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

Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais

DISSERTAÇÃO DE MESTRADO

Samuel Facchin

Porto Alegre
2005

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

Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais

Samuel Facchin

Dissertação de Mestrado apresentada como requisito parcial para obtenção do título de Mestre em Engenharia

Área de concentração: Controle de Processos

Orientador:
Prof. Dr. Jorge Otávio Trierweiler

Porto Alegre
2005

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Dissertação *Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais*, elaborada por Samuel Facchin, como requisito parcial para obtenção do Grau de Mestre em Engenharia.

Comissão Examinadora:

Dr. Lincoln Fernando Lautenschlager Moro

Dr. Luís Gustavo Soares Longhi

Dr. Mário C. Massa de Campos

"Há quem diga que todas as noites são de sonhos.
Mas há também quem garanta que nem todas, só as de verão.
Mas no fundo isso não tem muita importância.
O que interessa mesmo não são as noites em si, são os sonhos.
Sonhos que o homem sonha sempre.
Em todos os lugares, em todas as épocas do ano, dormindo ou acordado."
Sonhos – Sonhos de Uma Noite de Verão
William Shakespeare

Aos meus pais,

Por serem, ao mesmo tempo,

O vento que me impulsiona a singrar novos mares

E um porto seguro onde posso me abrigar durante a tormenta

Agradecimentos

Em primeiro lugar, gostaria de agradecer ao Engenheiro Químico Mor, seja qual for a denominação que os homens dêem para Ele, sem a Sua iluminação com certeza eu não poderia concluir esse trabalho.

Agradeço aos meus pais Vitor Hugo e Solange, por serem exemplo de honestidade e dedicação. Seu carinho, compreensão e incentivo foram decisivos durante toda essa longa jornada. Agradeço a meu irmão, Mateus, por me mostrar, da sua forma, de que quando temos um objetivo nada fica no nosso caminho.

Agradeço a Vanessa, muito mais que uma colega de trabalho, uma amiga verdadeira e ótima conselheira, é para mim um exemplo de dedicação e generosidade. Agradeço ao Farenzena, amigo para todas as horas, pelas dicas na elaboração dos algoritmos, os auxílios no desenvolvimento das simulações e principalmente pela amizade e horas de conversa sobre tudo um pouco. Agradeço também, a Carine, pela dedicação e auxílio no desenvolvimento desse trabalho, sempre pronta a ajudar com alegria e disposição, sem a sua ajuda com certeza esse trabalho não teria chegado ao fim.

Agradeço a todos os amigos do GIMSCOP, em especial a Letícia, o Ariel, o Ricardo, o Farina, parceiros de sala e que me aturaram durante o desenvolvimento desse trabalho, a Vanessa, o Farenzena, o Vinícius Machado, a Luciane, o Jorge, o Flávio e a Débora, sua amizade e incentivo foram o que me levaram a prosseguir, mesmo quando as coisas pareciam não caminhar para o lado certo.

Não poderia me esquecer de agradecer a Cíntia Silveira, pela amizade, conselhos, paciência em ouvir minhas lamentações e resmungos, compartilhar alegrias e expectativas durante todo esse trabalho.

Não poderia deixar de mencionar o pessoal da Esquadrilha Abutre, o Alexandre **Copat**, Alexandre **Zart**, Eduardo **Rech**, Guilherme "*Clark"* Mossmann,

Massimiliano “*Mássimo*” Fabricio, Nélson “*Shrek*” Felipe de Andrade Lopes, **Neyo** Frederico Schell Kruse e Vinícius “*Boneko*” Weissheimer Ribeiro. Podem ter certeza que a amizade de vocês e as *Santa-Ceias* foram fundamentais para a conclusão desse trabalho.

Agradeço ao Departamento de Engenharia Química da UFRGS e a todos os professores pela oportunidade de realizar o mestrado em um dos melhores Departamentos do Brasil. Em especial agradeço o professor Jorge, pela sua orientação durante todo esse trabalho e pela amizade sincera, desenvolvida durante esse período.

Finalmente, agradeço a COPESUL, pelo auxílio financeiro, indispensável para a realização desse trabalho, em especial ao Time de Controle e a Unidade de Olefinas da Planta 1. Agradeço pela atenção dos engenheiros Ricardo Abech e Odila Wonderlich dos Santos, e aos engenheiros e operadores da Olefinas 1, especialmente ao Engenheiro Manuel Jaime Hernández Albarrán, uma das pessoas mais dispostas a quebrar paradigmas que eu já conheci, e Andrea Cabral Farias, que gentilmente permitiram e apoiaram os testes realizados.

Resumo

A construção de um analisador virtual é sustentada basicamente por três pilares: o modelo, as variáveis que integram o modelo e a estratégia de correção/atualização do modelo. Os modelos matemáticos são classificados quanto ao nível de conhecimento do processo contido nele, indo de modelos complexos baseados em relações fundamentais e leis físico-químicas, denominados *white-box*, até modelos obtidos através de técnicas de análise multivariável, como técnicas de regressão multiváriavel e redes neurais, referenciados como *black box*.

O presente trabalho objetiva uma análise de dois dos pilares: os *modelos*, focando em modelos obtidos através das técnicas de redução de dimensionalidade do tipo PLS, e metodologias de *seleção de variáveis* para a construção dessa classe de modelos.

Primeiramente é realizada uma revisão das principais variantes lineares e não lineares da metodologia PLS, compreendendo desde o seu desenvolvimento até a sua combinação com redes neurais. Posteriormente são apresentadas algumas das técnicas popularmente utilizadas para a seleção de variáveis em modelos do tipo *black-box*, técnicas de validação cruzada e técnicas de seleção de dados para calibração e validação de modelos.

São propostas novas abordagens para os procedimentos de seleção de variáveis, originadas da combinação das técnicas de seleção de dados com duas metodologias de seleção de variáveis. Os resultados produzidos por essas novas abordagens são comparados com o método clássico através de casos lineares e não lineares.

A viabilidade das técnicas analisadas e desenvolvidas é verificada através da aplicação das mesmas no desenvolvimento de um analisador virtual para uma coluna de destilação simulada através do simulador dinâmico Aspen Dynamics®.

Por fim são apresentadas as etapas e desafios da implementação de um analisador virtual baseados em técnicas PLS em uma Torre Depropanizadora de uma central de matérias primas de um pólo petroquímico.

Palavras Chave: Analisador virtual, modelos empíricos, PLS, QPLS, seleção de variáveis, 1,3 Butadieno, Depropanizadora

Abstract

The construction of a virtual analyzer is sustained basically by three pillars: the model, the variables that integrate the model and the updating strategy of the model. The mathematical models are classified with relationship at the level of the process knowledge within it, going from complex models, based on fundamental relationships and physical-chemistries laws, called white-box, until models obtained through multivariable analysis techniques, as multiple linear regression and neural networks, also called as black box.

The focus of the present work is the analysis of two of the pillars: the models, specially the ones obtained by dimension reduction techniques, like PLS, and methodologies used in the development of this class of models.

Initially, a revision of the main linear and non linear variants of the PLS methodology is done, embracing since its development to its combination with neural networks. Later on, some popularly variables selection techniques for black-box models are explained, as well as some cross validation techniques and strategies for data selection for calibration and validation of models.

New approaches for variables selection procedures are proposed, originated by the combination of data selection strategies and two variables selection techniques. The results produced by those new approaches are compared with the classic method through linear and non linear case studies.

The viability of the analyzed and developed techniques is verified through the application of the same ones in the development of a virtual analyzer for a distillation column, simulated by the dynamic simulator Aspen Dynamics®.

The steps and challenges faced in the implementation of a virtual analyzer based on PLS technical for a Depropanizer Unit are finally presented.

Keywords: Virtual Analyzer, empirical models, PLS, QPLS, variable selection, 1,3 Butadiene, Depropanizer

Simbologia e Nomenclatura

13BD	Composto 1,3 Butadieno
a	Posto real do sistema
AG	Algoritmo genérico
AIC	*Akaike Information Criteria*
AV	Analisador Virtual
B	Matriz de regressores para modelos
BE	Método de seleção de variáveis*: Backward elimination*
BIC	*Bayesian Information Criteria*
BTPLS	Método de construção de modelos: *Box Tidwell Least Squares Regression*
CB	Condensado de baixa pressão
CCA	*Canonical Correlation Analysis*
Corte C_{3-}	Hidrocarbonetos com cadeia formada por 3 ou menos átomos de carbono
Corte C_{3+}	Hidrocarbonetos com cadeia formada por 3 ou mais átomos de carbono
Corte C_{4+}	Hidrocarbonetos com cadeia formada por 4 ou mais átomos de carbono
CS-PLS	Método de construção de modelos: *Centered Sigmoid Partial Least Squares*
d_{ij}^2	Norma quadrática entre as amostras *i* e *j*
EKF	Filtro de Kalman Estendido
E	Vetor residual da decomposição das entradas no PLS
F	Vetor residual da decomposição das saídas no PLS
FA	*Factor Analysis*
FC	Controlador de vazão
FS	Método de seleção de variáveis*: Forward Selection*
h	Índice referente a variável latente
i,*j*	Índices das linhas e colunas das matrizes
ICA	*Independent Canonical Analysis*
k	Número de variáveis candidatas a integrar um modelo
KS	Método de seleção de dados: Kennard-Stone
KSM	Método de seleção de dados*: Kennard-Stone Modificado*
LOO	*Leaving One Out*
m	Número de colunas das matrizes de variáveis explicativas
M	Matrizes de posto unitário resultantes da aplicação de PCA
MLR	Método de construção de modelos*: Multivariate least regression*
n	Número de linhas/amostras em uma matriz
n_1	Número de amostras no conjunto de calibração
n_2	Número de amostras no conjunto de validação
NNR	*Neural Network Regression*
p	Número de parâmetros a serem ajustados em um modelo
P	Matriz formada pelos vetores de projeção das variáveis explicativas
p_h	Vetor de projeção das variáveis explicativas do fator *h* (*loading vectors)*
PCA	Método de construção de modelos: *Principal Component Analysis*
PLS	Método de construção de modelos: *Partial Least Squares*
ppm	Partes por milhão
PR	Propeno líqudo refrigerante
PRESS	*Predictive Error Sum of Squares*
Q	Matriz formada pelos vetores de projeção das variáveis de resposta

q_h	Vetor de projeção das variáveis de resposta do fator *h* (*loading vectors)*
QPLS	Método de construção de modelos: *Quadratic Partial Least Squares*
R	Vazão de refluxo
R/F	Razão Refluxo/Carga
R^2	Coeficiente de correlação
$\overline{R}^2$	Coeficiente de correlação ajustado
RBF-PLS	Método de construção de modelos*: Radial Basis Function Partial Least Squares*
RMSE	*Root Mean Squared Error*
RMSEP	*Root Mean Squared Error of Prediction*
SRMP	Método de Seleção de Variáveis: *Stepwise Regression based on Model Prediction*
SSI	*Subspace Identification*
SSE	Soma quadrática do erro
S_{yy}	Soma quadrática das distâncias das amostras em relação a sua média
$\boldsymbol{T}$	Matriz formada pelos vetores de coordenadas das variáveis explicativas (*input score vectors)*
t_h	Vetores de coordenadas das variáveis explicativas (*input score vectors)*
TC	Controlador de Temperatura
$\boldsymbol{U}$	Matriz formada pelos vetores de coordenadas das variáveis de resposta (*output score vectors)*
u_h	Vetores de coordenadas das variáveis de resposta (*output score vectors)*
VB	Vapor de baixa pressão
$\boldsymbol{X}$	Matriz contendo os dados referentes as variáveis explicativas
$\boldsymbol{X_b}$	Matriz formada pelas variáveis que devem fazer parte do modelo
$\boldsymbol{X_{Cal}}$	Matriz das variáveis explicativas utilizadas para calibração de modelos
$\boldsymbol{X_{nb}}$	Matriz formada pelas variáveis que não devem fazer parte do modelo
$\boldsymbol{X_{Val}}$	Matriz das variáveis explicativas utilizadas para validação de modelos
$\boldsymbol{y}$	Matriz contendo os dados referentes as variáveis de resposta do modelo
$\boldsymbol{y_{Cal}}$	Matriz das variáveis de resposta utilizada para calibração de modelos
$\boldsymbol{y_{Val}}$	Matriz das variáveis de resposta utilizada para validação de modelos
w	Vetores pesos do procedimento PLS
$\boldsymbol{W}$	Matriz formada pelos vetores pesos do procedimento PLS
$\overline{y}$	Média entre medições da saída
y_i	Medida da saída para a amostra *i*
$\hat{y}_i$	Predição do modelo para a amostra *i*

Sumário

Lista de figuras

Lista de tabelas

Capítulo 1

Introdução

1.1 Motivação

O atual cenário econômico mundial fez com que o controle de qualidade dos produtos se tornasse algo imprescindível para o sucesso de qualquer negócio. Em industrias de processo, muitos dos índices que qualificam os produtos não podem ser monitorados de forma contínua, sendo necessária a realização de análises laboratoriais para a obtenção desses índices. Exemplos típicos dessa situação são as composições das correntes que deixam uma coluna de destilação e o índice de fluidez em reatores de polimerização.

A incapacidade da determinação desses parâmetros de forma contínua está associada a diversos fatores, dentre os quais se destacam a inexistência de instrumentação adequada para tal finalidade ou quando da existência de tal equipamento, esse geralmente apresenta elevado custo de aquisição e manutenção e, geralmente, inserem no sistema um atraso igual ao tempo de análise, caso dos analisadores em linha para colunas de destilação.

A necessidade do conhecimento desses índices de difícil medição em linha incentivou a pesquisa de alternativas que pudessem fornecer uma inferência desses parâmetros. Analisadores virtuais são algoritmos capazes de realizar tal inferência, utilizando para isso variáveis secundárias normalmente disponíveis em tempo real, como pressões, vazões, temperaturas, e um modelo matemático que relacione os valores dessas com o índice de qualidade que se queira estimar. A utilização de analisadores virtuais vai muito além da estimação de índices de qualidade de difícil aquisição. Eles podem estar associados a analisadores em linha, fornecendo o índice desejado de minuto em minuto, sendo utilizados também para manutenção preditiva do equipamento ou para racionalizar o uso do mesmo, uma vez que possuindo a estimativa de maneira contínua pode-se espaçar o tempo de análise realizada pelo analisador em linha, aumentando a vida útil do equipamento.

Esse trabalho tem foco na exploração sistemática das etapas que constituem a elaboração de um analisador virtual, procedendo-se comparações entre as diversas técnicas

disponíveis utilizando para isso simulações estacionárias e dinâmicas de uma unidade depropanizadora e um caso real de aplicação industrial para validar as conclusões obtidas.

1.2 Analisadores Virtuais

Analisadores Virtuais – AV's – são algoritmos capazes de estimar ou inferir variáveis de difícil aquisição de forma contínua, tais como composições em colunas de destilação, índice de fluidez em polímeros, entre outros, através da utilização de variáveis secundárias, também denominadas variáveis de processo ou auxiliares, que, tipicamente, são temperaturas, pressões e vazões, e de um modelo matemático.

