UM NOVO MÉTODO PARA SELEÇÃO DE VARIÁVEIS PREDITIVAS COM BASE EM ÍNDICES DE IMPORTÂNCIA DAS VARIÁVEIS

Juliano Zimmer

Universidade Federal do Rio Grande do Sul

Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre – RS

zimmer@producao.ufrgs.br

Michel J. Anzanello

Universidade Federal do Rio Grande do Sul

Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre – RS

anzanello@producao.ufrgs.br

RESUMO

A identificação das variáveis mais importantes para predição é relevante no controle de processos industriais. A regressão PLS (*Partial Least Squares*) vem sendo amplamente utilizada em procedimentos de seleção de variáveis por sua capacidade de lidar com variáveis em grande número, correlacionadas e ruidosas. Este artigo propõe um método para identificar o melhor subconjunto de variáveis de processo para a predição das variáveis de resposta. Indicadores de importância das variáveis foram desenvolvidos a partir de parâmetros da regressão PLS para guiar a eliminação das variáveis ruidosas e irrelevantes e testados em termos de seu desempenho. Ao ser aplicado em 5 bancos de dados industriais, o método utilizou 31% das variáveis originais e aumentou a acurácia de predição em 6% quando todas as variáveis são utilizadas. O método proposto também superou o desempenho de predição do tradicional método *Stepwise*.

PALAVARAS CHAVE. Seleção de variáveis, Regressão PLS, Predição

Área principal. IND - PO na Indústria

ABSTRACT

Variable selection is deemed a key step precise prediction in process control applications. The Partial Least Squares (PLS) regression has been widely used for that purpose due to its ability to handle a large number of correlated and noisy variables. This paper presents a method for selecting the most relevant variables aimed at predicting product variables. For that matter, variable importance indices are developed based on PLS parameters and used to guide the elimination of noisy and irrelevant variables. Variables are then systematically removed from the dataset, and the performance of the predictive model evaluated. When applied to 5 manufacturing datasets, the proposed method retained 31% of the original variables and yielded 6% more accurate predictions than using all original variables. Further, the proposed method also outperformed the traditional Stepwise method regarding prediction performance.

KEYWORDS: Variable selection. PLS regression. Prediction

Main area IND – OR in industry

1. Introdução

São vários os processos industriais que envolvem elevado número de variáveis correlacionadas e ruidosas: refino, processamento de petróleo, siderurgia, produção de alimentos, bem como os processos químicos em geral (processamento de polímeros, papéis e medicamentos, entre outros). Kourti & MacGregor (1995) e Montgomery (2004) ressaltam que, na presença de elevado número de variáveis de processo, é imperativo valer-se de métodos multivariados para o monitoramento do processo, uma vez que os métodos univariados podem levar a interpretações equivocadas. Por sua vez, Martin *et al.* (1999) destacam que o controle multivariado do processo tornou-se relevante por possibilitar um alerta preventivo sobre mudanças no processo, potenciais falhas, mal funcionamento e distúrbios no processo.

O elevado volume de informações coletadas de processos industriais, no entanto, pode inviabilizar o monitoramento preciso dos mesmos, visto que grande parte destas informações é inflada com ruído, colinearidade e dados faltantes (Kourti & MacGregor, 1995). Nesse contexto, constitui-se um desafio identificar um conjunto reduzido de variáveis que descrevam características do processo e viabilizem o monitoramento e controle do mesmo (Gauchi & Chagnon, 2001; Chong & Jun, 2005; Anzanello *et al.*, 2009).

Métodos para seleção de variáveis têm sido continuamente propostos na literatura (Lazraq *et al*, 2003; Gauchi & Chagnon, 2001; Anzanello *et al*, 2009; Chiang & Pell, 2004). Entre os métodos para seleção de variáveis com propósito de predição, destacam-se aqueles baseados em regressões PLS (*Partial Least Squares*). A regressão PLS consiste em uma análise multivariada que transforma as variáveis de resposta, **Y**, e de processo, **X**, em um número menor de componentes ou estruturas latentes. Seu uso na indústria é justificado por sua habilidade em lidar com um elevado número de variáveis de produto e resposta, dados com elevado nível de ruído, colinearidade, e observações incompletas (Wold *et al.*, 2001; Kourti & MacGregor, 1995).

Apesar do grande número de métodos com vistas à seleção de variáveis com propósitos de predição em PLS, não existe um método unânime. Nesse sentido, os trabalhos de Gauchi & Chagnon (2001), Chong & Jun (2005), Lazraq *et al.* (2003), Zhai *et al.* (2006) avaliam comparativamente o desempenho de algumas abordagens para seleção de variáveis com fins de predição baseadas em PLS. Entretanto, observa-se que nem todas as possibilidades de utilização dos parâmetros gerados pela regressão PLS foram exploradas e, por consequência, há espaço para abordagens mais eficientes com vistas à seleção de variáveis com propósito de predição. Complementarmente, setores e aplicações específicas ainda carecem de métodos mais robustos de seleção de variáveis.

