UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Juliano Zimmer

SELEÇÃO DE VARIÁVEIS PREDITIVAS COM BASE EM ÍNDICES DE IMPORTÂNCIA DAS VARIÁVEIS E REGRESSÃO PLS

Porto Alegre
2012

Juliano Zimmer

Seleção de variáveis preditivas com base em índices de importância das variáveis e regressão PLS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientador: Michel Jose Anzanello, *Ph.D.*

Porto Alegre

2012

Juliano Zimmer

Seleção de variáveis preditivas com base em índices de importância das variáveis e regressão PLS

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel Jose Anzanello*, Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. Carla Schwengber ten Caten
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professora Carla Schwengber ten Caten, Dr. (PPGEP/UFRGS)

Professor Danilo Marcondes Filho, Dr. (DEST/UFRGS)

Professora Liane Werner, Dr. (PPGEP/UFRGS)

AGRADECIMENTOS

Gostaria de expressar os meus sinceros agradecimentos a todos que contribuíram para a realização desta dissertação, de forma direta ou indireta.

Aos meus pais, Ingo e Dulci, pelos sacrifícios realizados para que eu pudesse chegar até aqui e pelos valores que procuraram transmitir ao longo da minha vida, dentre os quais destaco o apreço a educação como forma de crescimento pessoal.

A minha noiva Aline pela paciência, carinho e apoio incondicional.

Ao meu irmão, pela força, apoio e convivência ao longo de toda vida.

A UFRGS e a sua Escola de Engenharia, pela excelência no ensino.

Ao meu orientador, Prof. Michel Jose Anzanello, *Ph.D.*, pela dedicação, apoio e confiança depositada em mim.

Aos meus grandes amigos Piletti, Felipe, Diego Vinícius, Alexandre, Carlos Eduardo, Grazziotin, Paula e Sue pela convivência e aprendizado.

Aos professores e colegas do Programa de Pós-Graduação em Engenharia de Produção, por todo o conhecimento e apoio que recebi nos últimos dois anos.

ZIMMER, Juliano Zimmer ***Seleção de variáveis preditivas com base em índices de importância das variáveis e regressão PLS,*** 2012. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

A presente dissertação propõe métodos para seleção de variáveis preditivas com base em índices de importância das variáveis e regressão PLS (*Partial Least Squares*). Partindo-se de uma revisão da bibliografia sobre PLS e índices de importância das variáveis, sugere-se um método, denominado Eliminação Backward (EB), para seleção de variáveis a partir da eliminação sistemática de variáveis de acordo com a ordem definida por índices de importância das variáveis. Um novo índice de importância de variáveis, proposto com base nos parâmetros da regressão PLS, tem seu desempenho avaliado frente a outros índices reportados pela literatura. Duas variações do método EB são propostas e testadas através de simulação: (*i*) o método EBM (Eliminação *backward* por mínimos), que identifica o conjunto que maximiza o indicador de acurácia preditiva sem considerar o percentual de variáveis retidas, e (*ii*) o método EBDE (Eliminação *backward* por distância euclidiana), que seleciona o conjunto de variáveis responsável pela mínima distância euclidiana entre os pontos do perfil gerado pela eliminação das variáveis e um ponto ideal hipotético definido pelo usuário. A aplicação dos três métodos em quatro bancos de dados reais aponta o EBDE como recomendável, visto que retém, em média, apenas 13% das variáveis originais e eleva a acurácia de predição em 32% em relação à utilização de todas as variáveis.

Palavras-chave: Seleção de variáveis, regressão PLS, índice de importância das variáveis

ZIMMER, Juliano ***Selecting the most relevant predictive variables based on variable importance indices and PLS regression,*** 2012. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

ABSTRACT

This dissertation presents new methods for predictive variable selection based on variable importance indices and PLS regression. The novel method, namely Backward Elimination (BE), selects the most important variables by eliminating process variables according to their importance described by the variable importance indices. A new variable importance index is proposed, and compared to previous indices for that purpose. We then offer two modifications on the BE method: (*i*) the EBM method, which selects the subset of variables yielding the maximum predictive accuracy (i.e., the minimum residual index), and (*ii*) the EBDE, which selects the subset leading to the minimum Euclidian distance between the points generated by variable removal and a hypothetical ideal point defined by the user. When applied to four manufacturing data sets, the recommended method, EBDE, retains average 13% of the original variables and increases the prediction accuracy in average 32% compared to using all the variables.

Keywords: Variable selection, PLS regression, Variable importance indices.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

1.1 Considerações Iniciais

Tendo em vista a importância do monitoramento e controle da qualidade do produto final em processos industriais, torna-se vital elaborar modelos capazes de predizer as variáveis de produto que descrevem a qualidade do produto. Sendo assim, a identificação das variáveis de processo com maior influência sobre as variáveis de produto é fundamental para o preciso monitoramento e controle de processos industriais, dentre os quais se destacam as indústrias de refino e processamento de petróleo, siderurgia e produção de alimentos, bem como processos químicos em geral (polímeros, papel e medicamentos, entre outros) (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a; MONTGOMERY; RUNGER, 2009; PIERNA *et.* al, 2009; ANDERSEN; BRO, 2010).

A regressão PLS (*Partial Least Squares*) vem sendo amplamente utilizada para a seleção de variáveis em cenários com elevado número de variáveis, presença de variáveis altamente correlacionadas e ruidosas, e falta de dados. A regressão PLS captura a relação entre as variáveis de processo (independentes) e as de produto (dependentes) através da geração de um número reduzido de combinações lineares das variáveis independentes e dependentes, facilitando assim a manipulação dos dados e extração de informações relevantes (KOURTI e MACGREGOR, 1995; WOLD *et al.*, 2001; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a; ANDERSEN; BRO, 2010).).

Diversos autores utilizam os parâmetros oriundos da regressão PLS na geração de índices de importância das variáveis, os quais permitem avaliar a contribuição das variáveis de processo na explicação da variabilidade da variável de produto. Tais índices são usados para guiar a seleção das variáveis de processo, sinalizando quais variáveis devem ser adicionadas ou eliminadas do modelo preditivo final (WOLD *et al.*, 2001; LAZRAQ *et al.*, 2003; ANZANELLO *et al.*, 2009a; ERIKSSON; WOLD, 2010).

Sendo assim, o tema do presente trabalho é a seleção de variáveis preditivas a partir de índices de importância das variáveis gerados com base nos parâmetros oriundos da regressão PLS.

Esta dissertação é composta por três artigos abordando a seleção de variáveis com propósito de predição da variável de produto. No primeiro artigo é proposto e testado um método de seleção de variáveis a partir da Eliminação Backward (EB) de variáveis, com base em um novo índice de importância das variáveis de processo. O segundo artigo propõe duas variações do método proposto no primeiro artigo, EBM (Eliminação *backward* por mínimos) e EBDE (Eliminação *backward* por distância euclidiana), visando reduzir a subjetividade na escolha do conjunto de variáveis a ser usada no modelo de regressão. O mesmo artigo apresenta experimentos de simulação para avaliar o comportamento desses métodos frente a variações nos níveis de ruído e correlação entre as variáveis de processo, além de testar diferentes proporções entre o número de variáveis de processo e o número de observações. O terceiro artigo apresenta a comparação dos métodos EB, EBM e EBDE em bancos de dados reais, além de inserir um indicador alternativo para avaliação da acurácia de predição dos modelos gerados.

1.2 Objetivos

O objetivo principal do trabalho é propor métodos para seleção de variáveis de processo com fins de predição da variável de produto.

Como objetivos específicos têm-se:

- Apresentar a fundamentação teórica dos principais métodos de seleção de variáveis a partir de regressões PLS e de índices de importância das variáveis;
- Adaptar um método para seleção de variáveis com propósito de predição usando regressões PLS, comparando o mesmo com métodos propostos pela literatura;
- Desenvolver um novo índice de importância das variáveis a partir de parâmetros da regressão PLS;
- Propor variações do método para seleção de variáveis, visando melhorar a implementação da lógica de escolha do melhor subconjunto de variáveis de processo a predizerem a variável de produto.
- Avaliar a robustez dos métodos para seleção de variáveis propostos em bancos de dados afetados por distintos níveis de correlação e ruído das variáveis de processo, além de diferentes proporções entre o número de variáveis e de observações.

- Comparar os métodos para seleção de variáveis propostos ao aplica-los em bancos de dados reais, mensurando o desempenho através de indicadores de acurácia de predição e percentual de variáveis retidas no modelo final de predição.

1.3 Justificativa do Tema e dos Objetivos

São vários os processos industriais que envolvem centenas de variáveis altamente ruidosas e/ou correlacionadas. O controle da totalidade de tais variáveis, além de inviável em termos de tempo e custo, pode levar a resultados imprecisos. A inclusão de variáveis ruidosas e irrelevantes tende a reduzir a eficiência das ferramentas multivariadas utilizadas em controle de processo, levando a predições e classificações equivocadas. Além disso, monitorar as variáveis de produto isoladamente, sem considerar a relação das mesmas com as variáveis de processo, é pouco efetivo na presença de elevada colinearidade entre as variáveis de processo (KOURTI; MACGREGOR, 1995; MARTIN *et. al*, 1999; MONTGOMERY, 2004; GONZÁLEZ; SÁNCHEZ, 2010).

Diversos métodos para seleção de variáveis com propósitos de predição têm sido propostos na literatura. Contudo, a crescente capacidade de armazenamento de informações e dados demanda o aprimoramento contínuo de abordagens robustas e eficientes para seleção de variáveis, especialmente em aplicações caracterizadas por elevados níveis de ruído, colinearidade entre as variáveis e dados faltantes por conta da quebra de equipamentos de coleta (KOURTI; MACGREGOR, 1995; GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2009b., 2012; ANDERSEN; BRO, 2010).

Dentre as técnicas de análises multivariadas capazes de lidar com um elevado número de variáveis de processo e de produto, destaca-se a regressão PLS. Métodos de seleção de variáveis preditivas baseados em regressão PLS vêm sendo usados com sucesso por engenheiros e acadêmicos, justamente em função da sua robustez frente a bancos de dados mal condicionados (HÖSKULDSSON, 2001; KOURTI; MACGREGOR, 1995; WOLD *et al.*, 2001; FERRER *et al.*, 2008; PIERNA *et.* al, 2009; ANDERSEN; BRO, 2010).

Considerando as várias abordagens existentes para seleção de variáveis com fins de predição baseadas em PLS, observa-se que não há método unânime para tal finalidade. Tal afirmação é reforçada pelos estudos recentes que comparam o desempenho de algumas dessas abordagens, tanto em bancos de dados reais ou simulados [ver Gauchi e Chagnon, 2001; Lazraq *et al.*, 2003; Chong e Jun, 2005; Zhai *et al.*, 2006]. Observa-se que nem todas as possibilidades de comparação foram exploradas e há espaço para o desenvolvimento de métodos de seleção mais eficientes e robustos. Por eficiente, entende-se que os métodos de seleção devem garantir elevada capacidade preditiva da variável de produto, valendo-se para isso do menor número de variáveis retidas no modelo final. Por robustez, entende-se que o método deve ser consistente frente a alterações nos dados do processo, tais como variações na correlação entre as variáveis de processo, aumento dos níveis de ruído e disponibilidade de observações.

O uso dos parâmetros gerados pela regressão PLS para a criação de índices de importância das variáveis e integração a métodos de seleção de variáveis é uma das possibilidades exploradas na literatura [ver Wold *et al.*, 2001; Lazraq *et al.*, 2003; Anzanello *et al.*, 2009a,2009b, 2012; Erikson e Wold, 2010] . Diversos autores propuseram índices e realizaram estudos comparativos de desempenho de tais índices, aplicando-os para fins de predição e classificação das variáveis de produto. Contudo, constata-se a oportunidade de desenvolver e testar índices mais eficientes e robustos, os quais apoiem a seleção de variáveis para o modelo de predição.

De tal forma, justifica-se o desenvolvimento de métodos para seleção de variáveis com finalidade de predição a partir de índices de importância de variáveis e regressão PLS, tema dos três artigos dessa dissertação.

1.4 Procedimentos Metodológicos

Acerca do método de pesquisa utilizado, pode-se caracterizá-lo como de natureza aplicada, uma vez que o conteúdo teórico é explorado com vistas à solução de problemas

genéricos, e de objetivo exploratório, já que busca construir hipóteses para resolver os problemas a partir da sua análise (GIL, 1991; 2002).

A abordagem da pesquisa é quantitativa, em função das análises numéricas realizadas. O trabalho faz uso de procedimentos de pesquisa bibliográfica e de estudo de caso.

1.5 Estrutura da Dissertação

A dissertação está organizada em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos e as justificativas, bem como o método de pesquisa adotado. A delimitação e estrutura do trabalho completam o capítulo.

O segundo capítulo expõe o primeiro artigo, que introduz uma revisão da literatura acerca da regressão PLS e da seleção de variáveis em processos multivariados. Além disso, propõe uma adaptação do método proposto por Anzanello *et al.* (2009a), o método EB, para selecionar variáveis a partir de uma lógica de eliminação *backward* de variáveis guiada por índices de importância das variáveis. Nesse trabalho é proposto um novo índice, o qual é comparado a outros dois índices da literatura. Por fim, o método é comparado ao método *Stepwise* de seleção de variáveis. A avaliação de desempenho dos modelos gerados é realizada através do indicador de acurácia de predição RMSE (*Root Mean Square Error*) e do percentual de variáveis retidas no modelo final. O método proposto é aplicado em cinco bancos de dados reais.

O terceiro capítulo apresenta o segundo artigo, que explora uma fragilidade do método proposto no primeiro artigo: a subjetividade na escolha do conjunto de variáveis a partir da análise do perfil gráfico da acurácia de predição *versus* percentual de variáveis retidas. Tal artigo propõe dois novos métodos de seleção de variáveis com base em índices de importância, os métodos EBM e EBDE. Tais métodos diferenciam-se do primeiro por recomendarem um conjunto de variáveis com base em critérios de máxima acurácia de predição e distância entre os pontos gerados pela eliminação sistemática das variáveis. Tais métodos têm seu desempenho avaliado através de simulação, de modo a analisar a estabilidade dos mesmos frente a alterações nos níveis de correlação e ruído das variáveis de processo.

O próximo capítulo traz o terceiro artigo, que compara o desempenho dos métodos EB, EBM e EBDE, propostos nos dois primeiros artigos, em bancos de dados industriais. O artigo também traz como contribuição a proposição de um indicador de acurácia de predição da variável de produto (Q^2cum), mais estável que o RMSE, para avaliar comparativamente o desempenho dos três métodos. Tais indicadores são avaliados conjuntamente para indicar o método mais recomendado para a seleção de variáveis com propósitos de predição.

O quinto e último capítulo traz a conclusão do trabalho, na qual são avaliados os principais resultados frente aos objetivos almejados e as delimitações citadas. Essa seção traz ainda sugestões para desdobramentos futuros.

1.6 Delimitações do Estudo

Constituem restrições do presente estudo:

- A seleção de variáveis será estudada apenas através de métodos baseados em regressões PLS;
- Os bancos de dados usados para aplicação e simulação contêm apenas uma variável de produto (dependente), contínua e não discreta;
- As variáveis estudadas são selecionadas apenas com o objetivo de predição e não de classificação.

1.7 Referências Bibliográficas

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 24, p. 728-737, 2010.

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

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Identificação das variáveis mais relevantes para categorização de bateladas de produção: reduzindo a

variância do percentual de variáveis retidas. **Produto&Produção**, Porto Alegre, Brasil, v. 10, n.3, p.19-27, 2009b.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, , v. 218, p. 97-105, 2012.

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 78, p. 103-112, 2005.

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. **Journal of Chemometrics,** Bognor Regis, United Kingdom, v. 24, p. 779-789, 2010.

FERRER, A.; AGUADO, D.; VIDAL-PUIG, S.; PRATS, J.; ZARZO, M. PLS: A versatile tool for industrial process improvement and optimization. **Applied Stochastic Models in Business and Industry**, Malden, USA, v. 24, p. 551-567, 2008.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 171-193, 2001.