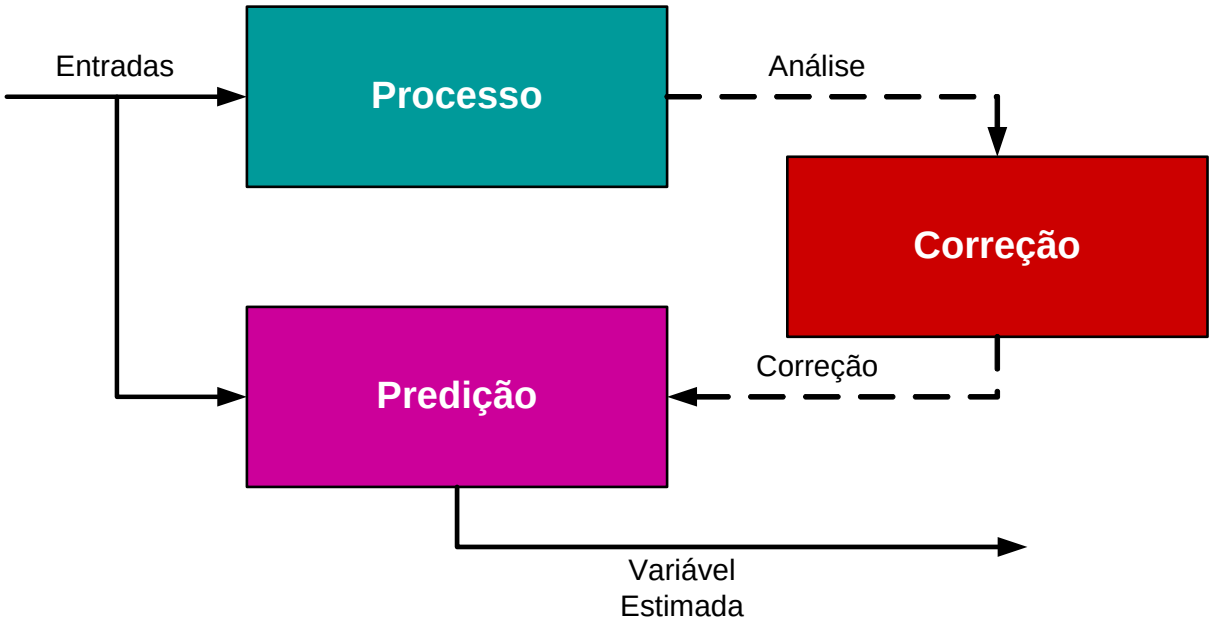


Figura 1.1: Estrutura básico de um analisador virtual

A Figura 1.1 apresenta os constituintes básicos de um sistema de analisador virtual. O bloco *predição* contém o modelo matemático que relaciona as variáveis de processo com o índice que se queira estimar, o bloco *correção* é composto por uma estratégia de adaptação do modelo, realizada através de medidas laboratoriais.

O desenvolvimento de AV's data da década de 1970 (THAM *et al.* 1991) com a publicação de diversos estudos para resolver os problemas de controle relacionados a variáveis com medidas infreqüentes. Diversas são as estratégias para a construção de AV's, sendo que o que diferencia as técnicas é a natureza do modelo utilizado. Analisadores virtuais baseados na técnica do Filtro de Kalman – FK - e Filtro de Kalman Estendido – EKF - são utilizados quando se possui um modelo simplificado do processo que relacione o parâmetro que se quer estimar e as variáveis de processo disponíveis. A utilização de tais estratégias está vinculada à necessidade de os estados, ou variáveis que se desejam inferir, serem completamente observáveis pelas variáveis secundárias. (THAM *et al.* 1991).

A estratégia acima se baseia em um modelo analítico e é, segundo FORTUNA *et al.* (2005), uma das duas maneira de se desenvolver analisadores virtuais. A segunda forma está baseada na utilização de modelos do tipo caixa preta ou caixa cinza. A utilização desses termos está relacionada à natureza puramente matemática, caixa preta, ou semi-empírica, caixa cinza, adotados por esses modelos.

Fortuna argumenta que em processos onde a complexibilidade é grande, como em processos existentes em refinarias, a construção de um modelo rigoroso demanda um tempo elevado e geralmente tem muitos parâmetros indeterminados. Por outro lado, esses processos geralmente apresentam uma grande quantidade de dados históricos armazenados, fazendo com que a utilização de estratégias caixa–preta seja favorecida. As estratégias contidas na classificação caixa-preta é bastante ampla, indo desde modelos simples de regressão multivariável, chegando a utilização de redes neuronais e neuro-fuzzy.

HYÖTYNIEMI (2001) apresenta um esquema, representado na Figura 1.2, que relaciona o tipo de sistema, a finalidade e a natureza do modelo a ser utilizado.

Figura 1.2: Espectro do sistema a ser modelado

Independentemente da estratégia adotada, há inúmeras vantagens em se utilizar um analisador virtual, dentre muitas, FORTUNA *et al.* (2005) destaca as seguintes:

- São uma alternativa de baixo custo frente a instrumentação cara;
- Podem trabalhar em paralelo à instrumentação, identificando falhas nesta;
- Podem ser facilmente implementadas nos sistemas atuais, sendo fácil o seu reajuste em função da mudança de parâmetros e
- Permitem a estimação em tempo real, eliminado os atrasos decorrentes de instrumentação, aumentado assim, a performance de sistemas de controle.

1.3 Desenvolvimento de Analisadores Virtuais

A elaboração de um analisador virtual é composta por um conjunto de procedimentos que envolvem desde a seleção do modelo a ser utilizado até a estratégia de correção do modelo que será adotada. A seguir são apresentadas brevemente essas etapas. O detalhamento dessas técnicas será apresentado nas seções posteriores desse documento.

1.3.1 Técnicas de Modelagem

Segundo DENN (1986), modelo matemático de um processo é um sistema de equações cuja solução, fornecidos dados de entrada específicos, representa as respostas do processo para o correspondente conjunto de entradas. O mesmo autor subdivide os modelos em três grupos principais:

- *Fenomenológico*: as equações do modelo derivam da utilização de teorias fundamentadas e princípios básicos da ciência;
- *Empíricos*: as equações são obtidas através da observação direta de um experimento e
- *Analogias*: utilizam as equações que descrevem um sistema dito análogo, com variáveis identificadas por analogia.

Os modelos fenomenológicos agregam conhecimentos fundamentais das leis da física e química, bem como princípios de conservação de massa, energia e quantidade de movimento. Esses são os modelos com maior capacidade extrapolativa, sua obtenção, no entanto, geralmente demanda um elevado esforço, especialmente em sistemas complexos.

Modelos empíricos, por sua vez, são caracterizados por não possuírem nenhuma base fenomenológica do sistema a ser modelado. Eles são o resultado da aplicação de métodos matemáticos, como técnicas de regressão, capazes de representar a relação entre as variáveis de entrada e saída de um processo. Sua capacidade extrapolativa é reduzida quando comparada a dos modelos fenomenológicos. Outro fator que deve ser levado em conta, quando se utiliza um modelo empírico baseado em técnicas de regressão, é que eles são capazes somente de modelar os dados e não modelar o processo como um todo (HYÖTYNIEMI, 2001).

Pode-se ainda definir uma outra classe de modelos, os denominados de semi-empíricos, cuja principal característica é possuir forma estabelecida por conhecimentos básicos do processo e a presença de alguns parâmetros desconhecidos, os quais são determinados com base em dados de processo ou experimentais. Um exemplo para essa classe de modelos é o utilizado por MOHR (2004) na elaboração de um analisador virtual para o índice de fluidez em reatores de polimerização.

1.3.2 Escolha das variáveis secundárias

A etapa de seleção de variáveis secundárias é uma das etapas mais críticas na elaboração de um analisador virtual, pois a utilização de variáveis não sensíveis termina por afetar a performance do estimador.

Em modelos fenomenológicos essa etapa é concluída conjuntamente com a escolha do modelo a ser utilizado, estando vinculado ao conjunto de equações desse. Em modelos empíricos, no entanto, existem diversas técnicas para selecionar as variáveis secundárias capazes de fornecer o melhor modelo. As estratégias de seleção de variáveis podem ser baseadas em algoritmos seqüenciais, onde a cada etapa uma variável é incluída/retirada do modelo, algoritmos de busca exaustiva, fazendo-se modelos com todas as combinações de variáveis possíveis, e algoritmos de busca aleatória, com a utilização de métodos baseados em algoritmos genéticos.

1.3.3 Estratégias de Adaptação

Todo e qualquer modelo, por mais capaz que seja de representar determinado processo, ao longo do tempo vai perdendo performance, ou seja, a diferença entre os valores reais e os preditos pelo modelo. Essa queda está relacionada a natureza variante dos processos industriais, que sofrem alterações como mudança de matéria prima, condições climáticas diversas, deterioração de equipamentos ao longo de uma campanha, entre outros.

A utilização de sistemas capazes de adaptar o modelo utilizado no analisador virtual passa a ser um fator crítico para que as inferências mantenham um elevado grau de confiança. Existem diferentes estratégias capazes de corrigir os modelos frente distúrbios não considerados por ele. A mais simples idéia é a adição de um *bias* que contenha a diferença existente entre a inferência e análises de laboratório. Essa estratégia é recomendada para sistemas com uma rotina de análise freqüente, caso contrário o *bias* poderá depreciar a qualidade da inferência gerada pelo modelo.

Estratégias mais sofisticadas visam alterar os parâmetros do modelo a cada novo resultado de análise de laboratórios. Dentro dessa categoria é possível incluir métodos de mínimos quadrados recursivos, Filtro de Kalman Estendido, métodos que utilizam janelas móveis entre outros.

1.4 Estrutura da Dissertação

Esta dissertação está estruturada de forma a conduzir uma análise sistemática das etapas de modelagem e seleção de variáveis para a construção de um analisador virtual baseado em técnicas de regressão multivariável.

O segundo capítulo faz uma descrição das técnicas de regressão multivariável de forma cronológica. Nele são detalhados os principais métodos estatísticos de regressão. São tratadas questões como linearidade e não linearidade dos modelos, bem como algumas estratégias para a adaptação dos modelos.

O terceiro capítulo foca as diferentes estratégias disponíveis para a seleção de variáveis em modelos empíricos. Novamente aqui é realizada uma breve revisão dos trabalhos publicados nessa área. São detalhados os métodos de busca exaustiva, métodos seqüenciais e métodos de seleção aleatória de variáveis. Para modelos onde a redução de dimensionalidade

é aplicável, tipicamente análise de componentes principais (PCA) e mínimos quadrados parciais (PLS), é realizada uma descrição das técnicas de validação cruzada para a determinação da real dimensão de um sistema posto deficiente.

O quarto capítulo apresenta inicialmente uma breve introdução das técnicas clássicas de controle de composição em colunas de destilação, com o objetivo de ressaltar a importância do conhecimento desses valores para a operação satisfatória de uma torre de destilação. É realizada também uma breve revisão dos trabalhos publicados no campo do desenvolvimento de analisadores virtuais para colunas de destilação utilizando técnicas do tipo PCA/PLS.

Ainda no capítulo 4 é apresentada uma aplicação sistemática das técnicas apresentadas no segundo e terceiro capítulo para o desenvolvimento de um analisador virtual de uma unidade depropanizadora simulada em ambiente ASPEN DYNAMICS®. A simulação é utilizada para avaliar diferentes alternativas para a utilização da técnica PLS, como métodos de linearização das composições através de logaritmos, técnicas de PLS não lineares e performance das estratégias de seleção de variáveis em um ambiente mais próximo aos encontrados no ambiente industrial.

O quinto capítulo apresenta a implementação de um analisador virtual em uma unidade depropanizadora industrial. Uma metodologia para o planejamento de testes em unidades industriais, utilizando-se simulações estacionárias e dinâmicas. São discutidos os principais desafios encontrados durante a elaboração, bem como apresentados os resultados diretos e indiretos da implantação do analisador na unidade.

O sexto capítulo trata das conclusões obtidas durante o trabalho e apresentará sugestões para trabalhos futuros na área de analisadores virtuais e métodos de regressão multivariável.

Em função desse trabalho focar dois dos aspectos fundamentais na elaboração de um analisador virtual, modelos empíricos e seleção de variáveis, é adotada uma estrutura de dissertação diversa da tradicional, onde o capítulo de revisão bibliográfica está difundido nos capítulos correspondentes aos assuntos específicos.

Capítulo 2

Técnicas de Análise Multivariável

A utilização de modelos baseados em técnicas de regressão multivariável é uma ferramenta extremamente útil em qualquer campo da ciência, sobretudo em ambiente industrial, onde o tempo para o desenvolvimento de um modelo rigoroso em boa parte das aplicações não é justificável ou mesmo factível.

A Tabela 2.1 apresenta, de forma sucinta, a evolução das técnicas de análise multivariável. Pode-se notar que esse campo é bastante fértil, tendo sua origem ainda no século XIX e sendo continuamente aprimorado até os dias de hoje.

Tabela 2.1: Visão simplificada do campo de análise multivariável

	Comunidade	Palavras Chave	Época
MLR *Multiple Linear Regression*	Todas	Ajuste	1800
PCA *Principal Component Analysis*	Quase todas	Compressão de Dados	1930
PLS *Partial Least Squares*	Quimiometria	Calibração de modelos	1970
CCA *Canonical Correlation Analysis*	Estatísticos	Variáveis canônicas	1960
FA *Factor Analysis*	Ciências Sociais	Fatores ocultos	1930
ICA *Independent Canonical Analysis*	Redes Neuronais	Reconhecimento de Padrões	1990
SSI *SubSpace Identification*	Teóricos de Controle	Identificação de Sistemas	1990
NNR *Neural Network Regression*	Quase todas	Aprendizagem	1980

As seções desse capítulo apresentarão as principais técnicas de regressão multivariável disponíveis atualmente. A seqüência de apresentação seguirá a cronologia do desenvolvimento das técnicas. A seção final é constituída por uma breve introdução das diferentes estratégias disponíveis para a adaptação de modelos .

2.1 Regressão Linear Multivariável

A técnica de Regressão Linear Multivariável – MLR – é uma extensão da regressão linear simples para o caso onde se possuem múltiplas variáveis de entrada ou explicativas. Segundo GELADI e KOWALSKI (1986) o problema de MLR pode ser estabelecido da seguinte forma. Dado um conjunto de variáveis explicativas $\boldsymbol{X}$, composto por m colunas e n linhas e um vetor contendo a variável de resposta $\boldsymbol{y}$, composto por 1 coluna e n linhas deseja-se estabelecer uma relação linear entre elas. Matematicamente essa relação pode ser expressa por:

$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots + b_m \cdot x_m + e \tag{2.1}$$

Pode-se também utilizar uma notação condensada:

$$\begin{aligned} y &= \sum_{j=1}^{m} b_j \cdot x_j + e \\ y &= x^T \cdot b + e \end{aligned} \tag{2.2}$$

As Eq. 2.1 e 2.2 são aplicáveis quando o número de amostras é igual a 1, ou seja, $n = 1$ (KOWASKI e GELADI, 1986). Para situações onde $n > 1$, as equações assumem uma notação matricial:

$$y = X \cdot b + e \tag{2.3}$$

Na equação 2.3, $\boldsymbol{y}$ é o vetor que contém os valores da variável de resposta, composto por n linhas e 1 coluna, $\boldsymbol{X}$ é a matriz de variáveis explicativas, formada por n linhas e m colunas, b é o vetor de regressores, ou seja, os parâmetros que permitem obter uma estimativa de $\boldsymbol{y}$ a partir dos dados contidos na matriz $\boldsymbol{X}$, e e é o vetor de resíduos, considerando que os dados contidos na matriz $\boldsymbol{X}$ e no vetor $\boldsymbol{y}$ estão escalonados.

A Eq. 2.3 pode ser representada graficamente através da Figura 2.1.

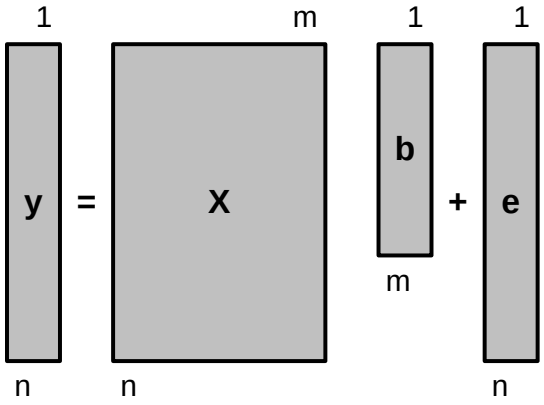


Figura 2.1: Representação gráfica do problema MLR

GELADI e KOWALSKI (1986) argumentam que em um problema MLR podem ocorrer três situações distintas:

- $m > n$: Existem mais variáveis do que amostras. Neste caso, existe um infinito número de soluções para o vetor b, sendo que todas satisfazem a equação 2.3. Essa situação deve ser evitada;
- $m = n$. O número de amostras é igual ao número de variáveis. Essa situação dificilmente será encontrada na prática. Todavia, ela fornece uma única solução para b, assumindo que a matriz $\boldsymbol{X}$ não é posto deficiente. Isso permite que se escreva a equação :

$$e = y - X \cdot b = 0 \tag{2.4}$$

  e é denominado de vetor residual.