Este artigo apresenta um método para seleção de variáveis de processo com propósito de predição. Para tanto, os parâmetros gerados pela regressão PLS dão origem a índices de importância das variáveis de processo, os quais identificam as variáveis mais relevantes para explicação da variabilidade na variável de resposta. Inicia-se então um processo de eliminação de variáveis do tipo *backward*, sendo a ordem de eliminação definida pelo índice de importância. O desempenho do modelo resultante após cada eliminação de variável é avaliado por intermédio do indicador RMSE (*root mean square error*). Por fim, o método proposto foi comparando com o tradicional método *Stepwise*. O artigo inova ao adaptar o método de seleção proposto por Anzanello *et al.* (2009), desenvolvido com propósito de classificação, para a seleção de variáveis com fins de predição. O artigo também desenvolve um novo índice de importância com base nos parâmetros oriundos da regressão PLS, além de testar outro proposto em Anzanello *et al.* (2009) e ainda não utilizado em contexto de predição.

O artigo está organizado em 4 Seções, além desta introdução. A revisão bibliográfica é apresentada na Seção 2, abordando os fundamentos da regressão PLS e métodos para seleção de variáveis. A Seção 3 descreve os procedimentos metodológicos do trabalho, enquanto que a Seção 4 apresenta os resultados obtidos. Por fim, tem-se a conclusão do trabalho na Seção 5.

2. Fundamentação teórica

2.1 Regressão PLS

A regressão PLS é usada para modelar a relação entre a matriz **X** (composta por variáveis de processo) e **Y** (composta por variáveis de produto), como tradicionalmente seria feito em uma regressão linear múltipla. Diferentemente da regressão linear múltipla, a regressão PLS permite analisar dados com forte correlação, elevados níveis de ruído, múltiplas variáveis de resposta e mais variáveis do que observações. Adicionalmente, a regressão PLS gera um conjunto de parâmetros que fornecem informações sobre a estrutura e comportamento de **X** e **Y**, o que corrobora para sua ampla aplicação em procedimentos de seleção de variáveis (Wold *et al*., 2001).

Considere uma matriz **X**, de dimensão ($K \times N$), e uma matriz **Y**, de dimensão ($M \times N$), na qual K denota o número de variáveis de processo, M o número de variáveis de resposta e N o número de observações. O vetor $\mathbf{x}_i$ ($x_{i1}, x_{i2}, \ldots, x_{ik}$) representa a observação i para cada variável de processo k, enquanto que o vetor $\mathbf{y}_i$ ($y_{i1}, y_{i2}, \ldots, y_{im}$) representa a observação i para cada variável de resposta m.

Um dos propósitos da regressão PLS é retratar a variabilidade das variáveis originais através de A combinações lineares de tais variáveis. Para tanto, a regressão PLS gera variáveis latentes $\mathbf{t}_a$ (a=1,2,...,A), as quais são usadas com propósitos de predição e controle de processo (Wold *et al*., 2001). Além de serem em número reduzido, geralmente de duas a cinco, as variáveis $\mathbf{t}_a$ são ortogonais entre si (Wold *et al*., 2001; Anzanello *et al*., 2009).

Para a escolha do número de componentes a a serem mantidos no modelo, avalia-se a significância em termos de predição de cada componente; a inclusão de componentes no modelo é interrompida quando os componentes deixam de ser significativos (Wold *et al*., 2001). Wold *et al*. (2001) e Höskuldosson (2001) sugerem o uso da técnica de validação cruzada (CV), a qual destaca-se por sua praticidade e robustez, para definir o número de componentes a serem retidos. Adicionalmente, pode-se optar pelo Algoritmo Inferencial de Lazraq & Cleroux (2001) ou pelo método de minimização da média quadrada do erro na predição de Denham (2000). Ressalta-se ainda que os componentes retidos descrevem grande parte da variância das variáveis **X** e **Y** e da covariância entre ambas (Anzanello *et al*., 2009).