GIL, A. C. (2002). **Como elaborar projetos de pesquisa**. 4 ed. São Paulo - Atlas.

GIL, A. C. (1991). **Como elaborar projetos de pesquisa**. 3 ed. São Paulo - Atlas.

GONZÁLEZ, I.; SÁNCHEZ, I. Variable selection for multivariate statiscal process control. **Journal of Quality Technology**, Milwaukee, v. 42, n.3, p. 242-259, 2010.

HÖSKULDSSON, A. Variable and subset selection in PLS regression. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 55, 23-38, 2001.

LAZRAQ, A.; CLÉROUX, R.; GAUCHI, J.-P. Selecting both latent and explanatory variables in the PLS1 regression model. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 66, 117-126, 2003.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 28, p. 3-21, 1995.

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, Amsterdam, Holland, v. 23, p. 35-44, 1999.

MONTGOMERY, D. C. **Introdução ao controle estatístico da qualidade**. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A, 2004. 513 p.

PIERNA, J. A. F.; ABBAS, O.; BAETEN, V.; DARDENNE, P. A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica Acta**, Amsterdam, Holland, v. 642, p. 89-93, 2009.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 109-130, 2001.

ZHAI, H. L.; CHEN, X. G.; HU, Z. De. A new approach for the identification of important variables. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 80, p. 130-135, 2006.

2 Primeiro Artigo: Um novo método para seleção de variáveis preditivas com base em índices de importância

Juliano Zimmer

Michel Jose Anzanello

Artigo enviado para publicação na revista Produção (ABEPRO)

Resumo

O grande volume de variáveis coletadas em processos industriais impõe dificuldades no controle e monitoramento de tais processos. A regressão PLS (*Partial Least Squares*) vem sendo amplamente utilizada em procedimentos de seleção de variáveis por sua capacidade de operar com variáveis em grande número, correlacionadas e ruidosas. Este artigo propõe um método, Eliminação Backward (EB), para identificar o melhor subconjunto de variáveis de processo para a predição das variáveis de produto. Indicadores de importância das variáveis são desenvolvidos a partir de parâmetros da regressão PLS e guiam a eliminação das variáveis irrelevantes. Tais índices são então testados em termos de seu desempenho. Ao ser aplicado em 5 bancos de dados industriais, o método utilizando o índice recomendado reteve apenas 31% das variáveis originais e aumentou a acurácia de predição do conjunto de teste em 6%. O método proposto também superou a acurácia do método *Stepwise*, tradicionalmente utilizado em procedimentos de seleção com propósitos de predição.

Palavras-chave: Seleção de variáveis, Regressão PLS, Indicador de importância das variáveis.

A new framework for predictive variable selection based on variable importance indices

Abstract

The large volume of process variables collected from manufacturing applications has jeopardized process control activities. The Partial Least Squares (PLS) regression has been widely used for variable selection 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 outperformed the traditional Stepwise method regarding prediction accuracy.

Keywords: Variable selection, PLS regression, Variable importance indices.

2.1 Introdução

Diversos processos industriais envolvem elevado número de variáveis correlacionadas e ruidosas para seu controle e monitoramento. Exemplos de tais processos incluem refino e processamento de petróleo, siderurgia e produção de alimentos, bem como processos químicos em geral (processamento de polímeros, papéis e medicamentos, entre outros). 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; ANDERSEN; BRO, 2010). Nesse contexto, engenheiros e técnicos de produção são desafiados a identificar um conjunto reduzido de variáveis relevantes que descrevam características do processo e viabilizem o monitoramento e controle do mesmo (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a; ANDERSEN; BRO, 2010).

Métodos para seleção de variáveis têm sido continuamente propostos na literatura (LAZRAQ *et al.*, 2001, 2003; GAUCHI; CHAGNON, 2001; ANZANELLO *et al.*, 2009ª, 2009b; CHIANG; PELL, 2004; PIERNA *et al.*, 2009). Entre os métodos 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 produto e de processo em um número reduzido de combinações lineares. Seu amplo uso na indústria decorre de sua habilidade em lidar com um elevado número de variáveis de produto, múltiplas variáveis de produto, dados com elevado nível de ruído, colinearidade e observações

incompletas (KOURTI; MACGREGOR, 1995; WOLD *et al.*, 2001; FERRER *et al.*, 2008; ANDERSEN; BRO, 2010).

Apesar do grande número de métodos para seleção de variáveis com propósitos de predição em PLS, não existe um método unânime para tal finalidade (LAZRAQ *et al.*, 2003). Essa condição é justificada pelas características peculiares dos processos industriais, conforme pode ser visto nos estudos comparativos de Gauchi e Chagnon (2001), Lazraq *et al.* (2003), Chong e Jun (2005) e Zhai *et al.* (2006). Observa-se ainda que diversas possibilidades de utilização dos parâmetros gerados pela regressão PLS permanecem inexploradas e, por consequência, há espaço para abordagens mais eficientes para seleção de variáveis. Complementarmente, setores e aplicações específicas ainda carecem de métodos mais robustos para predição das variáveis de produto, especialmente quando as variáveis de processo apresentam elevada correlação.

Este artigo apresenta um método para seleção de variáveis de processo, EB, 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 produto. 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 de regressão resultante após cada eliminação de variável é avaliado por intermédio do indicador RMSE (*Root Mean Square Error*). Por fim, o método proposto é comparando com o tradicional método *Stepwise*.

O artigo inova ao adaptar o método de seleção proposto por Anzanello *et al.* (2009a), desenvolvido com propósito de classificação, para a seleção de variáveis com fins de predição através da regressão PLS. O artigo também desenvolve um novo índice de importância com base nos parâmetros oriundos de tal regressão, além de testar outro índice de importância proposto por Anzanello *et. al* (2009a) ainda não utilizado em contexto de predição. A comparação com o método *Stepwise* também aparece como contribuição relevante, visto que o confronto de distintos métodos para seleção de variáveis auxilia na identificação dos métodos mais adequados em aplicações específicas.

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.

2.2 Fundamentação teórica

As seções seguintes apresentam os fundamentos da regressão PLS e parâmetros utilizados no método proposto, bem como abordagens para seleção de variáveis em contexto de predição.

2.2.1 Regressão PLS

A regressão PLS relaciona a matriz **X** (composta por variáveis de processo *x*) à matriz **Y** (composta por variáveis de produto *y*), permitindo analisar dados com forte correlação, elevados níveis de ruído e desequilíbrio entre o número de variáveis e observações. Tal regressão 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).

Ferrer *et al.* (2008) ressaltam que poucas ferramentas de análise estatística possuem a versatilidade da regressão PLS, a qual tem oferecido suporte em aplicações de diferentes naturezas, como discriminação e classificação de observações, modelagem e análise de processo e identificação de desvios. Suas aplicações não estão restritas a áreas industriais, mas também são verificadas em setores de negócios (avaliação de desempenho e comportamento humano), finanças e marketing (preferência por marcas, satisfação e fidelidade dos clientes), e áreas de ciências sociais (FERRER *et al.*, 2008; ESPOSITO-VINZI *et al.*, 2007).

Os fundamentos matemáticos da regressão PLS são agora apresentados. 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 (produto) e N o número de observações. O vetor $\mathbf{x}_i$ ($x_{i1}, x_{i2}, ..., x_{ik}$) representa a observação i para cada variável de processo k, enquanto que o vetor $\mathbf{y}_i$ ($y_{i1}, y_{i2}, ..., y_{im}$) representa a observação i para cada variável de resposta m. A regressão PLS gera A variáveis latentes (combinações lineares) $\mathbf{t}_a$ (a=1,2,...,A) a partir das variáveis originais, 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, ou seja, reduzem os problemas

oriundos da elevada colinearidade das variáveis originais (WOLD *et al*., 2001; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2009b).

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

As variáveis latentes $\mathbf{t_a}$ são combinações lineares independentes das variáveis *x* com coeficientes $\mathbf{w}_a$ (w_{1a},w_{2a}, ...,w_{ka}), conforme a equação (1). 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; FERRER *et al*., 2008; ANZANELLO *et al.*, 2009a).

$$\boldsymbol{t}_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{ka}x_{ik} = \boldsymbol{w}_a'\boldsymbol{x}_i \quad (1)$$

Da mesma forma, geram-se as variáveis latentes $\mathbf{u}_a$ (a=1,2,...,A), que são combinações lineares das variáveis *y*. O vetor $\mathbf{c}_a$ (c_{1a},c_{2a}, ...,c_{ma}) representa o peso de cada variável de produto *m* no componente *a* (WOLD *et al.*, 2001; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2009b).

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

De acordo com Ferrer *et al.* (2008) e Anzanello *et al*. (2009a), os vetores $\mathbf{w}_a$ e $\mathbf{c}_a$ são selecionados de forma a maximizar a covariância entre os componentes $\mathbf{t}_a$ e $\mathbf{u}_a$. Além disso, tais componentes aglutinam informações sobre as observações e suas semelhanças em relação ao modelo (WOLD *et al.*, 2001). Wold *et al.* (2001) afirmam ainda que $\mathbf{w}_a$ e $\mathbf{c}_a$ fornecem informações sobre como as variáveis se combinam para formar a relação quantitativa entre **X** e **Y**, sinalizando as variáveis *x* de maior relevância (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 $\mathbf{t_a}$, pode-se reconstituir a matriz **X** com valores reduzidos dos resíduos e_{ik} (WOLD *et al.*, 2001), conforme a equação (3). (Entre parênteses é apresentada a representação matricial do procedimento).

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

Por sua vez, a predição das variáveis de resposta *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 equação (5), onde $w^*_{ka} = w_{ka}(p_{ka}w_{ka})^{-1}$ e f_{im} são os resíduos da predição, os quais podem ser utilizados para diagnóstico da qualidade do modelo (WOLD *et al.*, 2001; ANZANELLO *et al.*, 2009a).

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

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

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

A qualidade da predição gerada pelo modelo pode ser avaliada através da soma do resíduo médio, RMSE = $\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$, onde y_i é o valor observado de y_i e $\hat{y}_i$ é o valor estimado a partir da regressão PLS (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009).

2.2.2 Abordagens para seleção de variáveis

A presença de um grande número de variáveis de produto e processo tem incentivado engenheiros e pesquisadores a buscarem modelos compostos por um número reduzido de variáveis com vistas à redução dos custos de coleta de dados, aumento da precisão das informações geradas e maior possibilidade de aplicações práticas (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009; PIERNA *et al.*, 2009; ANDERSEN; BRO, 2010).

O uso de abordagens explanatórias (por exemplo, gráficos de normalidade) com escolha manual das variáveis pode tornar-se impraticável quando as variáveis são numerosas e correlacionadas (GAUCHI; CHAGNON, 2001; ANDERSEN; BRO, 2010). Além disso, um modelo de regressão com bom ajuste aos dados não necessariamente conduz a boas predições, evidenciado por situações de *overfitting* ou quando o processo apresenta alterações no intervalo decorrido entre a construção do modelo e sua efetiva aplicação (HÖSKULDSSON, 2001; LAZRAQ *et al.*, 2003; CHONG; JUN, 2005; PIERNA *et al.*, 2009; ANDERSEN; BRO, 2010).

Dentre as abordagens para seleção de variáveis aplicadas no contexto de regressões lineares múltiplas, o método *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 na regressão, realizada com base em um teste estatístico de significância de cada variável (MONTGOMERY; RUNGER, 2009). Apesar de amplamente difundido, o desempenho do método *Stepwise* é afetado por variáveis correlacionadas e ruidosas (GAUCHI; CHAGNON, 2001).

Métodos mais robustos vêm sendo propostos para a seleção de variáveis em aplicações preditivas de PLS (HÖSKULDSSON, 2001; GAUCHI; CHAGNON, 2001; LAZRAQ *et al.*, 2003; CHONG; JUN, 2005; ZHAI *et al.*, 2006; PIERNA *et al.*, 2009). Gauchi e 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 cada variável *x* em cada componente a. O método BQ, por sua vez, sistematicamente elimina a variável associada ao menor coeficiente da regressão PLS e registra Q^2_{cum} para avaliar a qualidade da predição a cada eliminação. Por fim, o conjunto de variáveis que maximiza o Q^2_{cum} é escolhido. Já o AG, baseado num critério de busca, retém um número reduzido de variáveis e conduz a bons resultados na predição, porém apresenta alta variabilidade e requer demasiado processamento computacional.

Com propósitos semelhantes, Chong e Jun (2005) comparam o desempenho de três métodos para seleção de variáveis: método PLS-VIP (*variable importance in the projection*), regressão Lasso (*least absolute shrinkage and selection operator*) e regressão *Stepwise*. Experimentos simulados em cenários com alta colinearidade apontaram o método PLS-VIP como o mais adequado para previsões. Pierna *et al.* (2009) desenvolveram um método para seleção de variáveis espectrais baseado na regressão PLS e remoção *backward* a partir do desempenho de predição do modelo, mensurado através do RMSE. O método proposto manteve ou aumentou a capacidade de predição ao ser aplicado em dois bancos de dados com múltiplas variáveis de produto. O método assemelha-se ao aqui proposto, porém não faz o uso de indicadores de importância das variáveis para escolher a variável a ser eliminada a cada iteração (ao invés disso, utiliza uma parte do banco de dados para iterativamente identificar a variável que menos colabora com o RMSE). Além disso, o método em Pierna *et al.* (2009) não é comparado com outros métodos propostos pela literatura, o que dificulta a conclusão sobre seu desempenho.

A proposição de índices de importância das variáveis também tem encontrado elevada aplicação em procedimentos de seleção; tais índices atuam como guias no processo de eliminação ou inclusão sistemática de variáveis no modelo. Wold *et al.* (2001) desenvolveram um índice de importância das variáveis, VIP, a partir do coeficiente modificado de peso w^*_{ka} e da fração de variância explicada pelo componente a em $\boldsymbol{Y}$, R^2_{Ya}. Esse índice foi testado em Lazraq *et al.* (2003) e Anzanello *et al.* (2009a). Outros índices com propósitos semelhantes podem ser obtidos em Eriksson e Wold (2010).

Por sua vez, Anzanello *et al.* (2009a) propuseram um método para seleção de variáveis de processo para fins de classificação das variáveis de produto, 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 o subconjunto que maximiza a acurácia de classificação. No método proposto neste artigo, as variáveis são sistematicamente eliminadas com base em novos índices de importância, porém com objetivo de predição (e não de classificação).

Dos cinco índices de importância testados por Anzanello *et al.* (2009a) 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 pelo componente *a* (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 mede 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.

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

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

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

2.3 Método proposto

O método proposto, EB, é operacionalizado em cinco etapas: (*i*) divisão do banco de dados em conjuntos de treino e teste; (*ii*) aplicação da regressão PLS no conjunto de treino e geração de índices de importância das variáveis; (*iii*) predição da variável de produto *y* para o conjunto de treino e eliminação das variáveis irrelevantes e ruidosas; (*iv*) construção de um gráfico para identificação do melhor subconjunto de variáveis e (v) validação das variáveis

selecionadas no conjunto de teste. Enfatiza-se que o método proposto assume as variáveis de produto *y* como contínuas e, por isso, adaptações podem ser necessárias para uso com variáveis de produto discretas. Os passos propostos são detalhados na sequência.

2.3.1 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. Inicia-se dividindo 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 é utilizado para avaliação da capacidade predita do modelo gerado. Recomenda-se usar uma proporção de 3:2 entre as observações de N_{tr} e N_{ts}, respectivamente (ANZANELLO *et al.*, 2009a, DEHNAM, 2000).

2.3.2 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. Os parâmetros de interesse incluem os coeficientes de regressão b_{mk}, pesos w_{ka} e o percentual de variação em **Y** explicada pelo componente *a*, R^2_{Ya}. Tais parâmetros são utilizados para gerar os índices de importância das variáveis de processo.

