsymbol{x_l}), i = 1,2, \dots, n$, representa um produto/local de reposição, descrito por j variáveis e um único modelo de reposição; as variáveis $\boldsymbol{x}$ descrevem as características de cada produto, enquanto a variável $\boldsymbol{y}$ denota o modelo de reposição mais adequado (indicado por meio de uma classe).

Tabela 1 – Estrutura da Matriz de Dados

Fonte: Elaborado pelos autores.

Observações	Variável 1	Variável 2	...	Variável *J*	Modelo de Reposição
Produto 1/Loja 1	-	-	-	-	-
...	-	-	-	-	-
Produto n/Loja 2	-	-	-	-	-

3.2. Classificação de dados e criação dos grupos de treino e teste

Com o auxílio da equipe de consultores especializada em soluções de distribuição, o comportamento e características específicas dos produtos foram vinculados ao modelo de reposição mais adequado. Para tal estudo, foram considerados produtos e locais de reposição já consolidados na empresa, objetivando uma base mais sólida para o aprendizado do método.

Na sequência, os dados foram divididos em um grupo de treino ($\boldsymbol{T_r}$), contendo n_{Tr} observações, e um grupo de teste ($\boldsymbol{T_s}$), contendo n_{Ts} observações, de uma forma que $n_{Tr} + n_{Ts} = n$. Conforme Chong *et al.* (2007), é adequado manter 60% das observações no grupo de treino.

3.3. Aplicação de ACP no grupo de treino e geração do Índice de Importância de Variáveis (IIV)

A fim de caracterizar a relação entre as variáveis e os métodos de reposição de cada produto/local, foi aplicada ACP no grupo de treino, gerando pesos "*w*" para cada variável. O peso está relacionado com a importância da variável na explicação da variabilidade presente nos dados; assume-se, conforme Anzanello et al. (2009), que variáveis com maior peso absoluto denotam variáveis com maior importância na explicação de variância dos dados.

Com base nos pesos gerados pela ACP, foram estimados os índices de importância das variáveis (IIV) de acordo com a equação (8), adaptada de Rossini *et al.* (2012). O módulo dos pesos $\boldsymbol{a}$ de cada variável foi somado para cada componente retido até Δ, a fim de verificar

quais variáveis mantinham a maior variância, visto que essas são mais adequadas para aplicações de classificação de dados (ANZANELLO, *et al.*, 2009).

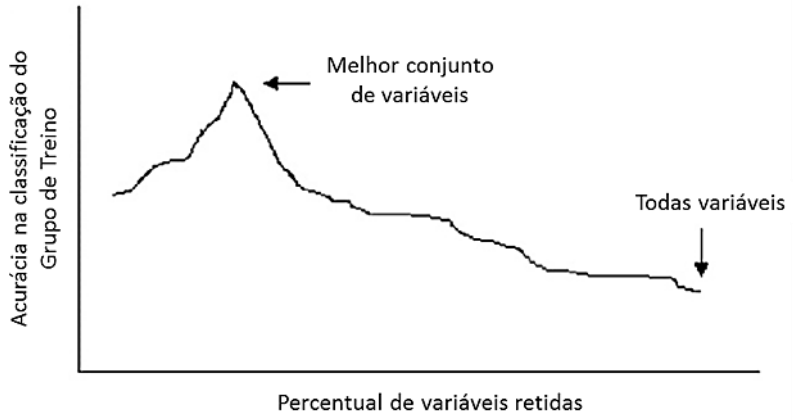
$$\mathbf{IIVj} = \sum_{j=1}^{\Delta} |a_{ij}|, i = 1, \dots, n \quad (8)$$

3.4. Seleção das variáveis mais relevantes para classificação dos produtos

Nessa etapa, objetivou-se definir as variáveis mais relevantes para classificação dos produtos em categorias de reposição. Para tanto, as variáveis foram organizadas em ordem decrescente de IIV, as observações descritas pelo conjunto completo de variáveis classificadas utilizando kNN e ADL, e a acurácia de classificação (razão entre o número de observações classificadas corretamente e o número total de observações) do grupo de treino calculada. Após o cálculo da acurácia, a variável menos relevante (com menor IIV) foi retirada da base de dados e o processo de classificação reiniciado, até restar apenas uma variável. Tal sistemática buscou a geração de um gráfico confrontando o número de variáveis retidas e a acurácia para cada método de classificação.