As variáveis latentes $\mathbf{t}_a$ são combinações lineares independentes das variáveis **X** com coeficientes $\mathbf{w}_a$ ($w_{1a}, w_{2a}, \ldots, w_{ka}$). O vetor $\mathbf{w}_a$ representa o peso da variável de processo k no componente a, sendo importante ressaltar que também leva em conta a influência das variáveis de produto (Wold *et al*., 2001; Anzanello *et al*., 2009).

$$t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{ja}x_{ij} = \mathbf{w}_a'\mathbf{x}_i \quad (1)$$

Da mesma forma, geram-se as variáveis latentes $\mathbf{u}_a$ (a=1,2,...,A), que são combinações lineares para as variáveis **Y**. O vetor $\mathbf{c}_a$ ($c_{1a}, c_{2a}, \ldots, c_{ma}$) representa o peso de cada variável de produto m no componente a (Wold *et al*., 2001; Anzanello *et al*., 2009).

$$u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + \ldots + c_{ma}y_{im} = \mathbf{c}_a'\mathbf{y}_i \quad (2)$$

De acordo com Anzanello *et al*. (2009) os vetores $\mathbf{w}_a$ e $\mathbf{c}_a$ são selecionados de forma a maximizar a covariância nos componente de processo ($\mathbf{t}_a$) e produto ($\mathbf{u}_a$) do modelo PLS. Além disso, pode-se afirmar que $\mathbf{t}_a$ e $\mathbf{u}_a$ aglutinam informações sobre as observações e suas semelhanças em relação ao modelo (Wold *et al*., 2001). Complementarmente, Wold *et al*. (2001) asseguram que $\mathbf{w}_a$ e $\mathbf{c}_a$ fornecem informação sobre como as variáveis se combinam para formar a relação quantitativa entre **X** e **Y**, sinalizando quais variáveis **X** são mais importantes (maiores valores de $\mathbf{w}_a$).

Multiplicando o vetor de cargas das variáveis de processo, $\mathbf{p}_a$ (p_{1a},p_{2a}, ...,p_{ka}), pelo vetor $\mathbf{t}_a$ obtém-se um bom resumo da matriz **X**, com valores pequenos para os resíduos, conforme apresentado na equação a seguir, onde e_{ik} são os resíduos de **X** (Wold *et al*., 2001).

$$X_{ik} = \sum_a t_{ia} p_{ak} + e_{ik} \qquad (\mathbf{X} = \mathbf{TP}' + \mathbf{E}) \tag{3}$$

Por sua vez, um bom resumo de **Y** pode ser obtida pela multiplicação de $\mathbf{u}_a$ pelos coeficientes $\mathbf{c}_a$ (Wold *et al*., 2001).

$$Y_{im} = \sum_a u_{ia} c_{am} + g_{im} \qquad (\mathbf{Y} = \mathbf{UC}' + \mathbf{G}) \tag{4}$$

Por fim, os coeficientes da regressão PLS podem ser reescritos como apresentado na Eq.5, onde $w^*_{ka} = w_{ka}(p_{ka} w_{ka})^{-1}$ (Wold *et al*., 2001; Anzanello *et al.*, 2009).

$$b_{mk} = \sum_a c_{ma} w^*_{ka} + f_{im} \qquad (\mathbf{B} = \mathbf{W} * \mathbf{C}') \tag{5}$$

Substituindo-se as equações anteriores pode-se chegar ao formato tradicional do modelo de regressão (Wold *et al*., 2001).

$$Y_{im} = \sum_a b_{mk} x_{ik} + f_{im} \qquad (\mathbf{Y} = \mathbf{XB} + \mathbf{F}) \tag{6}$$

Como na regressão linear múltipla, os resíduos f_m são usados para fins de diagnóstico da qualidade do modelo. Quanto menores os valores dos resíduos de **Y,** melhor a qualidade do modelo. Gráficos de probabilidade normal podem ser usados para verificar a presença de pontos extremos (*outliers*) na relação entre **X** e **Y** (Wold *et al*., 2001). Por sua vez, o índice RMSE = $\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$, onde y_i é o valor observado de **Y** e $\hat{y}_i$ é o valor estimado a partir da regressão PLS, pode ser usado para avaliar a qualidade de predição do modelo de regressão (Gauchi & Chagnon, 2001; Montgomery & Runger, 2009).

2.2 Seleção de Variáveis

A presença de um grande número de variáveis **X** e/ou **Y** tem incentivado engenheiros e pesquisadores a buscarem modelos compostos por um número de variáveis que viabilize aplicações práticas (Gauchi & Chagnon, 2001; Lazraq *et al*., 2003; Chiang & Pell, 2004; Chong & Jun, 2005; Montgomery & Runger, 2009). Ao mesmo tempo em que se selecionam variáveis de processo adequadas para propósitos de predição, procura-se reduzir custos de coleta de dados e facilitar o uso do modelo (Gauchi & Chagnon, 2001; Montgomery & Runger, 2009).

