PRODUÇÃO ONLINE
REVISTA CIENTÍFICA ELETRÔNICA DE ENGENHARIA DE PRODUÇÃO
ISSN 1676-1901

PROGRAMAÇÃO DE PRODUÇÃO VIA SELEÇÃO DE VARIÁVEIS E SIMULAÇÃO DE MONTE CARLO

PRODUCTION SCHEDULING THROUGH VARIABLE SELECTION AND MONTE CARLO SIMULATION

Marco Antonio Campetti* E-mail: marcocampetti@terra.com.br
Michel J. Anzanello* E-mail: anzanello@producao.ufrgs.br
Guilherme V. Etcheverry* E-mail: getcheverry@producao.ufrgs.br
* Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS

Resumo: Este artigo apresenta uma sistemática de programação de produção que combina seleção de variáveis para clusterização e Simulação de Monte Carlo em cenários multiprodutos. A sistemática é composta por quatro etapas: (*i*) estruturação e validação de dados, (*ii*) seleção de variáveis para clusterização de produtos com características físicas similares, (*iii*) simulação de demanda através de Simulação de Monte Carlo, e (*iv*) análise de resultados e aderência da sistemática à realidade. A sistemática proposta foi aplicada em uma empresa do segmento alimentício com grande variedade de produtos ofertados, permitindo simplificar a programação da produção e maximizar os lucros decorrentes da comercialização dos produtos.

Palavras-chaves: Seleção de variáveis. Clusterização. Simulação de Monte Carlo.

Abstract: This paper proposes a production scheduling approach which combines variable selection for clustering and Monte Carlo simulation in multiproduct scenarios. A four steps approach is proposed to perform: (i) data collection, (ii) variable selection for clustering products with similar physical characteristics, (iii) demand simulation using Monte Carlo simulation, and (iv) analysis of results and adherence of systematic to real scenarios. The proposed methodology was applied in a food industry with a wide variety of products, making production planning easier and maximizing profits.

Keywords: Variable Selection, clustering, Monte Carlo Simulation

1 INTRODUÇÃO

O conceito de valor para o cliente vai muito além de qualidade e preço justo, para justificar a escolha de um produto, se baseia em uma série de outras dimensões que caracterizam tal produto. Em muitos segmentos, tais dimensões são apenas qualificadoras, pois apenas permitem que o produto seja percebido pelo consumidor, mas não necessariamente escolhido. A agregação de diversas dimensões no momento da escolha cria um cenário complexo e de incerteza para as empresas. Esta incerteza, em parte, é gerada pela variedade de produtos ofertados, os quais requerem esforços extras da área de operações, em particular, do

planejamento de produção (KIM e MAUBORGNE, 2005; COSTA, 2007).

Desta forma, torna-se cada vez mais necessário que as empresas conheçam detalhadamente o mercado, bem como a si próprias, para melhor utilizar os recursos produtivos, quando a capacidade produtiva se aproxima de seu limite. No caso de empresas diversificadas, torna-se necessário priorizar um determinado produto em detrimento de outro, o que pode ser facilitado através da formação de famílias de produtos com características semelhantes e alocação dessas famílias aos recursos disponíveis. A geração de tais famílias usualmente contemplam suas características físicas e as necessidades de processamento dos diversos modelos de produtos (CHOPRA e MEINDL, 2003; RAFAELI, 2009).

Neste contexto, a utilização de uma sistemática estruturada de identificação das variáveis mais relevantes para a inserção dos produtos em famílias de acordo com suas demandas produtivas e características físicas torna-se fundamental para aprimoramento da programação de produção. A utilização de todas as variáveis disponíveis em procedimentos de clusterização tende a reduzir a qualidade dos grupos formados, visto que variáveis ruidosas e irrelevantes comprometem a eficiência dos algoritmos de clusterização, conduzindo a alocações equivocadas. Tais agrupamentos podem então ser integrados a outras ferramentas (por exemplo, simulação de eventos discretos) que viabilizem a geração de cenários voltados à maximização do desempenho dos recursos produtivos (HAIR JR et al.. 2003).

Este artigo propõe uma sistemática de seleção de variáveis com vistas à formação de famílias de produtos com demandas produtivas similares, a qual é integrada a uma ferramenta de simulação com o intuito de aprimorar a programação da produção diária. A primeira etapa é constituída por um procedimento de seleção em conjunto com um índice para avaliação da qualidade do agrupamento gerado. A cada variável omitida, a qualidade do agrupamento é avaliada e a variável omitida de menor índice é eliminada. O número de *clusters* é inicialmente estimado através de clusterização hierárquica (apoiada em dendograma), e a efetiva inserção dos produtos às famílias é realizada através do algoritmo *k-means* (ferramenta de clusterização do tipo não-hierárquica) (HAIR JR. et al., 2003). Os grupos gerados pelas variáveis selecionadas são então analisados em cenários produtivos diversos através da Simulação de Monte Carlo (SMC).

O artigo está estruturado como segue: a seção 2 apresenta o referencial

teórico acerca dos fundamentos de clusterização, seleção de variáveis e SMC. As seções 3 e 4 apresentam, respectivamente, o método proposto e os resultados. Por fim, a seção 5 traz as conclusões do estudo e sugestões de trabalhos futuros.

2 REFERENCIAL TEÓRICO

Em diversos cenários produtivos e gerenciais, é mais oportuno gerir produtos e recursos através da definição de grupos homogêneos. A construção destes grupos requer ferramentas apropriadas para identificar a similaridade entre os objetos agrupados, dentre as quais se destaca a clusterização (HAIR JR. et al., 2003; RODRIGUES e SELLITTO, 2009; CHEZNIAN e SUBASH, 2011).

Clusterização é o processo de alocação em grupos de objetos com características similares, de tal forma que objetos alocados a outros grupos apresentem características distintas. Tal similaridade é usualmente mensurada através de métricas apropriadas, destacando-se as medidas de distância entre as observações (MIMAROGLU e ERDIL; SANTHISREE e DAMODARAM, 2011).

A clusterização permite a abstração e interpretação de grandes quantidades de dados pela construção de um significado comum não aparente para cada grupo ou *cluster*. Apesar de não aparente, o objetivo da técnica é revelar o agrupamento natural que existe em uma série de dados (KASHEF e KAMEL; JAIN, 2010). Por vezes, a alocação de recursos em cenários produtivos é mais eficiente através do agrupamento, visto que determinadas características dos agrupamentos formados demandam abordagens específicas (ANZANELLO e FOGLIATTO, 2011).

Existem dois procedimentos tradicionais de clusterização, os hierárquicos e os não hierárquicos. A diferença entre os métodos está na forma como as observações são alocadas aos grupos. Os procedimentos hierárquicos constroem os agrupamentos através de árvore hierárquica (dendograma), avaliando progressivamente a similaridade entre os grupos e observações. Os procedimentos não-hierárquicos, por sua vez, alocam observações em uma único movimento baseado nas distâncias entre as observações (HAIR JR et al., 2003).

Os métodos hierárquicos apresentam relações de hierarquia entre agrupamentos formados em estágios subsequentes, isto é, os resultados de um estágio anterior de agrupamento são considerados no estágio seguinte. Dividem-se

em dois grupos: os aglomerativos e os divisivos, diferenciando-se pela sequência de execução. Enquanto procedimentos aglomerativos consideram cada observação como um agrupamento individual, os divisivos consideram um único agrupamento contendo todas as observações. À medida que o procedimento é executado, os aglomerativos diminuem o número de *clusters*, através de agrupamento por maior semelhança. Já nos divisivos, o processo é inverso: a partir do aglomerado inicial, observações são extraídas por critérios de diferença, formando agrupamentos menores e mais homogêneos (HAIR JR. et al, 2003; SANTHISREE e DAMODARAM, 2011).