- $m < n$. Existem mais amostras do que variáveis. Esse fato não permite obter uma solução exata para b. Pode-se, no entanto, se obter uma solução através da minimização do comprimento do vetor residual e na Eq. 2.5:

$$e = y - X \cdot b \tag{2.5}$$

O método mais difundido para a solução da Eq. 2.5 é o método de mínimos quadrados, cuja solução pode ser obtida da seguinte forma:

$$b = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y \tag{2.6}$$

A análise da equação aponta para a principal limitação da técnica MLR: a inversa de $X^T \cdot X$ pode não existir. A esse problema dá-se o nome de colinearidade, determinante nulo, singularidade, entre outros. Uma possível solução para esse caso é a utilização de um menor número de variáveis explicativas, de modo que $m \leq n$.

2.2 Análise de Componentes Principais – PCA

A presença de dados altamente correlacionados é, como mencionado anteriormente, uma das principais limitações do método MLR. A colinearidade pode ser evitada através da utilização de uma transformação da matriz original ***X***. Essa transformação nada mais que a decomposição em diversas matrizes de posto unitário, que possuem a característica de serem ortogonais entre si. Matematicamente pode-se representar a decomposição da matriz ***X*** como a soma de r matrizes de posto unitário, como apresentado na Eq. 2.7:

$$X = M_1 + M_2 + M_3 + \ldots + M_r \tag{2.7}$$

As matrizes de posto unitário ***M*** podem ser escritas como produtos de dois vetores, um vetor t_h, denominado de vetor de coordenada (*score vector*), e um vetor p_h, denominado de vetor de projeção (*loading vector*).

$$X = t_1 \cdot p_1^T + t_2 \cdot p_2^T + \ldots + t_h \cdot p_h^T \tag{2.8}$$

A Eq. 2.8 pode ser rescrita na sua forma matricial, para uma representação condensada:

$$X = T \cdot P^T \tag{2.9}$$

Na Eq. 2.9, $\boldsymbol{P^T}$ é constituída pelos vetores p^T dispostos em linhas e $\boldsymbol{T}$ é constituído pelos vetores t dispostos em colunas. A Figura 2.2 apresenta a representação gráfica da Eq. 2.8, sendo o parâmetro h o número de componentes principais considerados.

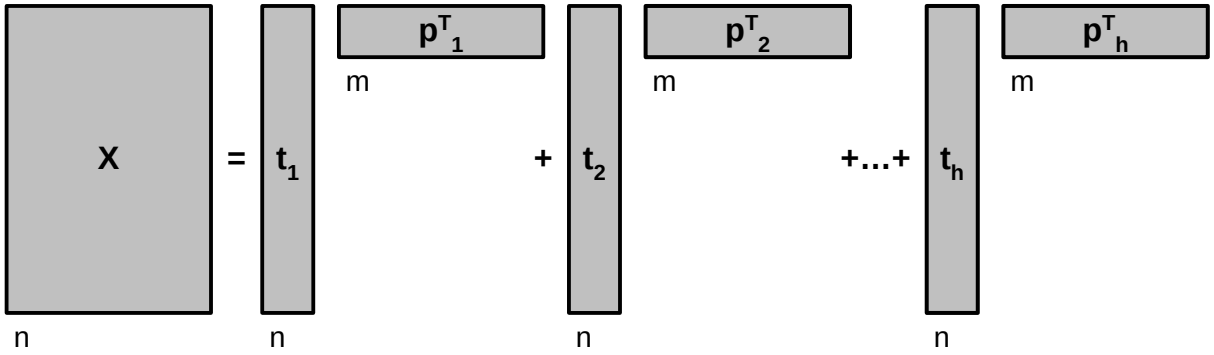


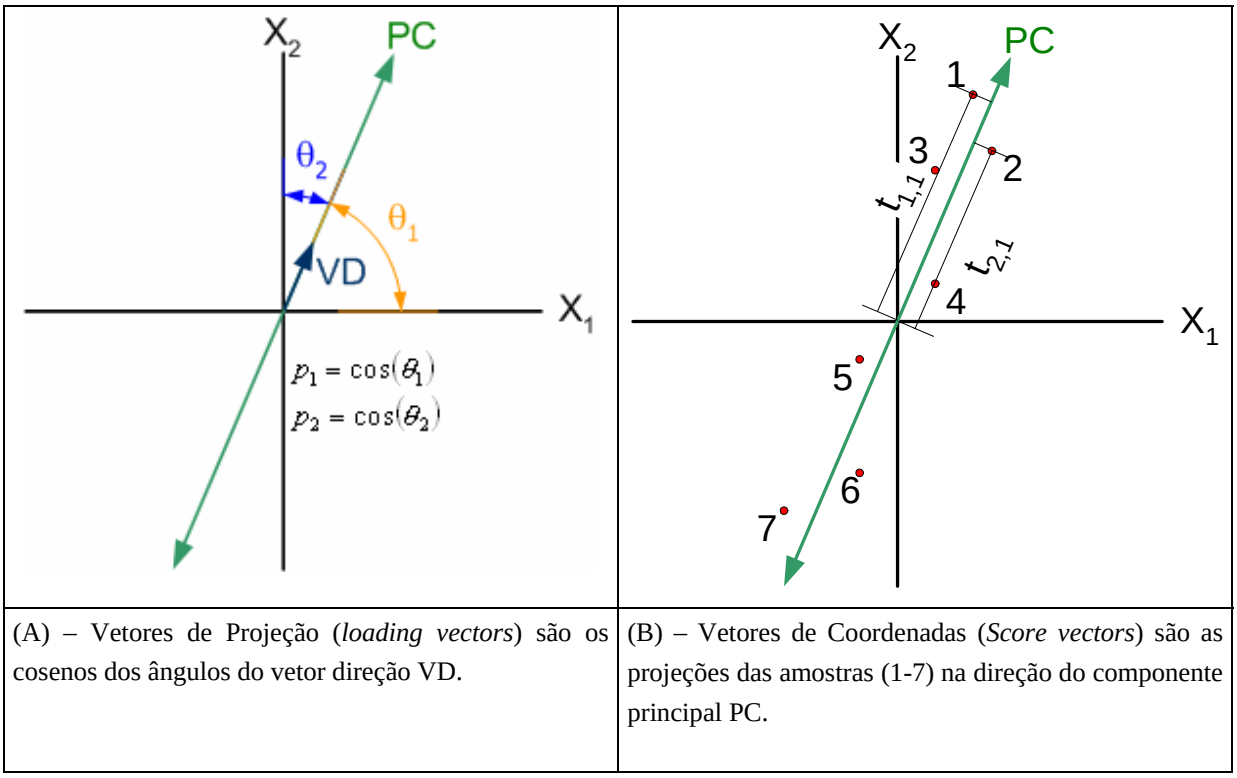
Figura 2.2: Representação gráfica de um problema PCA

O significado dos vetores t_h e p_h é melhor entendido através da representação gráfica de um problema PCA bidimensional. A escolha do caso bidimensional é puramente didática, sendo que o caso multidimensional é uma extensão desse. A decomposição em componentes principais nada mais é do que a obtenção de relações lineares entre os dados de entrada, ou seja, componentes principais são as retas obtidas através de regressões lineares entre as variáveis explicativas. No caso da Figura 2.3, o componente principal contido na Figura 2.3 (A) é a reta que melhor se ajusta aos dados contidos na Figura 2.3 (B). O vetor linha p^T_h contém os cosenos diretores do componente principal. O vetor coluna t_h contém as coordenadas dos dados originais sobre o componente principal.

A obtenção dos vetores t_h e p_h é conseguida através do algoritmo NIPALS – *Nonlinear Iterative Partial Least Squares* (BAFFI *et al.*, 1999b), o qual é o mais difundido método de execução do método PCA. A Tabela 2.2 apresenta o algoritmo NIPALS.

Tabela 2.2: Algoritmo NIPALS para PCA

Passo	Sumário do Passo	Computação
0	Escalonar a matriz $\boldsymbol{X}$	
1	Escolher um vetor de $\boldsymbol{X}$ como aproximação inicial para t_h	
2	Cálcular p_h	$p_h^T = \left(t_h^T \cdot X\right) / \left(t_h^T \cdot t_h\right)$
3	Normalizar p_h	$p_h^T = p_h^T / \left\| p_h^T \right\|$
4	Calcular t_h	$t_h^T = \left(X \cdot p_h\right) / \left(p_h^T \cdot p_h\right)$
5	Verificar convergência de t_h, se positiva, avançar para 6, caso contrário voltar para 2	
6	Obter matriz residual $\boldsymbol{F}$	$F = X - t_h \cdot p_h^T$
7	Para mais componentes, igualar $\boldsymbol{X}$ a $\boldsymbol{F}$ e voltar ao passo 1	$X = F$



(A) – Vetores de Projeção (*loading vectors*) são os cosenos dos ângulos do vetor direção VD.

(B) – Vetores de Coordenadas (*Score vectors*) são as projeções das amostras (1-7) na direção do componente principal PC.

Figura 2.3: A: visualização do vetor p_h; B: representação do vetor t_h

O procedimento NIPALS extrai os componentes principais um a um, de forma iterativa, sendo que a obtenção do segundo componente principal depende do primeiro, o terceiro depende do segundo e assim até se obter o número necessário de componentes principais. Deve se atentar que o número de componentes principais a é sempre menor ou igual ao número de variáveis m contidas na matriz $\boldsymbol{X}\left(a \leq m\right)$.

A idéia do método PCA, segundo van den BERG (2003) pode ser resumida através de três itens:

- PCA determina fatores de uma tabela de dados formando novas ‘pseudo-variáveis’ (também chamadas de componentes principais) a partir de combinações lineares das variáveis originais;
- Os fatores são selecionados de modo a explicar o máximo de informação (variância) possível dos dados;
- As novas variáveis eliminam informação redundante dos dados originais, agindo também como filtro de ruído.

A eliminação de informação redundante e ação de filtragem está relacionada com a capacidade do método PCA de capturar a real dimensão (posto) do sistema considerado. Uma vez que os primeiros componentes principais detém a maior quantidade de informação dos dados, é possível se descartar os últimos componentes, pois as informações neles contidos estarão principalmente relacionadas a ruídos. Técnicas para a determinação do real número de componentes a ser considerados em sistemas com redução de dimensionalidade serão tratados no capítulo 3.

Graficamente o processo de redução de dimensionalidade em um problema PCA pode ser representado através da Figura 2.4.

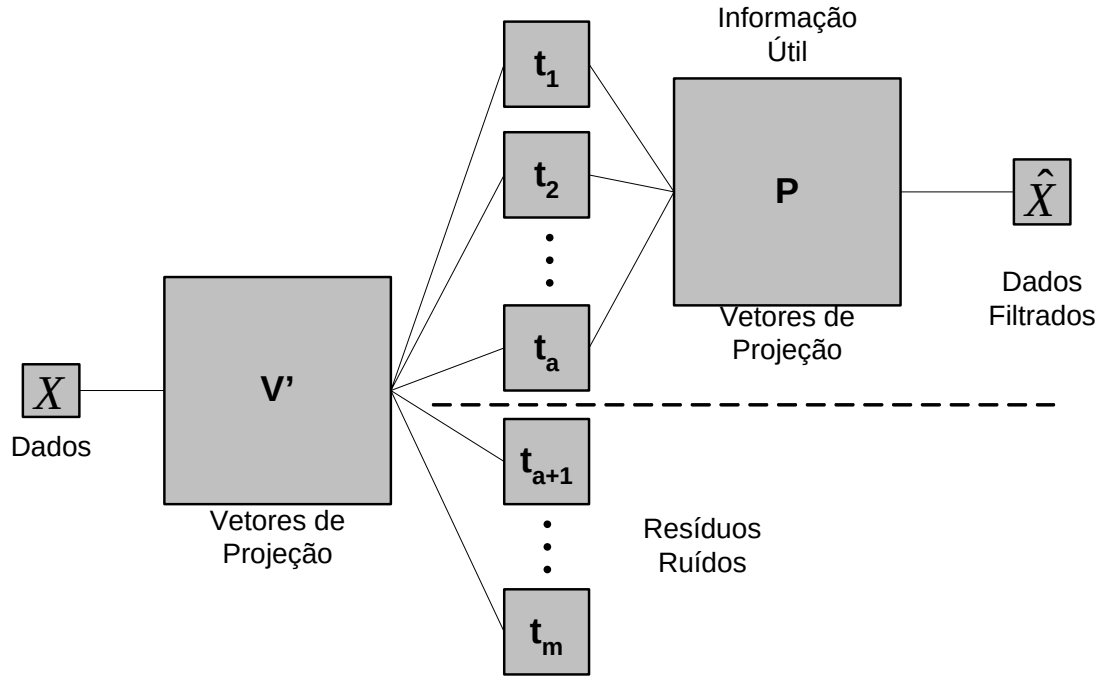


Figura 2.4: Esquema para redução de dimensionalidade.

No esquema apresentado na Figura 2.4, a informação útil contida na matriz $\boldsymbol{X}$ é capturada pelos primeiros *a* componentes principais, os demais componentes contém mais ruído do que informação e, portanto, podem ser descartados, originando assim um novo conjunto de dados filtrados $\hat{X}$.

Da maneira como os fatores são calculados, a informação útil contida na matriz ***X*** é acumulada em ordem decrescente nos componentes principais, ou seja, a maior parte da informação (variância) é capturada pelo primeiro componente principal, o segundo conterá mais informação que o terceiro, mas menos que o primeiro, e assim por diante.

Uma característica da matriz ***T***, formada pelos vetores t_h, é que suas colunas são ortogonais, sendo assim, sua inversa é estável, ao contrário de matrizes que apresentam colinearidade entre as colunas. Visando fazer uso dessa característica foi desenvolvido um método de regressão que faz uso da matriz ***T***, ao invés da matriz original ***X***. Esse método é denominado de Regressão de Componentes Principais – PCR – e será detalhado na seção seguinte.

2.3 Regressão de Componentes Principais - PCR

A idéia central do método PCR é utilizar a transformação da matriz ***X***, obtida através da aplicação de PCA, para se obter uma relação linear com a variável de resposta desejada, ***y***. A grande vantagem da técnica de PCA, como mencionado anteriormente, é que o seu resultado, ou seja, a matriz ***T*** é ortogonal e com isso sua inversão é estável. É possível ainda a extensão da técnica PCR para o caso de múltiplas variáveis de resposta, bastando realizar a regressão entre cada uma dessas variáveis e a matriz ***T***.

A obtenção do modelo é conseguida novamente pela aplicação do método de mínimos quadrados, sendo representada matematicamente através da Eq. 2.10:

$$\hat{B} = \left(T^T \cdot T\right)^{-1} \cdot T^T \cdot y \tag{2.10}$$

O modelo obtido apresentará a seguinte forma:

$$y = T \cdot \hat{B} + E \tag{2.11}$$

A principal desvantagem do método PCR é que o modelo obtido não pode ser aplicado diretamente aos dados de processo, sendo necessária a aplicação da técnica PCA nas variáveis de entrada para se obter a matriz ***T***, utilizada para a obtenção da predição dos valores de ***y***. Outra característica marcante, é que não pode-se garantir que a matriz resultante da projeção em componentes principais possua a melhor covariância com a matriz de variáveis de resposta ***y***, garante-se apenas que a matriz ***T*** contém a maior variância de ***X***.