No entanto, a escolha de um subconjunto de variáveis relevantes normalmente constitui-se em desafio. O uso de abordagens explanatórias (por exemplo, gráficos de normalidade) com escolha manual das variáveis pode tornar-se impraticável quando o número de variáveis é elevado e elas são altamente correlacionadas. Tais sistemáticas não asseguram que a predição gerada será a melhor possível (Gauchi & Chagnon, 2001). Além disso, um modelo de regressão com bom ajuste aos dados não necessariamente conduz a boas predições, evidenciado por situações de *overfitting* ou quando a aleatoriedade do processo muda entre a construção do modelo e a predição (Höskuldsson, 2001; Lazraq *et al*., 2003; Chong & Jun, 2005).

Dentre os métodos para seleção de variáveis aplicadas a regressões lineares múltiplas, o *Stepwise* é possivelmente o mais amplamente difundido (Montgomery & Runger, 2009). O método também vem sendo usado para a seleção de variáveis em regressões PLS com propósito

de predição (Gauchi & Chagnon, 2001; Chong & Jun, 2005; Zhai *et al.*, 2006). Sua operacionalização ocorre através da sistemática adição ou remoção de variáveis no modelo de regressão, realizada com base em um teste estatístico de significância de cada variável.

Métodos mais robustos que o *Stepwise* vêm sendo desenvolvidos para a seleção de variáveis em aplicações com propósito de predição via regressão PLS (Höskuldsson, 2001; Gauchi & Chagnon, 2001; Lazraq *et al.*, 2003; Chong & Jun, 2005; Zhai *et al.*, 2006; Anzanello *et al.*, 2009). Gauchi & Chagnon (2001) comparam 20 métodos de seleção baseados em diferentes critérios de avaliação, incluindo ajuste do modelo e capacidade de predição. Dentre os métodos, destacam-se o BCOR (*backward correlations*), BQ (*backward* Q^2_{cum}) e algoritmo genético (AG). O método BCOR usa os parâmetros da regressão PLS para rodar uma sequência de eliminação de variáveis a partir da significância dos coeficientes de correlação PLS de cada variável **X** em cada componente a. O método BQ, por sua vez, sistematicamente elimina a variável associada ao menor coeficiente da regressão PLS, registrando o valor Q^2_{cum} para medir a qualidade da predição a cada eliminação. Por fim, o conjunto de variáveis que maximiza o Q^2_{cum} é escolhido. Já o AG, utilizado para identificar as variáveis mais relevantes a serem utilizadas na regressão PLS, retém um número reduzido de variáveis e conduz a bons resultados na predição, porém apresenta alta variabilidade e requer demasiado processamento computacional. Por fim, Höskuldsson (2001) usou intervalos de variáveis para selecionar variáveis em dados de infravermelho para a regressão PLS.

A proposição de índices de importância das variáveis tem encontrado elevada aplicação em procedimentos de seleção; tais índices atuam como guias no processo de eliminação sistemática de variáveis. Wold *et al.* (2001) desenvolveram um índice de importância das variáveis, VIP (*variable importance in the projection*), a partir do coeficiente modificado de peso w^*_{ka} e da fração de variância explicada pelo componente a em **Y**, R^2_{Ya}. Esse índice foi testado em Lazraq *et al*. (2003). Com propósito semelhante, Chong & Jun (2005) comparam o desempenho de três métodos para seleção de variáveis: método VIP, regressão Lasso (*least absolute shrinkage and selection operator*) e regressão *Stepwise*. Os estudos utilizaram experimentos simulados em cenários com alta colinearidade; o uso combinado da regressão PLS com o índice VIP leva a melhores resultados.

Por sua vez, Anzanello *et al.* (2009) propuseram um método para seleção de variáveis de processo para fins de classificação das variáveis de resposta, a partir do uso combinado de índices de importância das variáveis e técnicas de mineração de dados. Através de um processo de eliminação do tipo *backward*, as variáveis com o menor índice de importância são sequencialmente removidas do conjunto de variáveis retidas. O desempenho de classificação é avaliado a cada iteração, sendo escolhido ao término o subconjunto que maximiza tal desempenho. No método proposto neste trabalho, as variáveis são sistematicamente eliminadas com base em novos índices de importância, porém com objetivo de predição (e não classificação).

Dos cinco indicadores de importância testados por Anzanello *et al.* (2009) destacam-se o v_w, o v_k e o v_b. O índice v_w é baseado no indicador VIP proposto por Wold *et al.* (2001) [ver equação (7)], amplamente usado para seleção de variáveis visando predição. O índice v_k, na equação (8), é uma variação do índice VIP e ainda não foi aplicado com propósitos de predição, sendo gerado com base nos pesos w_{ka} e na fração da variação de **Y**, R^2_{Ya}, explicada por cada componente a = 1,...,A. O índice v_b, na equação (9), define a importância da variável de processo k com base no coeficiente b_{mk} da regressão PLS, o qual mensura a magnitude da relação entre **X** e **Y**. Esses índices são combinados na seção 3 para a geração de um novo índice de importância das variáveis.