Nos procedimentos não hierárquicos, o ponto de partida é a definição do número de agrupamentos a serem gerados (*k*, número de *clusters*). Ao iniciar o procedimento, são geradas aleatoriamente *k* sementes, isto é, grupos a partir dos quais são calculados seus valores médios (centroides) e, então, a distância Euclidiana entre as sementes e as observações é calculada pela equação (1). O método busca, aleatória e iterativamente, distribuir as observações a *k* grupos, de forma que a distância total entre os dados de um grupo e o seu respectivo centroide, para todos os grupos, seja minimizada. Na etapa seguinte, as observações são realocadas aos *k* grupos de acordo com maior proximidade aos centroides iniciais, que na sequência são recalculados. Este processo iterativo acontece até que as realocações não sejam mais necessárias, gerando o valor mínimo na função objetivo (HAIR JR et al., 2003; LIU et al., 2009).

$$D_{\text{Eucliadeana}} = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \quad (1)$$

De forma geral, os métodos não hierárquicos apresentam melhores resultados em relação a dados atípicos e desempenho de clusterização do que os métodos hierárquicos, porém tais resultados estão diretamente ligados à qualidade dos dados analisados e ao número de agrupamentos a ser gerado. Deve-se salientar que cada cenário gerado é independente dos outros, e deriva somente do número de *clusters* definido inicialmente (HAIR JR. et al., 2003; SANTHISREE e DAMODARAM, 2011).

Uma prática recomendada na formação de agrupamentos é a combinação de procedimentos, fazendo uso das virtudes de cada um dos métodos acima descritos. Normalmente, inicia-se a busca pela identificação da quantidade ideal de *clusters*

valendo-se de um procedimento hierárquico. Uma vez definido o número de *clusters* a ser considerado, executa-se o processo de agrupamento através do procedimento não hierárquico. Esta combinação aprimora o procedimento de clusterização (HAIR JR. et al, 2003). Aplicações recentes de ferramentas de clusterização podem ser encontradas em sistemas produtivos, estudos de *marketing*, estudos biológicos e filtragem de correio eletrônico do tipo *spam* (LIU et al. 2009; MIMAROGLU e ERDIL; ANZANELLO e FOGLIATTO; CAI e SUN; CHEZHIAN e SUBASH; MOHAMMAD, 2011).

Concluído o processo de clusterização, é fundamental avaliar a qualidade do agrupamento gerado (HAIR JR. et al., 2003). Uma forma de medição consiste no cálculo do Índice Silhouette (IS), através da equação (2), a qual é composta por dois elementos: um referente à distância média entre a observação e demais observações alocadas ao mesmo *cluster a(j)*, e outro referente à distância média entre a observação em questão e as observações no agrupamento mais próximo *b(j)*. O IS oscila entre [-1; +1]; quanto mais próximo de +1, maior é a qualidade do agrupamento. Observações com índice próximo a +1 indicam adequada alocação ao *cluster* de tais observações, visto que as mesmas estão distantes dos outros *clusters*; valores próximos de -1 indicam que a observação foi alocada ao *cluster* errado (MIMAROGLU e ERDIL; ANZANELLO e FOGLIATTO, 2011).

$$IS\ (j) = \frac{b(j) - a(j)}{Max\{b(j), a(j)\}} \qquad (2)$$

Cabe ressaltar que, a escolha das variáveis utilizadas na formação dos agrupamentos deve estar diretamente relacionada aos objetivos da análise. Recomenda-se parcimônia na escolha das variáveis de agrupamento, visto que a inserção de variáveis irrelevantes no procedimento pode comprometer a qualidade dos agrupamentos gerados (NAVEIRO e PEREIRA FILHO, 1992; MIMAROGLU e ERDIL; ANZANELLO e FOGLIATTO; MOHAMMAD, 2011).

A seleção de variáveis de processo é fundamental para controle de processos produtivos, assim como para caracterização correta de produtos (ANZANELLO, 2009). Da mesma forma, grande dificuldade existente na avaliação de bens se deve à heterogeneidade de variáveis e atributos que os caracterizam (STEINER et al., 2008). A eficácia na seleção de variáveis para caracterização dos objetos responde por importante parcela do sucesso no agrupamento formado. Desta forma, os objetivos da clusterização não podem ser separados da seleção de variáveis (HAIR

JR. et al., 2003), uma vez que a escolha de uma variável significa selecionar uma dimensão específica do objeto em estudo em detrimento a outras (SENRA et al., 2007). Villanueva (2006) define seleção de variáveis como a identificação de subconjunto de variáveis que levam a resultados satisfatórios no reconhecimento de padrões em bases de dados consistindo de elevado volume de informações.

Há duas abordagens clássicas para os métodos de seleção de variáveis: *Forward Selection* e *Backward Elimination*. A abordagem *Forward Selection* parte da incorporação progressiva das variáveis ao modelo, enquanto que a sistemática *Backward Elimination* consiste em um método regressivo, partindo do grupo total de variáveis candidatas, e então eliminando sequencialmente aquelas tidas como irrelevantes. Ambos os procedimentos de seleção de variáveis têm provado eficiência e estabilidade. Uma variação das abordagens acima, o *Stepwise*, baseia-se na inserção e remoção alternada das variáveis ao modelo de acordo com a contribuição das mesmas para desempenho do agrupamento (GUYON e ELISSEEFF, 2003; ANZANELLO e FOGLIATTO; 2011).

Variáveis de natureza qualitativa requerem atenção adicional no que tange a seus escores ou pesos, pois podem comprometer a precisão da representação das observações. De acordo com a ponderação resultante entre os pesos, diferentes variáveis podem ficar em evidência em um mesmo cenário (ANZANELLO e FOGLIATTO, 2011). Como alternativa para este inconveniente, pode-se usar mão da premissa de que variáveis com grandes variações apresentam maior poder de clusterização (STEINLEY e BRUSCO, 2008), requerendo mesmo ou maior nível de atenção do que as demais variáveis.

Dentro do campo de seleção de variáveis, são possíveis duas abordagens: filtragem e envoltória (NAGATANI et al., 2010). Na abordagem de filtragem, a ideia central é uma pré-seleção e exclusão de variáveis irrelevantes, segundo critérios definidos pelo usuário. Já na abordagem envoltória, a escolha das variáveis faz parte de um algoritmo de aprendizagem que demanda recursos computacionais. Este algoritmo usa de uma taxa de reconhecimento que busca uma característica preditora fornecida pelo usuário. Esta busca tem por objetivo encontrar o menor subgrupo de variáveis que melhor caracteriza o conjunto geral de dados de acordo com a característica preditora (GUYON e ELISSEEFF, 2003; VILLANUEVA, 2006; NAGATANI et al., 2010; HORTA e ALVES, 2012).

Em termos de desempenho, os envoltórios apresentam capacidade de generalização maior, mas a um custo maior. Por outro lado, os métodos de filtragem têm custos menores e maior facilidade de operacionalização, podendo comprometer o desempenho da seleção resultante. O uso misto de abordagens de seleção, isto é, fazer uso de metodologia de filtragem como pré-processamento e então uso de técnica envoltória ou embutida, eliminar origens de ruído e de sobreajustamento através de filtragem, para então usar mecanismo de melhor desempenho, o envoltório (GUYON e ELISSEEFF, 2003; VILLANUEVA, 2006)

Através de uso de metodologia de seleção de variáveis, Costa Filho e Poppi (2002) constataram significativa melhora nos resultados em modelos multivariados. Nagatani et al. (2010) atribuem melhoria de desempenho em subconjunto selecionados à redução de complexidade dos modelos gerados pela escolha correta das variáveis mais relevantes. Entretanto, é difícil medir o desempenho de todos os subconjuntos de variáveis possíveis. No mesmo sentido, Senra et al. (2007) propõem a necessidade de uma análise prévia por parte de especialistas das variáveis disponíveis, antes mesmo de definição do método.