Com base no gráfico gerado (ilustração genérica na Figura 4), determinou-se o conjunto de variáveis responsável pela maior acurácia de classificação no grupo de treino. Tal conjunto foi escolhido com base na máxima acurácia.

Figura 4 – Perfil hipotético da acurácia do conjunto de treinamento após a eliminação de variáveis.
Fonte: ANZANELLO *et al.* (2009).



Objetivando não ancorar o algoritmo de classificação e os seus resultados em um único grupo de treino, visto que existiam produtos e locais com características distintas, os dados foram aleatoriamente embaralhados nas porções de $\boldsymbol{T_r}$ e $\boldsymbol{T_s}$ após a coleta de indicadores e os passos acima descritos executados novamente. O processo foi repetido 100 vezes e, ao final de cada

iteração, a acurácia e número de variáveis retidas foram armazenados, permitindo a montagem de uma base de dados mais sólida e assertiva quanto às variáveis selecionadas e a acurácia de cada um dos modelos.

3.5 Análise de frequência das variáveis e comparação dos métodos

Apoiando-se no conjunto de variáveis retidas para cada método de classificação em cada interação, foram criados histogramas para representar a frequência com que cada variável foi retida. Dessa forma, foi possível identificar as variáveis que oferecem maior contribuição na alocação de produtos aos modelos de reposição.

Para a comparação dos métodos, foram utilizados cinco critérios de avaliação, conforme a Tabela 2. O indicador de acurácia foi considerado como "maior é melhor", pois quanto maior a acurácia da classificação, melhor o desempenho que o modelo trará para os produtos/locais de reposição. O critério de número de variáveis com frequência igual a 100% e o número médio de variáveis retidas foram considerado como "menor é melhor", visto que quanto menos variáveis, menor o esforço de coleta e menor o trabalho computacional de manipulação de dados. Além disso, são contemplados dois indicadores de desvio, do tipo "menor é melhor", para avaliar a variação no comportamento dos dados e métodos ao longo das 100 iterações.

Tabela 2 – Avaliação comparativa dos métodos de classificação

Fonte: Elaborado pelos autores.

Critério	Tipo	kNN	ADL
Acurácia média de classificação (%)	Maior é melhor	-	-
Desvio da Acurácia de classificação (%)	Menor é melhor	-	-
Nº médio de variáveis retidas	Menor é melhor	-	-
Desvio do Nº médio de variáveis retidas	Menor é melhor	-	-
Nº de variáveis com frequência = 100%	Menor é melhor	-	-

4. Resultados

Esta seção aplica o método proposto aos produtos gerenciados por uma ferramenta de VMI de uma empresa de consultoria em *Supply Chain.* A empresa foco do estudo atua na área de tecnologia da informação aplicada à gestão da cadeia de suprimentos e demanda, possuindo quatro grandes áreas de atuação: planejamento avançado e sequenciamento de produção, gestão da demanda, gestão da distribuição e gestão da visibilidade, nas quais

oferece *softwares* e consultoria de negócios para auxiliar as empresas a aperfeiçoarem a gestão e operacionalização de seus processos. Está presente em todo o território nacional, com sede em São Leopoldo/RS e escritório de negócios em São Paulo/SP. Ao longo de dez anos de atuação, realizou mais de 250 projetos na área de *Supply Chain*, os quais apresentam *softwares* operando em três continentes.

Na área de gestão de distribuição, possui uma de suas soluções mais avançadas, o VMI (*Vendor Managed Inventory* – Estoque Gerenciado pelo Fornecedor). Nessa solução colaborativa, empresas das áreas de Cosméticos, Bens de Consumo, Alimentação Humana e Pet, Pneumáticos e Produtos Farmacêuticos realizam a gestão de abastecimento em parceria com todo o seu canal indireto e parte do seu canal direto no Brasil, reabastecendo automaticamente, em média, 10000 diferentes SKUs em 150 cidades nos 26 estados brasileiros. Frente a essa complexidade, a solução busca a melhoria contínua em suas análises e algoritmos, focando na melhor adequação dos cálculos a cada tipo de produto, empresa e região. Uma das principais dificuldades das empresas, nesse contexto, é determinar, junto à consultoria, quais os melhores algoritmos, métodos e parâmetros para cada um de seus produtos. Isso se deve ao fato de cada produto possuir diferentes públicos alvos, comportamentos de demanda e estratégias de comercialização e distribuição em cada região do Brasil.