$$v_w = \frac{\sum_{a=1}^{A}(w^*_{ka})^2 R^2_{Ya}}{\max_{k \in K}\left(\sum_{a=1}^{A}(w^*_{ka})^2 R^2_{Ya}\right)} \quad k = 1, \ldots, K \tag{7}$$

$$v_k = \frac{\sum_{a=1}^{A} |w_{ka}| R_{Ya}^2}{\max_{k \in K} (\sum_{a=1}^{A} |w_{ka}| R_{Ya}^2)} \quad k = 1, \ldots, K \tag{8}$$

$$v_b = \frac{\sum_{m=1}^{M} |b_{mk}|}{\max_{k \in K} (\sum_{m=1}^{M} |b_{mk}|)} \quad k = 1, \ldots, K \tag{9}$$

3. Método proposto

O método proposto é operacionalizado em cinco etapas. As quatro primeiras etapas visam selecionar variáveis com propósitos de predição, sendo adaptadas de Anzanello *et al.* (2009). Naquele estudo, indicadores de importância das variáveis foram integrados a técnicas de mineração de dados para selecionar variáveis com propósitos de classificação de bateladas de produção. Neste artigo, geram-se distintos indicadores de importância das variáveis a partir dos parâmetros oriundos da regressão PLS, os quais são combinados a uma lógica *backward* de eliminação de variáveis. A quinta e última etapa corresponde à comparação do desempenho do método utilizando os diferentes índices frente ao método *Stepwise* (amplamente citado por sua relevância teórica e prática). Enfatiza-se que o método proposto assume as variáveis de produto **Y** como contínuas; adaptações podem ser necessárias para uso com variáveis de produto discretas. Os passos propostos são detalhados na sequência.

Etapa 1: Dividir o banco de dados em conjuntos de treino e teste

Considere as matrizes **X** e **Y**, introduzidas na Seção 2, com *N* observações, *K* variáveis de processo e uma variável de produto. Divide-se aleatoriamente as observações do banco de dados em um conjunto de treino *tr*, com N_{tr} observações, e um conjunto de teste *ts* com N_{ts} observações, tal que $N_{tr} + N_{ts} = N$. As variáveis relevantes são identificadas a partir do conjunto de treino. Já o conjunto de teste representa novas observações que são usadas para avaliação da capacidade de predição das variáveis selecionadas previamente. Recomenda-se usar uma proporção de 3:2 entre as observações de N_{tr} e N_{ts}, respectivamente (Anzanello *et al.*, 2009).

Etapa 2: Aplicar a regressão PLS no conjunto de treino e gerar índices de importância das variáveis

Para evitar efeitos de escala nos resultados sugere-se normalizar os dados antes de aplicar a regressão. A regressão PLS fornece os parâmetros b_{mk}, peso w_{ka} e o percentual de variação de **Y** explicada por cada componente *a*, R_{Ya}^2, utilizados para gerar os índices de importância para cada variável de processo.

Três são os indicadores de importância usados no presente método. O índice v_w [ver equação (7)] foi escolhido por ser baseado no índice VIP, amplamente usado para seleção de variáveis com propósitos de predição (Wold *et al.*, 2001). O índice v_k [ver equação (8)], elaborado por Anzanello *et al.* (2009), é aqui utilizado de forma inovadora com a finalidade de guiar a escolha das variáveis de processo mais significativas para predição de **Y**. Por fim, com base nas equações (7) e (9), propõe-se um novo índice, v_{bw}, apresentado na equação (10), o qual considera três parâmetros da regressão PLS para definir a importância da variável *k*: o coeficiente de regressão b_{mk}, os pesos w_{ka} e, a fração da variação de **Y**, R_{Ya}^2, explicada por cada componente $a = 1,\ldots,A$.

$$v_{bw} = \frac{\sum_{a=1}^{A} |w_{ka}| R_{Ya}^2}{\max_{k \in K} (\sum_{a=1}^{A} |w_{ka}| R_{Ya}^2)} \frac{\sum_{m=1}^{M} |b_{mk}|}{\max_{k \in K} (\sum_{m=1}^{M} |b_{mk}|)} \quad k = 1, \ldots, K. \tag{10}$$

Etapa 3: Predizer os valores de **Y** *para o conjunto de treino e eliminar as variáveis irrelevantes e ruidosas*

Uma primeira predição é feita valendo-se de *K* variáveis de processo, e o desempenho da predição é medido através do RMSE. Na sequência, remove-se do conjunto de treino a variável com o menor índice de importância da variável, roda-se a regressão PLS com as $K-1$ variáveis de processo, e registra-se novo RMSE. Repete-se o processo, removendo a variável com menor índice e aplicando a regressão PLS para predizer **Y**, até que reste apenas uma variável de processo.