A simulação é a estruturação de um modelo que visa representar uma operação ou situação do mundo real. Este modelo utiliza diversos parâmetros, detalhando o sistema em análise com determinada fidelidade. O intuito destas técnicas é suportar decisões quando a realização de pilotos ou testes reais é inviável, sejam por questões de segurança, financeira, recursos tecnológicos ou temporais (AMANIFARD et al., 2011). Entretanto, a qualidade das análises geradas por modelos simulados, assim como seus resultados, está diretamente ligada à qualidade dos dados de entrada e estruturação do modelo (CATELLI; SARAIVA JÚNIOR et al., 2010).

Dentre as técnicas de simulação disponíveis na literatura, destaca-se a Simulação de Monte Carlo (SMC). Esta técnica é baseada na geração de números aleatórios e probabilidade de ocorrência de valores associados ao fenômeno em análise. Em casos de difícil modelagem ou formulação, dados de entrada podem ser substituídos e representados por padrões estatísticos, sobre os quais a SMC é aplicada (ZAPATA et al., 2004; SARAIVA JÚNIOR et al. 2010).

A SMC é operacionalizada através de um processo iterativo, onde são gerados, aleatória e sucessivamente, N valores de uma variável de entrada

específica, aplicados ao modelo em análise, resultando em uma distribuição de probabilidade com média e desvio padrão de ocorrências para o evento (modelo) estudado (ZAPATA et al., 2004). Variáveis aleatórias são então geradas e rebatidas contra a função de distribuição acumulada. Tal conversão é repetida por um número elevado de vezes, de forma que os valores gerados possam representar a frequência de ocorrência do fenômeno em análise. Os dados gerados são então inseridos na modelagem de interesse, e cenários alternativos são avaliados de acordo com o propósito da análise.

Em termos práticos, os resultados obtidos para uma variável aleatória não devem condicionar/influenciar ou ser condicionados/influenciados pelos resultados de outras variáveis aleatórias. Faz-se necessário, também, conhecer precisamente as distribuições de probabilidade dos dados de entrada do sistema modelado (ZAPATA et al., 2004).

Dentre as vantagens da SMC, destaca-se que os dados de entrada podem apresentar qualquer tipo de distribuição, podendo analisar cenários de forma ágil (alterando somente dados de entrada) e, devido a não existência de um algoritmo único para SMC, pode-se ajustar o procedimento de simulação da forma mais oportuna à situação. Complementarmente, Pamplona e Silva (2005) e Saraiva Júnior et al. (2010) entendem a SMC como um método numérico estocástico universal para a solução de problemas matemáticos, propondo sua utilização ao gerenciamento de riscos.

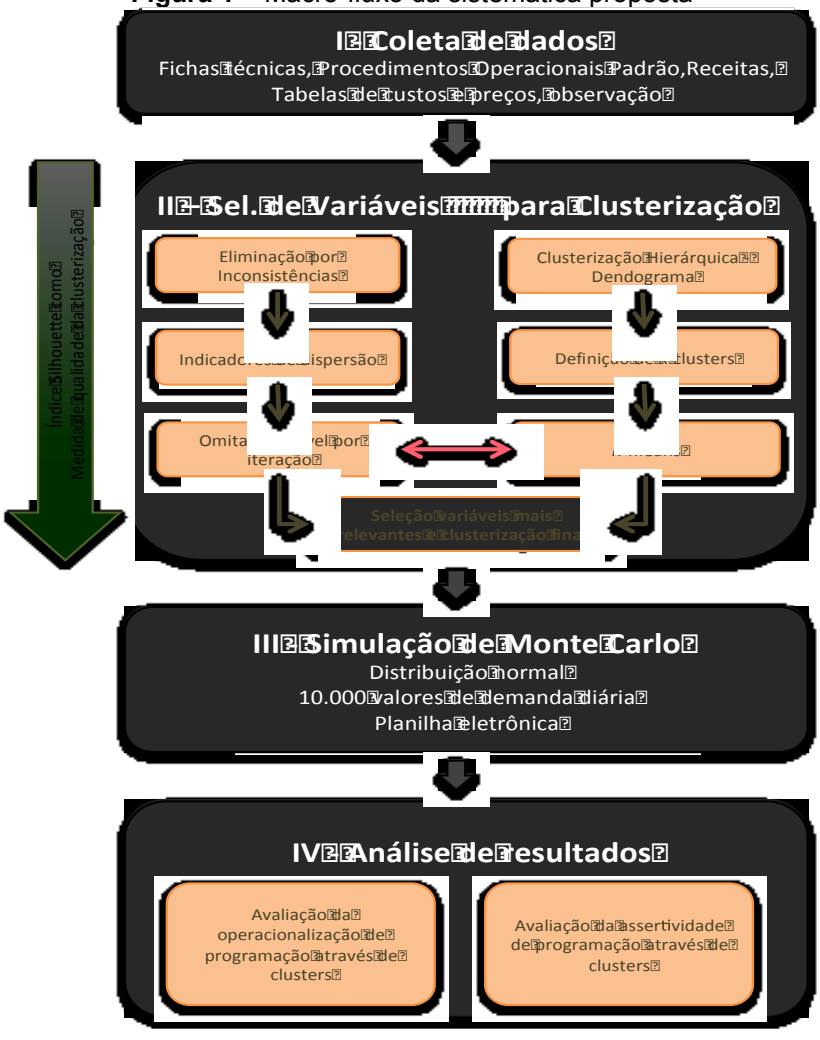
A SMC vem sendo usada nas mais diversas áreas. Zapata et al. (2004) utilizaram a técnica para avaliar a confiabilidade de um sistema de distribuição de energia elétrica. Através deste, foi possível constatar a versatilidade e flexibilidade do sistema. Já Saraiva Júnior et al. (2010) constataram a utilidade da SMC ao utilizá-la no auxílio à definição de mix de produção de plásticos. Através da simulação foram feitas análises econômico-financeiras relativas à utilização de recursos produtivos, as quais viabilizaram definir a melhor política de mix de produtos. Como ponto forte do método, além de flexibilidade e simplicidade de aplicação, os autores ressaltam a possibilidade de utilização de conceitos de áreas de conhecimento distintas de forma integrada.

3 MÉTODO

Este artigo se constitui em uma pesquisa aplicada de abordagem quantitativa, orientado à solução de um problema específico, o qual utiliza ferramentas multivariadas para seleção de variáveis e simulação. O objetivo geral do estudo é a simplificação do processo de programação da produção. Para isto são avaliadas e quantificadas as variáveis que têm influência sobre a caracterização do processo.

A metodologia proposta é dividida em quatro etapas: (*i*) coleta de dados, (*ii*) seleção de variáveis para clusterização de observações em grupos, (*iii*) aplicação da SMC nos grupos gerados, e (*iv*) análise dos resultados obtidos. Essas etapas são apresentadas na Figura 1 e detalhadas na sequência.

Figura 1 – Macro-fluxo da sistemática proposta



3.1 Coleta de dados

Na primeira etapa, se realiza a coleta dos dados que descrevem as características gerenciais e operacionais dos modelos de produtos analisados. Tais variáveis permitem a formação de grupos de produtos com necessidades de processamento similares. Exemplos incluem variáveis associadas à forma de consumo do produto em análise, estocagem, características de ordem financeira, processos de preparo, recursos operacionais necessários e tipo de produto, entre outras. Tais variáveis podem ser coletadas de bancos de dados históricos da empresa analisada e oriundas de opiniões de especialistas.

Para dados de ordem financeira, processos de preparo, recursos operacionais e estocagem, os dados são coletados de procedimentos operacionais padrão e tabelas de custo padrão. As dimensões de consumo e tipo de produto são extraídas diretamente da análise de propriedades físicas dos produtos.

3.2 Seleção de variáveis para clusterização

Esta etapa é subdividida em três passos. O primeiro passo aplica um procedimento hierárquico de clusterização sobre os dados com intuito de estimar o número adequado de *clusters*. O segundo passo refere-se a uma pré-seleção de variáveis através de indicadores de variabilidade e opiniões de especialistas, com objetivo de reduzir o número de variáveis candidatas. Por fim, ocorre a etapa de seleção de variáveis para clusterização definitiva, utilizando uma abordagem de omissão de uma variável por iteração. Esses passos são agora detalhados.