Três são os indicadores de importância usados no presente método. O índice v_w 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, elaborado por Anzanello *et al.* (2009a), 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 (8) e (9), propõe-se um novo índice, v_{bk}, apresentado na equação (10), o qual integra 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^2_{Ya}, explicada pelo componente *a* (a = 1,...,*A*).

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

2.3.3 Etapa 3: Predizer a variável de produto 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 predição de *y*, até que reste apenas uma variável de processo.

2.3.4 Etapa 4: Construir um gráfico para identificar o melhor subconjunto de variáveis

O RMSE calculado a cada eliminação de variável é relacionado ao percentual de variáveis retidas através de um gráfico, como apresentado na Figura 2.1. A escolha do melhor conjunto de variáveis é feita subjetivamente através de uma solução de conjunto entre o menor percentual de variáveis retidas e o menor valor do indicador RMSE.

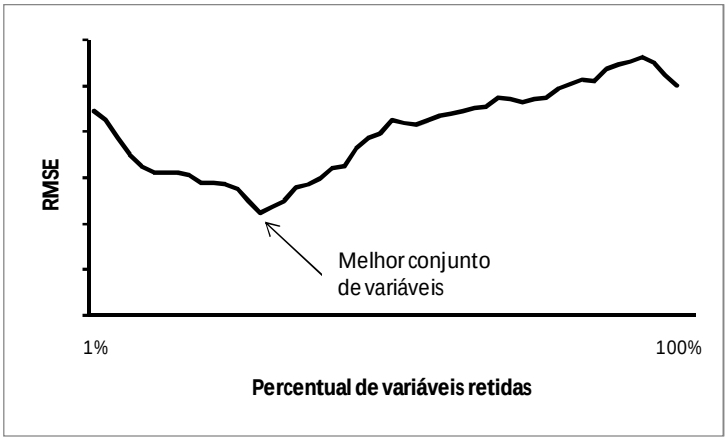

Figura 2.1 – Perfil hipotético de RMSE à medida que as variáveis de processo são eliminadas do conjunto de treino.

2.3.5 Etapa 5: Testar o subconjunto de variáveis selecionado no conjunto de teste

Fazendo uso do conjunto de variáveis selecionado na etapa anterior, aplica-se a regressão PLS para o conjunto de observações de teste e, por fim, estima-se o desempenho da predição via RMSE.

2.4 Resultados e Discussão

O desempenho do método EB, juntamente com os três índices, foi comparado frente ao método *Stepwise*. Para tanto, aplicou-se o método *Stepwise* no conjunto de treino; as variáveis selecionadas são inseridas como independentes em um modelo PLS. Comparou-se então o desempenho do método EB utilizado os índices v_w, v_k e v_{bk}, e o método *Stepwise* em termos de RMSE e percentual de variáveis retidas.

Para aplicação do método proposto e avaliação do seu desempenho foram utilizados os cinco bancos de dados em Gauchi e Chagnon (2001), os quais também constam nos trabalhos de Lazraq *et al.* (2003) e Anzanello *et al.* (2009a). 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, é apresentado na Tabela 2.1. O banco de dados ADPN, com 71 observações, é procedente de um processo intermediário da produção de nylon. As 262 observações do LATEX foram extraídas de um processo de manufatura de látex. Os dados de OXY, com 30 observações, correspondem ao processo de produção do óxido de titânio, o qual é usado na mistura de tintas. O processo SPYRA refere-se à etapa de fermentação para a produção de antibiótico. Por fim, o banco GRANU é proveniente de um processo de emulsões anti-espuma utilizado na indústria de papel.

A regressão PLS foi aplicada ao conjunto de treino de cada banco de dados. Foram retidos 3 componentes *a* da regressão PLS para cada banco de dados através de validação-cruzada [ver WOLD *et al*. (2001)], resultando nos seguintes R^2_{Ya}'s: ADPN, 94%, LATEX, 77%, OXY, 94%, SPIRA, 71% e GRANU, 86%.

A Figura 2.2 apresenta o perfil de RMSE à medida que as variáveis são eliminadas do conjunto de treino para o banco de dados OXY ao aplicar-se o método EB com o índice v_{bk}.

A escolha do melhor conjunto de variáveis considerou uma solução de compromisso entre o menor percentual de variáveis retidas e o menor RMSE. O método proposto reteve apenas 23% das variáveis, gerando um RMSE de 0,208 (valor próximo ao menor valor possível de RMSE). A utilização de todas as variáveis conduz a um RMSE de 0,257, representando um uma redução de 19% no RMSE do conjunto de treino para o banco OXY, o que representa um aumento de acurácia preditiva . A mesma lógica foi aplicada aos demais bancos de dados e índices de importância das variáveis.

Tabela 2.1 - Bancos de dados considerados.

Banco de dados	Número de variáveis de processo	Número de observações		
		Conjunto de treino	Conjunto de teste	Total de observações
ADPN	**100**	**57**	**14**	**71**
LATEX	**117**	**210**	**52**	**262**
OXY	**95**	**18**	**12**	**30**
SPIRA	**96**	**115**	**29**	**144**
GRANU	**78**	**23**	**6**	**29**

Fonte: elaborado pelos autores.

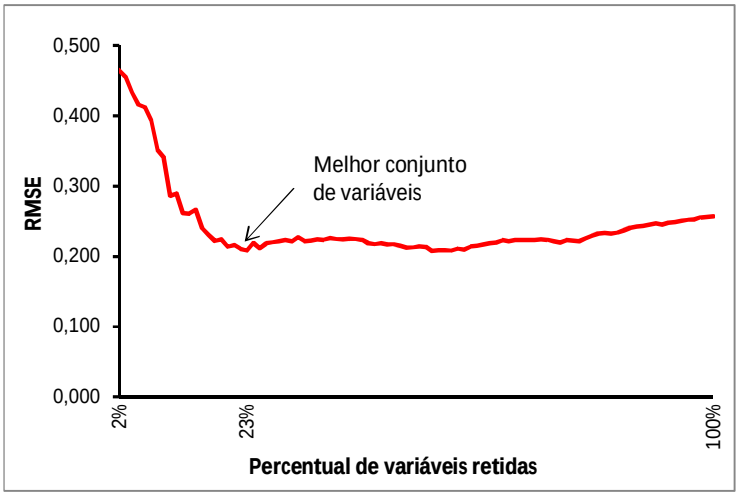

Figura 2.2 – Desempenho da predição no conjunto de treino do processo OXY usando o método EB com o índice $\boldsymbol{v_{bk}}$.

A comparação do desempenho dos distintos índices de importância nas porções de treino e teste é apresentada na Tabela 2.2. O método EB com o índice v_{bk} apresentou desempenho superior de predição no conjunto de treino de todos os processos analisados, com destaque para o processo ADPN, onde se verificou acurácia 26% superior ao segundo melhor índice. O RMSE médio para os cinco bancos é 0,494, o que incrementa a acurácia de predição média em 6% (o RMSE médio utilizando todas as variáveis é 0,525) valendo-se de 31% das variáveis originais (em média sobre todos os bancos). Tal método e índice também conduziram aos melhores resultados para as predições da porção de teste, com RMSE médio igual a 0,440. O método EB com os índices v_k e v_w alternam seus desempenhos de acordo com o banco analisado, conduzindo a maiores valores de RMSE e retendo mais variáveis do que o índice v_{bk}.

Tabela 2.2 – Desempenho dos métodos de seleção de variáveis para os conjuntos de treino e teste.

	RMSE para o conjunto de treino				**Variáveis retidas (%)**				**RMSE para o conjunto de teste**			
Processo	$\mathbf{v_k}$	$\mathbf{v_w}$	$\mathbf{v_{bk}}$	**Stepwise**	$\mathbf{v_k}$	$\mathbf{v_w}$	$\mathbf{v_{bk}}$	**Stepwise**	$\mathbf{v_k}$	$\mathbf{v_w}$	$\mathbf{v_{bk}}$	**Stepwise**
ADPN (100)	1,110	1,216	0,961	1,102	56%	36%	37%	31%	1,051	1,228	0,968	1,225
LATEX (117)	0,602	0,586	0,546	0,588	7%	16%	24%	15%	0,594	0,573	0,549	0,610
OXY (95)	0,227	0,225	0,208	0,213	0%	0%	23%	0%	0,126	0,121	0,089	0,172
SPIRA (96)	0,160	0,157	0,156	0,161	57%	47%	41%	10%	0,148	0,147	0,147	0,182
GRANU (78)	0,654	0,638	0,601	0,669	28%	37%	28%	6%	0,742	0,455	0,448	0,978
Média	0,551	0,564	0,494	0,547	30%	27%	31%	13%	0,532	0,505	0,440	0,633

Fonte: elaborado pelos autores.

Por fim, o método EB com o índice selecionado (v_{bk}) é comparado ao tradicional método *Stepwise*. O método proposto conduz a predições mais precisas tanto para a porção de treino (RMSE=0,494 contra RMSE=0,547 do método *Stepwise*), quanto para a porção de teste (RMSE=0,440 contra RMSE=0,633 do método *Stepwise*). Essa última vantagem é expressiva, pois, apesar do método proposto reter mais variáveis do que o *Stepwise*, conduz a predições 44% mais acuradas do que o modelo gerado pela seleção tradicional.

2.5 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. O presente artigo apresentou um método de seleção das variáveis de processo mais relevantes com vistas à predição das variáveis de produto.

O método EB proposto se apoia nas seguintes etapas: (1) divisão dos bancos de dados compostos por variáveis de processo e produto 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_{bk}; (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 para seleção do subconjunto recomendado e (5) predição da variável de produto para o conjunto de teste usando tal subconjunto de variáveis.

O novo método EB com o novo índice de importância de variáveis, v_{bk}, foi comparado ao método fazendo uso dos índices v_w e v_k, também gerados com base nos parâmetros da regressão PLS. O método EB com o novo índice v_{bk} apresentou a melhor acurácia de predição de **Y**, quando comparado ao método com os índices v_k e v_w e ao método tradicional *Stepwise*. Além disso, o método EB com o índice v_{bk} reteve um percentual de variáveis inferior ao obtido com o método EB com os índices v_k e v_w. Ao valer-se do índice v_{bk}, o método proposto utilizou 31% das variáveis para predição, gerando uma 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 eliminar a lógica de seleção do melhor subconjunto de variáveis a partir da análise gráfica entre desempenho preditivo e percentual de variáveis retidas no modelo final. 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 também é um estudo relevante. 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.

2.6 Referências Bibliográficas

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 24, p. 728-737, 2010.

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

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Identificação das variáveis mais relevantes para categorização de bateladas de produção: reduzindo a variância do percentual de variáveis retidas. **Produto&Produção**, Porto Alegre, Brasil, v. 10, n.3, p.19-27, 2009b.

CHIANG, L. H.; PELL, R. J. Genetic algorithms combined with discriminant analysis for key variable identification. **Journal of Process Control**, Amsterdam, Holland, v. 14, p. 143-155, 2004.

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 78, p. 103-112, 2005.

DENHAM, M. C. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 14, p. 351-361, 2000.

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. **Journal of Chemometrics,** Bognor Regis, United Kingdom, v. 24, p. 779-789, 2010.

ESPOSITO-VINZI V.; CHIN, W.; HENSELER, J.; WANG, W. **Handbook of Partial Least Squares:** Concepts, Methods and Applications in Marketing and Related Fields. 1st. ed. Berlin: Springer, 2007. 850 p.

FERRER, A.; AGUADO, D.; VIDAL-PUIG, S.; PRATS, J.; ZARZO, M. PLS: A versatile tool for industrial process improvement and optimization. **Applied Stochastic Models in Business and Industry**, Malden, USA, v. 24, p. 551-567, 2008.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 171-193, 2001.

HÖSKULDSSON, A. Variable and subset selection in PLS regression. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 55, 23-38, 2001.

LAZRAQ, A.; CLÉROUX, R. The PLS multivariate regression model: testing the significance of successive PLS components. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 15, p. 523–536, 2001.

LAZRAQ, A.; CLÉROUX, R.; GAUCHI, J.-P. Selecting both latent and explanatory variables in the PLS1 regression model. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 66, 117-126, 2003.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 28, p. 3-21, 1995.

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, Amsterdam, Holland, v. 23, p. 35-44, 1999.

MONTGOMERY, D. C. **Introdução ao controle estatístico da qualidade**. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A, 2004. 513 p.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A, 2009. 493 p.

PIERNA, J. A. F.; ABBAS, O.; BAETEN, V.; DARDENNE, P. A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica Acta**, Amsterdam, Holland, v. 642, p. 89-93, 2009.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 109-130, 2001.

ZHAI, H. L.; CHEN, X. G.; HU, Z. De. A new approach for the identification of important variables. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 80, p. 130-135, 2006.

3 Segundo Artigo: Desempenho de dois novos métodos de seleção de variáveis de processo na presença de ruído e multicolinearidade

Juliano Zimmer

Michel Jose Anzanello

Artigo enviado para publicação na revista Produção (ABEPRO)

Resumo

Este artigo aprimora o método EB (Eliminação Backward) de Zimmer e Anzanello (2011) com vistas à eliminação do aspecto subjetivo na seleção do melhor subconjunto de variáveis para propósitos de predição, originalmente operacionalizado através de análise gráfica. Os novos métodos geram, assim como o anterior, o índice de importância das variáveis v_{bk} a partir dos parâmetros gerados pela regressão PLS (*Partial Least Squares*), o qual é usado para sinalizar a ordem de eliminação das variáveis do modelo. A cada iteração a acurácia de predição é mensurada, fazendo uso do indicador RMSE (*root mean square error*). Na sequência, o conjunto de variáveis de processo é selecionado através de duas sistemáticas. A primeira, que corresponde ao método EBM (Eliminação *backward* por mínimos), seleciona o conjunto que minimiza o erro de predição (RMSE), ao passo que o segundo, EBDE (Eliminação *backward* por distância euclidiana), seleciona o conjunto que minimiza a distância euclidiana entre os pontos gerados pela eliminação das variáveis e um ponto hipotético definido pelo usuário como ideal (caracterizado pelo reduzido valor de RMSE e reduzido percentual de variáveis retidas). Estudos de simulação permitem avaliar as vantagens e desvantagens de cada método, bem como sua robustez frente à presença de ruído nas variáveis de processo, correlação e distintas proporções entre número de observações e variáveis.

Palavras-chave: Seleção de variáveis, Regressão PLS, Indicador de importância das variáveis.

Abstract

This paper improves the method proposed in Zimmer and Anzanello (2011) by eliminating the subjective graphical analysis used to select the best subset of predictive variables. The new method uses the variable importance index v_{bk} to guide the backward variable elimination; after each removal, prediction accuracy is measured through the RMSE until a single variable is left. The recommended subset of variables is selected based on two approaches: (*i*) the EBM, which selects the subset leading to the minimum prediction error index (RMSE), and (*ii*) the EBDE, which selects the set that minimizes the Euclidian distance between the points generated by variable elimination and a hypothetical ideal point defined by the user. The ideal point is characterized by reduced RMSE and reduced percentage of retained variables. Simulation experiments evaluate the performance of each approach as different levels of variable noise, correlations and proportions between the number of observations and variables are considered.

Keywords: Variable selection, PLS regression, Variable importance indices.

3.1 Introdução

O monitoramento e controle de processos industriais tem como base a identificação das variáveis de processo com maior influência sobre as variáveis de produto. Isso é particularmente importante em processos industriais com elevado número de variáveis, onde selecionar um conjunto reduzido de variáveis pode levar à redução dos custos de coleta, aumento da precisão das informações geradas e maior possibilidade de aplicações práticas (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009; PIERNA *et al.*, 2009; ANDERSEN; BRO, 2010; GONZÁLEZ; SÁNCHEZ, 2010).

Para lidar com o elevado número de variáveis de processo e o consequente volume de informações coletadas, faz-se necessário considerar métodos capazes de lidar com ruído, colinearidade e dados faltantes (KOURTI; MACGREGOR, 1995; ANZANELLO *et al.*, 2009a; ANDERSEN; BRO, 2010). Em processos industriais com variáveis de processo correlacionadas, monitorar as variáveis de produto separadamente, sem considerar a relação das mesmas com as variáveis de processo, pode ser inadequado e pouco efetivo. Da mesma forma, monitorar variáveis redundantes pode representar desperdício de tempo e recursos

financeiros. Ao mesmo tempo, desconsiderar uma variável relevante pode reduzir a capacidade de identificar quando o processo sai de controle (GONZÁLEZ; SÁNCHEZ, 2010).