Graficamente o procedimento utilizado pelo método PCR pode ser representado pelo esquema na Figura 2.5.

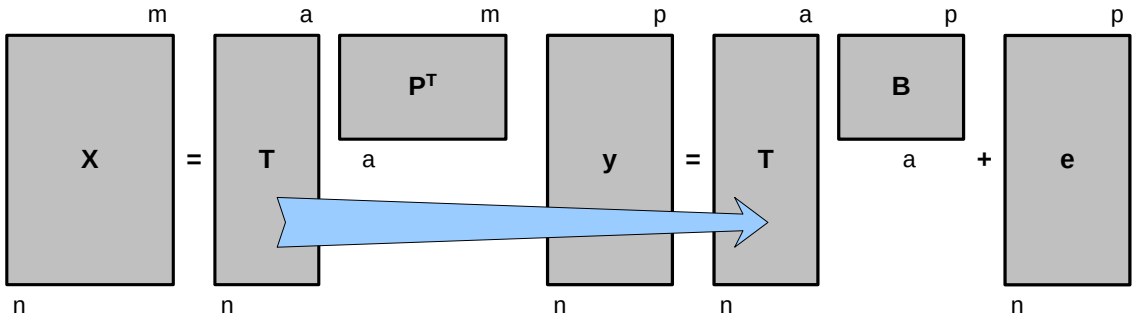


Figura 2.5: Esquema simplificado para um problema PCR

2.4 Mínimos Quadrados Parciais – PLS

Como mencionado na seção anterior, o método PCR não é capaz de garantir que a matriz ***T*** possua a melhor covariância com a matriz de variáveis de resposta ***y***. Isso ocorre principalmente em função da decomposição gerada pela técnica PCA ser aplicada somente na matriz das variáveis explicativas ***X***. O método de mínimos quadrados parciais – PLS – utiliza uma decomposição semelhante a da PCA, porém a faz simultaneamente nas matrizes ***X*** e ***y***, gerando dois novos espaços, ***T*** e ***U***, que possuem a máxima covariância possível.

O método PLS foi inicialmente desenvolvido por WOLD (1966) no fim da década de 1960 (GELADI e KOWALSKI, 1986) para aplicações no campo da economia. Os primeiros trabalhos na área de quimiometria foram desenvolvidos pelos grupos de WOLD e MARTENS (1983) no fim da década de 1970 após uma aplicação inicial feita por KOWALSKI *et al*. (1982).

Inicialmente foi proposta uma metodologia independente para a obtenção dos blocos ***T*** e ***U***, ou seja, basicamente era realizado um procedimento PCA para a matriz ***X*** e outro para a matriz ***y***. Com os blocos resultantes das duas decomposições, ***T*** obtido de ***X*** e ***U*** obtido de ***y***, é realizado o mapeamento da relação linear entre esses.

$$u_h = b_h \cdot t_h + e \tag{2.12}$$

A metodologia descrita acima não garante a melhor covariância entre os blocos ***X*** e ***y***, pois sua decomposição é realizada em separado. Um modo de contornar esse problema é permitir que os blocos troquem informação, aumentando dessa forma, a covariância. O algoritmo para a realização de tal procedimento pode ser obtido em GELADI e KOWALSKI (1986).

Essa modificação no algoritmo NIPALS tem como principal desvantagem o fato que o bloco ***T*** passe a ser não ortogonal, pois houve uma alteração na ordem de cálculo do procedimento PCA. Para garantir que o bloco ***T*** passe a ser ortogonal, o bloco ***P*** é inicialmente substituído por um bloco de pesos ***W***. Posteriormente, o bloco ***P*** é calculado.

Os vetores ***w***, que constituem a matriz ***W***, são obtidos de forma a aumentar a covariância entre os vetores t_h e u_h, porém esse aumento de covariância faz com que os vetores t_h deixem de ser ortogonais, característica desejável para fins de estabilidade do método de mínimos quadrados. Para solucionar esse problema, após verificada a convergência da relação entre t_h e u_h, um bloco ***P*** é calculado, de modo a garantir a ortogonalidade entre os vetores que formam a matriz ***T***.

O algoritmo NIPALS modificado (BAFFI *et al.*, 1999b), com a inclusão do bloco de pesos, ***W***, para o PLS é apresentado na Tabela **2.3**.

Tabela 2.3: Algoritmo NIPALS modificado para PLS

Passo	Descrição	Computação
0	Escalonar ***X*** e ***y***	
1	Atribuir u_h igual a primeira coluna de ***y***	
2	Calcular ***w***	$w^T = \left(u^T \cdot X\right) / \left(u^T \cdot u\right)$
3	Normalizar ***w***	$w = \left(w\right) / \left\|w\right\|$
4	Calcular ***t***	$t = \left(X \cdot w\right) / \left(w^T \cdot w\right)$
5	Calcular ***q***	$q^T = \left(t^T \cdot y\right) / \left(t^T \cdot t\right)$
6	Normalizar ***q***	$q = q / \left\|q\right\|$
7	Calcular ***u***	$u = \left(y \cdot q\right) / \left(q^T \cdot q\right)$
8	Verificar convergência de ***u***, se positiva avançar para 9, caso contrário retornar para 2	
9	Calcular ***p***	$p^T = \left(t^T \cdot X\right) / \left(t^T \cdot t\right)$
10	Estabelecer relação entre ***t*** e ***u***	$b = \left(t^T \cdot u\right) / \left(t^T \cdot t\right)$
11	Calcular a matriz residual de ***X***	$E = X - t \cdot p^T$
12	Calcular a matriz residual de ***y***	$F = y - b \cdot t \times q^T$
13	Caso sejam necessários mais fatores, substituir ***X*** por ***E*** e ***y*** por ***F*** e repetir os passos de 1 a 13	

O algoritmo do modo que é apresentado na Tabela 2.3 foi proposto por BAFFI *et al*. (1999a). Variações sobre a forma desse podem ocorrer, basicamente no que se refere a verificação de convergência. Outros autores, como GELADI e KOWALSKI (1986), utilizam o vetor ***t*** para fins de convergência.

Graficamente a decomposição das matrizes ***X*** e ***y*** fornecida pelo método PLS pode ser representado pelo esquema simplificado contido na Figura 2.6.

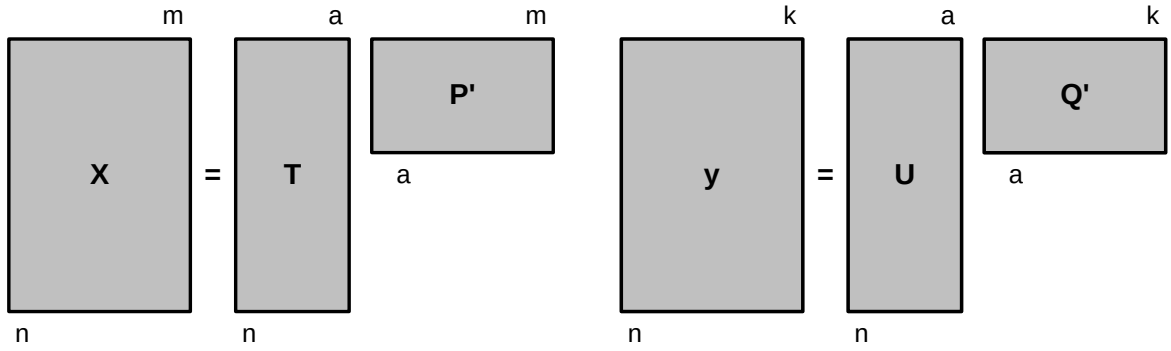


Figura 2.6: Esquema simplificado para um problema PLS

O modelo PLS é obtido através do mapeamento da relação linear existente entre os blocos ***U*** e ***T***. Em uma primeira análise o modelo obtido através da aplicação da técnica PLS também não poderia ser aplicado de forma direta, a partir de novos dados. GELADI e KOWALSKI (1986) apresentam um procedimento para a obtenção de estimativas a partir de um conjunto novo de dados através do método PLS, esse procedimento é divido em duas etapas. A primeira é responsável pela projeção dos novos dados no espaço formado pelos vetores p_h. Esse procedimento é repetido para cada uma das direções consideradas na obtenção do modelo. A seqüência de operações envolvidas nesse etapa está contida na Tabela 2.4.

Tabela 2.4: Algoritmo para etapa de estimação - Operações em X

Passo	Descrição	Computação
1	Obter o vetor de dados projetados ***t***	$t = X \cdot w$
2	Atualizar a matriz de resíduos ***E***	$E = X - t \cdot p^T$
3	Substituir ***X*** por ***E*** e avançar para o próximo fator	$X = E$

Uma vez que os dados foram projetados em todas as direções contidas no modelo, realiza-se a etapa de predição das variáveis de saída através da Eq.2.13.

$$\hat{y} = \sum_{h=1}^{a} b_h \cdot t_h \cdot q_h^T \tag{2.13}$$

Na Eq. 2.12 o subscrito *h* se refere a cada um dos *a* fatores, ou direções, considerados.

É possível se obter as predições de forma mais direta, sem a necessidade do procedimento iterativo contido na Tabela 2.4. Esse procedimento é apresentado por BAFFI *et al*. (1999b). A Eq, 2.14 sumariza a obtenção das predições $\hat{y}$ diretamente através das matrizes ***P***, ***W*** e ***Q***.

$$\hat{y} = X \cdot W \cdot \left(P^T \cdot W\right)^{-1} \cdot Q' \tag{2.14}$$

BAFFI *et al*. (1999b) ressaltam ainda que se o número de fatores mantidos no modelos PLS, *a*, for igual ao número de variáveis contidas na matriz ***X***, *m*, o modelo PLS converge para um modelo MLR.

Aplicações da técnica PLS são encontradas nas mais diversas áreas, desde métodos de análise química, como pode ser visto em BLANCO *et al.* (2000) e DING *et al*. (1998) entre outros, controle de processos, pode-se citar KRUGER *et al*. (2001) e KOURTI e MacGREGOR (1994) e no desenvolvimento de analisadores virtuais, KANO *et al*. (2002) e KOMULAINEN *et al*. (2004).

2.5 Técnicas de PLS Não-Lineares

LI *et al.* (2001) enunciam que as técnicas baseadas em modelos PLS se mostram como ferramentas poderosas para o desenvolvimento de modelos na área de modelagem e controle estatístico de processo quando os dados apresentam um elevado índice de colinearidade ou correlação. No entanto, quando o processo apresenta comportamento não-linear, a técnica perde eficiência.

CHIANG *et al.* (2001) argumentam que para sistemas onde o grau de não-linearidade não é muito elevado, a utilização de um PLS linear com um maior número de direções é capaz de obter um bom ajuste. No entanto, para casos onde a não-linearidade é elevada é necessário se utilizar uma técnica de PLS não-linear. Essa seção apresentará algumas estratégias de PLS não-lineares desenvolvidas ao longo das últimas décadas.

2.5.1 Modificação do Algoritmo NIPALS para PLS Não-Lineares

Técnicas de PLS não-lineares se caracterizam por substituírem a relação linear entre ***T*** e ***U***, por um funcional qualquer. Desse modo, a Eq. 2.12 passa a ser escrita na forma da Eq. 2.15.

$$u_h = f(t_h) + e \tag{2.15}$$

O resultado dessa modificação faz com que seja necessário adaptar o algoritmo NIPALS para a estimação dos parâmetros da função $f(\cdot)$. WOLD *et al.* (1989) enunciaram que o funcional que relaciona os vetores u_h e t_h pode ser qualquer um, desde que esse seja contínuo e diferenciável em relação a t_h.

A principal modificação do algortimo NIPALS está relacionada ao procedimento de cálculo dos vetores pesos ***w***. BAFFI *et al.* (1999a) enunciam que sendo os pesos ***w*** calculados a partir da covariância entre os vetores ***u*** e a matriz de entrada ***X***, a utilização de uma relação não-linear para relacionar cada par de variáveis u_h e t_h afeta o cálculo dos pesos ***w***.

Para contornar os efeitos dessa relação, WOLD (1989) propôs um sistema de atualização dos vetores ***w*** através de um procedimento Newton-Raphson da relação interna entre u_h e t_h. Basicamente esse procedimento é uma expansão de primeira ordem em série de

Taylor do funcional $f(\cdot)$, com posterior solução do problema formado em relação a Δw. A síntese desse processo é apresentada a seguir.

Partindo de uma relação não-linear entre u_h e t_h:

$$u_h = f(t_h) + e = f(X, w_h, c_h) + e \tag{2.16}$$

onde $f(\cdot)$ é uma função contínua e diferenciável em relação a w_h e c_h, sendo w_h os pesos para a *h-ésima* variável latente e c_h os parâmetros da função $f(\cdot)$ para a *h-ésima* variável latente.

A Eq. 2.16 pode ser aproximada através da linearização de Newton-Raphson:

$$\hat{u} = f_{00} + \left.\frac{\partial f}{\partial c}\right|_{00} \Delta c + \left.\frac{\partial f}{\partial w}\right|_{00} \Delta w \tag{2.17}$$

onde:

$$\begin{aligned} &f_{00} = \hat{u} = f(t, c) && \text{(a)} \\ &\left.\frac{\partial f}{\partial c}\right|_{00} \Delta c = \sum_i \left.\frac{\partial f}{\partial c_i}\right|_{00} \Delta c_i && \text{(b)} \\ &\left.\frac{\partial f}{\partial w}\right|_{00} \Delta w = \sum_m \left.\frac{\partial f}{\partial w_m}\right|_{00} \Delta w_m && \text{(c)} \end{aligned} \tag{2.18}$$

O cálculo das derivadas nas Eq. 2.18 b e 2.18 c pode ser realizado numericamente no ponto ***t*** correspondente a $f_{00} = \hat{u} = f(t, c)$, uma vez que ***t*** é conhecido e é assumido que $f(\cdot)$ é diferenciável em relação a ***w*** e ***c***. Daí, os únicos parâmetros desconhecidos na aproximação são Δc_i e Δw_m.

As correções de Δw_m podem ser calculadas da seguinte forma:

1. f_{00}, $\partial f / \partial c$ e $\partial f / \partial w$ são inicialmente agrupados em um matriz $Z = \begin{bmatrix} f_{00} & \partial f / \partial c & \partial f / \partial w \end{bmatrix}$
2. Um vetor coluna ***v***, normalizado a unidade, é projetado em ***Z*** e ***u***

   $v = \dfrac{Z^T \cdot u}{u^T \cdot u}$, correspondendo a $Z = u \cdot v^T$

   $v = \dfrac{v}{\|v\|}$
3. Um vetor coluna ***s*** é então calculado, $s = Z \cdot v$
4. O vetor ***u*** é projetado em *s*:

$b = \dfrac{s^T \cdot u}{s^T \cdot s}$, correspondendo a $u = b \cdot s$

5. Os valores de $\begin{bmatrix} b & v \end{bmatrix}$ correspondentes a $\partial f / \partial w_m$ são atribuídos a Δw_m
6. Finalmente, $\boldsymbol{w}$ é atualizado, $w = w + \Delta w$, e normalizado a unidade.

Com base nesse método de atualização dos vetores pesos $\boldsymbol{w}$ é possível reescrever o algoritmo de NIPALS contido na Tabela 2.3 para a sua variante não-linear apresentada na Tabela 2.5.