No primeiro passo, identifica-se o número recomendado de *clusters*, *k*, a serem formados através de um dendograma (ferramenta típica em procedimentos hierárquicos de clusterização), no qual se visualiza agrupamentos progressivos das observações (HAIR JR. et al., 2003). Tal valor é utilizado como parâmetro de entrada na clusterização não-hierárquica *k-means* (HAIR JR. et al., 2003; SANTHISREE e DAMODARAM, 2011). Na sequência, agrupam-se as observações utilizando-se todas as variáveis através do algoritmo *k-means*. A qualidade da clusterização gerada é medida através do Índice Silhouette (IS) (ANZANELLO e FOGLIATTO, 2011), o qual será utilizado como valor de referência para avaliar

aprimoramentos nos procedimentos de clusterização decorrentes da seleção de variáveis.

O segundo passo realiza-se uma pré-seleção de variáveis, visando dois objetivos: (*i*) redução do número de variáveis que serão investigadas e, consequentemente, (*ii*) redução do número de iterações realizadas nos passos seguintes. O segundo objetivo é mais sutil, porém, de acordo com qualidade dos dados coletados, pode se tornar fundamental, visto que avalia a consistência e qualidade dos dados e variáveis candidatas. Tal seleção pode ser feita de duas formas: através de opinião de especialistas quanto à consistência das variáveis candidatas, ou utilizando-se indicadores de variabilidade das variáveis coletadas. Tais técnicas podem ser realizadas em conjunto ou individualmente.

O indicador de variabilidade utilizado neste trabalho é o coeficiente de variação (CV), calculado pela razão entre desvio padrão e média. Da mesma forma, outras medidas, como amplitude e variância, podem indicar o poder de clusterização de uma variável. O princípio é simples: variáveis que apresentam os maiores valores de amplitude, variância, desvio-padrão e coeficiente de variação tendem a ter melhor desempenho de clusterização. A cada eliminação de variáveis do grupo de candidatas, é repetida a clusterização através do *k-means* e calculado o IS médio para o agrupamento gerado. Este procedimento é executado até que o IS resultante do agrupamento seja inferior ao anteriormente calculado. Nesse instante, inicia-se um procedimento exaustivo de seleção das melhores variáveis remanescentes através do procedimento de omissão de uma variável por iteração (GUYON e ELISSEEFF, 2003; STANLEY e BRUSCO, 2008).

O procedimento de omissão de uma variável por iteração visa identificar o menor conjunto possível de variáveis relevantes para a formação dos grupos de produtos, sem perder qualidade no agrupamento. Nesse procedimento, uma variável é momentaneamente omitida a cada iteração e uma sistemática de clusterização (do tipo *k-means*) é realizada. A cada omissão de variável, a qualidade de clusterização gerada pela ausência daquela variável é medida através do IS. A variável responsável pelo maior IS ao ser omitida é eliminada do banco de dados, visto que os resultados do agrupamento são melhores quando tal variável não é incluída na análise.

Na sequência, o mesmo procedimento de omissão de uma variável por vez é

executado sobre o conjunto de variáveis remanescentes. Esse processo é repetido até restar apenas uma variável. O procedimento acima pode ser repetido para um intervalo de valores de *k* (número de *clusters*) considerado adequado por especialistas de processo, caso seja diferente daquele encontrado com o procedimento hierárquico acima descrito.

3.3 Simulação baseada nos grupos gerados

Nesta etapa, utiliza-se a SMC para identificação do mix de produção que maximize a receita, ao menor custo médio de mercadoria vendida possível, através da minimização de escassez e perdas. Tal função-objetivo, apresentada na Equação (3), é testada nos grupos de produtos formados na etapa anterior.

As simulações são realizadas em planilha eletrônica. A Tabela 1 mostra a estrutura das simulações para cada grupo formado no passo anterior, onde números aleatórios são gerados e então convertidos em demanda diária via SMC. Através da variação do lote de produção diário, pode-se avaliar o melhor cenário para minimizar escassez e perdas e maximizar vendas e receitas.

Tabela 1 – Estrutura das simulações

Iteração	Produção				Itens disponíveis				Demanda				Venda				Não Venda				Perda			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	130	38	21	160	130	38	21	160	86	40	13	104	86	38	13	104	0	2	0	0	44	0	8	56
2	130	38	21	160	174	38	29	216	115	28	10	131	115	28	10	131	0	0	0	0	15	10	11	29
3	130	38	21	160	145	48	32	189	103	25	26	138	103	25	26	138	0	0	5	0	27	13	0	22
4	130	38	21	160	157	51	21	182	101	25	24	99	101	25	24	99	0	0	3	0	29	13	0	61
9.997	130	38	21	160	172	52	25	240	63	21	16	69	63	21	16	69	0	0	0	0	67	17	5	91
9.998	130	38	21	160	197	55	26	251	79	39	18	123	79	38	18	123	0	1	0	0	51	0	3	37
9.999	130	38	21	160	181	38	24	197	91	32	16	121	91	32	16	121	0	0	0	0	39	6	5	39
10.000	130	38	21	160	169	44	26	199	124	26	13	95	124	26	13	95	0	0	0	0	6	12	8	65

Os cenários testados, Conservador, Misto e Agressivo, utilizam o tamanho do lote de produção diária como variável de entrada e demanda diária como variável aleatória, conforme a Equação (3). Tal relação representa o lucro (*L (c)*) gerado pela venda dos produtos, receita (*V* x *P*), custo de escassez (*ÑV* x *P*) e custo de mercadoria disponível (*D* x *C*), onde V significa quantidade de venda efetiva, *P* preço de venda, *ÑV* não-venda ou diferença entre demanda e disponibilidade de mercadoria (nos casos em que a demanda é maior), *D* a quantidade de itens

disponíveis e *C* o custo de produção destes. Para cada *cluster*, são calculados valores de preço e custo médios, ponderados de acordo com a participação de cada produto no histórico de vendas.

$$L\,(c) = V \times P - \tilde{N}V \times P - D \times C \qquad (3)$$

3.4 Análise e revisão de resultados

O objetivo principal de agrupar produtos em famílias é facilitar o processo de programação da produção para uma grande quantidade de produtos com diferentes características. A programação através de famílias deve gerar resultados semelhantes, em relação à maximização de vendas e minimização de perdas, aos resultados obtidos caso o procedimento fosse realizado considerando os produtos individualmente.

4 RESULTADOS E DISCUSSÃO

A sistemática proposta foi aplicada em uma empresa do segmento alimentício que dispõe de aproximadamente 80 produtos, entre doces e salgados, quentes e resfriados, bebidas e alimentos. Alguns destes produtos são fabricados a partir de ingredientes base, enquanto que outros são preparados através da combinação de insumos e matérias-primas pré-manufaturados. Há, ainda, um pequeno grupo (5% sobre o total de produtos ofertados), no qual a fabricação é terceirizada, sendo realizada apenas a comercialização pela empresa. Para oferecer tais produtos, são necessários mais de 250 itens de matéria prima e insumos. Tais materiais podem ser específicos de um produto ou compartilhados por diversos.

Além da variabilidade entre produtos, insumos e matérias-primas no que tange a volumes, pesos e condições de armazenamento, outro ponto crítico é o prazo de validade ou tempo para consumo. O prazo de validade médio dos produtos é de 2,4 dias, ou seja, após o preparo ou fabricação, os produtos devem ser consumidos no prazo médio máximo de 2,4 dias, sendo que mais de 80% dos produtos devem ser consumidos no dia de fabricação ou preparo. Este cenário faz com que o planejamento de produção seja crucial para minimizar perdas e desperdícios sem implicar em perdas de vendas por escassez.