A etapa inicial do método proposto define as variáveis a serem coletadas para inclusão dos produtos em categorias associadas a modelos específicos de reposição de estoque. Para tal, foram levantadas variáveis quantitativas (dados de venda, dados logísticos, dados de previsão de demanda e demais dados de mercado presentes na ferramenta) e subjetivas (nível de serviço desejado, estoque máximo parametrizado, algoritmo do estoque de segurança, entre outros dados que impactam a gestão de estoque e distribuição). Tais variáveis foram apresentadas aos consultores e gestores da área de Gestão de Distribuição – responsáveis pelo sistema VMI – e esses retiraram da análise, por meio de uma reunião expositiva, variáveis com pouco histórico ou baixa consistência de informações. As variáveis remanescentes são apresentadas na Tabela 3.

Na sequência, foram escolhidos os produtos e locais de reposição que teriam seus valores coletados para abastecer o método. Em conjunto com a equipe de consultores, optou-se por analisar parte da gama de produtos de uma empresa de bens de consumo (produtos de higiene pessoal e doméstica, limpeza e alimentos), além dos produtos voltados a consumidores finais

de uma empresa fabricante de pneus. Quanto aos locais de reposição, foram selecionadas todas as 26 capitais além de outras cidades com perfil de demanda e consistência de dados satisfatória. É importante ressaltar que nem todos os produtos são reabastecidos em todos os locais, devido a características regionais de mercado. Esse conjunto de análise gerou uma base de dados com 2000 registros.

Tabela 3 – Variáveis selecionadas para análise

Fonte: Sistema VMI da empresa em estudo.

Nº	Variável	Descrição da variável
1	Sell-out (R$)	Representa o resultado financeiro de um determinado produto em uma loja.
2	Sell-out (Un.)	Representa o número de unidades transacionadas de um determinado produto em uma loja.
3	Preço Médio	Representa o preço de venda de um determinado produto em uma loja.
4	Cancelamento (Un.)	Representa o número de vendas canceladas de um determinado produto em uma loja.
5	Bonificação (Un.)	Representa o número de unidades de produto que foram "vendidas" sem custo em uma loja.
6	Devolução (Un.)	Representa o número de devoluções de um determinado produto em uma loja.
7	Sell-out Médio (Un.)	Representa, em média, o número de unidades comercializadas de um produto quando existe uma venda para ele.
8	Dias com venda	Representa o número de dias em que um produto foi comercializado em uma loja.
9	Dias úteis sem venda	Representa o número de dias em que um produto não foi comercializado em uma loja.
10	Frequência (%)	Representa a frequência de vendas de um determinado produto em uma loja.
11	Lote Múltiplo (Un.)	Representa o lote de ressuprimento de estoque de um determinado produto para uma loja.
12	Venda Máxima (Un.)	Representa a maior quantidade vendida de um produto em um único dia em uma loja.
13	Número Transações	Representa o número de compras realizadas em uma loja que continham um determinado produto.
14	Base Clientes	Representa o número de clientes finais que adquirem o produto em uma determinada loja.
15	ABC	Representa a classificação ABC de um produto de acordo com seu volume de vendas.
16	VMD	Representa o número médio de unidades vendidas dentro de um período.
17	Desvio	Representa o desvio médio de unidades vendidas dentro de um período.
18	CV	Representa o coeficiente de variação (desvio/média) de um produto em uma loja.
19	Cobertura Objetivo	Representa o número de dias que se objetiva manter de estoques para cada produto em uma loja.
20	Nível de Serviço	Representa o coeficiente k da curva normal que representa o nível de serviço que se objetiva manter para cada produto em uma loja.
21	Lead Time	Representa o número de dias que um determinado produto demora a chegar a uma loja.
22	Desvio LT	Representa o desvio médio em dias que um determinado produto demora a chegar a uma loja.
23	Dias de Ruptura	Representa o número de dias que um produto manteve-se sem estoques em uma loja.
24	Cobertura Média (Dias)	Representa o número médio de dias de estoque para um produto em uma loja.