Dentre os métodos multivariados de análise estatística destaca-se a regressão PLS (*Partial Least Squares*), a qual reduz um conjunto de variáveis de resposta (produto) e de processo em um número diminuído de combinações lineares ou estruturas latentes. Métodos para seleção de variáveis baseados em regressões PLS têm sido vastamente desenvolvidos e aplicados na indústria, notadamente em função da sua capacidade em tratar um grande número de variáveis de produto e processo, dados faltantes, colinearidade e ruído (KOURTI; MACGREGOR, 1995; WOLD *et al.*, 2001; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2011; ANDERSEN; BRO, 2010).

Para orientar a seleção das variáveis mais relevantes para monitoramento do processo, índices de importância das variáveis têm sido propostos na literatura [ver Wold *et al.* (2001), Lazraq *et al.* (2003), Anzanello *et al.* (2009a), Eriksson e Wold (2010) e Zimmer e Anzanello (2011)]. O trabalho de Zimmer e Anzanello (2011) propõe e avalia o desempenho do método EB (Eliminação Backward) com três índices de importância das variáveis, gerados a partir de parâmetros da regressão PLS, os quais são usados como referência na eliminação *backward* das variáveis do modelo de predição. O melhor conjunto de variáveis considera uma solução de compromisso entre o menor percentual de variáveis retidas e o melhor resultado em termos de precisão de predição (RMSE), subjetivamente realizada através de análise gráfica. O método considerando os três índices foram testados em 5 bancos de dados industriais e foram avaliados em termos do percentual de variáveis retidas e da acurácia de predição. Por fim, o método proposto foi comparado ao método *Stepwise*.

Apesar do método EB em Zimmer e Anzanello (2011) ter apresentado desempenho satisfatório ao ser aplicado em bancos de dados reais, considera-se pouco prática a seleção subjetiva do melhor conjunto de variáveis através de análise gráfica. Tal sistemática de seleção de conjuntos de variáveis pode conduzir a resultados distintos quando avaliados por especialistas diversos, comprometendo a consistência do método. De tal forma, torna-se necessário o desenvolvimento de uma sistemática estruturada (ou seja, não dependente de avaliação subjetiva) para identificação do subconjunto de variáveis a ser retido. Complementarmente, vê-se como pertinente simular e avaliar o comportamento do novo

método em cenários afetados por diferentes níveis de ruído, colinearidade e proporção entre variáveis de processo e observações.

Este artigo propõe dois novos métodos alternativos para identificação do subconjunto de variáveis a ser retido com base no método proposto em Zimmer e Anzanello (2011). Ao invés de identificar o subconjunto de variáveis mais relevantes com base na análise do gráfico "percentual de variáveis retidas versus precisão de predição", pretende-se (*i*) selecionar o subconjunto que conduz ao mínimo erro de predição, independente do percentual de variáveis retido; e (*ii*) selecionar o conjunto que minimiza a distância euclidiana de cada ponto gerado pela eliminação sistemática de variáveis proposta em Zimmer e Anzanello (2011) em relação a um ponto ideal (de percentual de variáveis retidas versus desempenho em termos de predição) definido pelo usuário. A principal vantagem da alternativa (*ii*) está na minimização da possibilidade de sobreajuste (*overfitting*) do modelo, visto que tal sistemática penaliza soluções que retenham elevado percentual de variáveis.

O artigo está organizado em 4 seções, além desta introdução. Os conceitos da regressão PLS e de métodos para seleção de variáveis são apresentados na Seção 2. A seguir tem-se a descrição do método proposto na Seção 3, enquanto que a Seção 4 apresenta a simulação e os resultados obtidos. A conclusão do trabalho na Seção 5 encerra o artigo.

3.2 Fundamentação teórica

A regressão PLS relaciona a matriz de variáveis de processo **X** (variáveis independentes) com a matriz de variáveis de produto **Y** (variáveis dependentes), reduzindo a quantidade de variáveis de processo e produto a um número pequeno combinações lineares, as quais são usadas com propósitos de predição e controle de processo. A regressão PLS (composta por *x*) à matriz **Y** (composta por *y*), permitindo analisar dados com forte correlação, elevados níveis de ruído e desequilíbrio entre o número de variáveis e observações. Tal regressão gera um conjunto de parâmetros, que propiciam informações sobre a estrutura e comportamento de **X** e **Y**, o que contribui para o uso combinado com métodos de seleção de variáveis (WOLD *et al.*, 2001; ADBI, 2003).

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

(produto) e *N* o número de observações. O vetor $\mathbf{x}_i$ (x_{i1}, x_{i2}, ..., x_{ik}) representa a observação *i* para cada variável de processo *k*, enquanto que o vetor $\mathbf{y}_i$ (y_{i1}, y_{i2}, ..., y_{im}) representa a observação *i* para cada variável de resposta *m*. A regressão PLS gera *A* variáveis latentes (combinações lineares) $\mathbf{t}_a$ (a=1,2,...,A), conforme apresentado na equação (1). O vetor $\mathbf{w}_a$ (w_{1a},w_{2a}, ...,w_{ka}) representa os coeficientes das combinações lineares independentes das variáveis *x*. (WOLD *et al.*, 2001; ADBI, 2003; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2009b). O número de componentes a serem mantidos no modelo é obtido em função da capacidade de predição de cada componente *a*, sendo que a validação cruzada é uma das técnicas mais utilizadas para suportar essa decisão. (WOLD *et al.*, 2001; HÖSKULDOSSON, 2001).

$$\boldsymbol{t}_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{ka}x_{ik} = \boldsymbol{w}'_a\boldsymbol{x}_i \tag{1}$$

As variáveis latentes $\mathbf{u}_a$ (a=1,2,...,A) são combinações lineares das variáveis *y*. O peso de cada variável de produto *m* no componente *a* é representado pelo vetor $\mathbf{c}_a$ (c_{1a},c_{2a}, ...,c_{ma}) (WOLD *et al.*, 2001; ADBI, 2003; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a).

$$\boldsymbol{u}_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + \ldots + c_{ma}y_{im} = \boldsymbol{c}'_a\boldsymbol{y}_i \tag{2}$$

A matriz **X** pode ser reconstituída através da multiplicação do vetor de cargas das variáveis de processo, $\mathbf{p}_a$ (p_{1a},p_{2a}, ...,p_{ka}), pelo vetor $\mathbf{t_a}$, (WOLD *et al.*, 2001; ADBI, 2003), onde e_{ik} é o termo de erro.

$$x_{ik} = \sum_a t_{ia}p_{ak} + e_{ik} \tag{3}$$

A multiplicação do vetor $\mathbf{u_a}$ pelos coeficientes $\mathbf{c}_a$ gera a matriz Y, com resíduos de predição reduzidos g_{im}. (WOLD *et al.*, 2001):

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

A partir do coeficiente alterado $w^*_{ka} = w_{ka}(p_{ka}w_{ka})^{-1}$, pode-se definir o coeficiente b_{mk} da regressão PLS. A qualidade do modelo pode ser avaliada pelos resíduos da predição f_{im} (WOLD *et al.*, 2001; ADBI, 2003; ANZANELLO *et al.*, 2009a).

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

A regressão PLS tem sido amplamente utilizada na seleção de variáveis em processos industriais com vistas à predição das características do produto final [ver Forina *et al*., 1999; Sarabia *et al*., 2001, Lindgren *et al*., 1994, Wold *et al.*, 2001; Gauchi e Chagnon, 2001; Lazraq *et al.*, 2003, Meiri e Zahavi, 2006; Ozturk *et al.*, 2006; Olafsson *et al.*, 2008; Zimmer e Anzanello, 2011). Wold *et al.* (2001), por exemplo, utilizou a regressão PLS para seleção de variáveis em processo processos de reciclagem de papel, e Gauchi e Chagnon (2001) e Lazraq *et al*. (2003) em diversos processos químicos.

A avaliação do desempenho de diferentes métodos para seleção de variáveis é tema do trabalho de diversos autores [ver Almoy, 1996; Baumann *et al*, 2002; Chong e Jun, 2005; Höskuldsson, 2001; Lazraq *et al.*, 2003; Zhai *et al.*, 2006; Pierna *et al.*, 2009]. Dentre estes trabalhos, destaca-se o artigo de Chong e Jun (2005) que avaliou o desempenho de três métodos distintos de seleção de variáveis para predição, simulando alterações nos dados em termos de colinearidade das variáveis, nível de ruído e razão de variáveis relevantes.

Anzanello *et al.* (2009a) criaram um *framework* para identificar um conjunto de variáveis visando classificar as variáveis de produto. Para tanto, uma rotina de eliminação de variáveis do tipo *backward* é definida, na qual um índice de importância das variáveis sinaliza qual seria a variável a ser eliminada em cada iteração. A acurácia de classificação é avaliada a cada iteração, sendo escolhido o subconjunto que maximiza a acurácia de classificação.

O uso de índices de importância das variáveis para conduzir o processo de eliminação ou inclusão sistemática de variáveis no modelo é tema recorrente na literatura. Wold *et al.* (2001) desenvolveram um índice de importância das variáveis, VIP, a partir do coeficiente modificado de peso w_{ka}^{*} e da fração de variância explicada pelo componente a em $\boldsymbol{Y}$, R_{Ya}^{2}. Esse índice foi testado em Lazraq *et al*. (2003) e Anzanello *et al.* (2009a). Outros índices com propósitos semelhantes podem ser obtidos em Eriksson e Wold (2010).

O trabalho de Zimmer e Anzanello (2011) propõe uma adaptação do método proposto por Anzanello *et al* (2009a), o método EB (Eliminação *Backward*), para selecionar as variáveis de processo para predição das variáveis de produto. 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 produto. Inicia-se então um processo de eliminação de variáveis do tipo *backward*, sendo a ordem de eliminação definida pelo índice de importância. A cada iteração gera-se o modelo de predição com as variáveis de processo retidas, calcula-se e faz-se o registro do indicador de acurácia de predição RMSE (*Root Mean Square Error)*. A seleção do conjunto de variáveis final é realizada através da análise do gráfico de valores de RMSE versus o percentual de variáveis retidas no modelo preditivo a cada iteração. Zimmer e Anzanello (2011) testam três índices de importância de variáveis em termos de seu desempenho. O método EB utilizando o índice recomendado, v_{bk}, reteve apenas 31% das variáveis originais e aumentou a acurácia de predição do conjunto de teste em 6% (avaliado por intermédio do indicador RMSE - *Root Mean Square Error*).

A equação 6 apresenta o índice, v_{bk}, o qual integra 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^2_{Ya}, explicada pelo componente a (a = 1,...,*A*).

$$v_{bk} = \frac{\sum_{a=1}^{A}|w_{ka}|R^2_{Ya}}{\max_{k\in K}(\sum_{a=1}^{A}|w_{ka}|R^2_{Ya})} \frac{\sum_{m=1}^{M}|b_{mk}|}{\max_{k\in K}(\sum_{m=1}^{M}|b_{mk}|)} \; k = 1, \dots, K. \qquad (6)$$

Embora o método EB, de Zimmer e Anzanello (2011), tenha apresentado desempenho aceitável, avalia-se que a escolha do melhor conjunto de variáveis através de análise gráfica pode ser pouco prática e levar a distintos quando aplicados por especialistas diversos, comprometendo a consistência do método. As Figuras 3.1 e 3.2 ilustram casos reais onde a escolha do melhor conjunto não é unânime, sendo dependente do que o especialista priorizar: percentual de variáveis retidas ou capacidade preditiva (menor valor de RMSE) do modelo. Sendo assim, considera-se pertinente desenvolver uma sistemática estruturada para selecionar subconjunto de variáveis a ser retido para predição da variável de produto.

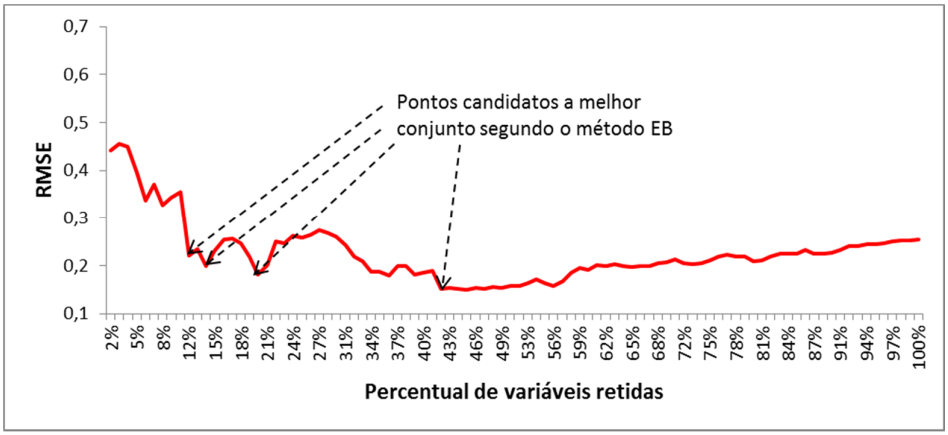

Figura 3.1 – Exemplo 1 do perfil de RMSE versus variáveis retidas no modelo final para uma aplicação real do método EB

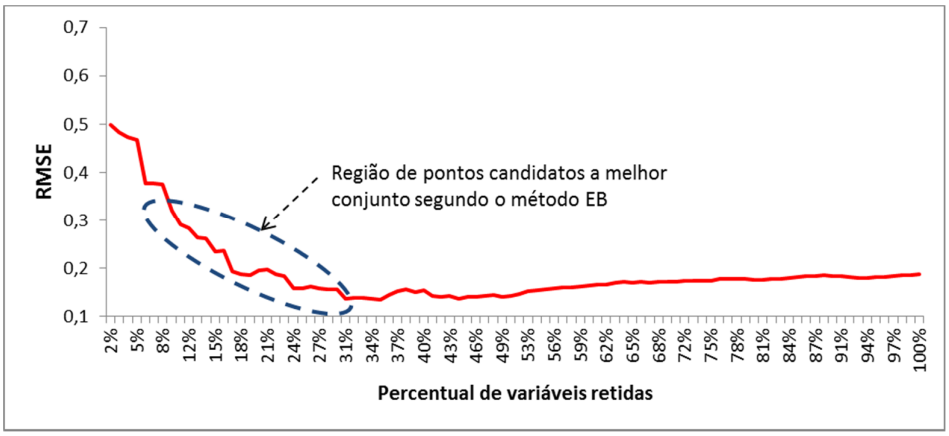

Figura 3.2 – Exemplo 2 do perfil de RMSE versus variáveis retidas no modelo final para uma aplicação real do método EB

3.3 Método proposto

As três etapas do método são detalhadas a seguir.

3.3.1 Etapa 1: Aplicação da regressão PLS no conjunto de treino e geração do índice de importância das variáveis

Considere as matrizes **X** e **Y**, introduzidas na Seção 2, com *N* observações, *K* variáveis de processo e uma variável de produto. O primeiro passo é separar aleatoriamente o banco de

dados em duas partes, um conjunto de treino, para identificar as variáveis mais relevantes, e um conjunto de teste, para avaliar a capacidade preditiva do modelo a ser gerado. Sugere-se uma proporção de 3:2 entre o conjunto de treino e teste, respectivamente (ANZANELLO *et al.*, 2009a, DENHMAN, 2000). Antes de aplicar a regressão PLS, recomenda-se normalizar os dados. Os parâmetros gerados pela regressão PLS permitem calcular o índice de importância das variáveis de processo v_{bk} conforme a equação (6)

3.3.2 Etapa 2: Predição da variável de produto y para o conjunto de treino e eliminação das variáveis irrelevantes e ruidosas

Na primeira iteração gera-se a predição de *y* para o conjunto de treino e avalia-se a acurácia da predição gerada pelo modelo, calculada através da soma do resíduo médio, 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 predito pela regressão PLS (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009; CHONG; JUN, 2005).