4.1 Coleta de dados

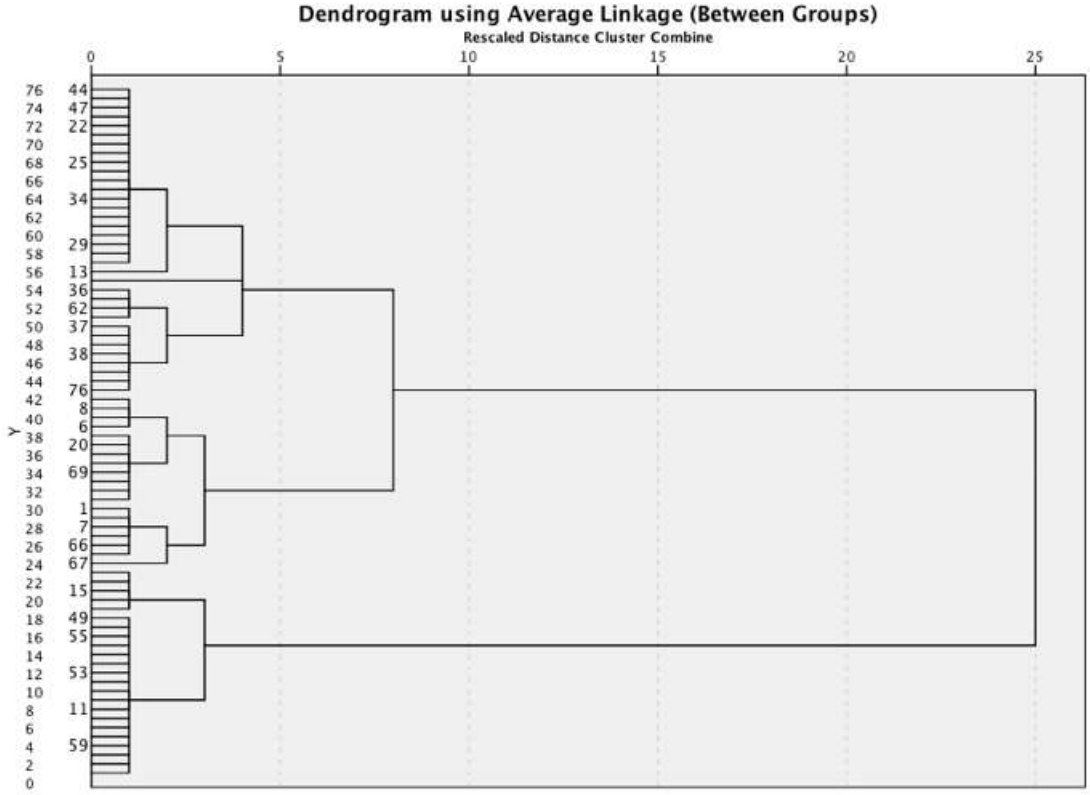
Os dados coletados foram organizados em uma matriz descritiva dos produtos, a qual apresenta variáveis que descrevem os produtos em análise com vistas ao seu agrupamento. Inicialmente foram levantadas 45 variáveis para 76 produtos.

Em relação aos dados de vendas, foi coletado um histórico de 90 dias para todos os produtos. Estes dados são importantes em duas frentes. A primeira diz respeito à operacionalização da simulação, visto que a variável aleatória é a demanda. De tal forma, é imprescindível conhecer profundamente o comportamento desta variável para correta geração de números aleatórios e incidência de demanda. Em segundo plano, percebe-se a demanda como potencial variável relevante para o procedimento de clusterização.

4.2 Seleção de variáveis para clusterização

A etapa de seleção de variáveis para clusterização foi iniciada através de procedimento hierárquico e opiniões de especialistas (proprietária da empresa, coordenadora e duas colaboradoras), identificando-se como sendo quatro o número adequado de *clusters*, conforme mostra o dendograma da Figura 2. Neste dendograma, os produtos a serem agrupados encontram-se no eixo vertical, enquanto que o eixo horizontal representa a distância entre os centros dos *clusters* formados. Cada ramificação indica um *cluster* e as observações que o compõem. Embora uma primeira análise tenha sugerido três como o número adequado de grupos, tal valor foi ajustado para quatro com base na opinião de especialistas de processo: a proprietária da empresa e mais três colaboradoras.

Figura 2 - Processo Hierárquico de Clusterização: identificação de nº de *clusters*



Na Tabela 2 é apresentada a alocação preliminar dos produtos aos *clusters* resultantes do processo hierárquico sobre as 45 variáveis iniciais. Percebe-se que o *Cluster* 1 concentra mais da metade das observações (41 produtos), ao passo que o *cluster* 4 contém apenas seis produtos. Tais agrupamentos poderão alterados, já que as etapas subsequentes do método realizarão novas clusterizações com base nas variáveis selecionadas.

Tabela 2 – Alocação dos produtos aos *clusters* (via dendograma)

	Clusters																	
	1									2			3				4	
Produtos	1	2	3	4	5	6	7	8	9	10	11	12	27	28	29	30	14	15
	13	19	20	48	49	50	51	52	53	21	22	23	31	32	33	34	16	17
	54	55	56	57	58	59	60	61	62	24	25	26	35	36	37	38	18	43
	63	64	65	66	67	68	69	70	71	44	45	46	39	40	41	42		
	72	73	74	75	76					47								

A etapa de clusterização não hierárquica considera as 45 variáveis iniciais. Primeiramente foram identificadas e eliminadas dez variáveis com possível inconsistência na opinião de especialistas, restando assim 35 variáveis. Dentre tais

inconsistências, ressalta-se a ausência de dados para determinados produtos, não consenso entre os respondentes, irrelevância da variável e magnitudes duvidosas das variáveis (provavelmente em decorrência da inserção de informações equivocadas nos bancos de dados). A cada execução do *k-means*, cada uma das dez variáveis apontadas como inconsistentes foi eliminada e o IS recalculado. O IS inicial de 0,3723 com as 45 variáveis sinaliza baixa qualidade nos agrupamentos. O IS de 0,7693, alcançado após a exclusão das dez variáveis indica que tais variáveis contribuíam negativamente para a qualidade da clusterização, confirmando a opinião dos especialistas. A Tabela 3 apresenta a evolução do IS a cada eliminação das variáveis tidas como inconsistentes.

Tabela 3 – Evolução do IS após cada eliminação de variáveis inconsistentes

Nº variáveis	45	44	43	42	41	40	39	38	37	36	35
IS	0,3723	0,3885	0,4036	0,4073	0,4187	0,4288	0,3921	0,4068	0,4259	0,4368	0,7693

Concluída a primeira fase de eliminação, calcula-se o CV das 35 variáveis remanescentes, conforme a equação (4). Tal coeficiente é utilizado como ordenador de sequência de exclusão de variáveis. A variável com menor CV a cada rodada é eliminada, partindo-se da premissa que variáveis com maior dispersão permitem agrupamentos mais precisos, como afirmando por Stanley e Brusco (2008). A cada eliminação, os produtos são reagrupados em quatro *clusters* e o IS recalculado.

$$Coeficiente\ de\ variação = \frac{\hat{\sigma}}{\hat{\mu}}, \begin{cases} \hat{\sigma} = desvio\ padrão \\ \hat{\mu} = média\ aritmética \end{cases}$$

$$Coeficiente\ de\ variação = \frac{\hat{\sigma}}{\hat{\mu}}, \begin{cases} \hat{\sigma} = desvio\ padrão \\ \hat{\mu} = média\ aritmética \end{cases}$$

(4)

O uso do CV possibilita a redução de 35 para 10 variáveis candidatas. Na Tabela 4 são apresentados os valores de CV para as 25 variáveis eliminadas nesta etapa, bem como o IS gerado. Percebe-se um incremento na qualidade de clusterização à medida que as variáveis com menor variabilidade são eliminadas. Contudo, quando a variável 31 é retirada, o IS apresenta decréscimo significativo, indicando que o processo de eliminação de omissão de uma variável por iteração deve começar.