Tabela 2.5: Algoritmo NIPALS para PLS Não-linear

Passo	Descrição	Computação
0	Centrar e escalonar $\boldsymbol{X}$ e $\boldsymbol{y}$	
1	Igualar u a uma coluna de $\boldsymbol{y}$	
2	Projetar as colunas de $\boldsymbol{X}$ em $\boldsymbol{u}$	$w^T = u^T \cdot X / u^T \cdot u$
3	Normalizar $\boldsymbol{w}$	$w = w / \lVert w \rVert$
4	Calcular o vetor $\boldsymbol{t}$	$t = X \cdot w / w^T \cdot w$
5	Realizar regressão não-linear	$c \xleftarrow{ajuste} [u = f(t) + h]$
6	Calcular a predição de $\boldsymbol{u}$	$r = f(t, c)$
7	Projetar as colunas de $\boldsymbol{y}$ em $\boldsymbol{r}$	$q^T = r^T \cdot y / r^T \cdot r$
8	Normalizar $\boldsymbol{q}$	$q = q / \lVert q \rVert$
9	Calcular novo valor para $\boldsymbol{u}$	$u = y \cdot q / q^T \cdot q$
10	Atualizar o vetor de pesos $\boldsymbol{w}$ a partir do procedimento descrito anteriormente	
11	Normalizar o vetor de pesos	$w = w / \lVert w \rVert$
12	Calcular o novo vetor $\boldsymbol{t}$	$t = X \cdot w / w^T \cdot w$
13	Verificar convergência em $\boldsymbol{t}$. Se positiva avançar para 14, caso contrário voltar para 5	
14	Realizar regressão não-linear	$c \xleftarrow{ajuste} [u = f(t) + h]$
15	Calcular a predição de $\boldsymbol{u}$	$r = f(t, c)$
16	Calcular o vetor $\boldsymbol{p}$	$p^T = t^T \cdot X / t^T \cdot t$
17	Calcular a matriz residual das entradas	$E = X - t \cdot p^T$
18	Calcular a matriz residual das saídas	$F = y - r \cdot q^T$
19	Caso sejam necessárias direções adicionais, $\boldsymbol{X}$ e $\boldsymbol{y}$ devem ser substituídos por $\boldsymbol{E}$ e $\boldsymbol{F}$ e os passos 1-19 devem ser repetidos	

Posteriormente, BAFFI *et al*. (1999) realizaram uma revisão no método de atualização dos pesos $\boldsymbol{w}$ proposto por WOLD *et al.* (1989) e concluíram que além de complicado, o procedimento apresentava convergência lenta quando os dados fossem mal estruturados. Os autores propuseram 3 variações para o procedimento de atualização dos pesos $\boldsymbol{w}$, as quais eles denominaram de PLS-A, PLS-B e PLS-C.

O ponto de partida para a elaboração das três variantes foi a Eq. 2.17. Essa foi rescrita em sua forma matricial da seguinte maneira:

$$\hat{u} = Z \cdot \Delta v \tag{2.19}$$

onde:

$$\begin{aligned} Z &= \begin{bmatrix} f_{00} & \dfrac{\partial f}{\partial c} & \dfrac{\partial f}{\partial w} \end{bmatrix} \quad (\text{a}) \\ \Delta v &= \begin{bmatrix} 1 \\ \Delta c \\ \Delta w \end{bmatrix} \qquad\qquad (\text{b}) \end{aligned} \tag{2.20}$$

As correções Δw_k podem ser calculadas diretamente a partir da regressão de ***u*** em ***Z***:

$$\Delta v = \left(Z^T \cdot Z\right)^{-} \cdot Z^T \cdot u \tag{2.21}$$

Na Eq, 2.21 $\left(Z^T \cdot Z\right)^{-}$ é a pseudo-inversa da matriz $\left(Z^T \cdot Z\right)$. Baffi argumentou ainda que se as correções Δc_i não forem utilizadas, a aproximação linear pode ser reduzida a :

$$\hat{u} = f_{00} + \left.\frac{\partial f}{\partial w}\right|_{00} \Delta w \tag{2.22}$$

A Eq. 2.22 pode ser rescrita em sua notação matricial:

$$\hat{u} = Z \cdot \Delta v \tag{2.23}$$

onde:

$$\begin{aligned} Z &= \begin{bmatrix} f_{00} & \dfrac{\partial f}{\partial w} \end{bmatrix} \quad (\text{a}) \\ \Delta v &= \begin{bmatrix} 1 \\ \Delta w \end{bmatrix} \qquad (\text{b}) \end{aligned} \tag{2.24}$$

Novamente, a obtenção das correções Δw_k pode ser obtida através da resolução da Eq. 2.25:

$$\Delta v = \left(Z^T \cdot Z\right)^{-} \cdot Z^T \cdot u \tag{2.25}$$

Seguindo na linha de simplificações da Eq. 2.17, chega-se a última modificação do procedimento de atualização dos vetores pesos ***w*** proposto por BAFFI. Observando-se que a

diferença *e* existente entre os valores de ***u*** obtidos a partir de $u = y \cdot q$ e os gerados pela relação não-linear, $\hat{u} = f(t,c)$, pode ser expressa por:

$$e = u - \hat{u} \tag{2.26}$$

Utilizando-se a aproximação de Newton-Raphson para representar a Eq. 2.26 obtêm-se:

$$e = u - \hat{u} = u - f_{00} = \left.\frac{\partial f}{\partial w}\right|_{00} \cdot \Delta w \tag{2.27}$$

Na Eq. 2.27 os únicos parâmetros desconhecidos são os fatores Δw_k. Utilizando uma representação similar a adotada anteriormente pode-se agrupar as derivadas parciais $\partial f / \partial w_m$ em uma matriz ***Z*** e a diferença ***e*** pode ser escrita como $e = Z \cdot \Delta w$ e as correções Δw_k podem ser obtidas diretamente através da resolução da Eq. 2.28.

$$\Delta w = \left(Z^T \cdot Z\right)^{-} \cdot Z^T \cdot e \tag{2.28}$$

Essas três maneiras de resolver o problema de atualização dos pesos receberam as denominações de PLS-A (Eq. 2.21), PLS-B (Eq. 2.25) e PLS-C (Eq. 2.28). A última variante, PLS-C, foi denominada de PLS baseado no erro. BAFFI *et al.* (1999a) argumentam que essa última modificação é a que se mostra mais precisa, uma vez que a diferença entre ***u***, obtida através de $u = y \cdot q$, e o valor de $\hat{u}$, gerado através da relação não-linear $\hat{u} = f(t,c)$, está relacionada somente com os pesos ***w*** através dos fatores de correção Δw.

Nos algoritmos PLS-A e PLS-B os fatores de correção Δw são obtidos através do vetor Δv. No entanto, apesar de com isso a diferença entre ***u*** e $\hat{u}$ ser relacionada não somente a ***w***, mas também a c e f_{00}, somente a porção correspondente a ***w*** é utilizada, dessa forma a fração relacionada a c e $\hat{u}$ é desconsiderada.

Tendo como base os resultados obtidos por BAFFI *et al*. (1999a) ao longo desse trabalho foi utilizado o algoritmo PLS-C para a elaboração de estratégias não-lineares da técnica PLS. As subseções a seguir irão descrever algumas estratégias de PLS não-linear que utilizam o procedimento acima descrito para a obtenção dos modelos.

2.5.2 Q-PLS – Mínimos Quadrados Parciais Quadrático

WOLD *et al.* (1989) propuseram a substituição do mapeamento linear existente entre os blocos ***T*** e ***U*** por uma relação quadrática. WOLD et al. argumentam que para mapear uma relação não-linear entre ***X*** e ***y*** é necessário que a relação entre ***T*** e ***U*** seja também não-linear. Dentre os funcionais não-lineares mais simples estão os polinômios e, dentre esses, o mais simples é o polinômio quadrático. A Eq. 2.29 apresenta a relação quadrática entre os vetores ***t*** e ***u***.

$$u = c_0 + c_1 \cdot t + c_2 \cdot t^2 + h \tag{2.29}$$

BAFFI *et al.* utilizaram o mapeamento não-linear descrito pela Eq. 2.29 juntamente com o procedimento de atualização PLS-C para elaborar sua estratégia de PLS Quadrático. A utilização da relação quadrática permite que a Eq. 2.17 seja expressa explicitamente através da Eq. 2.30, onde m corresponde a cada uma das colunas da matriz $\boldsymbol{X}$.

$$u = f_{00} + \underbrace{\Delta c_0 + \Delta c_1 \cdot t + \Delta c_2 \cdot t^2}_{\left.\frac{\partial f}{\partial c}\right|_{00} \cdot \Delta c} + \underbrace{\sum_{m=1}^{M} \left(c_1 + 2 \cdot c_2 \cdot t\right) \cdot x_m \cdot \Delta w_m}_{\left.\frac{\partial f}{\partial w}\right|_{00} \cdot \Delta w} \tag{2.30}$$

A atualização dos pesos $\boldsymbol{w}$ a partir do algoritmo PLS-C para o caso de relação quadrática entre $\boldsymbol{t}$ e $\boldsymbol{u}$ é representada por:

$$\Delta w = \left(Z^T \cdot Z\right)^{-} \cdot Z^T \cdot e \tag{2.31}$$

Para esse caso a matriz $\boldsymbol{Z}$ passa a ter a seguinte estrutura:

$$Z = \left[\left(c_1 + 2 \cdot c_2 \cdot t\right) \cdot x_m\right] \tag{2.32}$$

2.5.3 Box - Tidwell PLS

LI *et al.* (2001) propuseram uma relação não-linear extremamente flexível para o mapeamento da relação entre os vetores $\boldsymbol{t}$ e $\boldsymbol{u}$. Essa relação é uma modificação da função proposta por Box e Tidwell (1962):

$$\begin{aligned} y(x) &= \beta_0 + \beta_1 \cdot r(x) \\ r(x) &= \begin{cases} x^{\alpha}, & \alpha \neq 0 \\ \ln(x), & \alpha = 0 \end{cases} \quad x > 0 \end{aligned} \tag{2.33}$$

O conjunto de Eq. 2.33 somente é válido para valores positivos da variável independente $\boldsymbol{x}$, o que limita sua utilização para o mapeamento entre aos vetores $\boldsymbol{t}$ e $\boldsymbol{u}$, pois estes sofrem um processo de escalonamento que torna suas médias nulas.

LI *et al.* (2001) propuseram uma modificação da Eq. 2.33 para que ela pudesse ser utilizada para qualquer valor real de $\boldsymbol{x}$. A modificação proposta, apresentada na Eq. 2.34, contorna o problema do domínio, mas traz uma penalização. Como o parâmetro δ é binário, ou seja, pode assumir valor 0 ou 1, tem-se que resolver o problema duas vezes, uma vez com $\delta = 0$ e outra com $\delta = 1$.

$$\begin{aligned} y(x) &= \beta_0 + \beta_1 \cdot r(x) \\ r(x) &= \begin{cases} \left[\text{sgn}(x)\right]^{\delta} \cdot |x|^{\alpha}, & \alpha \neq 0 \\ \left[\text{sgn}(x)\right]^{\delta} \cdot \ln(|x|), & \alpha = 0 \end{cases} \quad \delta = 0 \quad ou \quad \delta = 1 \end{aligned} \tag{2.34}$$

Analisando-se a Eq. 2.34 pode-se notar que essa não é válida quando $x = 0$ e $\alpha \leq 0$. Surge daí uma limitação no valor dos parâmetros, pois α é restrito a valores positivos ($\alpha > 0$). Uma alternativa clássica para se incluir essa restrição no problema de estimação de parâmetros é fazer $\alpha = \nu^2$ para qualquer valor real de ν.

Os parâmetros $\alpha, \beta_0, \beta_1, e\, \delta$ não podem ser estimados em uma só etapa, sendo necessária uma seqüência de passos, onde em cada etapa um dos parâmetros é obtido. A seguir é apresentada a seqüência necessária para a correta estimação dos parâmetros para as transformações Box-Tidwell.

O procedimento se inicia pela expansão da função contida na Eq. 2.34. É válido se observar que o parâmetro α foi substituído por ν^2.

$$\begin{aligned} & y = g\left(r, \beta_0, \beta_1, \delta, \nu^2\right) \\ & \approx \beta_0 + \beta_1 \cdot \left(\operatorname{sgn}(x)\right)^{\delta} \cdot |x| + (\nu - \nu_0) \cdot \left\{ \frac{dg\left(r, \beta_0, \beta_1, \delta, \nu^2\right)}{d\nu} \right\}\Bigg|_{\nu=\nu 0} \\ & = \beta_0 + \beta_1 \cdot \left(\operatorname{sgn}(x)\right)^{\delta} \cdot |x| + 2 \cdot (\nu - 1) \cdot \beta_1 \cdot \left(\operatorname{sgn}(x)\right)^{\delta} \cdot \ln\left(|x|\right) \end{aligned} \tag{2.35}$$

onde $\gamma = 2 \cdot (\nu - 1) \cdot \beta_1$ e $z = \left(\operatorname{sgn}(x)\right)^{\delta} \cdot \ln\left(|x|\right)$.

Os parâmetros $\alpha, \beta_0, \beta_1, e\, \delta$ são então estimados através da resolução dos sistemas 2.36 a - 2.36 c:

$$\hat{\beta}_1 = \underset{\beta_0, \beta_1, \delta}{\arg\min} \sum_{i=1}^{N} \left[y_i - \left\{ \beta_0 + \beta_1 \cdot \left(\operatorname{sgn}(x_i)\right)^{\delta} \cdot |x_i| \right\} \right]^2 \quad (a)$$

$$\hat{\gamma} = \underset{\beta_0, \beta_1, \delta}{\arg\min} \sum_{i=1}^{N} \left[y_i - \left\{ \beta_0 + \beta_1 \cdot \left(\operatorname{sgn}(x_i)\right)^{\delta} \cdot |x_i| + \gamma \cdot \left(\operatorname{sgn}(x_i)\right)^{\delta} \cdot |x_i| \cdot \ln\left(|x_i|\right) \right\} \right] \quad (b) \tag{2.36}$$

$$\alpha^* = \left\{ \frac{\hat{\gamma}}{2 \cdot \hat{\beta}_1} + 1 \right\}^2 \quad (c)$$

Finalmente com a estimativa de α^* é possível se resolver o problema contido na Eq. 2.37:

$$\left[\beta_0^* \quad \beta_1^* \quad \delta^* \right] = \arg\min \sum_{i=1}^{N} \left[y_i - \left\{ \beta_0 + \beta_1 \cdot \left(\operatorname{sgn}(x_i)\right)^{\delta} \cdot |x_i|^{\alpha^*} \right\} \right] \quad (d) \tag{2.37}$$

A resolução das Eqs. 2.36 a – 2.36 c e 2.37 são otimizações quadráticas e podem ser resolvidas via mínimos quadrados. Como mencionado anteriormente, uma vez que o parâmetro δ pode assumir valores 0 ou 1 é preciso se decompor o problema em dois, um com $\delta = 0$ e outro com $\delta = 1$. Além disso, visando garantir estabilidade numérica, é preciso eliminar do subproblema representado pela Eq. 2.36 b os valores de x_i menores que zero.

LI *et al.* (2001) sugerem ainda a adoção de um limite ρ , de tal modo que somente os valores de $x_i > \rho$ sejam mantidos na resolução da Eq. 2.36 b. Os autores aconselham também utilizar limites máximos e mínimos para o parâmetro α^* para manter o algoritmo estável. Com a adoção de limites, os valores de α^* seriam truncados com base na Eq. 2.38.

$$\begin{aligned} \alpha^* &= \alpha_{\min} \quad se \left\{ \frac{\hat{\gamma}}{2 \cdot \beta_1} + 1 \right\}^2 < \alpha_{\min} \\ \alpha^* &= \alpha_{\max} \quad se \left\{ \frac{\hat{\gamma}}{2 \cdot \beta_1} + 1 \right\}^2 > \alpha_{\max} \end{aligned} \tag{2.38}$$

O procedimento para a utilização das transformações de Box-Tidwell com funcional para o PLS é análogo ao apresentado na subseção do PLS quadrático, tendo como diferença a necessidade de se realizarem dois problemas, uma para cada valor de δ. A forma do funcional para o caso de ele ser utilizado para mapear a relação entre os vetores t e os vetores u é apresentada na Eq. 2.39

$$\hat{u} = \beta_0^* + \beta_1^* \cdot [\mathrm{sgn}(t)]^{\delta^*} \cdot |t|^{\alpha^*} \tag{2.39}$$

No mesmo trabalho LI *et al*. (2001) propuseram uma extensão da Eq. 2.39, incluindo uma parcela linear, obtendo a Eq. 2.40.

$$\hat{u} = \beta_0^* + \beta_1^* \cdot t + \beta_2^* \cdot \{\mathrm{sgn}(t)\}^{\delta^*} \cdot |t|^{\alpha^*} \tag{2.40}$$

A metodologia para a determinação dos parâmetros $\alpha^*, \beta_0^*, \beta_1^*, \beta_2^*, \delta^*$ é similar a adotada para a obtenção dos parâmetros $\alpha^*, \beta_0^*, \beta_1^*, \delta^*$ e pode ser encontrada com mais detalhes em LI *et al.* (2001).