Na sequência, remove-se do conjunto de treino a variável com o menor índice de importância, aplica-se a regressão PLS no conjunto de treino 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 no conjunto de treino para predição de *y*, até que reste apenas uma variável de processo.

3.3.3 Etapa 3: Escolha do melhor conjunto de variáveis e aplicação de regressão PLS no conjunto de testes

A última etapa tem duas variações para identificação do melhor subconjunto de variáveis a ser retido:

- Método EBM – seleciona-se o subconjunto de variáveis que minimiza o valor do RMSE;
- Método EBDE – seleciona-se o subconjunto de variáveis que minimiza a distância euclidiana em relação a um ponto ideal, definido como o menor o percentual

possível de variáveis retidas (igual a 1/K) e menor valor possível de RMSE (igual a zero).

A Figura 3.3 ilustra um perfil hipotético de RMSE versus percentual de variáveis retidas no modelo final, no qual pode-se observar o resultado da aplicação dos métodos EBM e EBDE.

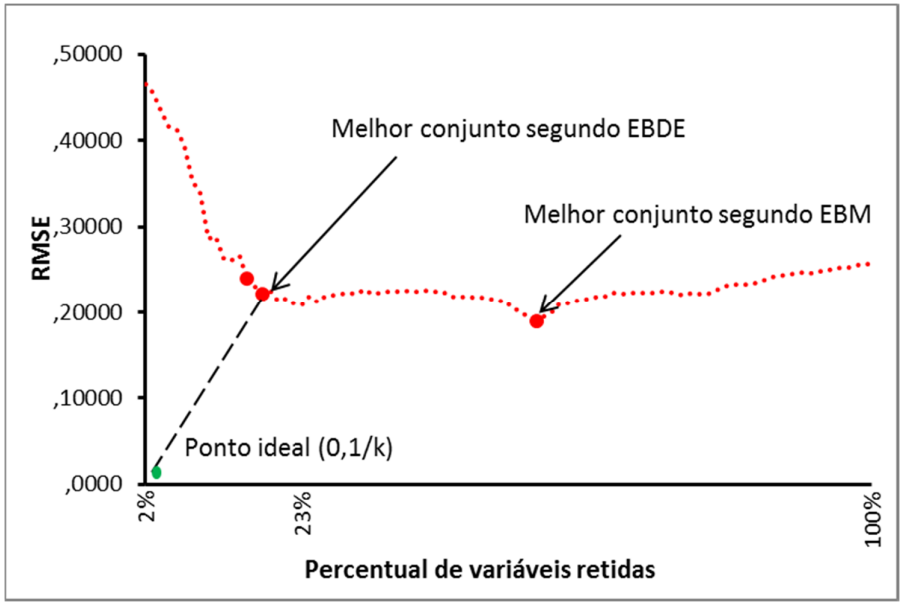

Figura 3.3 – Perfil hipotético de RMSE versus variáveis retidas no modelo final, com a ilustração dos pontos escolhidos pelos métodos EBM e EBDE

Por fim, as variáveis selecionadas são testadas no conjunto de teste, que representa novas observações a serem modeladas e preditas pelo modelo PLS gerado.

3.4 Simulação e resultados

As alternativas para identificação do subconjunto de variáveis a serem retidas (etapa 3) são simuladas considerando diferentes cenários de variância do erro, correlação entre as variáveis e proporção entre número de observações e número de variáveis.

Os parâmetros usados como input da simulação são baseados em um banco de dados real, extraído de Gauchi e Chagnon (2001), correspondente ao processo de produção do óxido de titânio, que é usado na mistura de tintas. Tal banco de dados possui 95 variáveis de processo, uma variável de produto e 30 observações. Assume-se que o processo possa ser representado por um modelo PLS, conforme equação (7).

$$y_i = \sum_a b_k x_{ik} + \varepsilon_i \tag{7}$$

onde b_k é o coeficiente da regressão PLS e $\varepsilon_i \sim N(0, \sigma^2)$.

A matriz X é gerada a partir de uma distribuição multinormal com média μ_k para as Kvariáveis de processo e matriz de correlação τ, sendo μ_ke τ estimados com base no banco de dados real.

A estimativa de b_k é oriunda da aplicação da regressão PLS sobre os dados originais. A variância do erro σ^2 é estimada pela soma do resíduo de predição SRP_A do modelo PLS gerado, conforme a equação (8), onde A denota o número de componentes retidos e n é o número total de observações (DENHAM, 2000; ANZANELLO *et. al*, 2009b) .

$$\sigma^2 = \frac{SRP_A}{n - A - 1} \tag{8}$$

Foram simulados 3 níveis de variância de erro, 3 níveis de correlação entre as variáveis e 2 níveis de proporção entre número de observações e número de variáveis de processo, conforme apresentado na Tabela 3.1. Os 18 $(= 3 \times 3 \times 2)$ cenários foram repetidos 100 vezes cada. Os níveis nominais para a variância do erro (σ^2) e correlação (τ) são extraídos do banco de dados real.

O desempenho das simulações é apresentado na Tabela 3.2. Verifica-se desempenho superior da acurácia de predição (RMSE) do método EBM frente ao EBDE para todas as combinações de níveis de fatores, porém com base na retenção de um percentual superior de variáveis. Percebe-se ainda menor variabilidade (desvio-padrão) do RMSE para o método EBM.

Tabela 3.1 – Fatores e níveis da simulação

Fatores	**Níveis**
Variância do erro	$0{,}5\sigma^2$; σ^2; $2\sigma^2$
Correlação entre as variáveis	$\tau;\tau^2;\tau^3$
Proporção entre número de observações e número de variáveis de processo	0,5; 5

Fonte: elaborado pelos autores.

Tabela 3.2– Desempenho dos métodos EBM e EBDE no conjunto de teste para cada nível simulado

Proporção de observações p/ variáveis	Correlação entre variáveis	Variação no Erro	RMSE		RMSE (Desvio Padrão)		RMSE Todas Variáveis		% Variáveis Retidas		% Variáveis Retidas (Desvio Padrão)	
			EBM	EBDE	EBM	EBDE	EBM	EBDE	EBM	EBDE	EBM	EBDE
0,5	τ	$0{,}5\sigma^2$	0,109	0,211	0,027	0,108	0,140	0,140	61%	61%	23%	54%
0,5	τ	σ2	0,106	0,214	0,024	0,036	0,133	0,133	67%	8%	21%	3%
0,5	τ	2σ2	0,114	0,212	0,025	0,034	0,145	0,145	63%	53%	22%	39%
0,5	τ^2	$0{,}5\sigma^2$	0,083	0,199	0,018	0,029	0,093	0,093	71%	7%	20%	4%
0,5	τ^2	σ2	0,086	0,196	0,016	0,028	0,097	0,097	66%	6%	21%	4%
0,5	τ^2	2σ2	0,087	0,201	0,021	0,031	0,095	0,095	74%	8%	20%	4%
0,5	τ^3	$0{,}5\sigma^2$	0,074	0,245	0,016	0,034	0,084	0,084	12%	12%	5%	5%
0,5	τ^3	σ2	0,077	0,255	0,017	0,034	0,088	0,088	13%	13%	5%	5%
0,5	τ^3	2σ2	0,080	0,252	0,018	0,033	0,091	0,091	12%	12%	4%	4%
5	τ	$0{,}5\sigma^2$	0,116	0,238	0,009	0,026	0,134	0,134	8%	8%	1%	1%
5	τ	σ2	0,119	0,245	0,010	0,024	0,137	0,137	8%	8%	2%	2%
5	τ	2σ2	0,120	0,247	0,010	0,024	0,138	0,138	8%	8%	2%	2%
5	τ^2	$0{,}5\sigma^2$	0,133	0,223	0,006	0,011	0,144	0,144	9%	9%	3%	3%
5	τ^2	σ2	0,133	0,226	0,006	0,011	0,144	0,144	9%	10%	3%	3%
5	τ^2	2σ2	0,134	0,226	0,007	0,013	0,144	0,144	9%	9%	3%	3%
5	τ^3	$0{,}5\sigma^2$	0,089	0,305	0,009	0,013	0,097	0,097	16%	16%	3%	3%
5	τ^3	σ2	0,090	0,304	0,010	0,011	0,098	0,098	16%	16%	3%	45%
5	τ^3	2σ2	0,093	0,306	0,009	0,011	0,101	0,101	16%	16%	3%	3%

Fonte: elaborado pelos autores.

Quanto ao efeito dos fatores, observa-se que aumentos na proporção entre observações e variáveis de processo reduzem o percentual de variáveis retidas, visto que maior volume de informação é oferecido à regressão PLS e, por consequência, variáveis mais relevantes e em menor número são identificadas pelo índice de importância. Incrementos nos níveis de ruído não afetam significativamente a acurácia e percentual de variáveis retidas, fato associado à

robustez da regressão PLS frente a cenários ruidosos. A redução da correlação entre as variáveis de processo, por sua vez, aumenta a precisão de predição. Tal fato está associado ao melhor desempenho da regressão PLS na geração dos parâmetros e, por consequência, na obtenção dos índices de importância das variáveis.

O ponto ideal para o método EBDE é definido com coordenadas no eixo x=1/K, onde K=95 variáveis de processo no modelo usado na simulação. No eixo y deseja-se minimizar o RMSE, então o valor é zero.

As Figuras 3.4 e 3.5 ilustram dois casos simulados da Tabela 3.2: o primeiro considera proporção igual a 0,5, correlação τ e ruído igual a $0{,}5\sigma^2$; o segundo considera proporção igual a 5, correlação τ e ruído σ^2. A dispersão dos pontos na Figura 3.4 corrobora a alta variabilidade no percentual de variáveis retidas e no RMSE gerada pelo EBM e pelo EBDE. Já a Figura 3.5, descrevendo uma situação caracterizada por maior proporção entre número de observações e variáveis, apresenta menor variabilidade.

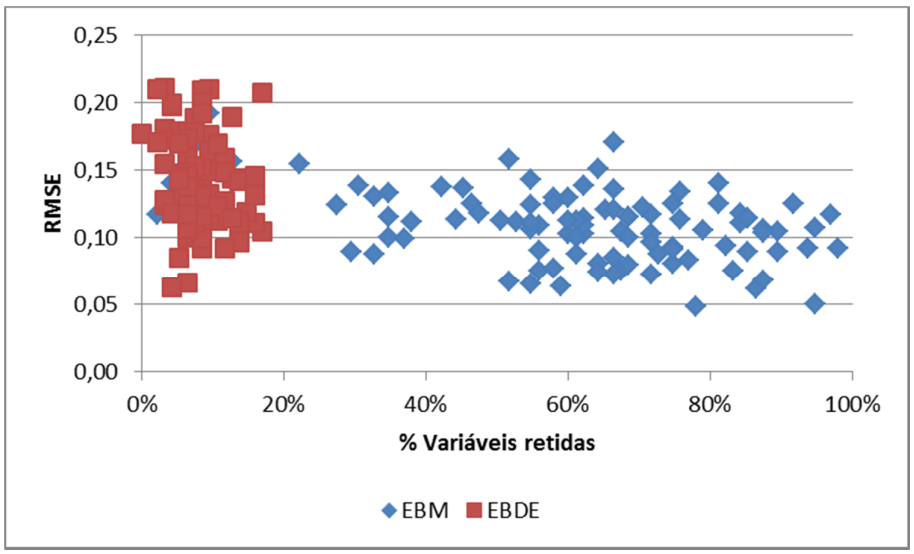

Figura 3.4 – Diagrama de dispersão para os fatores e níveis [0,5 τ $0{,}5\sigma^2$]

A Tabela 3.3 apresenta a média de desempenho de EBM e EBDE no conjunto de teste dos dados simulados, considerando todos os fatores e níveis. O método EBM apresenta maior acurácia de predição (valor do RMSE 58% menor) quando comparado ao método EBDE.

Quando se compara a acurácia considerando todas as variáveis no modelo (RMSE = 0,117), o método EBM (RMSE = 0,102) alcança um incremento de 12% na acurácia, enquanto que o método EBDE (RMSE = 0,239), representa uma perda de 105%. No entanto, o percentual de variáveis retidas para o método EBDE é significativamente menor do que o resultado gerado pelo EBM, sendo o desvio-padrão de ambos é muito próximo.

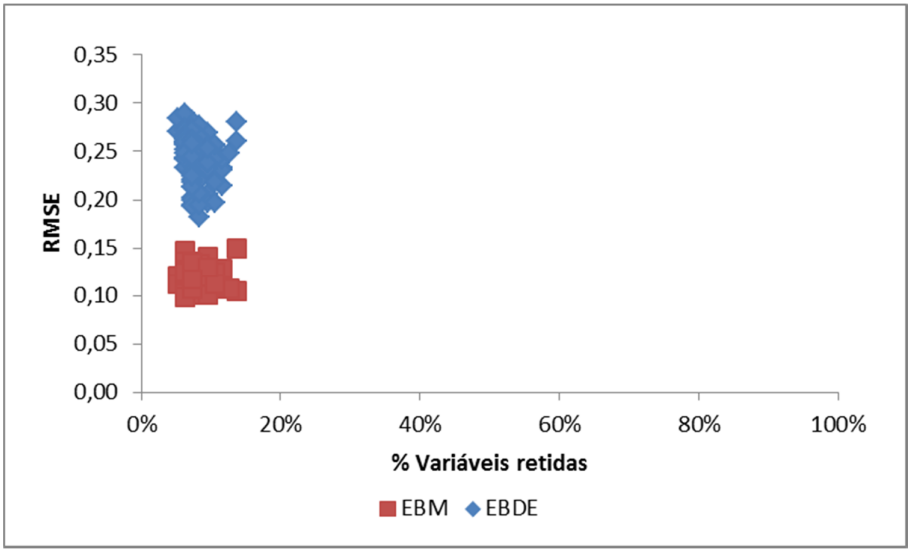

Figura 3.5 – Diagrama de dispersão para os fatores e níveis [5 τ σ^2]

Tabela 3.3 – Resumo do desempenho dos métodos no conjunto de teste dos dados simulados

	EBM	EBDE
Acurácia de predição (RMSE)	0,102	0,239
Desvio padrão da Acurácia de predição (RMSE)	0,014	0,028
Variáveis retidas (%)	30%	15%
Desvio padrão das Variáveis Retidas (%)	9%	10%
Acurácia de predição para todas variáveis (RMSE)	0,117	0,117

Fonte: elaborado pelos autores.

Portanto, o método EBM conduz a uma maior acurácia de predição à custa de um maior percentual de variáveis retidas no modelo final. O método EBDE por sua vez, retém menor percentual de variáveis, mas apresenta menor acurácia de predição. A opção por

determinada alternativa (EBM ou EBDE) recai sobre a disponibilidade de recursos para coleta de grande volume de variáveis frente à criticidade de acurácia das predições.

3.5 Conclusões

Este artigo apresentou modificações no método de Zimmer e Anzanello (2011) para eliminar o aspecto subjetivo na seleção do melhor subconjunto de variáveis com base na análise gráfica originalmente proposta. A partir dos parâmetros da regressão PLS, os novos métodos geram o índice de importância das variáveis $v_{bw,}$ o qual é usado para sinalizar a ordem de eliminação das variáveis do modelo. A cada iteração a acurácia de predição é mensurada, fazendo uso do indicador RMSE. Na sequência, o conjunto de variáveis com boa capacidade de predição da variável de produto é selecionado através de duas sistemáticas. A primeira, que corresponde ao método EBM, seleciona o conjunto que minimiza o erro de predição (RMSE), ao passo que a segunda, que corresponde ao método EBDE, seleciona o conjunto que minimiza a distância euclidiana entre o ponto ideal (menor valor possível de RMSE e % de variáveis retidas) e o ponto gerado após a eliminação de cada variável.