Tabela 4 – Eliminação de variáveis por coeficiente de variação

Variável	**13**	**12**	**22**	**6**	**9**	**5**	**8**	**20**	**26**	**25**	**11**	**28**	**30**
CV	15%	19%	24%	25%	26%	27%	27%	28%	31%	33%	37%	40%	40%
IS	0,7736	0,7785	0,7787	0,7836	0,7882	0,7950	0,8002	0,8002	0,8008	0,8014	0,8015	0,8027	0,8035
Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Variável	**10**	**7**	**4**	**24**	**23**	**21**	**27**	**2**	**15**	**16**	**3**	**35**	**31**
CV	40%	42%	45%	45%	47%	49%	49%	57%	57%	61%	62%	72%	73%
IS	0,8036	0,8040	0,8044	0,8054	0,8069	0,8069	0,8093	0,8094	0,8096	0,8113	0,8114	0,8129	0,6353
Ordem	14	15	16	17	18	19	20	21	22	23	24	25	26

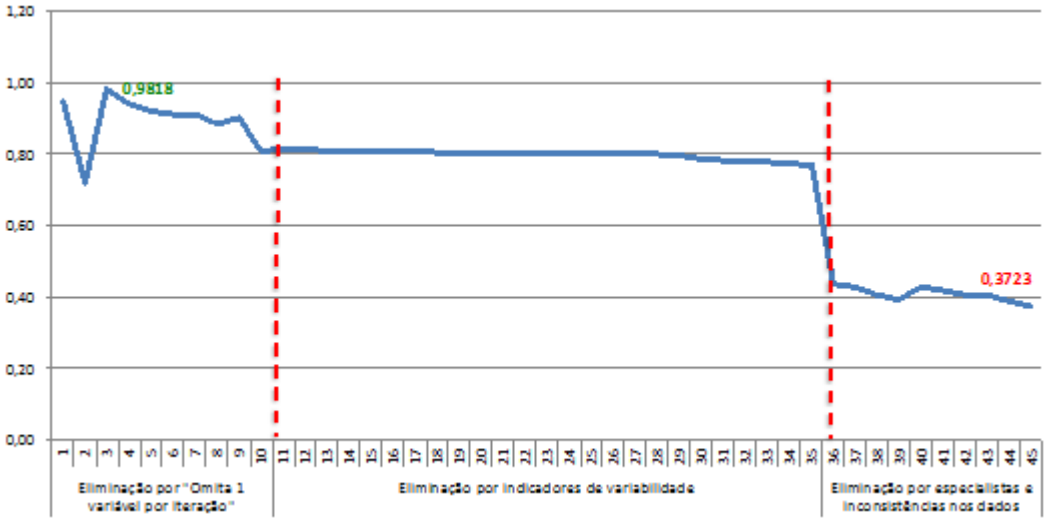
O procedimento iterativo de omissão de uma variável por vez foi inicializado com as variáveis restantes: 31, 18, 14, 17, 29, 19, 32, 1, 34 e 33, sendo os valores médios de IS para cada iteração apresentados na Tabela 5.

Tabela 5 – IS's gerados pelo procedimento de omissão de uma variável por vez

	Índice Silhouette (IS)									
Variáveis	31	0,6353	0,5597	0,4372	0,5292	0,5868	0,6900	0,8141	0,9818	0,4935	0,9456
	18	0,8133	0,9566	0,9670	0,9724	0,9761	0,9786	0,9809	0,9818	0,9456	
	14	0,8138	0,9571	0,9675	0,9729	0,9767	0,9791	0,9814	0,9818		
	17	0,8145	0,8201	0,8382	0,9733	0,9771	0,9795	0,9818			
	29	0,8139	0,9575	0,9679	0,9733	0,9771	0,9795				
	19	0,8141	0,9576	0,9681	0,9734	0,9772					
	32	0,8153	0,9589	0,9695	0,9748						
	1	0,7868	0,9606	0,9711							
	34	0,8202	0,9657								
	33	0,9554									
	Média	0,8083	0,8993	0,8858	0,9099	0,9118	0,9213	0,9396	0,9818	0,7196	0,9456

O maior valor de IS médio é alcançado quando três variáveis (31, 18 e 14) são retidas. É possível identificar que há certa complementariedade e talvez interação entre as variáveis, pois a exclusão de uma pode gerar resultados diferentes em outras. Exemplo disto é a variável 1, que na primeira iteração é a segunda que mais contribui para a qualidade do agrupamento (o IS é reduzido para 0,7868 quando essa variável é omitida da clusterização). Por outro lado, na terceira iteração, esta mesma variável torna-se a mais dispensável, com o valor de IS chegando a 0,9711 quando ela é omitida.

Figura 3 – Evolução da qualidade de clusterização medida pelo IS



A Figura 3 apresenta a evolução da qualidade das clusterizações em razão da eliminação das variáveis nos três passos. Tal redução facilita o processo de caracterização e apropriação dos produtos, inclusive novos, às famílias. A redução de variáveis chegou a mais de 93%, passando das 45 originais para 3 (variáveis 31, 18 e 14, em ordem decrescente de importância).

Na Tabela 6 são apresentados os *clusters* e os produtos que os compõem. Abaixo do número de cada produto são mostrados os percentuais de participação em volume de cada produto em seu respectivo *cluster*.

Tabela 6 – Alocação final de produtos aos *clusters*

	Clusters																			
	1					2					3					4				
Produtos	1	2	3	5	6	13	21	22	23	24	4	36	37	38	39	9	10	11	12	14
	42%	24%	11%	1%	1%	3%	4%	3%	4%	8%	22%	2%	9%	10%	2%	1,1%	32%	2%	9%	2%
	7	8	19	20	63	25	26	27	28	29	40	41	42	43	62	15	16	17	18	48
	6%	6%	2%	1%	0,2%	5%	3%	4%	6%	5%	2%	0%	1%	2%	18%	2%	1%	1%	2%	7%
	64	65	66	67	68	30	31	32	33	34	75	76				49	50	51	52	53
	1%	0,4%	0,2%	0,5%	1%	7%	2%	3%	2%	2%	26%	6%				0,1%	3%	12%	1%	0,4%
	69	71	72	73		35	44	45	46	47						54	55	56	57	58
	0,6%	0,2%	2%	0,2%		4%	5%	4%	7%	5%						0,5%	0,8%	0,4%	0,1%	0,4%
						70	74									59	60	61		
						7%	7%									9%	1%	14%		

Os agrupamentos gerados foram avaliados por especialistas, tendo sido considerados satisfatórios em termos de similaridades dos produtos inseridos em cada grupo. Em termos práticos, os *clusters* 2, com 22 produtos - Salgados - e o *cluster* 3, com 12 produtos - Sanduíches - são formados somente por alimentos,

enquanto que os grupos 1, com 19 produtos - Bebidas e Doces Elaborados - e o *cluster* 4, com 23 produtos - Bebidas e Doces de simples elaboração - são formados por alimentos e bebidas. A empresa agrega produtos em famílias pela natureza destes, por exemplo, há famílias de bebidas quentes e geladas, industrializadas e naturais, alimentos doces e salgados, quentes e frios. No total, são nove as famílias utilizadas habitualmente pela empresa para agrupar os produtos.

Através da sistemática utilizada neste estudo, nenhuma destas famílias foi segregada, isto é, produtos pertencentes à mesma família foram alocados ao mesmo *cluster*. Isto indica que a utilização da sistemática identificou similaridades não aparentes sem conflitar com aquelas oriundas do conhecimento técnico da empresa. Este ponto se torna importante, pois em caso de oscilação de demanda específica de um produto, os demais produtos da mesma família (utilizada pela empresa) podem absorver a variação de demanda, minimizando o desvio entre valor estimado e executado para o *cluster* em questão.