Embora LI *et al.* (2001) argumentarem que esse funcional confere a técnica PLS grande flexibilidade, sua utilização para o desenvolvimento de analisadores virtuais não se mostrou possível em função de problemas de convergência durante os processos de obtenção dos parâmetros do modelo.

2.5.4 Outras Estratégias PLS Não-Lineares

Além das estratégias apresentadas nas subseções anteriores há um número razoável de proposições para o funcional não-linear que relacione ***u*** e ***t***. QIN e McAVOY (1992) propuseram a utilização de uma rede neuronal de base sigmoidal composta por uma camada escondida. Os autores argumentam que qualquer função contínua pode ser aproximada com a desejada acuracidade por esse funcional. WILSON *et al.* (1997) propuseram a utilização de uma rede neuronal com função de base radial como funcional não-linear entre as variáveis latentes.

Ambos os procedimento utilizam uma rede neuronal de uma camada escondida para mapear a relação existente entre cada uma das variáveis latentes extraídas pelo PLS.

BAFFI *et al.* (1999b) propuseram a utilização do procedimento por eles desenvolvido, PLS baseado no erro, para a atualização dos vetores de peso para ambas as funções de ativação. O procedimento detalhado pode ser encontrado em seu trabalho, contendo, inclusive, as derivações das funções de ativação sigmoidal e de base radial fundamentais para a implementação do algoritmo.

2.6 Técnicas para adaptação de modelos

Em processos industriais é comum que ocorram mudanças ao longo do tempo (QIN 1997). Essas mudanças podem ser, por exemplo, desativação catalítica, acúmulos e decaimento de eficiência. Nesse cenário é necessário que se utilize uma estratégia para que os modelos se adaptem a essas mudanças, mantendo a qualidade de suas respostas.

As técnicas existentes para aperfeiçoar os modelos estão baseadas na existência de análises *off-line*, sendo que a cada novo dado disponível é realizada a etapa de correção/adaptação do modelo. Existem diferentes maneiras de se proceder a atualização de um modelo, desde uma simples atualização de *bias* até procedimentos mais sofisticados como as técnicas de Filtro de Kalman. Nas próximas subseções serão apresentadas de forma breve algumas dessas estratégias, enfocando suas vantagens e desvantagens.

2.6.1 Correção de BIAS

A forma mais simples de se corrigir qualquer modelo é simplesmente adicionar as respostas um termo que contenha a diferença entre a predição do modelo em um instante *i* e o real valor nesse mesmo instante, geralmente determinado por uma análise *off-line*. Essa diferença é denominada de *bias*.

Um modelo que se utilize dessa ferramenta para aprimorar suas respostas é formulado tipicamente da seguinte forma:

$$\hat{y}(t) = \hat{y}_p(t) + y_{BIAS} \tag{2.41}$$

Na Eq. 2.41 $\hat{y}(t)$ é a resposta fornecida pelo modelo corrigido no instante t , $\hat{y}_p(t)$ é a predição do modelo no instante t sem a correção e y_{BIAS} é a correção do modelo baseada na diferença entre o valor de $\hat{y}(k)$ e o valor determinado *off-line* y_{OL} no instante k multiplicado por um fator de escala α, o qual é uma função da confiança que se tem na medida laboratorial, sendo tipicamente um valor levemente menor que 1, podendo ser calculado por algum algoritmo mais sofisticado que leva em conta acertos passados.

$$y_{BIAS} = \alpha\left(y_{OL}(k) - \hat{y}(k)\right) \tag{2.42}$$

Em processos industriais, as análises *off-line* normalmente são realizadas em freqüência reduzida, podendo variar de 1 a 3 vezes por dia. Desse fato resulta uma das principais limitações dessa estratégia, pois como a correção é feita através da soma de um termo constante entre as análises, o erro será carregado até a próxima análise. A grande vantagem dessa técnica é a sua simplicidade de implementação.

2.6.2 Mínimos Quadrados Recursivos

O método de mínimos quadrados recursivos é a técnica mais empregada para a estimação de parâmetros em tempo real (MacGREGOR e DAYAL, 1996). O processo é fundamentado na atualização das matrizes de variância das entradas ($X^T \cdot X$) e covariância entre as entradas e saídas ($X^T \cdot Y$).

O procedimento é realizado de forma a cada novo dado disponível, os dados antigos são exponencialmente desconsiderados através da atualização das matrizes de variância e covariância utilizando-se as Eqs. 2.43 e 2.44.

$$\left(X^T \cdot X\right)_t = \lambda_t \cdot \left(X^T \cdot X\right)_{t-1} + x_t^T \cdot x_t \tag{2.43}$$

$$\left(X^T \cdot Y\right)_t = \lambda_t \cdot \left(X^T \cdot Y\right)_{t-1} + x_t^T \cdot y_t \tag{2.44}$$

Nas Eqs. 2.43 e 2.44 x_t e y_t são os vetores contendo os novos dados no tempo *t*, $\left(X^T \cdot X\right)_t$ e $\left(X^T \cdot Y\right)_t$ são as matrizes de variância e covariância atualizadas para o tempo *t*. A cada nova amostra disponível, os dados anteriores nas matrizes de covariância são desconsiderados de forma exponencial através de um fator de esquecimento λ_t ($0 < \lambda_t \leq 1$) e os dados novos são adicionados. O parâmetro λ_t é que controla o esquecimento dos dados antigos, sendo que quando ele é igual a unidade não há desconsideração dos dados antigos.

Essa estratégia permite que o modelo se adapte as novas condições de processo, ao invés de simplesmente adicionar um valor de *bias* constante. Algoritmos para a utilização desse procedimento em modelos do tipo MLR e PLS podem ser encontrados em QIN (1997), MacGREGOR e DAYAL (1996), entre outros.

2.6.3 Filtro de Kalman Estendido – EKF

O filtro de Kalman Estendido é uma das técnicas mais utilizadas para a estimação de estados no meio industrial (BARATTI *et al.*, 1995). Em função de sua larga aplicação descrições detalhadas do procedimento podem ser encontras em muitas referências, GELB (1974), BROWN e HWANG (1992), entre outros, sendo que será apresentado aqui somente uma breve descrição dos procedimentos envolvidos no algoritmo EKS, a qual foi baseada na versão apresentada por OISIOVICI e CRUZ (2000).

Assume-se que o processo a ser observado possui um vetor de estados $x(t) \in \Re^n$ e é descrito pelo modelo contínuo de processo apresentado na Eq. 2.45.

$$\dot{x}(t) = f\left(x(t), u(t)\right) + w(t) \tag{2.45}$$

O vetor de ruídos do processo *w* contém todas as perturbações que agem sobre o sistema e não são descritas deterministicamente. Essas são compostas, por exemplo, por dinâmicas não modeladas e entradas corrompidas por ruído ou não medidas. O modelo pode ser apresentado na sua forma discreta.

$$x_{k+1} = \bar{f}\left(x_k, u_k\right) + w_k \tag{2.46}$$

Assume-se que o ruído do processo é uma variável aleatória, com média zero e distribuição Gaussiana, com covariância Q_k e é independente dos ruídos do processo ou dos estados do sistema que ocorreram em instantes passados.

$$p\left(w_k\right) \sim \mathrm{N}\left(0, Q_k\right), \tag{2.47}$$

$$E\left[w_k \quad w_i'\right] = \begin{cases} Q_k, & i = k, \\ 0, & i \neq k, \end{cases} \tag{2.48}$$

$$E\left[w_k \quad x_i'\right] = 0, \quad \forall i, k. \tag{2.49}$$

As medidas adquiridas no instante *k* são agregadas ao vetor *p*-dimensional de observação z_k, o qual é relacionado com o estado do sistema através de

$$z_k = \bar{h}\left(x_k\right) + v_k. \tag{2.50}$$

O vetor v_k é o ruído da medida e é assumido que ele possui as mesmas características do vetor w_k, cuja variância é representada por R_k:

$$p\left(v_k\right) \sim \mathrm{N}\left(0, R_k\right). \tag{2.51}$$

Novamente assume-se que o vetor v_k é independente dos ruídos das medidas de todos os passos anteriores e é independente do ruído de processo:

$$E\left[v_k \quad v_i^T\right] = \begin{cases} R_k, & i = k, \\ 0, & i \neq k, \end{cases} \tag{2.52}$$

$$E\left[w_k \quad v_i^T\right] = 0, \quad \forall i, k. \tag{2.53}$$

Os erros das estimativas *a priori* e *a posteriori* são definidos, respectivamente, por:

$$e_{k|k-1} \equiv x_k - x_{k|k-1}, \tag{2.54}$$

$$e_{k|k} \equiv x_k - x_{k|k}. \tag{2.55}$$

A covariância da estimativa do erro *a priori* e *a posteriori* são representadas através de:

$$P_{k|k-1} = E\left[e_{k|k-1} \quad e_{k|k-1}^T\right], \tag{2.56}$$

$$P_{k|k} = E\left[e_{k|k} \quad e_{k|k}^T\right] \tag{2.57}$$

O EKF é inicializado com $x_{0|0} = x_0$ e $P_{0|0} = P_0$, e então o procedimento opera recursivamente, executando um único ciclo a cada nova medida disponível. Cada iteração propaga a estimativa do instante da última medida até o instante atual. O processo de propagação é formado por duas etapas: atualização e predição.

Etapa de Atualização

As equações de atualização são responsáveis pela incorporação da nova medida a estimativa *a priori* visando obter uma estimativa *a posteriori* melhorada. A estimativa dos estados *a posteriori* $x_{k|k}$ é calculada como uma combinação linear da estimativa *a priori* $x_{k|k-1}$ e uma diferença ponderada entre a medida atual z_k e a predição da medida:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k \cdot \left[z_k - \overline{h}\left(\hat{x}_{k|k-1}\right)\right], \tag{2.58}$$

onde

$$L_k = P_{k|k-1} \cdot \overline{H}_k^T \left(\overline{H}_k \cdot P_{k|k-1} \cdot \overline{H}_k^T + R_k\right)^{-1}, \tag{2.59}$$

$$\overline{H}_k = \left(\frac{\partial \overline{h}(x)}{\partial x}\right)_{x=\hat{x}_{k|k-1}}. \tag{2.60}$$

A matriz de ganhos L_k é determinada de forma a minimizar a covariância do erro *a posteriori*. As literaturas citadas anteriormente podem fornecer mais detalhes desse procedimento.

A matriz de covariância é atualizada através de:

$$P_{k|k} = \left(I - L_k \cdot \overline{H}_k\right) \cdot P_{k|k-1}. \tag{2.61}$$

Segundo MOHR (2004), a Eq. 2.61 não é a melhor forma para o cálculo da matriz de covariância sob o ponto de vista de estabilidade numérica. Em seu trabalho foi utilizada uma alternativa numericamente melhor, apresentada na Eq. 2.62.

$$P_{k|k} = \left(I - L_k \cdot \overline{H}_k\right) \cdot P_{k|k-1} \cdot \left(I - L_k \cdot \overline{H}_k\right)^T + L_k \cdot R_k \cdot L_k \tag{2.62}$$

Etapa de Predição

As equações da etapa de predição são responsáveis pela propagação do estado corrente e as estimativas das covariâncias dos erros para se obter as estimativas *a priori* para o próximo passo.

O estado e a matriz de covariância no próximo instante de amostragem são estimados através de:

$$\hat{x}_{k+1|k} = \bar{f}\left(\hat{x}_{k|k}, u_k\right), \tag{2.63}$$

$$P_{k+1|k} = \overline{F}_k \cdot P_{k|k} \cdot \overline{F}_k^T + Q_k, \tag{2.64}$$

onde,

$$\overline{F}_k = \left(\frac{\partial \bar{f}(x,u)}{\partial x}\right)_{x=\hat{x}_{k|k}, u=u_k} \tag{2.65}$$

Capítulo 3

Seleção de Variáveis

A etapa de seleção de variáveis é de fundamental importância na construção de qualquer modelo, independente de sua tipologia. Ela é fortemente responsável pelo sucesso, ou não, da modelagem.

Em estruturas fenomenológicas, esse etapa está associada à etapa de seleção do modelo, pois esse é quem vai determinar quais variáveis serão utilizadas. Em sistemas baseados em análises multivariáveis, no entanto, deve-se analisar de forma sistemática quais variáveis deverão compor o modelo, para que ele possa representar o mais fielmente possível os dados.

O campo de estudo de seleção de variáveis para modelos tipo caixa-preta é hoje largamente explorado, basta ver a diversidade de artigos publicados nessa área, nos mais diversos ramos da ciência, por exemplo, podem-se citar os trabalhos de LEARTI *et al.* (2002), GEASTAL *et al.* (2004), ZAMPROGNA *et al*. (2005), entre outros.

Além da seleção do melhor subconjunto de variáveis explicativas, em modelos com capacidade de redução de dimensionalidade, tais como PCA. PCR e PLS se faz necessário também a determinação do número ótimo de variáveis latentes, ou fatores, a serem utilizados.

Esse capítulo descreve algumas técnicas existentes para a seleção de variáveis para modelos baseados em análise multivariável, discutindo os seus pontos fortes e fracos. São também apresentados alguns métodos utilizados na determinação do número ótimo de fatores a serem utilizados em técnicas de redução de dimensionalidade.

Aqui neste capítulo também é apresentada uma modificação para o procedimento de validação cruzada, proposto neste trabalho. Este procedimento é comparado com algoritmos clássicos utilizados para essa finalidade.

3.1 Critérios de Seleção de Variáveis

Todo modelo que for considerado será, de alguma forma, errado. (BROWNE, 2000). Essa afirmação acarreta que procurar pelo modelo correto é uma impossibilidade. Por isso, ao invés de procurar-se por esse modelo perfeito e inexistente, deve-se procurar um modelo que seja uma aproximação adequada da realidade. Segundo BROWNE (2000) um bom modelo deve ser de claro entendimento e ser adequado para a situação estudada.

A seleção de variáveis que irão compor um modelo é realizada através da comparação entre os diferentes modelos gerados mediante as combinações das variáveis disponíveis. Essa comparação deve ser realizada de forma sistemática e, para isso, foram desenvolvidos diversos índices capazes de quantificar o ajuste do modelo aos dados experimentais.

QI e ZHANG (2001) dividiram os diferentes métodos de seleção de modelos/variáveis em duas classes, a primeira denominada *in sample*, utiliza uma série de índices de ajuste no conjunto de dados utilizados para gerar os modelos. A segunda utiliza um conjunto independente de dados, não utilizados na obtenção do modelo, para o cálculo de outros índices de ajuste. Essa metodologia recebeu a denominação de *out of sample*.

3.1.1 Índices de Ajuste

Como mencionado anteriormente, a determinação do modelo que mais se ajusta aos dados é feita através da utilização de índices de ajuste. Existem diversos tipos de índices, mas basicamente todos utilizam a soma quadrática do erro, SSE, como medida do desvio dos dados fornecidos pelo modelo e os originais, ou seja,

$$SSE = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 . \tag{3.1}$$

O mais difundido índice para medir a capacidade de ajuste de um modelo é o coeficiente de determinação, normalmente referenciado como R^2. Esse índice estabelece uma relação entre o erro originado pelo modelo e a distância de cada ponto a média do conjunto de calibração, expressa através da S_{yy}, que é dado por:

$$S_{yy} = \sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2 \tag{3.2}$$

Complementando-se em relação a 1 a relação entre *SSE* e S_{yy} obtemos a expressão final para o cálculo do R^2, ou seja,

$$R^2 = 1 - \frac{SSE}{S_{yy}} . \tag{3.3}$$

MONTGOMERY e PECK (1982) argumentam que o R^2 não é um índice utilizável para a determinação do melhor conjunto de variáveis, pois, em função de sua natureza, ele apresenta crescimento assintótico para a unidade a medida que novas variáveis são incluídas ao modelo.