A coleta dos dados foi realizada diretamente no banco de dados do sistema, o qual já possuía as informações organizadas por produto e local de reposição, cabendo apenas uma extração simples e um filtro nos locais e produtos desejados.

Os modelos de reposição (classes) nos quais os produtos foram classificados (Tabela 4) foram determinados em conjunto com os consultores em consistência com as ferramentas e algoritmos já existentes no sistema de VMI em estudo.

Tabela 4 – Modelos de Reposição de Estoques (classes) considerados

Fonte: Sistema VMI da empresa em estudo.

Classe	Modelo
1	**Reposição Contínua Clássica**: Estoque de segurança calculado, supondo que a venda dos produtos obedeçam a uma distribuição normal e utilização do ponto de pedido.
2	**Reposição Push & Pull (Método do Mínimo/Máximo):** níveis de reposição são cadastrados e o algoritmo busca deixar o estoque sempre entre os níveis (momento de realizar a reposição e estoque máximo desejado)
3	**Reposição Periódica com Segurança Cadastrada**: reposição baseada em intervalo de tempo determinado com o estoque de segurança parametrizado em dias de acordo com a venda média diária do produto.
4	**Reposição Contínua com Segurança Cadastrada:** reposição baseada em ponto de pedido calculado com o estoque de segurança parametrizado em dias de acordo com a venda média diária do produto.
5	**Reposição Periódica baseada na previsão:** reposição baseada na previsão de vendas (*forecast)* do produto com intervalo de reposição cadastrado.

Na sequência, aplicou-se a ACP para definição dos pesos de cada variável em cada componente principal gerado. Foram retidos dois componentes principais com base no percentual de variância explicada pelos mesmos (83%), e análise do *Scree Graph*. O IIV calculado através dos pesos e equação (8) é apresentado na Tabela 5. Com base nesses valores, recalculados a cada interação, as variáveis foram organizadas em ordem decrescente de IIV, de modo a facilitar o processo iterativo de remoção de variáveis.

Tabela 5 – IIV das variáveis em estudo

Fonte: Elaborado pelos autores.

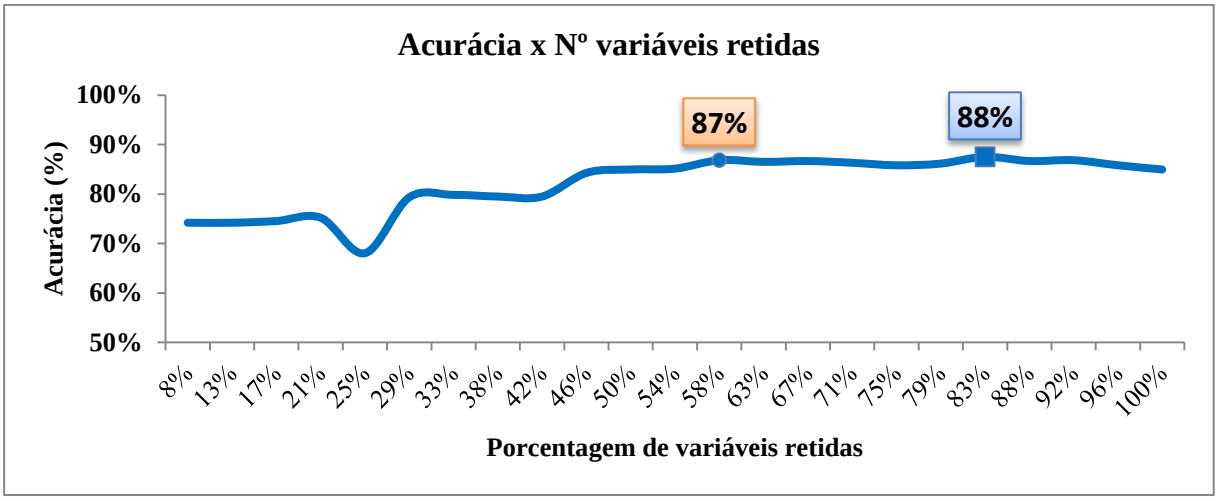
Variável	IIV
Sell-out Médio (Un.)	3,625
Sell-out (Un.)	3,625
VMD	3,560
Desvio	3,464
Dias com venda	3,447
Frequência (%)	3,447
Dias úteis sem venda	3,446