A simulação dos dois métodos permitiu constatar que o método EBM é mais recomendado para situações nas quais se deseja priorizar a capacidade de predição, visto que conduz a uma acurácia de predição 12% superior em relação ao modelo com todas as variáveis, retendo 30% das variáveis. O segundo método, EBDE, é mais recomendado quando objetiva-se um percentual ainda menor de variáveis retidas (visto que a média foi de 15%), aceitando-se a obtenção de predições menos precisas.

Sugere-se, entre possíveis estudos futuros, comparar o método EBM e EBDE com relação a outros métodos para seleção de variáveis propostos na literatura, e propor e testar novos índices de importância das variáveis a serem usados pelos métodos EBM e EBDE para a eliminação das variáveis. Outro possível tema de interesse é a avaliação de como alterações na definição do ponto ideal mudam os resultados do método EBDE.

3.6 Referências Bibliográficas

ABDI, H. **Partial Least Squares (PLS) Regression**. Encyclopedia of Social Sciences Research Methods. Thousand Oaks, Sage. 2003

ALMOY, T. A simulation study on comparison of prediction methods when only a few components are relevant. **Computanional Statistics & Data Analysis**, v. 21, p. 87-107, 1996.

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 24, p. 728-737, 2010.

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

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Identificação das variáveis mais relevantes para categorização de bateladas de produção: reduzindo a variância do percentual de variáveis retidas. **Produto&Produção**, Porto Alegre, Brasil, v. 10, n.3, p.19-27, 2009b.

BAUMANNN, K.; ALBERT, H.; VON KORFF, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 16, p. 339-350, 2002.

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 78, p. 103-112, 2005.

DENHAM, M. C. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 14, p. 351-361, 2000.

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. **Journal of Chemometrics,** Bognor Regis, United Kingdom, v. 24, p. 779-789, 2010.

FERRER, A.; AGUADO, D.; VIDAL-PUIG, S.; PRATS, J.; ZARZO, M. PLS: A versatile tool for industrial process improvement and optimization. **Applied Stochastic Models in Business and Industry**, Malden, USA, v. 24, p. 551-567, 2008.

FORINA, M.; CASOLINO, C.; MILLAN, C., Iterative predictor weighting (IPW) PLS: a technique for elimination of useless predictors in regression problems. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 13, p. 164-184, 1999.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 171-193, 2001.

GONZÁLEZ, I.; SÁNCHEZ, I. Variable selection for multivariate statiscal process control. Journal of Quality Technology, Milwaukee, v. 42, n.3, p. 242-259, 2010.

HÖSKULDSSON, A. Variable and subset selection in PLS regression. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 55, 23-38, 2001.

LAZRAQ, A.; CLÉROUX, R.; GAUCHI, J.-P. Selecting both latent and explanatory variables in the PLS1 regression model. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 66, 117-126, 2003.

LINDGREN, F.; GELADI, P.; RANNAR, S.; WOLD, S. Interactive variable selection (IVS) for PLS: Part 1. Theory and algorithms. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 8, p. 349-363, 1994.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 28, p. 3-21, 1995.

MEIRI, R.; ZAHAVI, J. Using simulated annealing to optimize the feature selection problem in marketing applications. **European Journal of Operational Research**, v. 171, p. 842-858, 2006.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A, 2009. 493 p.

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. **European Journal of Operational Research**, v. 187, p. 1429-1448, 2008.

OZTURK, A.; KAYALIGIL, S. OZDEMIREI, N. Manufacturing lead time estimation using data mining. **European Journal of Operational Research**, v. 187, p. 1429-1448, 2008.

PIERNA, J. A. F.; ABBAS, O.; BAETEN, V.; DARDENNE, P. A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica Acta**, Amsterdam, Holland, v. 642, p. 89-93, 2009.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 109-130, 2001.

ZHAI, H. L.; CHEN, X. G.; HU, Z. De. A new approach for the identification of important variables. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 80, p. 130-135, 2006.

ZIMMER, J.; ANZANELLO, M.J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Revista Produção**. São Paulo. Aguardando publicação, 2011.

4 Terceiro Artigo: Comparação de desempenho de três métodos de seleção de variáveis com fins de predição

Juliano Zimmer

Michel Jose Anzanello

Artigo enviado para publicação na revista Gestão e Produção

Resumo

Este artigo propõe a aplicação dos métodos EB (Eliminação Backward), EBM (Eliminação *backward* por mínimos) e EBDE (Eliminação *backward* por distância euclidiana), propostos em Zimmer e Anzanello (2011, 2012) em quatro bancos de dados reais, a fim de avaliar seu desempenho em termos de acurácia de predição e percentual de variáveis retidas no modelo final. Para mensurar a precisão de predição da variável de produto, além do tradicional indicador RMSE (*root mean square error*), este trabalho introduz o indicador Q^2cum, proposto em Gauchi e Chagnon (2001). Os três métodos EB, EBM e EBDE geram um índice de importância das variáveis com base nos parâmetros da regressão PLS; a cada iteração, uma variável é retirada do modelo de regressão de acordo com o menor índice de importância das variáveis, repetindo-se a medição da acurácia de predição através do Q^2cum e do RMSE. Para a definição do melhor subconjunto de variáveis a ser retido, cada método tem a sua sistemática: (*i*) o método EB proposto em Zimmer e Anzanello (2011), que seleciona o conjunto de variáveis com base em uma análise subjetiva do perfil de acurácia de predição *versus* percentual de variáveis retidas; (*ii*) o método EBM, que seleciona o conjunto que maximiza o desempenho do indicador de acurácia independente do percentual de variáveis retidas, proposto em Zimmer e Anzanello (2012); e (*iii*) e o método EBDE, que identifica o conjunto que minimiza a distância euclidiana entre um ponto ideal hipotético e os pontos gerados pela eliminação de cada variável (também proposto em Zimmer e Anzanello, 2011). Ao ser aplicado em quatro bancos de dados reais, o método recomendado, EBDE, faz uso de apenas 13% das variáveis originais, aumentando a acurácia de Q^2cum em 32% quando comparado à utilização de todas as variáveis.

Palavras-chave: Seleção de variáveis, Regressão PLS, Indicador de importância das variáveis.

Abstract

This paper evaluates the performance of three methods presented in Zimmer and Anzanello (2011, 2012), EB, EBM and EBDE, in four manufacturing data sets. Besides the traditional RMSE (root mean square error), this paper introduces the Q^2cum as a second index for measuring predictive accuracy as variables are removed one-by-one according to the order defined by the variable importance index proposed in Zimmer and Anzanello (2011). Each of the tested methods presents a different approach for identify the best subset of variables: (*i*) the BE method subjectively selects the subset that compromises the prediction performance and the percent of retained variables, as proposed in Zimmer and Anzanello (2011); (*ii*) the EBM method selects the subset leading to the minimum prediction error index (RMSE); and (*iii*) the EBDE selects the subset that minimizes the Euclidian distance between the points generated by variable elimination and a hypothetical ideal point defined by the user [propositions (*ii*) and (*iii*) are presented in Zimmer and Anzanello (2012)]. When applied to four manufacturing data sets, the recommended method, EBDE, retains average 13% of the original variables and increases the prediction accuracy in average 32% compared to using all the variables.

Keywords: Variable selection, PLS regression, Variable importance indices.

4.1 Introdução

Considerando a importância do monitoramento e controle da qualidade do produto final em processos industriais, torna-se vital elaborar modelos capazes de predizer as variáveis de produto que desdobram a qualidade do produto. A qualidade do produto final é determinada pelas variáveis de processo e, por isso, faz-se necessário identificar um conjunto reduzido de variáveis relevantes que descrevam características do processo e viabilizem o monitoramento e controle dos mesmos. Contudo, em processos que envolvem elevado número de variáveis, tais como a indústria química, siderúrgica, de alimentos e petroquímica, a seleção e o monitoramento preciso do processo podem ser inviabilizados caso não sejam

usadas técnicas de análises multivariadas habilitadas a tratar de ruído, colinearidade e dados faltantes (KOURTI; MACGREGOR, 1995; GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a; ANDERSEN; BRO, 2010).

A regressão PLS (*Partial Least Squares*) destaca-se entre as analises multivariadas justamente por conseguir tratar um grande número de variáveis de produto e processo, dados faltantes, colinearidade e ruído e, por isso, vem sendo usada como base de métodos de seleção de variáveis (KOURTI; MACGREGOR, 1995; WOLD *et al.*, 2001; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2012; ANDERSEN; BRO, 2010).

Índices para a quantificação da importância das variáveis podem ser gerados a partir da regressão PLS, indicando as variáveis de processo que melhor explicam a variabilidade da variável de produto (Anzanello *et al*., 2009a). Por isso, diversos autores propuseram métodos para seleção de variáveis fazendo uso de índices de importância [ver Wold *et al.* (2001), Lazraq *et al.* (2001, 2003), Anzanello *et al.* (2009a, 2012), Eriksson e Wold (2010), Zimmer e Anzanello (2011, 2012)]. Anzanello *et al.* (2009a) propõem uma abordagem para seleção de variáveis a partir de índices de importância com fins de classificação da variável de produto; a eliminação *backward* das variáveis do modelo de classificação segue a ordem definida pelos índices de importância. Zimmer e Anzanello (2011) adaptam o método de Anzanello *et al.* (2009a) para fins de predição da variável de produto, criando o método EB (Eliminação Backward) e propondo um novo índice de importância da variável, v_{bk}, e comparando o desempenho desse índice frente a outros 2 índices e ao método *Stepwise* em 5 bancos de dados reais. A escolha do melhor conjunto nesse método é realizada subjetivamente, considerando uma solução de compromisso entre a capacidade de predição (medida através do indicador *root mean square error* – RMSE) e o percentual de variáveis retidas no modelo de predição.

Zimmer e Anzanello (2012) aprimoram o trabalho de Zimmer e Anzanello (2011) através da proposição de dois métodos distintos para a seleção do conjunto de variáveis a serem usadas na predição. O primeiro método, EBM (Eliminação *backward* por mínimos), prioriza o conjunto que maximiza o indicador de acurácia de predição (menor valor de RMSE), não observando o percentual de variáveis retidas no modelo resultante; o segundo, EBDE (Eliminação *backward* por distância euclidiana), seleciona o subconjunto que minimiza a distância euclidiana entre os pontos gerados pela eliminação *backward* das variáveis (visto

que o processo iterativo de eliminação gera um perfil associando precisão de predição ao percentual de variáveis retidas) a um ponto hipotético, definido pelo usuário como ideal (caracterizado por um reduzido percentual de variáveis retidas e reduzido valor de RMSE).

Tanto o trabalho de Zimmer e Anzanello (2011) quanto o artigo de Zimmer e Anzanello (2012) fazem uso do indicador RMSE para avaliar a capacidade de predição do modelo gerado e guiar a seleção final do conjunto de variáveis. Gauchi e Chagnon (2001), que comparam 20 métodos distintos de seleção de variáveis em seu estudo, afirmam que outros indicadores de acurácia de predição podem ser mais estáveis do que o RMSE em situações onde os bancos de dados apresentam variáveis altamente correlacionadas e ruidosas. Dentre tais indicadores, destaca-se a métrica Q^2cum, que relaciona o indicador de predição, PRESS, que é a soma quadrada dos desvios da predição da variável dependente do modelo aplicado sobre dados que não foram usados na construção do modelo de regressão, e outro indicador, RSS, que é a soma quadrada dos desvios da predição da variável dependente do modelo aplicado sobre dados que foram usados na construção do modelo de regressão.

Este artigo insere o indicador de precisão de predição Q^2cum nas abordagens de seleção de variáveis propostas em Zimmer e Anzanello (2011; 2012), como alternativa ao indicador RMSE. Os três métodos valem-se de parâmetros gerados pela regressão PLS para calcular um índice de importância das variáveis de processo, o qual sinaliza as variáveis mais relevantes para explicação da variabilidade na variável de produto. Através de um processo *backward*, as variáveis são eliminadas de acordo com a ordem definida pelos índices de importância. O conjunto recomendado de variáveis é definido com base nas proposições de Zimmer e Anzanello (2011, 2012), sendo elas: (*i*) método EB - seleção do conjunto de variáveis com base em uma análise subjetiva do perfil de acurácia de predição *versus* percentual de variáveis retidas; (*ii*) método EBM - seleção do conjunto de variáveis que maximiza a acurácia de predição (maior valor de Q^2cum ou menor valor de RMSE) independente do percentual de variáveis retidas; e (*iii*) método EBDE - seleção do conjunto de variáveis que minimiza a distância euclidiana entre um ponto ideal hipotético (valor igual a zero para o RMSE e o menor percentual possível de variáveis retidas) e os valores de RMSE gerados pela eliminação de cada variável. Os três métodos são aplicados em quatro bancos de dados de processos reais.

A principal contribuição do artigo está na comparação de 3 diferentes métodos de seleção de variáveis para predição com base em índice de importância de variáveis em dados reais, o que permite avaliar qual deles é mais recomendado em qual caso. A outra contribuição está no uso do indicador de acurácia de predição Q^2cum, mais robusto do que o RMSE utilizado em Zimmer e Anzanello (2011; 2012).

O artigo está organizado em cinco seções. A fundamentação teórica, com uma breve descrição dos fundamentos da regressão PLS e dos métodos para seleção de variáveis consta na Seção 2. A próxima seção descreve os procedimentos metodológicos. Os resultados são apresentados e discutidos na Seção 4; a conclusão encerra o artigo na Seção 5.

4.2 Fundamentação teórica

A regressão PLS relaciona a matriz de variáveis de processo **X** (variáveis independentes) com a matriz de variáveis de produto **Y** (variáveis dependentes), reduzindo a quantidade de variáveis de processo e produto a um número pequeno combinações lineares, as quais são usadas com propósitos de predição e controle de processo. A regressão PLS (composta por *x*) à matriz **Y** (composta por *y*), permitindo analisar dados com forte correlação, elevados níveis de ruído e desequilíbrio entre o número de variáveis e observações. Tal regressão gera um conjunto de parâmetros, que propiciam informações sobre a estrutura e comportamento de **X** e **Y**, o que contribui para o uso combinado com métodos de seleção de variáveis (WOLD *et al.*, 2001; ADBI, 2003).

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 (produto) e N o número de observações. O vetor $\mathbf{x}_i$ (x_{i1}, x_{i2}, ..., x_{ik}) representa a observação i para cada variável de processo k, enquanto que o vetor $\mathbf{y}_i$ (y_{i1}, y_{i2}, ..., y_{im}) representa a observação i para cada variável de resposta m. A regressão PLS gera A variáveis latentes (combinações lineares) $\mathbf{t}_a$ (a=1,2,...,A), conforme apresentado na equação (1). O vetor $\mathbf{w}_a$ (w_{1a},w_{2a}, ...,w_{ka}) representa os coeficientes das combinações lineares independentes das variáveis x. (WOLD *et al.*, 2001; ADBI, 2003; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a, 2009b). O número de componentes a serem mantidos no modelo é obtido em função da capacidade de predição de cada componente a, sendo que a validação cruzada é uma das

técnicas mais utilizadas para suportar essa decisão. (WOLD *et al.*, 2001; HÖSKULDOSSON, 2001).

$$\boldsymbol{t}_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{ka}x_{ik} = \boldsymbol{w}'_a\boldsymbol{x}_i \quad (1)$$

As variáveis latentes $\mathbf{u}_a$ (a=1,2,...,A) são combinações lineares das variáveis y. O peso de cada variável de produto m no componente a é representado pelo vetor $\mathbf{c}_a$ (c_{1a},c_{2a}, ...,c_{ma}) (WOLD *et al.*, 2001; ADBI, 2003; FERRER *et al.*, 2008; ANZANELLO *et al.*, 2009a).