Para contornar esse problema foi desenvolvido o R^2 Ajustado, que nada mais é que o coeficiente R^2 acrescido de um termo de penalização decorrente da complexidade do modelo. A inclusão desse termo faz com o coeficiente R^2 Ajustado não cresça necessariamente com a inclusão de uma nova variável.

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-p} \right) \cdot \left(1 - R^2\right) \tag{3.4}$$

A Eq. 3.4 representa a formulação do R^2 Ajustado, sendo n o número de amostras testadas e p o número de variáveis contidas no modelo.

Seguindo essa linha de aplicar penalidades em função da complexidade do modelo, foram desenvolvidos diversos índices, sendo que, segundo QI e ZHANG (2001), destancam-se dois em função de sua larga utilização.

O primeiro deles é o AIC, desenvolvido por AKAIKE (1974), que é compreendido por duas partes, como pode ser visto na Eq. 3.5. A primeira parte da equação responde pela capacidade do modelo em reproduzir os dados, enquanto que a segunda imprime uma penalidade em função da dimensão do modelo.

$$AIC = \log\left(\frac{SSE}{n} \right) + \frac{2 \cdot p}{n} \tag{3.5}$$

Outro parâmetro utilizado com freqüência para seleção de variáveis é o *Bayesian Information Criteria*, BIC, que apresenta uma modificação no termo de penalidade da complexidade do modelo, conforme pode ser observado na Eq. 3.6.

$$BIC = \log\left(\frac{SSE}{n} \right) + \frac{p \cdot \log(n)}{n} \tag{3.6}$$

QI e ZHANG (2001) apresentam adaptações das formulações originais do AIC e do BIC através das Eq. 3.7 e Eq. 3.8, onde d é um parâmetro empírico e, segundo GRANGER (1993), para modelos não-lineares deve assumir valores maiores que a unidade ($d > 1$).

$$AIC = \log\left(\frac{SSE}{n} \right) + \frac{2 \cdot p^d}{n} \tag{3.7}$$

$$BIC = \log\left(\frac{SSE}{n} \right) + \frac{p^d \cdot \log(n)}{n} \tag{3.8}$$

QI e ZHANG (2001) estudaram diversas variações das Eq. 3.7 e Eq. 3.8, realizando comparações desses frente a outros índices classicamente utilizados para a determinação da capacidade preditiva de modelos, ou seja, sua capacidade de estimar a variável de resposta em pontos fora do conjunto de calibração.

Dentre os índices utilizados para esse fim, os mais citados na literatura são a PRESS e a RMPSE. PRESS é a abreviação de *Predictive Sum of Squares*, que nada mais é que uma modificação da *SSE* (Eq. 3.1), onde é utilizado um conjunto independente de amostras, isto é, um conjunto de dados diferente do empregado no ajuste, para determinar a capacidade preditiva do modelo.

$$PRESS = \frac{\sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2}{n_2} \tag{3.9}$$

Na Eq. 3.9 n_2 é o número de observações contidas no conjunto de teste ou validação. A RMPSE é o acrônimo para *Root Mean Prediction Squared Error*. Esse índice nada mais é que a raiz quadrada da PRESS.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2}{n_2}} \tag{3.10}$$

Em seus trabalhos QI e ZHANG (2001) utilizaram alguns dos índices acima descritos para avaliar a performance dos métodos de seleção de variáveis *in sample* e *out of sample* para a construção de modelos capazes de inferir o comportamento de índices do mercado financeiro dos Estados Unidos da América.

A conclusão obtida pelos autores foi de que os modelos obtidos através da seleção de variáveis utilizando os métodos *in sample* foram muito inferiores aos obtidos através dos métodos *out of sample*. Em relação aos índices utilizados, os autores concluíram que parece não haver relação entre os melhores modelos obtidos com as técnicas *in sample* e *out of sample*, quando o critério de performance adotado para avaliar os modelos é o mesmo.

As conclusões obtidas por QI e ZHANG (2001) enfatizam a necessidade da utilização de procedimentos capazes de avaliar a qualidade das inferências produzidas pelos modelos em regiões/amostras não utilizadas no desenvolvimento do mesmo, esses procedimentos são também denominados de Validação Cruzada e, alguns deles, são apresentados na próxima subseção.

3.1.2 Validação Cruzada

Quando se deseja avaliar a capacidade preditiva de um modelo é necessário utilizar um conjunto de dados diferente daqueles utilizados na etapa de ajuste (calibração/construção) do modelo. Esse procedimento é denominado de Validação Cruzada.

A idéia da técnica de validação cruzada é dividir o conjunto original de *n* observações em dois novos conjuntos, um para calibração do modelo, composto por n_1 amostras e outro para validação ou teste do modelo, contendo n_2 amostras, de tal forma que $n = n_1 + n_2$.

O procedimento pode ser sumarizado através dos seguintes passos:

1. Geram-se os conjuntos de calibração e validação
2. Constrói-se o modelo a partir dos dados no subconjunto de calibração
3. Utiliza-se o conjunto de validação para calcular algum dos índices de performance, a PRESS por exemplo
4. Repete-se os passos 1 a 3 **B** vezes, somando-se os resultados da PRESS a cada etapa

O número **B** de repetições é função da estratégia utilizada para a geração dos conjuntos de calibração e validação. O modelo selecionado é aquele em que a soma da PRESS nas **B** etapas for o menor.

HÖSKULDSSON (1996) apresenta o algoritmo onde a geração dos subconjuntos se dá de forma aleatória, sendo que o conjunto de validação é composto por 10 a 20% dos dados originais. Em função do mecanismo estocástico utilizado na divisão das amostras o autor recomenda a adoção de um valor elevado para B, tipicamente igual a 100.

O procedimento LOO, abreviação *Leaving One Out*, é outra metodologia para a implementação do algoritmo de validação cruzada. Nesse procedimento, partindo-se de um conjunto de *n* amostras, geram-se *n* pares de conjuntos de calibração e validação. O processo é repetido *n* vezes, de modo que cada uma das observações seja utilizada no conjunto de validação uma única vez. A PRESS de cada modelo é obtida pela soma do erro quadrático de cada uma das *n* amostras.

Para esclarecer o conceito da metodologia LOO é utilizado o exemplo contido na Figura 3.1. Para esse caso, um conjunto de dados contendo nove amostras é utilizado. Na primeira etapa do procedimento, a amostra 1 é movida do conjunto original para o conjunto de validação. Calibra-se o modelo com as amostras restantes e a capacidade preditiva do modelo gerado é testada com a utilização da amostra 1.

Na próxima etapa, a amostra 2 é movida para o conjunto de validação. As demais oito amostras são utilizadas na calibração do modelo. Novamente, o modelo obtido tem sua capacidade preditiva avaliada pela amostra não utilizada na calibração, nesse caso a número 2. O procedimento é repetido até que cada uma das 9 amostras presentes seja utilizada uma única vez para a avaliação do modelo.

A capacidade preditiva final do modelo é conseguida pela soma de cada uma das contribuições individuais e depois comparadas com modelos gerados com diferentes conjuntos de variáveis.

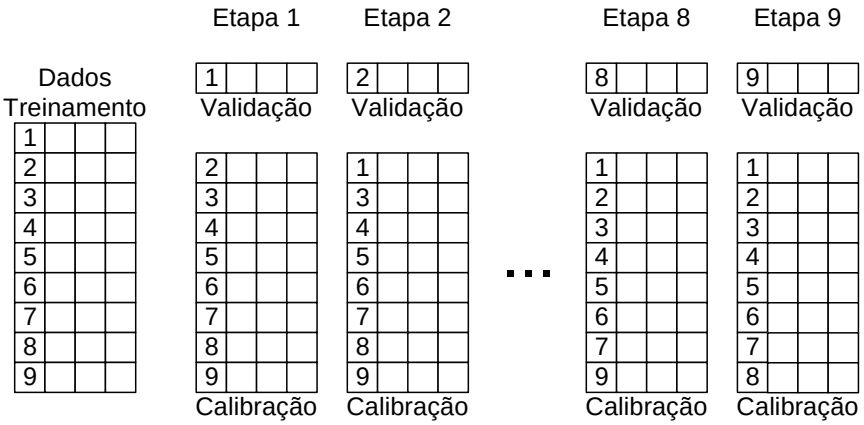


Figura 3.1: Esquema simplificado da segregação dos dados nos conjuntos de calibração e validação para o procedimento LOO

Teoricamente o método LOO é capaz de extrair o máximo de informação e, portanto, é o melhor. (LI *et al.*, 2002). No entanto para um conjunto grande de variáveis o procedimento acaba demandando um elevado esforço computacional. Em função disso alguns autores argumentam que é necessária somente a divisão dos dados em um número bem menor de subconjuntos de calibração e validação. Os pares formados por um subconjunto de calibração e um subconjunto de validação recebem a denominação de bloco. LI *et al.* (2002) recomendam a adoção de 4 a 6 blocos para a realização do procedimento de validação cruzada.

LI e colaboradores (2002) avaliaram o impacto da utilização de um número menor de blocos para o procedimento de validação cruzada. Nesse estudo os autores utilizaram a validação cruzada com 5, 10 e 100 blocos para um conjunto de 100 amostras.

OS resultados obtidos mostraram que quando o método LOO gerou um aumento de performance no procedimento de validação cruzada, esse foi marginal, não justificando o elevado esforço computacional demandado. Os autores concluíram que os resultados obtidos com a divisão dos dados em 5 blocos geraram resultados satisfatórios. Tendo como base esses resultados, durante os processos de validação cruzada nesse trabalho será adotado o número de 5 blocos de calibração e validação.

3.2 Metodologias para Seleção de Variáveis

O procedimento de seleção variáveis pode ser visto como um problema de otimização, onde a função objetivo consiste na minimização ou maximização de algum dos critérios apresentados na seção anterior.

A seguir são apresentadas algumas das estratégias encontradas na literatura para a determinação de forma automatizada do melhor modelo para um determinado conjunto de

dados. Serão apresentados procedimentos baseados em mecanismos de busca exaustiva, algoritmos seqüenciais e metodologias heurísticas.

3.2.1 Método de Busca Exaustiva

O método de busca exaustiva testa todas as possíveis combinações entre um conjunto de variáveis candidatas. Os modelos são gerados para cada um dos subconjuntos de variáveis e, através da utilização de algum critério, é escolhido o modelo que mais se adapta aos dados.

Apesar de esse método ser capaz de identificar o "melhor" subconjunto de variáveis seu custo computacional se mostra extremamente elevado, especialmente para casos onde há um grande número de variáveis candidatas (MONTGOMERY e PECK, 1982).

O número de modelos a serem avaliados por essa metodologia, supondo que exista um termo constante β_0 em todos os modelos, e que o número de variáveis candidatas seja *K*, é 2^K. Dessa forma, se o conjunto de variáveis candidatas for composto por 4 elementos, será necessário se avaliar 16 modelos. A Tabela 3.1 ilustra os possíveis modelos existentes para o caso de 4 variáveis candidatas.

Tabela 3.1: Modelos possíveis para 4 variáveis candidatas.

Número de Variáveis	Estruturas
0	β_0
1	$[x_1], [x_2], [x_3], [x_4]$
2	$[x_1 \quad x_2], [x_1 \quad x_3], [x_1 \quad x_4]$ $[x_2 \quad x_3], [x_2 \quad x_4], [x_3 \quad x_4]$
3	$[x_1 \quad x_2 \quad x_3], [x_1 \quad x_2 \quad x_4]$ $[x_1 \quad x_3 \quad x_4], [x_2 \quad x_3 \quad x_4]$
4	$[x_1 \quad x_2 \quad x_3 \quad x_4]$

Para um conjunto de 5 variáveis candidatas seria necessário avaliar 32 modelos, 6 variáveis seriam 64 modelos e assim por diante, sendo que o número de modelos dobraria a cada nova variável introduzida no conjunto de variáveis candidatas a compor o modelo.

3.2.2 Métodos Seqüenciais ou Stepwise

Em função da impossibilidade de avaliar os modelos gerados da técnica de busca exaustiva, foram desenvolvidos diversos métodos para avaliar apenas um pequeno número de subconjuntos de variáveis (MONTGOMERY e PECK, 1982).

Uma família desses métodos é a dos Métodos Seqüenciais ou *Stepwise*. A característica dessa classe é avaliar os efeitos das variáveis através da adição ou remoção de apenas uma variável em cada etapa, existindo três principais variações da metodologia.

A primeira variante é denominada de Seleção por Adição (*Forward Selection)* onde são adicionadas seqüencialmente as variáveis candidatas ao modelo, a segunda é a Seleção por Eliminação (*Backward Elimination)*, que por sua vez parte de um modelo com todas as variáveis e procede a eliminação seqüencial das variáveis. A terceira modificação é uma mistura das duas técnicas anteriores. Sendo que a cada adição de variável no conjunto é realizado um procedimento para verificar se não há variáveis redundantes. Essa metodologia recebe a denominação de *Stepwise Regression*.

Adição Seqüencial (Forward Selection, FS)

O procedimento de adição seqüencial de variáveis parte de um modelo onde há nenhuma variável. Criam-se então modelos com cada uma das variáveis candidatas e a variável que apresentar melhor desempenho em relação ao um determinado índice de performance é adicionada ao conjunto de variáveis que irão compor o modelo.

Na segunda etapa de seleção são criados modelos contendo combinações da variável previamente incluída e cada uma das variáveis não selecionadas anteriormente. A variável que gerar o melhor modelo em combinação com aquela introduzida na etapa anterior é pré-selecionada para compor o modelo. A variável pré-selecionada é então submetida a um teste F para verificar se a sua adição acarreta melhora significativa ao modelo dentro de uma margem de confiança.

Em caso positivo, a variável é adicionada ao conjunto que irá fazer parte do modelo e repete-se o procedimento para as demais variáveis. Isso é repetido até que a adição de uma nova variável não traga melhoria significativa ao modelo ou se todas as variáveis tiverem sido incluídas.

O algoritmo dessa metodologia pode ser representado através do diagrama de blocos contido na Figura 3.2.

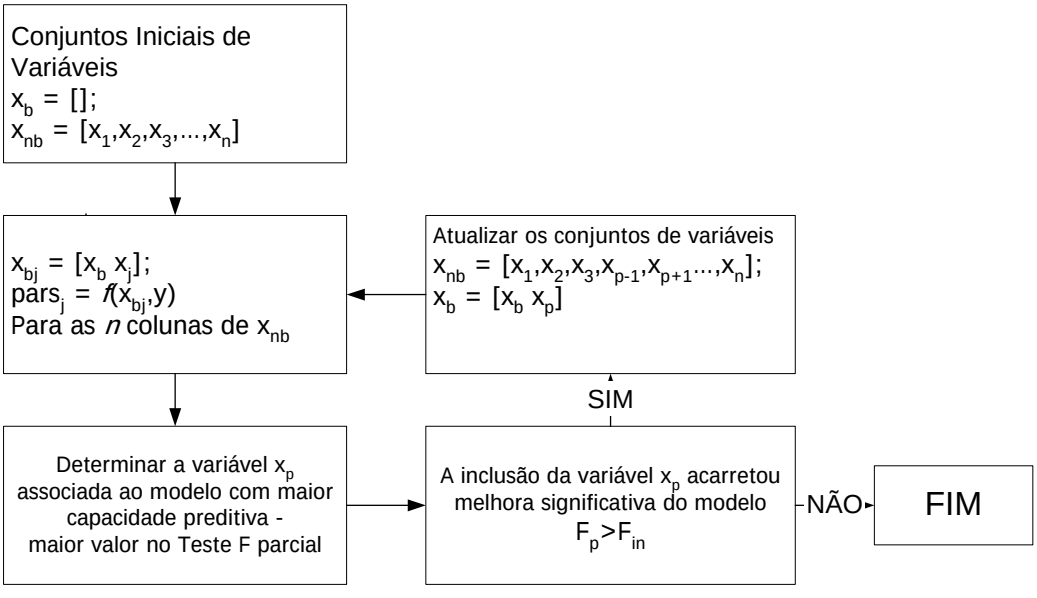


Figura 3.2: Representação do algoritmo da metodologia *Forward Selection*

Seleção por Eliminação (Backward Elemination, BE)

A idéia por trás do método *Backward Elemination* é obter o melhor subconjunto de variáveis através da sucessiva remoção de variáveis do modelo. O procedimento se inicia pela construção de um modelo contendo todas as *k* variáveis candidatas. A segunda etapa consiste em gerar modelos com *k*-1 variáveis. Verifica-se entre os modelos gerados qual não apresentou piora significativa nos resultados, através de um teste F, por exemplo. A variável que foi excluída do modelo pré-selecionado é removida do modelo.