4.3 Simulação

Para realização da simulação de demanda dos produtos agrupados em *clusters*, foram utilizados os dados históricos de 90 dias de demanda individual de cada produto. Inicialmente, foram levantados preço e custo unitário de obtenção de cada produto e, num segundo momento, foram calculados a média e o desvio padrão de demanda por produto. Estas informações foram utilizadas para obter os parâmetros de preço, custo unitário de obtenção, média e desvio padrão de cada *cluster* através de ponderação pela participação de cada produto nas vendas totais de seu respectivo *cluster*, conforme composições e percentuais apresentados na Tabela 6.

A Tabela 7 apresenta a média e o desvio padrão gerados pela simulação de cada grupo e os percentuais de participação de cada *cluster* no histórico de vendas em unidades, assim como os dados de preço e custo unitário ponderados de cada *cluster*. Para obtenção dos dados de média e desvio padrão de demanda diária, foram simulados 10.000 valores seguindo uma distribuição normal para cada *cluster*, conforme estrutura apresentada na Tabela 1.

Tabela 7 – Dados das simulações

			Cluster			
			1	2	3	4
SMC		Média (un)	101	29	17	123
		Desvio Padrão (un)	22,39	6,05	5,19	23,06
Histórico de vendas		Participação %	37,4%	10,8%	6%	45,8%
		Custo médio ponderado ($)	0,66	1,87	1,14	1,10
		Preço médio ponderado ($)	3,71	5,98	3,31	3,58

Estes dados foram inseridos na Equação (3) em três cenários distintos de lotes de produção diária: conservador, agressivo e misto, para os quatro *clusters*. O cenário conservador visa minimizar perdas, mesmo que implique em aumento de escassez de produtos. Já no cenário agressivo, o objetivo é realizar todo potencial de demanda maximizando a receita, sem considerar o risco de incremento das perdas. Por último, com o cenário misto, buscou-se encontrar o limite entre maximização das vendas, sem incorrer em maiores perdas. Em termos quantitativos, no cenário conservador foi considerado um lote de produção diária referente a 70% dos valores de média apresentados na Tabela 7, no cenário misto os lotes de produção diária foram as médias da Tabela 7, e no agressivo foram considerados valores 30% superiores a estes. Tais cenários foram definidos em acordo com os especialistas.

A Tabela 8 apresenta um resumo com os valores de lote de produção em cada cenário, bem como média e desvio padrão de *L (c)*. Percebe-se que uma redução no lote de produção diária ocasiona perdas consideráveis em *L (c)*, já que este recua mais de 66% do cenário conservador em relação ao cenário misto (de $593 para $199), introduzindo ainda maior variabilidade (desvio padrão de $111). Em contrapartida, o cenário mais agressivo incrementa *L (c)* médio em menos de 1%, também aumentando a variabilidade (desvio padrão de $119).

Como segundo plano de análise, a sistemática auxilia em uma melhor compreensão sobre os custos de perdas e de escassez inerentes a cada cenário. Na Tabela 8 são apresentados tais valores. Percebe-se que, apesar do baixo custo de perda no cenário conservador, o custo de escassez é equivalente a *L (c)*, lembrando que o custo de escassez representa a receita adicional não realizada devido à falta de produtos. Nota-se que mesmo em um cenário mais agressivo, o custo de escassez é inevitável. Tal comportamento é oposto ao custo de perda, que

representa prejuízo referente aos produtos disponíveis não vendidos e consequentemente descartados.

Tabela 8 – Comparação dos cenários produtivos simulados

	Conservador					Misto					Agressivo				
	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total
Produção diária (un)	70	20	11	86	187	100	29	16	123	268	130	38	21	160	349
L(c) médio ($)	95	24	4	76	199	243	90	21	239	593	263	87	22	223	595
Desvio padrão de L(c) ($)	73	33	14	76	111	50	22	10	49	74	76	36	15	83	119
Custo médio de perdas ($)	1	0	0	1	2	6	4	2	10	22	20	16	6	40	82
Custo médio de escassez ($)	119	58	19	4	200	34	15	7	4	60	4	1	2	4	11

4.4 Análise e verificação de resultados

Para avaliar os benefícios da simulação de demanda com base nos *clusters* gerados foi realizada uma SMC para demanda individual dos produtos. Para todos os produtos foi rodada a SMC, com base em suas distribuições de probabilidade individuais. Em ambos os cenários – produtos agrupados em *clusters* e individualmente - foi utilizada a distribuição normal e os resultados finais comparados.

Na Tabela 9 são apresentados os dados de demanda simulada para os produtos individualmente e para os grupos formados, lembrando que o objetivo principal de agrupar os produtos é aprimorar o processo de programação da produção sem, no entanto, gerar resultados financeiros inconsistentes àqueles obtidos pela avaliação individual dos produtos. Os produtos inseridos nos *clusters* 2 e 3 são caracterizados por significativas oscilações nos seus históricos de vendas (o que é amortecido pela simulação com base nos grupos, os quais apoiam-se em médias de demanda e desvios para cálculo), fator que explica a divergência entre dados de demanda simulados agrupada e individual. Por outro lado, os resultados dos *clusters* 1 e 4 são satisfatórios, pois estes representam as maiores vendas em volume e retorno financeiro, além de apresentar maior regularidade nas vendas.

Tabela 9 – Média e desvio padrão da simulação agrupada e individual

	SMC Agrupados				**SMC Individual**			
	1	2	3	4	1	2	3	4
Média (un)	101	29	17	123	106	37	19	130
Desvio Padrão (un)	22,39	6,05	5,19	23,06	14,2	7,4	6,1	23,6

Dois aspectos interferem no desempenho da sistemática ao simular-se cenários produtivos para itens clusterizados: (*i*) representatividade dos *clusters* nos volumes totais de vendas e (*ii*) representatividade dos produtos nas vendas totais dos *clusters*. Oscilações de demanda pontuais e específicas de determinados produtos, como sazonalidades, promoções ou ações específicas afetam diretamente estes dois parâmetros. Caso tais alterações sejam verificadas, é preciso ajustar os percentuais de participação dos produtos nos seus respectivos *clusters* (e destes no total) antes de proceder à SMC.

A Tabela 10 compara os dados simulados no cenário misto com os dados reais de vendas do mês de setembro de 2012. O erro absoluto médio foi de 3 unidades (+4 unidades no *cluster* 1 e –1 unidade no *cluster* 3), que representa um desvio de 1,03%.

Tabela 10 – Comparação da simulação em cenário misto com a demanda real

	SMC Mista				Dados reais Set/2012				Erro %			
	1	2	3	4	1	2	3	4	1	2	3	4
% vendas	37,4%	10,8%	6,0%	45,8%	36,6%	10,8%	6,6%	46,0%	2,2%	0,0%	-9,1%	-0,4%
Quantidade simulada (un)	109	31	18	132	105	31	19	132	3,8%	0,0%	-5,3%	0,0%

5 CONCLUSÃO

Este artigo apresentou uma sistemática combinando método de seleção de variáveis para clusterização e Simulação de Monte Carlo (SMC) com vistas ao aprimoramento do processo de programação de produção. Como primeiro passo foi realizada a coleta e estruturação de dados de produtos, variáveis e demanda. A segunda etapa inicia pela identificação de quantidade adequada de *clusters* através de procedimento hierárquico de clusterização e dendograma. Na sequência, reduz-se a quantidade de variáveis para clusterização, em um primeiro momento através de indicador de dispersão e finalmente por um procedimento iterativo de omissão de uma variável por vez.

A qualidade das clusterizações durante a etapa de seleção de variáveis é medida através do Índice Silhouette. Uma vez definidas as variáveis de clusterização, os produtos são agrupados e a SMC é estruturada valendo-se dos grupos gerados. A simulação tem como objetivo avaliar cenários distintos de produção com vistas à maximização de lucros.