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

A matriz **X** pode ser reconstituída através da multiplicação do vetor de cargas das variáveis de processo, $\mathbf{p}_a$ (p_{1a},p_{2a}, ...,p_{ka}), pelo vetor $\mathbf{t_a}$, (WOLD *et al.*, 2001; ADBI, 2003). O termo de erro é representado pelo vetor e_{ik}.

$$x_{ik} = \sum_a t_{ia}p_{ak} + e_{ik} \quad (3)$$

A multiplicação do vetor $\mathbf{u_a}$ pelos coeficientes $\mathbf{c}_a$ gera a matriz Y, com resíduos de predição reduzidos g_{im}. (WOLD *et al.*, 2001):

$$\mathrm{y}_{im} = \sum_a u_{ia}c_{am} + g_{im} \qquad (\mathbf{Y} = \mathbf{UC}' + \mathbf{G}) \quad (4)$$

A partir do coeficiente alterado $w^*_{ka} = w_{ka}(p_{ka}w_{ka})^{-1}$, pode-se definir o coeficiente b_{mk} da regressão PLS. A qualidade do modelo pode ser avaliada pelos resíduos da predição f_{im} (WOLD *et al.*, 2001; ADBI, 2003; ANZANELLO *et al.*, 2009a).

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

Um das formas de avaliar a acurácia da predição gerada pelo modelo de regressão é pelo cálculo da soma do resíduo médio, RMSE = $\sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2}{\mathrm{N}}}$, onde y_i é o valor observado de y e $\hat{y}_i$ é o valor estimado a partir da regressão PLS (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; MONTGOMERY; RUNGER, 2009).

Para identificar um conjunto reduzido de variáveis preditivas, diversos autores propuseram métodos com base no uso da regressão PLS [ver Forina *et al.*, 1999; Sarabia *et al.*, 2001, Lindgren *et al.*, 1994, Wold *et al.*, 2001; Gauchi e Chagnon, 2001; Lazraq *et al.*, 2001, 2003, Meiri e Zahavi, 2006; Ozturk,*et al.*, 2006; Olafsson *et al.*, 2008; Zimmer e Anzanello, 2011]. A comparação entre os métodos existentes na literatura é realizada em alguns trabalhos [ver Almoy, 1996; Baumann *et al*, 2002; Chong e Jun, 2005; Höskuldsson, 2001; Lazraq *et al.*, 2003; Zhai *et al.*, 2006; Pierna *et al.*, 2009]. Gauchi e Chagnon (2001) comparam 20 métodos de seleção baseados em diferentes critérios de avaliação, incluindo ajuste do modelo e capacidade de predição. Um desses critérios é o Q^2cum, o qual leva em conta os diferentes passos da construção do modelo de PLS e, por isso, está associado a resultados mais estáveis frente a bancos de dados com variáveis altamente correlacionadas e ruidosas. A equação (6) apresenta a fórmula para o cálculo do Q^2cum, que relaciona o indicador de predição PRESS e o indicador RSS, onde *h* é o número de componentes da validação cruzada.

$$\mathbf{Q^2_{cum}} = 1 - \prod_{j=1}^{h} \frac{PRESS_j}{RSS_{j-1}} \tag{6}$$

O indicador $PRESS_j$ está associado a um componente *j* do modelo PLS, sendo PRESS = $\sum_{i=1}^{n_L}\left(y_i - \hat{y}_{(-i)}\right)^2$, onde $\hat{y}_{(-i)}$ é a predição de y_i quando y_i é deixado fora dos dados que foram usados para a construção do modelo. O indicador RSS_{j-1} está associado a um componente (*j-1)* do modelo PLS, sendo PRESS = $\sum_{i=1}^{n_L}\left(y_i - \hat{y}_{(i)}\right)^2$, onde $\hat{y}_{(i)}$ é a predição de *y* quando y_i faz parte dos dados que foram usados para a construção do modelo.

O critério do Q^2cum também é usado por Lazraq *et al.* (2003) para avaliar comparativamente o desempenho dos dois procedimentos inferenciais propostos no trabalho para selecionar preditores relevantes para a regressão PLS, em relação a outros 2 métodos de seleção de variáveis da literatura.

Anzanello *et al.* (2009a) propuseram um método para selecionar variáveis com vistas a classificação das variáveis de produto, a partir do uso de regressões PLS e índices de importância das variáveis. Tais índices são usados para orientar a eliminação sucessiva das variáveis do modelo, sendo que a cada iteração calcula-se o *y* pelo modelo de classificação e

registra-se a acurácia medida. O trabalho de Anzanello *et al*. (2009a) não é o único a fazer uso de índices de importância das variáveis para dirigir a rotina de supressão ou admissão sistemática de variáveis no modelo. Wold *et al.* (2001) desenvolveram um índice de importância das variáveis, VIP, a partir do coeficiente modificado de peso w_{ka}^{*} e da fração de variância explicada pelo componente a em $\boldsymbol{Y}$, R_{Ya}^{2}. Esse índice foi testado em Lazraq *et al*. (2003) e Anzanello *et al.* (2009a). Outros índices com propósitos semelhantes podem ser obtidos em Eriksson e Wold (2010).

O trabalho de Zimmer e Anzanello (2011) propõe uma adaptação do método proposto por Anzanello *et al* (2009a), o método EB (Eliminação Backward), para selecionar as variáveis de processo para predição das variáveis de produto. 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 produto. Inicia-se então um processo de eliminação de variáveis do tipo *backward*, sendo a ordem de eliminação definida pelo índice de importância. Tais índices são então testados em termos de seu desempenho. O método utilizando o índice recomendado, v_{bk}, reteve apenas 31% das variáveis originais e aumentou a acurácia de predição do conjunto de teste em 6% (avaliado por intermédio do indicador RMSE - *Root Mean Square Error*).

A equação 7 apresenta o índice, v_{bk}, o qual integra 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 pelo componente *a* (a = 1,...,*A*).

$$v_{bk} = \frac{\sum_{a=1}^{A}|w_{ka}|R_{Ya}^{2}}{\max_{k\in K}(\sum_{a=1}^{A}|w_{ka}|R_{Ya}^{2})}\ \frac{\sum_{m=1}^{M}|b_{mk}|}{\max_{k\in K}(\sum_{m=1}^{M}|b_{mk}|)}\ k = 1,\dots,K. \qquad (7)$$

O trabalho de Zimmer e Anzanello (2012) indica dois novos métodos para identificação do subconjunto de variáveis a ser retido, a partir de uma adaptação do método EB de Zimmer e Anzanello (2011). O método EBM (Eliminação *backward* por mínimos) prioriza o subconjunto de variáveis que minimiza o erro de predição, independente do percentual de variáveis retido. Por sua vez, o método EBDE (Eliminação *backward* por Distância Euclidiana) prioriza o conjunto que torna mínima a distância euclidiana de cada ponto gerado pela eliminação sistemática de variáveis em relação a um ponto ideal definido pelo usuário.

4.3 Método

O primeiro passo do método é dividir o banco de dados em uma parte de treino, que será usada para a construção do modelo e a seleção das variáveis e, outra parte de teste. Na sequencia, aplica-se a regressão PLS no conjunto de treino, gerando o índice v_{bk} de importância das variáveis. Daí inicia-se a rotina de eliminação de variáveis, na qual a variável com o menor valor para o índice de importância é eliminada a cada iteração, na qual também roda-se a regressão PLS para gerar o *y* e os indicadores de predição RMSE e Q^2cum. Repete-se o processo, removendo a variável com menor índice e aplicando a regressão PLS no conjunto de treino para predição de *y*, até que reste apenas uma variável de processo.

A próxima etapa é a que diferencia os três métodos deste trabalho.

(i) Para o método EB, a identificação do conjunto é realizada subjetivamente, pelo perfil gráfico da acurácia de predição (medida através do RMSE e do Q^2cum) e o percentual de variáveis retidas no modelo de predição;

(ii) Para o segundo método, EBM, o conjunto escolhido é aquele que maximiza o indicador de acurácia de predição (menor valor de RMSE ou maior valor de Q^2cum), desconsiderando o percentual de variáveis retidas no modelo resultante;

(iii) Por fim, no método EBDE seleciona-se a conjunto que minimiza a distância euclidiana entre um ponto hipotético, definido pelo usuário como ideal (caracterizado por um reduzido percentual de variáveis retidas e maior valor possível para o indicador de acurácia usado) e o percentual de variáveis retidas em cada iteração do método. Quando o indicador de acurácia for o Q^2cum o ponto ideal será o seu valor máximo, que é igual a 1 (um). No caso do RMSE, que mede o erro médio, o ponto ideal tem valor igual a zero.

Dois exemplos da aplicação dos três métodos são apresentados nas Figuras 4.1 e 4.2. A Figura 4.1 apresenta o gráfico hipotético considerando o indicador de acurácia de predição RMSE versus o percentual de variáveis retidas. A Figura 4.2 ilustra o gráfico considerando o Q^2cum como indicador de desempenho preditivo versus percentual de variáveis retidas no modelo final.

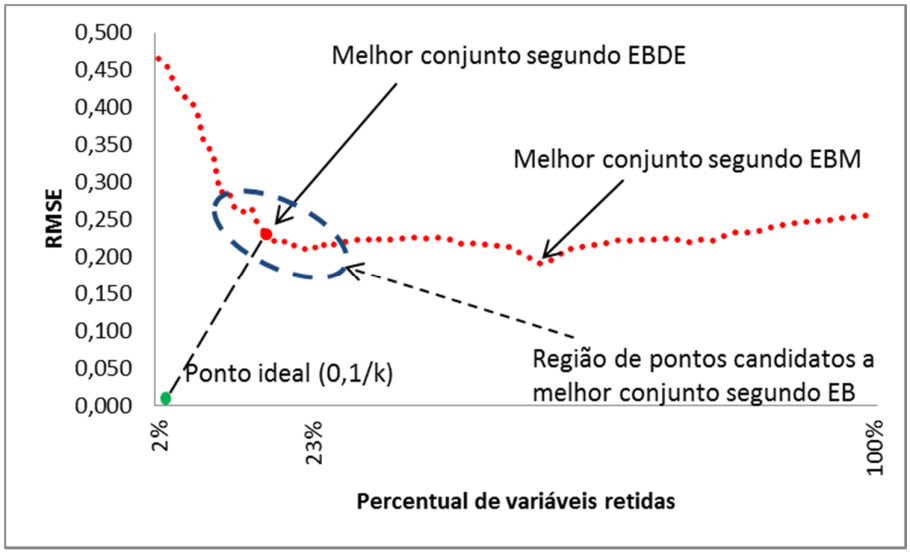

Figura 4.1 – Perfil hipotético de RMSE versus variáveis retidas no modelo final, com a ilustração dos pontos escolhidos pelos métodos EB, EBM e EBDE

Por fim, as variáveis selecionadas são testadas no conjunto de teste, que representa novas observações a serem modeladas e preditas pelo modelo PLS gerado.

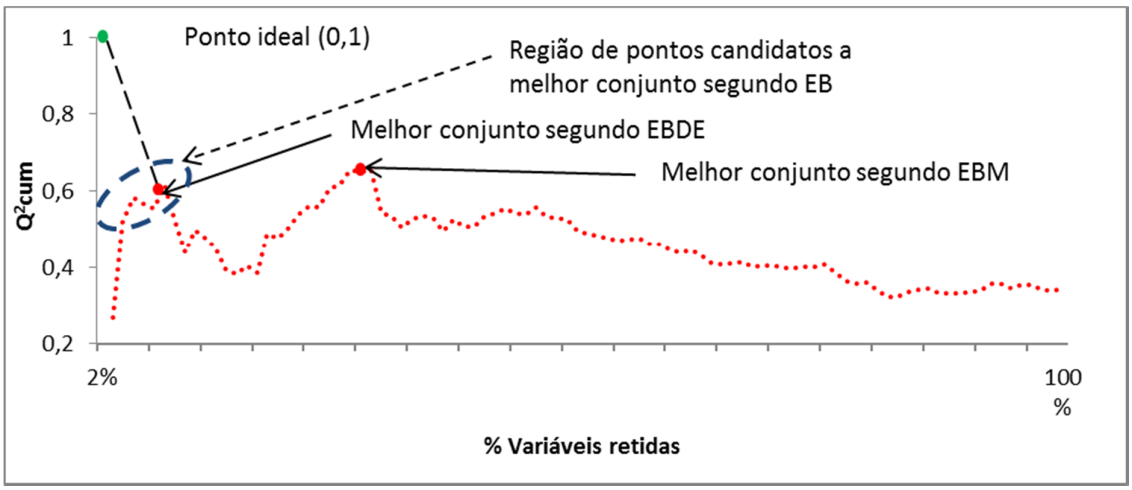

Figura 4.2 – Perfil hipotético de Q^2cum versus variáveis retidas no modelo final, com a ilustração dos pontos escolhidos pelos métodos EB, EBM e EBDE

4.4 Resultados

A Tabela 4.1 apresenta a estrutura dos quatro bancos de dados utilizados para a avaliação dos métodos propostos. Os bancos de dados são oriundos de Gauchi e Chagnon (2001), e também foram usados em Lazraq *et al*. (2003), Anzanello *et al.* (2009a) e Zimmer e Anzanello (2011). O banco de dados ADPN corresponde a um processo intermediário da produção de nylon; LATEX a um processo de manufatura de látex; os dados do banco OXY foram extraídos de processo de produção do óxido de titânio; os dados do banco SPYRA provêm de um processo de fermentação para a produção de antibiótico.

A regressão PLS foi aplicada ao conjunto de treino de cada banco de dados. As análises foram realizadas em MATLAB® versão 7.10. Foram realizadas 50 replicações para cada banco de dados, na qual as observações que compunham os conjuntos de treino e teste foram randomicamente alteradas para capturar eventuais variabilidades dentro do banco de dados. Foram retidos 3 componentes da regressão para cada banco de dados através de validação-cruzada [ver WOLD *et al*. (2001)], resultando nos seguintes R^2_{Ya}'s: ADPN, 94%, LATEX, 77%, OXY, 94% e SPIRA, 71%.

Tabela 4.1 - Bancos de dados analisados

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**

Fonte: elaborado pelos autores.

A Figura 4.3 ilustra o perfil de acurácia de predição gerado pela eliminação sistemática das variáveis utilizando o método EBDE, para o conjunto de treino do banco de dados OXY. À medida que as variáveis são eliminadas, o desempenho do indicador Q^2cum é avaliado e por fim, a eleição do melhor conjunto considera a menor distância euclidiana entre o ponto ideal (0% variáveis retidas, Q^2cum =1). O melhor conjunto obteve um Q^2cum de 0,613 (valor próximo ao maior valor possível de Q^2cum) com apenas 7% das variáveis. Se

comparado com o valor de 0,452 para Q^2cum quando utilizando todas as variáveis, o método proposto apresenta aumento de acurácia preditiva de 36% no conjunto de treino para o banco OXY. As demais iterações e banco de dados seguem a mesma lógica.

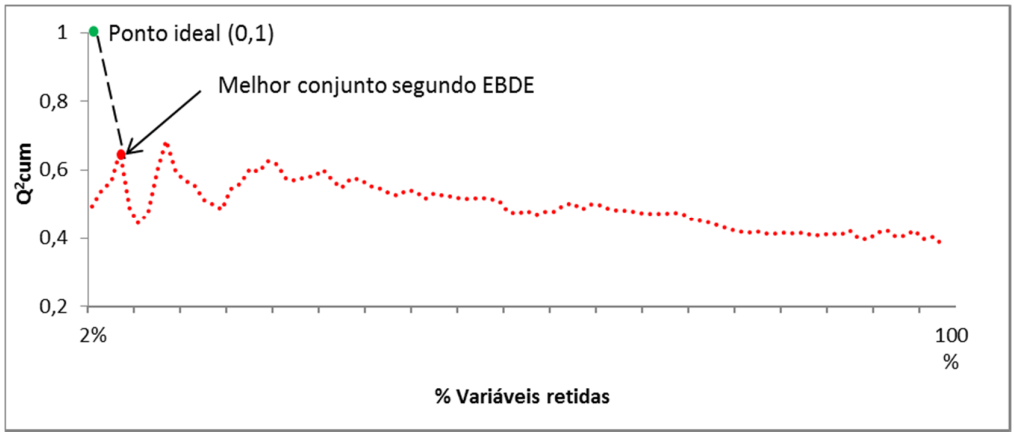

Figura 4.3 – Desempenho da predição no conjunto de treino do processo OXY usando o método EBDE.