Sell-out (R$)	3,347
Devolução (Un.)	3,324
Venda Máxima (Un.)	3,167
Desvio LT	3,047
Base Clientes	3,018
Nível de Serviço	3,008
Bonificação (Un.)	2,951
Número Transações	2,546
Cobertura Objetivo	2,372
Preço Médio	2,177
Lote Múltiplo (Un.)	1,942
ABC	1,722
CV	1,468
Lead Time	1,431
Cancelamento (Un.)	0,714
Dias de Ruptura	0,499
Cobertura Média (Dias)	0,230

Na sequência, partiu-se para um processo iterativo de remoção de variáveis e classificação de dados utilizando os métodos kNN e ADL. A cada variável retirada de acordo com a ordem estabelecida pelo IIV, uma nova classificação era realizada (sempre com base nas variáveis restantes) e a acurácia calculada. Esse procedimento gerou o gráfico da Figura 5, na qual se observa que 20 variáveis devem ser retidas ao utilizar-se a ferramenta kNN. Ressalta-se, porém, que a acurácia obtida com 20 variáveis (88%) é muito similar à obtida com 14 variáveis (87%), por vezes preferível por necessitar um menor esforço de coleta de dados e menor desempenho computacional. Para o estudo em questão, sempre foi considerado o valor máximo da acurácia, independente do número de variáveis retidas. Destaca-se, também, a redução significativa da acurácia a partir de 11 variáveis restantes, o que denota a dificuldade de classificação dos produtos com menos de 50% das variáveis originais. Tal procedimento foi repetido 100 vezes para cada ferramenta de classificação (KNN e ADL).

Figura 5 – Acurácia x número de variáveis retidas

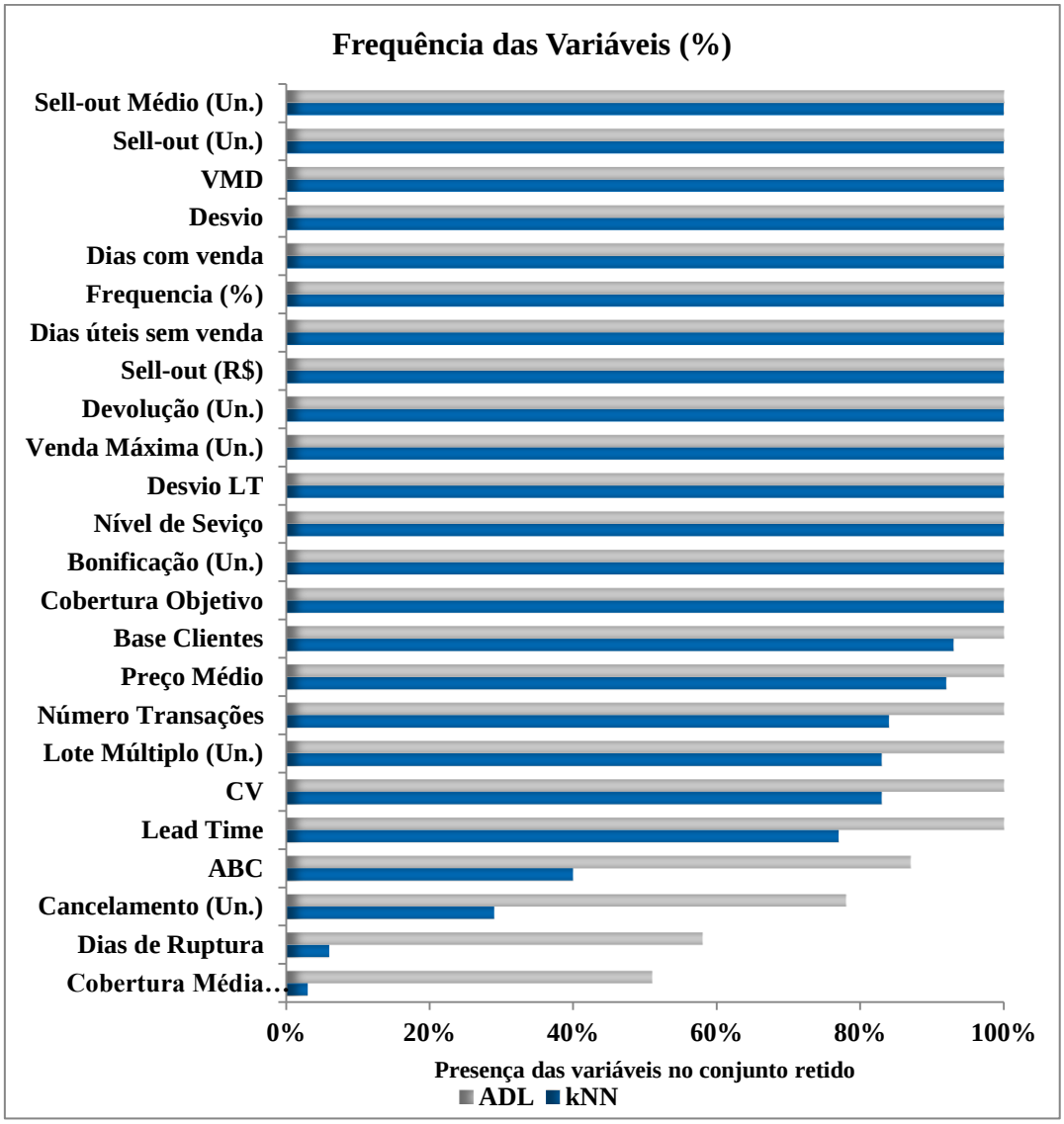
Fonte: Elaborado pelos autores.