Etapa 4: Construir um gráfico para identificar o melhor subconjunto de variáveis e testar essas variáveis no conjunto de teste

A partir dos valores de desempenho da predição de **Y** no conjunto de treino, gera-se um gráfico relacionando RMSE e o percentual de variáveis retidas, como apresentado na Figura 1. O subconjunto responsável pelo menor RMSE é selecionado para predição do conjunto de observações de teste, estimando-se o desempenho de predição via RMSE.

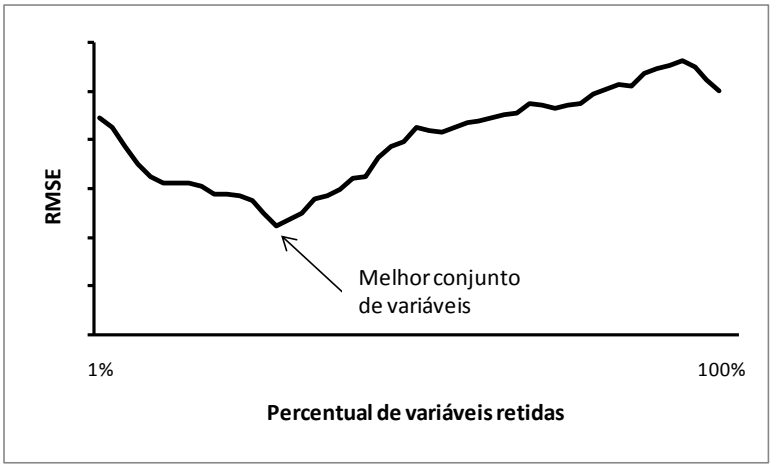

Figura 1 – Perfil hipotético de desempenho de predição à medida que variáveis são eliminadas do conjunto de treino

Etapa 5: Comparar o desempenho dos métodos de seleção de variáveis para predição de **Y**

Aplica-se o método *Stepwise* no conjunto de treino, gerando-se uma regressão PLS com as variáveis selecionadas. Compara-se então o desempenho do método proposto utilizado os índices v_w, v_k e v_{bw}, e o método *Stepwise* em termos de RMSE e percentual de variáveis retidas.

4. Resultados e Discussão

Para aplicação e avaliação do desempenho do método proposto, foram usados os cinco banco de dados em Gauchi & Chagnon (2001), os quais também constam nos trabalhos de Lazraq *et al.* (2003) e Anzanello *et al.* (2009). As análises foram realizadas em MATLAB® versão 7.10.

O número de variáveis de processo de cada banco de dados, assim como a divisão das observações em conjuntos de treino e teste, são apresentados na Tabela 1. O banco de dados ADPN é procedente de um processo intermediário da produção de nylon, enquanto que as observações do LATEX foram extraídas do processo de manufatura de látex. Os dados do banco OXY correspondem ao processo de produção do óxido de titânio, o qual é usado na mistura de tintas. Já SPIRA refere-se a um processo de fermentação para a produção de antibiótico. Por fim,

as observações do processo GRANU provêm de um processo de emulsões anti-espuma utilizado na indústria do papel.

Tabela 1 - Bancos de dados

Banco de dados	Número de variáveis de processo	Número de observações	
		Conjunto de treino	Conjunto de teste
ADPN	100	57	14
LATEX	117	210	52
OXY	95	18	12
SPIRA	96	115	29
GRANU	78	23	6

A regressão PLS foi aplicada ao conjunto de treino de cada banco de dados. Foram retidos 3 componentes da regressão PLS para cada banco de dados através de validação-cruzada, resultando nos seguintes R^2_{Ya}'s: ADPN, 94%, LATEX, 77%, OXY, 94%, SPIRA, 71%, GRANU, 86%.

A Figura 2 apresenta o RMSE no conjunto de treino para o banco de dados OXY ao aplicar-se o índice v_{bw}. A escolha do melhor conjunto de variáveis considerou uma solução de compromisso entre menor percentual de variáveis retidas e o menor valor para o RMSE. Retendo-se apenas 23% das variáveis, obtém-se um RMSE de 0,208, valor próximo ao mínimo valor possível de RMSE. Um modelo com todas as variáveis gera um RMSE de 0,257. A mesma lógica foi aplicada aos demais bancos de dados e índices de importância das variáveis.