A sistemática proposta apresentou uma relação esforço-desempenho/resultado satisfatória, quando bem ajustada à situação que se deseja avaliar. A necessidade de seleção de variáveis para clusterização ficou evidente para redução de esforços e recursos computacionais e aumento da qualidade dos agrupamentos formados. Quanto à SMC, ressalta-se que falhas de ajuste nos percentuais de composição dos *clusters* podem causar erros consideráveis, porém é fácil ajustá-lo quando cada *cluster* é simulado individualmente.

Por fim, a sistemática proposta tem como vantagens a facilidade e flexibilidade de geração de cenários alternativos de análise, evidenciando seu caráter prático. O desvio de 1% da simulação no cenário misto em relação às vendas reais demonstra o potencial do método quando aplicado ao planejamento de produção, visando a redução de perdas por desperdício (perecibilidade dos produtos) e por escassez de produtos para a venda.

Desdobramentos futuros incluem a análise da possível sinergia ou interferência entre as variáveis na etapa de seleção de variáveis, assim como o efeito que a escala e sua cardinalidade impõem à clusterização. Outro ponto consiste na investigação de como ajustar os parâmetros de participação dos produtos nos *clusters* e destes no total, quando em cenários de alta oscilação de demanda.

REFERÊNCIAS

AMANIFARD, N.; RAHBAR, B.; HESAN, M. numerical simulation of the mitral valve openning using smoothed particles hydrodynamics. **Proceedings of the World Congress in Engineering**, v. 3, july, 2011.

ANZANELLO, M. J. Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes. **Gestão e Produção**, São Carlos, v. 16, n. 4, p. 526-533, out./dez. 2009

ANZANELLO, M. J.; FOGLIATTO, F. S. Selecting the best variables for grouping mass-customized products involving worker's learning. **Int. J. Production Economics,** v. 130, p. 268–276, 2011. http://dx.doi.org/10.1016/j.ijpe.2011.01.009

CAI, Y.; SUN, Y. Spirit - Tree: hierarquical clustering analysis of millions of 16s rRNA pyrosequences in quasilinear computaional time. **Nucleic Acids Research**, v.39, n. 14, 2011. http://dx.doi.org/10.1093/nar/gkr349

CATELLI, A. **Controladoria**: uma abordagem da gestão econômica – GECON. São Paulo. Atlas, 2010.

CHEZNIAN, V. U.; SUBASH, T.; Hierarchical sequence clustering algorithm for data mining. **Proceedings of the World Congress on Engineering,** v.3, jul. 2011

CHOPRA, S.; MEINDL, P. **Gerenciamento da cadeia de suprimentos**: estratégia, planejamento e operação. São Paulo. Prentice Hall, 2003

COSTA FILHO, P. A.; POPPI, R. J. Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação de glicose, maltose e frutose. **Quim. Nova,** v. 25, n. 1, p. 46-52, 2002 http://dx.doi.org/10.1590/S0100-40422002000100009

COSTA, F. J. **A influência do valor percebido pelo cliente sobre comportamentos de reclamação e boca a boca**. Tese (Doutorado em Administração de Empresas). Fundação Getúlio Vargas - EAESP. 2007

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157-1182, 2003.

HAIR JR., J. F. et al. **Análise multivariada de dados**. Prentice-Hall. São Paulo. 2003

HORTA, R. A. M.; ALVES, F. J. C. Aplicação de técnicas de data mining para o entendimento da política de financiamento das empresas brasileiras. CONGRESSO ANPCONT, 4., 2012. **Anais eletrônicos.** Disponível em: http://www.anpcont.com.br/site/docs/congressoIV/04/MFC161.pdf

JAIN, A. K. Data clustering: 50 years beyond the K-means. **Pattern Recognition**, v. 31, p. 651-666, 2010. Disponível em: www.elsevier.com/locate/patrec http://dx.doi.org/10.1016/j.patrec.2009.09.011

KASHEF, R.; KAMEL, M. S. Cooperative clustering. **Pattern Recognition**, v. 43, p. 2315–2329, 2010. Disponível em: www.elsevier.com/locate/pr http://dx.doi.org/10.1016/j.patcog.2009.12.018

KIM, W. C.; MAUBORGNE, R. **A estratégia do oceano azul**. São Paulo. Campus, 2005.

LIU, M.; JIANG, X.; KOT, A. C. A multi-prototype clustering algorithm. **Pattern Recognition**, v. 42, p. 689-698, 2009. Disponível em: www.elsevier.com/locate/pr http://dx.doi.org/10.1016/j.patcog.2008.09.015

MIMAROGLU, S.; Erdil, E. Combining multiple clusterings using similarity graph. **Pattern Recognition**, v. 44, p. 694–703, 2011. Disponível em: www.elsevier.com/locate/pr http://dx.doi.org/10.1016/j.patcog.2010.09.008

MOHAMMAD, N. T. A fuzzy clustering aproach to filter spam. **Proceedings of the World Congress on Engineering,** v. 3, July, 2011.

NAGATANI, T.; OZAWA, S.; ABE, S. Fast variable selection by block addition and block deletion. **Journal of Intelligent Learning Systems and Applications**, v. 2, p. 200-211, 2010. http://dx.doi.org/10.4236/jilsa.2010.24023

NAVEIRO, R. M.; PEREIRA FILHO, I. C. A análise de grupamentos: uma contribuição à padronização do projeto. **Produção,** v. 2, n. II2, p. 157, março 1992.

PAMPLONA, E. O.; SILVA, W. F. Contribuição da simulação de Monte Carlo na projeção de cenários para gestão de custos na área de laticínios. In: CONGRESSO INTERNACIONAL DE CUSTOS, 9., 2005. **Anais...** Florianópolis, SC, Brasil .

RAFAELI, L. **Análise envoltória de dados como ferramenta para avaliação de desempenho relativo**. Dissertação (Mestrado em Engenharia). Universidade Federal do Rio Grande do Sul. 2009

RODRIGUES, D. M.; SELITTO, M. A. Análise do desempenho de fornecedores de uma empresa de manufatura apoiada em análise de aglomerados. **Produção,** v. 19, n. 1, p. p. 055-069, 2009.

SANTHISREE, K; DAMODARAM, A. SSM-DBSCAN and SSM-OPTICS: Incorporating a new similarity measure for density for density based clustering of web usage data. **International Journal on Computer Science and Engineering** (IJCSE) v. 3, n. 9, september 2011.

SARAIVA JÚNIOR, A. F.; RODRIGUES, M. V.; COSTA, R. P. Simulação de Monte Carlo aplicada à decisão de mix de produtos. **Produto e Produção,** v. 11, n. 2, p. 26-54, jun. 2010

SENRA, L. F. A. C. et al. Estudo sobre métodos de seleção de variáveis em DEA. **Pesquisa Operacional,** v.27, n.2, p.191-207, maio/agosto 2007.

STEINER, M. T. A. et al. Métodos estatísticos multivariados aplicados à engenharia de avaliações. **Gestão e Produção**, São Carlos, v. 15, n. 1, p. 23-32, jan./abr. 2008

STEINLEY, D.; BRUSCO, M. A new variable weighting and selection procedure for K-means cluster analysis. **Multivariate Behavioral Research**, v.43, n. 1, p. 77–108, 2008. http://dx.doi.org/10.1080/00273170701836695

VILLANUEVA, W. J. P. **Comitê de máquinas em predição de séries temporais**. Dissertação (Mestrado em Engenharia Elétrica e de Computação). Universidade Estadual de Campinas, 2006

ZAPATA, CARLOS J.; PIÑEROS, LUIS C.; CASTAÑO, DIEGO A. El método de simulación de Montecarlo en estudios de confiabilidad de sistemas de distribución eléctrica . **Scientia Et Technica**, v.10, n. 24, p. 55-60 mayo, 2004.

Artigo recebido em 20/05/2013 e aceito para publicação em 15/04/2014
DOI: http://dx.doi.org/10.14488/1676-1901.v14i2.1603