A Figura 6 traz a frequência de retenção das variáveis para o KNN e ADL. Percebe-se que um número significativo de variáveis (mais de 50%) foi retido em todas as replicações do método proposto. Dentre essas, algumas possuem relação direta com os modelos de reposição de estoque, sendo utilizadas diretamente nos cálculos e análises do sistema. Outras, porém, relacionam-se mais ao perfil de mercado do produto nos locais, impactando de forma indireta, mas significativa, na escolha de cada modelo de reposição. Consideraram-se as variáveis com frequência igual a 100% como fundamentais na escolha do modelo de reposição de um determinado produto, devendo ser controladas e estudadas em maior profundidade para definições de novos produtos e locais de comercialização.

Figura 6 – Histograma de presença de variáveis no conjunto retido

Fonte: Elaborado pelos autores.



Quanto aos métodos de classificação, nota-se um comportamento mais conservador do método ADL na eliminação de variáveis, retendo aproximadamente 85% das variáveis presentes em todas as iterações – o método de kNN armazenou cerca de 55%. Das variáveis retidas por esse, todas também estavam presentes no método ADL, evidenciando uma consistência entre os dois métodos. De tal forma, evidencia-se um desempenho superior do kNN para a classificação de dados, o qual apresentou maior acurácia e menor desvio desse indicador. Além de obter um melhor grau de acurácia, trabalha com um menor número de variáveis retidas nas medições e também possui menor quantidade de atributos com presença constante nas classificações. Isso permite que se opere com uma base de dados reduzida e,

mesmo assim, alcance-se um resultado satisfatório. A Tabela 6 resume os resultados obtidos por cada um dos métodos.

Tabela 6 – Avaliação comparativa dos métodos de classificação

Fonte: Elaborado pelos autores.

Critério	Tipo	kNN	ADL
Acurácia média de classificação (%)	Maior é melhor	89,41%	87,1%
Desvio da Acurácia de classificação (%)	Menor é melhor	0,94%	1,31%
Nº médio de variáveis retidas	Menor é melhor	19,8	22,58
Desvio do Nº médio de variáveis retidas	Menor é melhor	2,45	1,47
Nº de variáveis com frequência = 100%	Menor é melhor	14	20

Por fim, ressalta-se que o número elevado de variáveis retidas pode estar relacionado à significativa diferença de perfil de vendas e negócio entre os produtos de higiene, limpeza e alimentos aos produtos da empresa fabricante de pneus. Com a acurácia obtida, porém, o método mostra-se robusto o suficiente para ambas as realidades de negócio.