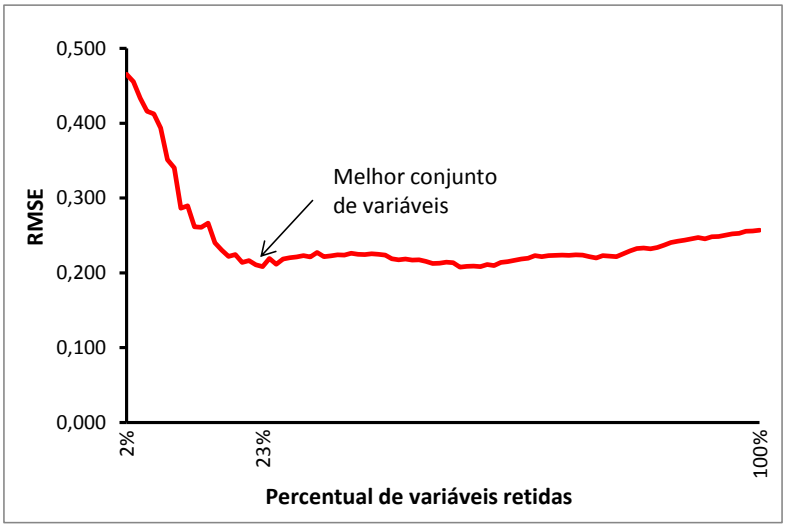

Figura 2 – Desempenho da predição no conjunto de treino do processo OXY usando o índice

A comparação do desempenho dos métodos de seleção de variáveis é apresentada na Tabela 2. Dentre os três índices integrados ao método proposto, o índice v_{bw} foi o que apresentou o melhor desempenho de predição (menores valores para RMSE para o conjunto de teste) e o menor percentual de variáveis retidas. O índice v_w obteve o segundo melhor desempenho, com acurácia 15% inferior ao índice v_{bw}, retendo 1% a mais de variáveis. Já o índice v_w usou 6% a mais de variáveis e alcançou acurácia 21% inferior relativamente ao índice v_{bw}. Ao se observar o desempenho dos métodos em cada banco de dados, o método com o índice v_{bw} obteve acurácia superior em relação aos demais índices nos processos ADPN, LATEX, OXY e GRANU, destacando-se que para o processo OXY alcançou acurácia 26% superior ao

segundo melhor valor. Em relação ao percentual de variáveis retidas, o índice v_{bw} apresenta os menores valores nos processos OXY e SPIRA, com o mesmo percentual que o índice v_k para o processo GRANU. Em relação aos índices v_k e v_w, o primeiro superou o segundo em termos de desempenho apenas no processo ADPN, além de reter menos variáveis nos processos LATEX e GRANU.

Quando comparado ao método *Stepwise*, o método proposto com os índices v_k, v_w e v_{bw} apresentou desempenho médio superior, apesar de reter um percentual maior de variáveis. Enfatiza-se que o método valendo-se do novo índice v_{bw} apresenta acurácia média 31% superior ao método *Stepwise*, retendo 16% a mais de variáveis. Além disso, o índice v_{bw} obteve maior acurácia em todos os processos, destacando-se o desempenho 49% superior no processo OXY e 54% para o processo GRANU. Ao seu tempo, o índice v_k excedeu em 16% a acurácia do método *Stepwise*, ainda que retenha 22% a mais de variáveis. Comparando processo a processo, o índice v_k alcançou maior acurácia do que o *Stepwise*, embora o percentual de variáveis tenha sido inferior apenas no processo LATEX.

Tabela 2 – Desempenho dos métodos de seleção de variáveis avaliados para todos os bancos de dados

	RMSE para o conjunto de treino				RMSE para o conjunto de teste				Variáveis retidas (%)			
Processo	v_k	v_w	v_{bw}	**Stepwise**	v_k	v_w	v_{bw}	**Stepwise**	v_k	v_w	v_{bw}	**Stepwise**
ADPN (100)	1,110	1,216	0,961	1,102	1,051	1,228	0,968	1,225	56%	36%	37%	31%
LATEX (117)	0,602	0,586	0,546	0,588	0,594	0,573	0,549	0,610	7%	16%	24%	15%
OXY (95)	0,227	0,225	0,208	0,213	0,126	0,121	0,089	0,172	39%	24%	23%	9%
SPIRA (96)	0,160	0,157	0,156	0,161	0,148	0,147	0,147	0,182	57%	47%	41%	10%
GRANU (78)	0,654	0,638	0,601	0,669	0,742	0,455	0,448	0,978	28%	37%	28%	6%
Média	0,551	0,564	0,494	0,547	0,532	0,505	0,440	0,633	37%	32%	31%	15%

Constata-se, a partir dos resultados e análises da Tabela 2, que o método proposto fazendo uso do novo índice v_{bw} obteve os melhores desempenhos de predição, tanto em relação aos índices v_k e v_w quanto ao método tradicional *Stepwise*. O melhor desempenho do índice v_{bw} é alcançado com um percentual de variáveis menor que o obtido com os índices v_k e v_w, apesar de ser maior do que o alcançado pelo método *Stepwise*. Ressalta-se ainda que a abordagem proposta utiliza 31% das variáveis originais para obter RMSE=0,494, um acréscimo de 6% na acurácia de predição relativamente ao RMSE=0,525 obtido quando utilizam-se todas as variáveis.