A Tabela 4.2 apresenta o desempenho médio dos métodos avaliados. O método EBM apresentou desempenho superior de predição (maior valor de Q^2cum e menor valor de RMSE) em todos os processos analisados, enquanto que o método EB e o método EBDE alternam seus desempenhos de acordo com o banco analisado. O método EBDE retém menor número de variáveis em todos os processos, à exceção do banco ONY (onde empata com o valor do método EB quando usando o Q^2cum). O percentual de variáveis retidas para o método EBM é sempre superior aos demais métodos.

Em termos de acurácia média considerando o Q^2cum, o método EBM é 3% superior ao segundo método EBDE, e 6% superior ao método EB. O desvio padrão da acurácia é muito próximo para os três métodos. Avaliando a precisão de predição do modelo com todas as variáveis (Q^2cum=0,381), percebe-se que o método EBM (Q^2cum de 0,518) representa um ganho médio de 36% em acurácia de predição. Por sua vez, o método EBDE incrementa a acurácia de predição média em 32%, enquanto que o EB aumenta o mesmo indicador em 28%. Em termos de variáveis retidas, o método EBDE retém os menores percentuais, 3% superior ao EB e 14% menor do que o EBM. O desvio padrão do % de variáveis retidas indica que o método EBDE possui menor variação, seguido do método EB.

Tabela 4.2– Desempenho médio dos métodos no conjunto de teste de cada banco de dados

	Q^2cum			Desv Q^2cum			Variáveis retidas (%)			Desv Variáveis retidas (%)			Q^2cum p/ conj. c/todas variáveis
Processo	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	
ADPN (100)	0,499	0,558	0,548	0,061	0,056	0,065	16%	24%	15%	12%	17%	6%	0,368
LATEX (117)	0,404	0,411	0,402	0,042	0,036	0,035	15%	23%	8%	11%	25%	5%	0,346
OXY (95)	0,595	0,640	0,616	0,089	0,081	0,083	12%	25%	12%	10%	21%	7%	0,420
SPIRA (96)	0,460	0,462	0,445	0,045	0,051	0,052	21%	37%	15%	11%	26%	8%	0,391
Média	0,490	0,518	0,503	0,059	0,056	0,059	16%	27%	13%	11%	23%	7%	0,381

	RMSE			Desv RMSE			Variáveis retidas (%)			Desv Variáveis retidas (%)			RMSE p/ conj. c/todas variáveis
Processo	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	**EB**	**EBM**	**EBDE**	
ADPN (100)	1,088	0,845	0,890	0,117	0,105	0,108	37%	67%	44%	11%	19%	12%	1,021
LATEX (117)	0,566	0,524	0,546	0,023	0,022	0,029	25%	41%	16%	9%	19%	5%	0,700
OXY (95)	0,217	0,170	0,199	0,061	0,044	0,048	21%	36%	14%	13%	20%	4%	0,182
SPIRA (96)	0,157	0,146	0,176	0,006	0,015	0,016	45%	49%	6%	8%	15%	2%	0,186
Média	0,507	0,421	0,453	0,052	0,047	0,050	32%	48%	20%	10%	18%	6%	0,522

Fonte: elaborado pelos autores.

Quando se mede a precisão de predição média fazendo uso do RMSE, o método EBM apresenta desempenho 8% superior ao método EBDE, e 20% superior ao método EB. O desvio padrão do RMSE também é semelhante para os três métodos. Quando se compara a acurácia do modelo com todas as variáveis (RMSE=0,522), constata-se que o método EBM (RMSE de 0,421) apresenta um incremento de 19% em acurácia de predição. Ao seu tempo, o método EBDE (RMSE de 0,453) aumenta a acurácia de predição média em 13%, enquanto que o EB (RMSE de 0,507) melhora o mesmo indicador em 3%. Os resultados considerando o % de variáveis retidas são semelhantes aos já expostos quando medido o Q^2cum, o método EBDE retém 20% de variáveis, 12% a menos que o método EB e 28% a menos que o método EBM.

Tendo em vista os resultados, o método EBM é recomendado quando se prioriza elevada capacidade de predição da variável de produto em detrimento à retenção de maior percentual de variáveis. O método EBDE, por sua vez, retém o menor percentual de variáveis dentre os métodos testados, combinado com satisfatória capacidade de predição (Q^2cum = 0,503, apenas 3% inferior ao melhor resultado possível de EBM = 0,518, mas ainda assim 32% superior ao valor obtido com todas as variáveis no modelo). Por fim, o método EB retém

um percentual intermediário de variáveis (média de 16%, considerando o Q^2cum), porém com uma acurácia de predição inferior em relação aos demais métodos. De tal forma, recomenda-se o uso do método EBDE por conta do reduzido percentual de variáveis retidas e satisfatório nível de acurácia de predição obtido. O uso do método EBM é recomendado quando a acurácia de predição é crítica em relação ao percentual de variáveis retidas.

4.5 Conclusões

Este artigo comparou o método EB proposto em Zimmer e Anzanello (2011) com os métodos EBM e EBDE de Zimmer e Anzanello (2012), através da aplicação em bancos de dados industriais, mensurando a acurácia de predição com o tradicional indicador RMSE e com um novo indicador, o Q^2cum. Os resultados foram avaliados em termos de percentual de variáveis retidas e da acurácia de predição usando os indicadores RMSE e Q^2cum. Em termos de sua operacionalização, os três métodos avaliados geram um índice de importância das variáveis com base nos parâmetros da regressão PLS. A cada iteração uma variável é retirada do modelo de regressão de acordo com o menor valor para o índice de importância das variáveis, e tem-se a medição da acurácia de predição, fazendo uso dos indicadores RMSE e Q^2cum. O método EB seleciona o melhor conjunto de variáveis manualmente com base em uma análise do gráfico do indicador de acurácia de predição versus percentual de variáveis retidas. O método EBM seleciona o conjunto que maximiza o valor de Q^2cum ou minimiza RMSE, e terceiro método EBDE, seleciona o conjunto que minimiza a distância euclidiana entre o ponto ideal (maior valor possível de Q^2cum ou menor de RMSE e menor % de variáveis retidas) e o ponto gerado após a eliminação de cada variável.

A comparação dos três métodos mostra que o método EBM conduz a uma maior acurácia de predição (Q^2cum 36% superior em relação ao modelo com todas as variáveis), com a contrapartida de reter um maior percentual de variáveis no modelo final (em média de 27% de variáveis). Por sua vez, o método EBDE faz uso de apenas 13% das variáveis (menor percentual de variáveis retidas dentre os métodos testados) para gerar uma acurácia 32% (Q^2cum) superior ao valor obtido quando todas as variáveis são utilizadas. Já o método EB possui um desempenho em termos de acurácia de predição inferior aos demais métodos, além da desvantagem já citada de ter parte de sua operacionalização dependente de aspectos

subjetivos. Sendo assim, recomenda-se o uso do método EBDE, dado o reduzido número de variáveis e o satisfatório nível de acurácia de predição. O uso do método EBM justifica-se apenas quando a acurácia de predição é crítica em relação ao percentual de variáveis retidas.

Como extensão da pesquisa descrita nesse trabalho sugere-se a adaptação dos métodos EBM e EBDE para uso com múltiplas variáveis de produto. Da mesma forma, sugere-se adaptar os métodos EBM e EBDE para a seleção de variáveis com a finalidade de controle estatístico multivariado de processo.

4.6 Referências Bibliográficas

ABDI, H. **Partial Least Squares (PLS) Regression**. Encyclopedia of Social Sciences Research Methods. Thousand Oaks, Sage. 2003

ALMOY, T. A simulation study on comparison of prediction methods when only a few components are relevant. **Computanional Statistics & Data Analysis**, v. 21, p. 87-107, 1996.

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 24, p. 728-737, 2010.

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

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Identificação das variáveis mais relevantes para categorização de bateladas de produção: reduzindo a variância do percentual de variáveis retidas. **Produto&Produção**, Porto Alegre, Brasil, v. 10, n.3, p.19-27, 2009b.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, p. 97-105, 2012.

BAUMANNN, K.; ALBERT, H.; VON KORFF, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 16, p. 339-350, 2002.

CHIANG, L. H.; PELL, R. J. Genetic algorithms combined with discriminant analysis for key variable identification. **Journal of Process Control**, Amsterdam, Holland, v. 14, p. 143-155, 2004.

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 78, p. 103-112, 2005.

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. **Journal of Chemometrics,** Bognor Regis, United Kingdom, v. 24, p. 779-789, 2010.

FERRER, A.; AGUADO, D.; VIDAL-PUIG, S.; PRATS, J.; ZARZO, M. PLS: A versatile tool for industrial process improvement and optimization. **Applied Stochastic Models in Business and Industry**, Malden, USA, v. 24, p. 551-567, 2008.

FORINA, M.; CASOLINO, C.; MILLAN, C., Iterative predictor weighting (IPW) PLS: a technique for elimination of useless predictors in regression problems. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 13, p. 164-184, 1999.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 171-193, 2001.

LAZRAQ, A.; CLÉROUX, R. The PLS multivariate regression model: testing the significance of successive PLS components. **Journal of Chemometrics**, Bognor Regis, United Kingdom, v. 15, p. 523–536, 2001.

LAZRAQ, A.; CLÉROUX, R.; GAUCHI, J.-P. Selecting both latent and explanatory variables in the PLS1 regression model. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 66, 117-126, 2003.

LINDGREN, F.; GELADI, P.; RANNAR, S.; WOLD, S. Interactive variable selection (IVS) for PLS: Part 1. Theory and algorithms. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 8, p. 349-363, 1994.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 28, p. 3-21, 1995.

MEIRI, R.; ZAHAVI, J. Using simulated annealing to optimize the feature selection problem in marketing applications. **European Journal of Operational Research**, v. 171, p. 842-858, 2006.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A, 2009. 493 p.

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. **European Journal of Operational Research**, v. 187, p. 1429-1448, 2008.

OZTURK, A.; KAYALIGIL, S. OZDEMIREI, N. Manufacturing lead time estimation using data mining. **European Journal of Operational Research**, v. 187, p. 1429-1448, 2008.

PIERNA, J. A. F.; ABBAS, O.; BAETEN, V.; DARDENNE, P. A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica Acta**, Amsterdam, Holland, v. 642, p. 89-93, 2009.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 58, p. 109-130, 2001.

ZHAI, H. L.; CHEN, X. G.; HU, Z. De. A new approach for the identification of important variables. **Chemometrics Intelligent Laboratory Systems**, Amsterdam, Holland, v. 80, p. 130-135, 2006.

ZIMMER, J.; ANZANELLO, M.J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Revista Produção**. São Paulo. Aguardando publicação, 2011.

ZIMMER, J.; ANZANELLO, M.J. Desempenho de dois novos métodos de seleção de variáveis de processo na presença de ruído e multicolinearidade. **Revista Produção**. São Paulo. Aguardando publicação, 2012.

5 Considerações Finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

O presente trabalho teve como principal objetivo propor métodos para seleção de variáveis de processo com o intuito de predizer a variável de produto.

A revisão da literatura dos três artigos permitiu atingir um dos objetivos específicos desta dissertação, que foi de apresentar a fundamentação teórica acerca dos principais métodos de seleção de variáveis a partir de regressões PLS e de índices de importância das variáveis.

O primeiro artigo abrangeu o objetivo específico de adaptar um método para seleção de variáveis com propósito de predição com base em regressões PLS No primeiro artigo também foi realizado a comparação do método proposto com outros da literatura, e teve-se a introdução de um novo índice de importância das variáveis a partir de parâmetros da regressão PLS, outro objetivo específico.

Variações do método para seleção de variáveis proposto no primeiro artigo são apresentadas no segundo artigo, com o intuito de aperfeiçoar a escolha do melhor subconjunto de variáveis, atendendo o quarto objetivo específico. O segundo artigo também encaminhou a execução do quinto objetivo específico, avaliando o desempenho da estabilidade dos métodos de seleção de variáveis para predição, frente a alterações no nível de colinearidade entre as variáveis, ruído e proporção entre o número de variáveis de processo e observações.

O terceiro artigo conduz o último objetivo específico, ao aplicar os métodos para seleção de variáveis propostos em bancos de dados reais, comparando o desempenho dos mesmos através de indicadores de acurácia de predição e percentual de variáveis retidas no modelo final de predição.

Portanto, infere-se que todos objetivos específicos foram alcançados e, por conseguinte, pode-se afirmar que o objetivo principal deste trabalho foi, igualmente, obtido.

O primeiro artigo apresentou a fundamentação teórica sobre regressões PLS e seleções de variáveis. Foi proposto um método, EB, para a seleção de variáveis a partir de uma rotina de eliminação de variáveis *backward*, guiada por índices de importância das variáveis. O conjunto final de variáveis é escolhido subjetivamente através da análise gráfica da acurácia de predição (mensurada através do RMSE) versus quantidade de variáveis presentes no modelo a cada iteração. Desenvolveu-se também um novo índice de importância das variáveis, que foi comparado em relação a outros dois índices da literatura e em relação ao método *Stepwise*. O método EB fazendo uso do novo índice v_{bk} superou os demais índices e o método *Stepwise*, alcançando uma acurácia de predição (mensurada através do RMSE) 6% superior ao valor obtido quando todas as variáveis são utilizadas, com apenas 31% das variáveis de processo.

O segundo artigo propôs um aprimoramento do método EB, eliminando a escolha do melhor subconjunto de variáveis a partir da análise gráfica. Dois novos métodos foram propostos, o primeiro, EBM, prioriza o subconjunto que minimiza o erro de predição (RMSE). O segundo método, EBDE, identifica o conjunto ótimo que minimiza a distância euclidiana entre o ponto ideal (menor valor possível de RMSE e % de variáveis retidas) e o ponto gerado após a eliminação de cada variável. Foi simulado o desempenho dos dois novos métodos em cenários com diferentes níveis de colinearidade, ruído e proporção entre as quantidades de variáveis e de observações. Os resultados demonstraram que método EBM é mais recomendado quando se deseja priorizar a acurácia de predição, tendo em vista que em média retendo 30% das variáveis obteve-se uma acurácia de predição 12% superior em relação ao modelo com todas as variáveis. Não obstante, o método EBDE deve ser utilizado em situações em que se deseja um percentual ainda menor de variáveis retidas (a média foi de 15%), e é possível habituar-se a predições menos precisas.

O terceiro e último artigo consolida o trabalho desenvolvido nos dois primeiros artigos ao comparar os três métodos de seleção de variáveis propostos, a partir da aplicação em quatro bancos de dados reais e da mensuração da acurácia de predição por dois indicadores distintos, o RMSE e o Q^2cum. Na média o método EBDE reteve 13% das variáveis, com uma acurácia de predição (Q^2cum) média 32% superior ao valor obtido com todas as variáveis no modelo. Esse resultado é apenas 3% inferior a acurácia média obtida

com o método EBM que, no entanto, necessitou 27% das variáveis para obter tal resultado. Sendo assim, os resultados permitem inferir que o método EBDE é o mais recomendado por reter o menor percentual de variáveis com boa capacidade de predição.

5.2 Sugestões para trabalhos futuros

Como extensões das proposições apresentadas nessa dissertação, sugere-se as seguintes pesquisas futuras:

a) Comparar o método EBM e EBDE com relação a outros métodos para seleção de variáveis propostos na literatura.

b) Propor e testar novos índices de importância das variáveis a serem usados pelos métodos EBM e EBDE para a eliminação das variáveis.

c) Adaptar os métodos EBM e EBDE para a utilização com múltiplas variáveis de produto.

d) Adaptar os métodos EBM e EBDE para a seleção de variáveis com a finalidade de controle estatístico multivariado de processo.

e) Adaptar os métodos EB, EBM e EBDE para uso com variáveis discretas.

f) Avaliar como alterações na definição do ponto ideal impactam os resultados do método EBDE.