5. Conclusões

Empresas cujas estruturas estão dispostas em cadeias de suprimento não podem mais apenas buscar um desempenho satisfatório dos seus processos internos. O desempenho de cada elo presente em uma cadeia é fundamental para o bom desempenho global dessa. O VMI busca passar a responsabilidade do gerenciamento de estoque dos clientes para o fornecedor, otimizando os níveis de estoque dentro da cadeia e aperfeiçoando o nível de serviço. Para o correto funcionamento dessa sistemática, é preciso que os modelos de reposição dos estoques para cada produto e local estejam bem definidos e parametrizados. Dentro desse contexto, a MVA, por meio de métodos com ACP, kNN e ADL auxilia na identificação das variáveis mais relevantes na identificação de um modelo de reposição para os produtos em análise.

Este artigo apresentou uma sistemática para classificação de produtos em modelos de reposição, identificando quais variáveis eram mais relevantes para tal, utilizando como base o *software* de VMI de uma empresa de consultoria especializada em *Supply Chain.* Para o estudo, foi necessária a montagem de uma base de dados de produtos em locais específicos de reposição. Isso permitiu a aplicação da ACP e levantamento da importância para cada variável. Após essa ponderação, os métodos de classificação foram executados, atribuindo classes às observações e medindo a acurácia e número de variáveis retidas no conjunto. As

variáveis retidas foram organizadas em um histograma de acordo com sua frequência, o que determinou sua importância na definição dos modelos de reposição.

Quando aplicado em um cenário prático, a ferramenta de classificação kNN obteve um melhor desempenho frente à ADL, gerando uma acurácia média de 90% - com um desvio de 0,9% - e retendo cerca de 80% das variáveis (55% em todas as 100 iterações). Já a ADL alcançou 87% de acurácia média – com desvio de 1,3%- porém retendo 95% das variáveis (83% em todas as iterações). Com tais resultados, foi possível identificar que 14 das 24 variáveis são fundamentais na determinação de um modelo de reposição de estoques para um produto em um local específico de reposição.

O estudo realizado utilizou como base apenas variáveis quantitativas e subjetivas referentes a duas empresas multinacionais de segmentos distintos (higiene pessoal e doméstica, alimentos e pneumáticos). Sugere-se, para estudos futuros, a consideração de variáveis qualitativas, tais como ação da concorrência, ações de marketing, perfil do consumidor alvo, objetivando uma análise segregada por empresa, a qual trará um viés mais estratégico e uma acurácia mais voltada à realidade e produtos de cada empresa.

6. Referências

ABDI, H., WILLIAMS, L. J. Principal Component Analysis. **WIREs Computational Statistics**, 2, 433-459, 2010.

ANZANELLO, M., ALBIN, L. S., CHAOVALITWONGSE, A. W. Selecting the best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratory Systems**, 97, 111-117, 2009.

BALLABIO, D., CONSONNI, V., MAURI, A., TODESCHINI, R. Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 3. Varible selection in classification. **Analytica Chimica Acta**, 657(2), 116 -122, 2010.

BARTHOLOMEW, D. J. The Interpretation of Multivariate Data. **International Encyclopedia of Education**, 3, 12-17, 2010.

CHOI, S-IL., OH, J., CHOI, C-HO., KIM, C. Input variable selection for feature extraction in classification problems. **Signal Processing**, 2011. Doi: http://dx.doi.org/10.1016/j.sigpro.2011.08.023

CHONG, I., ALBIN, S., JUN, C. A data mining approach to process optimization without an explicit quality function. **IIE Transactions**, 39(8), 795-804, 2007.

DAKHLAOUI, H., BARGAOUI, Z., BÁRDOSSY, A. Toward a more efficient Calibration Schema for HBV rainfall–runoff model. **Journal of Hydrology**, 444/445, 161–179, 2012.

DUDA, R. O., HART, P.E., STORK, D.G. Pattern Classification. New York: John Wiley and Sons, 2012.

FISHER, R. A. The Use of Multiple Measurements in Taxonomic Problems. **Annals of Eugenics**, 7(2), 179-188, 1936.

GAO Y., GAO, F. Edited AdaBoost by weighted kNN. **Neurocomputing**, 73(16-18), 3079-3088, 2010.

GERTHEISS, J., TUTZ, G. Feature selection and weighting by nearest neighbor ensembles. **Chemometrics and Intelligent Laboratiry Systems**, 99 (30-38), 2009.

GOLDRATT, E. M., COX, J. *The Goal*: a process of ongoing improvement. New York: North River Press, 1984.