5. Conclusões

Processos industriais caracterizados por elevado número de variáveis correlacionadas e ruidosas demandam métodos de seleção para assegurar boa capacidade de predição dos modelos gerados. Assim, o presente artigo apresentou um método para seleção de variáveis de processo mais relevantes com vistas à predição das variáveis de resposta.

O método proposto se apóia nas seguintes etapas: (1) divisão dos bancos de dados compostos por variáveis de processo e resposta em conjuntos de treino e teste; (2) aplicação da regressão PLS no conjunto de treino e geração dos índices de importância das variáveis v_w v_k, e v_{bw}; (3) predição dos valores de **Y** para o conjunto de treino e eliminação das variáveis com base nos índices de importância, registrando-se o desempenho preditivo via RMSE; (4) construção de um gráfico associando RMSE e percentual de variáveis retidas, seleção do subconjunto recomendado e predição da variável de resposta para o conjunto de teste usando tal subconjunto de variáveis; e (5) comparação do desempenho do método composto pelos três índices frente ao método *Stepwise*.

Um novo índice de importância de variáveis, v_{bw}, foi comparado aos índices v_w e v_k, também gerados com base nos coeficientes da regressão PLS. Tal índice apresentou melhor desempenho de predição de **Y** quando comparado aos índices v_k e v_w e ao método *Stepwise*. Além disso, o índice v_{bw} reteve um percentual de variáveis inferior ao obtido com os índices v_k e v_w, apesar de superior ao do método *Stepwise*. Analogamente, ao valer-se do índice v_{bw}, o método proposto utilizou 31% das variáveis para predição, com acurácia de predição 6% superior ao valor obtido quando todas as variáveis são utilizadas. Portanto, o método com o novo indicador de importância das variáveis é recomendado para aplicações que necessitam de elevada acurácia de predição a partir de um conjunto reduzido de variáveis.

Pesquisas futuras incluem a comparação do método de seleção de variáveis proposto e do novo indicador de importância das variáveis com relação a outros métodos para seleção de variáveis. Da mesma forma, sugere-se a utilização de outros indicadores de acurácia da predição de **Y** além do RMSE, para corroborar com os resultados e conclusões do presente artigo.

Referências

Anzanello, M. J.; Albin, S. L. e Chaovalitwongse, W. A. (2009), Selecting the best variables for classifying production batches into two quality levels, *Chemometrics Intelligent Laboratory Systems*, 97, 111-117.

Chiang, L. H. e Pell, R. J. (2004), Genetic algorithms combined with discriminant analysis for key variable identification, *Journal of Process Control*, 14, 143-155.

Chong, I.-G. e Jun, C.-H. (2005), Performance of some variable selection methods when multicollinearity is present, *Chemometrics Intelligent Laboratory Systems*, 78, 103-112.

Denham, M. C. (2000), Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction, *Journal of Chemometrics*, 14, 351–361

Gauchi, J.-P. e Chagnon, P. (2001), Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, *Chemometrics Intelligent Laboratory Systems*, 58, 171-193.

Lazraq, A. e Cléroux, R. (2001), The PLS multivariate regression model: testing the significance of successive PLS components, *Journal of Chemometrics*, 15, 523–536

Lazraq, A.; Cléroux, R.e Gauchi, J.-P. (2003), Selecting both latent and explanatory variables in the PLS1 regression model, *Chemometrics Intelligent Laboratory Systems*, 66, 117-126.

Kourti, T. e Macgregor, J. F. (1995), Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemometrics Intelligent Laboratory Systems*, 28, 3-21.

Martin, E. B.; Morris, A. J. e Kiparissides, C. (1999), Manufacturing performance enhancement through multivariate statistical process control, *Annual Reviews in Control*, 35-44.

Montgomery, D. C., *Introdução ao controle estatístico da qualidade*, LTC – Livros Técnicos e Científicos Editora S.A, Rio de Janeiro, 2004.

Montgomery, D. C. e Runger, G. C., Est*atística aplicada e probabilidade para engenheiros*, LTC – Livros Técnicos e Científicos Editora S.A, Rio de Janeiro, 2009.

Wold, S.; Sjöström, M. e Eriksson, L. (2001), PLS-regression: a basic tool of chemometrics, *Chemometrics Intelligent Laboratory Systems*, 58, 109-130.

Zhai, H. L.; Chen, X. G. e Hu, Z. De. (2006), A new approach for the identification of important variables, *Chemometrics Intelligent Laboratory Systems*, 80, 130-135.