HAIR JR, J. F., BLACK, W.C., BABIN, B. J., ANDERSON, R.E. Multivariate Data Analysis. New Jersey: Pearson, 2010.

HAN, J., KAMBER, M. *Data Mining*: Concepts and Techniques. San Francisco: Elsevier, 2006.

HARDLE, W., SIMAR, L. Applied Multivariate Statistical Analysis. **Metrika**, 64 (121-122), 2003

HE, S.-G., 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.

JOLLIFFE, I. T. Principal Component Analysis. (2nd ed.). New York: Springer, 2002.

KAZMIERCZAK NETO, E. Um estudo sobre a implementação de um sistema VMI em uma empresa do setor de cosméticos. Monografia de Especialização, Programa de Pós-graduação em Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

KITANI, E. C., THOMAZ, C. E. Análise de discriminantes lineares para modelagem e reconstrução de imagens de face. **Anais do XXVII Congresso da SBC**, Rio de Janeiro, 2007.

MALINOWSKI, E. R. Factor Analysis in Chemistry. New York: John Wiley and Sons, 2002.

MARTINEZ, A. M., KAK, A. C. PCA versus LDA. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 23(2), 228-233, 2001.

MOROPOLOU, A., POLIKRETI, K. Principal Component Analysis in monument conservation: three application examples. **Journal of Cultural Heritage**, 10(1), 73-81, 2009.

PARK, C. H., LEE, M. On applying linear discriminant analysis for multi-labeled problems. **Pattern Recognition Letters**, 29(7), 878–887, 2008.

RAO, K. R., LAKSHMINARAYANAN, S. Partial correlation based variable selection approach for multivariate data classification methods. **Pattern Recognition**, 42 (7-16), 2006.

RENCHER, A. C., CHRISTENSEN, W. F. Methods of Multivariate Analysis. New Jersey: Wiley, 2012.

ROSSINI, K., ANZANELLO, M., FOGLIATTO, F, Seleção de atributos em avaliações sensoriais descritivas. **Produção**, 22(3), 380-390, 2012.

SAHMER, K., QANNARI E. M. Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, 19 (141-145), 2008.

SHREVE, J., SCHNEIDER, H., SOYSAL, O. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. **Decision Support Systems**, 52(1), 247-257, 2011.

SILVA, G. R. Desenvolvimento de um modelo de simulação para avaliação de desempenho de uma Cadeia de Suprimentos multicamadas do ramo de mineração através da adoção da estratégia colaborativa VMI (Vendor Managed Inventory). Dissertação, Escola de Politécnica, Universidade de São Paulo, São Paulo, 2010.

SU, MING-YANG. Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification. **Journal of Network and Computer Applications**, 34, 722–730, 2011.

SYMS, C. Principal Component Analysis. Oxford: Academic Press, 2008.

TAVARES, S. Modelo de Estoque Gerenciado pelo Fornecedor (VMI) Aplicado ao Varejo de Materiais de Construção no Setor de Revestimentos Cerâmicos. Dissertação, Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2003.

WOLD, S., SJÖSTRÖM, M., ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and intelligent laboratory systems**, 58 (109-130), 2001.

XU, P., BROCK, N., PARRISH, R.S. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. **Computational Statistics & Data Analysis**, 53 (1674-1687), 2008.

YIN, R.K. Estudo de caso planejamento e métodos. 3ª Edição. Porto Alegre: Bookman, 2003.

Proposition of a variable selection framework for product replenishment

Abstract

Companies integrated in supply chains seek initiatives to improve the chain's overall performance. The VMI (Vendor Managed Inventory) enables better results when it comes to manage and balance stocks along the chain. For that matter, VMI frameworks must rely on well defined parameters and algorithms aimed at allocating products to replenishment local characteristics. This paper presents a method to classify products in replenishment categories based on Principal Component Analysis (PCA) along with two classification algorithms, k-Nearest Neighbor and Linear Discriminant Analysis. The model seeks to identify the most relevant variables for assigning products to the most appropriate replenishment model. When applied to a real situation, the proposed method yielded 90% classification accuracy when retaining average 55% of the original variables.

Key Words: SCM (Supply Chain Management), VMI (Vendor Managed Inventory), PCA (Principal Component Analysis), kNN (k-Nearest Neighbor), LDA (Linear Discriminant Analysis), MVA (Multivariate Data Analysis)