O procedimento é repetido até que a remoção de uma variável piore significativamente o modelo ou até que todas as variáveis tenham sido eliminadas. MONTGOMERY e PECK argumentam que se costuma preferir esse método ao *Forward Selection* em função da possibilidade de se verificar o efeito produzido por um modelo com todas as variáveis. O algoritmo do procedimento pode ser representado pelo diagrama de blocos da Figura 3.3.

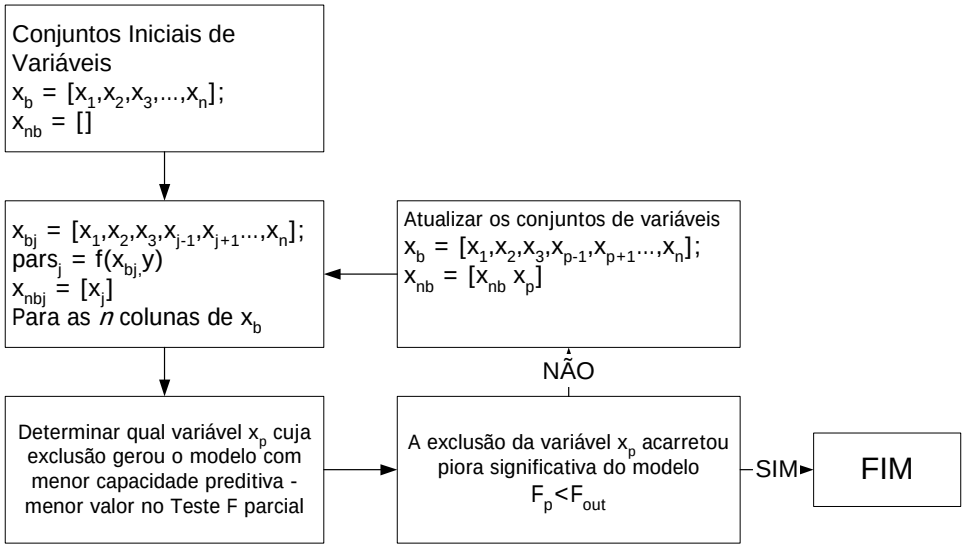


Figura 3.3: Representação do algoritmo do procedimento *Backward Elimination*

Stepwise Regression

Este procedimento é na verdade uma mistura dos anteriormente apresentados. O procedimento inicia-se como um *Forward Selection*, construindo-se modelos com cada uma das variáveis candidatas e selecionando-se aquela que resultar no melhor modelo. A diferença é que a partir da segunda etapa, a cada inclusão de variável é realizado o procedimento *Backward Elimination* para verificar se a variável adicionada não é redundante a uma outra previamente adicionada. A Figura 3.4 apresenta o algoritmo dessa metodologia.

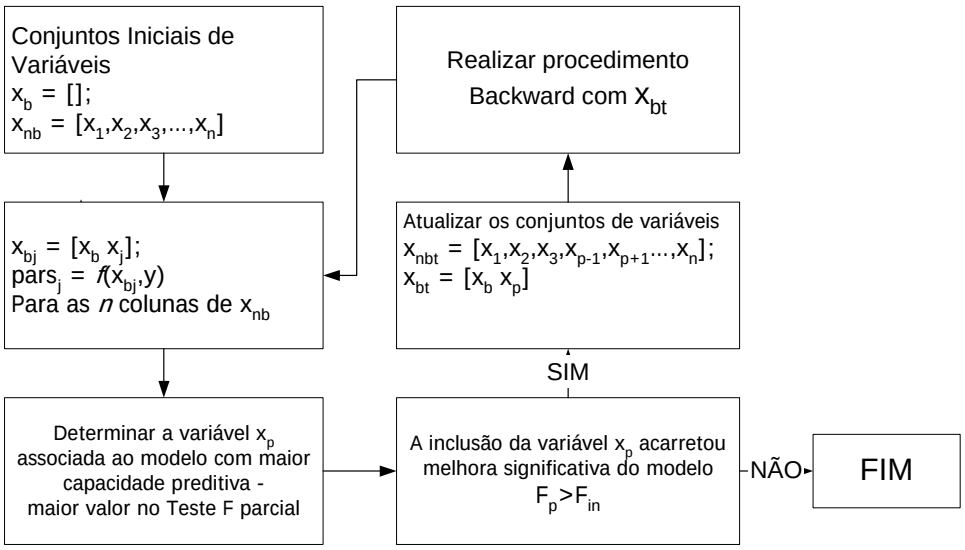


Figura 3.4: Representação do algoritmo do procedimento *Stepwise Regression*

Stepwise Regression Based on Model Predicition – SRMP

Esse método é uma variação do clássico procedimento *Forward Selection* e foi proposto por FINKLER (2003). A grande modificação em relação ao algoritmo original é o fato de que as variáveis são selecionadas de modo a gerarem o modelo com melhor capacidade preditiva.

Isso é conseguido através da utilização de validação cruzada, sendo que em cada etapa do procedimento *Forward Selection*, são gerados aleatoriamente ***B*** conjuntos de calibração e validação. Os modelos são construídos com os conjuntos de calibração e testados no conjunto de validação. Computa-se a PRESS para cada um dos ***B*** modelos gerados e estabelece-se a média.

Na primeira etapa, a variável que apresentar menor PRESS média é adicionada ao conjunto básico de variáveis, ou seja, ao conjunto de variáveis que farão parte do modelo. Na segunda etapa repete-se o procedimento, através da combinação da variável previamente selecionada e aquelas ainda não selecionadas. A variável que fornecer modelos com a menor PRESS média é pré-selecionada. Utiliza-se, então, o teste T para verificar se a adição de uma nova variável contribui significativamente para a melhoria do modelo, medida através do decréscimo da PRESS média.

O procedimento é repetido até que todas as variáveis tenham sido selecionadas ou até que a inclusão de uma nova variável ao conjunto básico não traga redução significativa da PRESS.

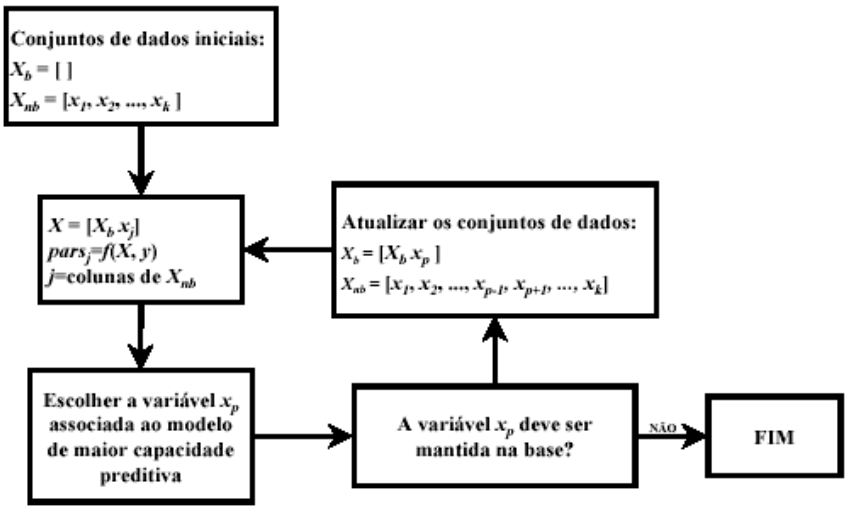


Figura 3.5: Representação do algoritmo do método SRMP

3.2.3 Algoritmos Genéticos

O método é assim denominado em função da sua origem, que buscava emular o comportamento evolutivo das espécies vivas. Ele tem sido largamente utilizado em problemas de otimização (HAN e YANG, 2004). Por ser um método estocástico ele não garante que a solução encontrada seja a ótima, porém a probabilidade dessa estar próxima ao ótimo global é alta.

Além da área de otimização essa técnica tem sido empregada com sucesso em diversas outras áreas, tais como seleção de componentes principais e análise de padrões (BARROS e RUTLEDGE, 1998; COWGILL e HARVEY, 1999; McSHANE *et al.*, 1999) e na área de design de produtos (NISHIMO *et al.*, 1994; TSUCHIYA *et al.*, 1996).

No campo de seleção de variáveis ele representa uma importante ferramenta de seleção de variáveis, especialmente quando o número de variáveis independentes passíveis de comporem o modelo é elevado.

A descrição sucinta da aplicação do algoritmo genético (AG) para a seleção de variáveis apresentada nesse documento está baseada, em grande parte, do trabalho de HAN e YANG (2004).

O procedimento é iniciado através da criação aleatória de uma população de subconjuntos, sendo que cada subconjunto contém uma combinação das k variáveis candidatas disponíveis. Na terminologia do AG esses subconjuntos são denominados de cromossomos e cada um dos genes que compõem o cromossomo é uma das variáveis. Os genes são a codificação binária para representar se a variável *i* está ou não presente no subconjunto, ou seja, se a posição i do cromossomo apresentar o número 1 significa que a variável *i* está presente nesse subconjunto, a presença do zero implica na inexistência da variável nesse conjunto.

Para exemplificar consideremos um conjunto formado por 9 variáveis candidatas a integrarem um modelo qualquer. Por exemplo, se o cromossomo for 001101101 as variáveis 3, 4, 6, 7 e 9 são incluídas, enquanto que as variáveis 1, 2, 5 e 8 não são incluídas.

Uma seleção aleatória ponderada é aplicada sobre a população original, onde a probabilidade de sobrevivência (seleção) de um determinado cromossomo está associada a uma resposta de uma função custo.

Dessa forma, cromossomos com uma resposta favorável apresentaram uma maior chance de serem selecionadas, transmitindo suas características para a próxima geração. Alguns cromossomos recessivos são removidos, enquanto que outros dominantes são repetitivamente selecionados.

A função custo pode ser representada por alguns dos índices de avaliação de modelos descritos nas seções anteriores, tais como o R^2 ajustado e a raiz quadrada do erro médio (RMSE).

A obtenção das próximas gerações é feita através de operadores genéticos entre os cromossomos selecionados. Esses operadores imitam o procedimento de transmissão de características próprio das espécies vivas, tais como cruzamento, mutação, clonagem.

O cruzamento é feito entre dois cromossomos selecionados, sendo que durante esse processo eles geram dois novos indivíduos através da combinação de partes dos seus genes. A mutação é conseguida através da alteração de alguns dos genes do cromossomo selecionado e a clonagem é simplesmente a cópia de um cromossomo da geração anterior para a nova população. Essas combinações são ilustradas na Figura 3.6.

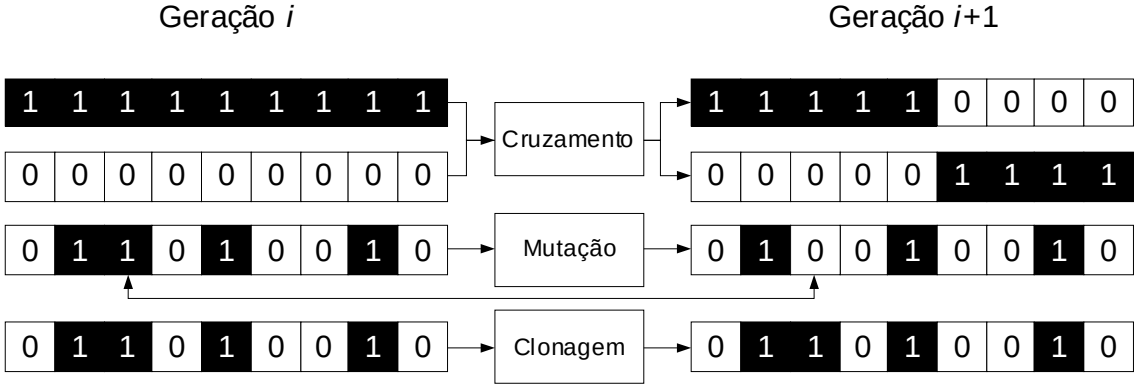


Figura 3.6: Operadores básicos utilizados no método de algoritmos genéticos

O mecanismo de escolha de qual gene sofrerá mutação é aleatório, sendo que a probabilidade de ocorrência desse fenômeno é um dos parâmetros que deve ser definido antes da execução do método.

Ao longo das gerações, a população tende a se concentrar em uma região de busca, causando o que na literatura se denomina *genetic drift* (BISCAIA Jr. E VIEIRA, 2004). Esse fenômeno

resulta em um efeito nefasto sobre o AG, pois tende a estagnar o algoritmo. Para evitar isso foi proposto um operador, chamado de reinicialização periódica, que após um determinado número de gerações armazena os melhores cromossomos, reinicializa aleatoriamente a população e insere nessa população os melhores indivíduos das gerações anteriores.

Outro operador interessante, sobretudo na área de problemas de otimização, é o *niching*. Esse processo tende a penalizar uma região onde sabe-se da existência de um mínimo local. Essa penalização é tanto maior quanto mais próximos dessa região estiverem os cromossomos.

Esses processos são repetidos entre os cromossomos selecionados da geração anterior até que a população original seja atingida. A avaliação da função peso é então aplicada a essa nova população e o processo é repetido.

O algoritmo continua a criação de novas gerações até que se atinja um critério de parada pré-definido. Esse critério pode ser um número de gerações pré-determinado, ou um valor do parâmetro de performance adotado, ou ainda através da similiradidade da estrutura dos cromossomos gerados.

A técnica de algoritmos genéticos para seleção de variáveis foi aplicada com sucesso para modelos MLR (WASSERMAN e SUDJIANTO, 1994; BROADHURST *et al.*, 1997; YIN *et al.*, 2002), modelos PLS (HASEGAWA e FUNATSU, 1997), entre outros.

3.3 Técnicas de Seleção de Dados para a Calibração e Validação de Modelos

Um fator tão importante quanto as variáveis que constituirão o modelo é a escolha de um conjunto de dados que contenha o máximo de informação possível. Esse procedimento é especialmente importante quando dados de processo serão utilizados para a determinação de um modelo (FLÅTEN e WALMSLEY, 2004).

Para automatizar esse processo foram desenvolvidos métodos capazes de, a partir de um conjunto de *n* amostras, gerar dois novos subconjuntos, sendo um utilizado para a calibração e outro para a validação do modelo.

Os seguintes métodos de seleção de dados são apresentados nesse trabalho: Seleção Aleatória, *D-Optimal Subset*, *y-Rank*, Algoritmo *Kennard-Stone*. Nessa seção também é proposta uma modificação do Algoritmo de *Kennard-Stone*, para que o mesmo considere também o vetor de saída na seleção dos dados.

As diferenças produzidas pelos algoritmos utilizados ao longo do desenvolvimento desse trabalho serão mostradas através de um exemplo bastante simples, sendo o objetivo mais didático do que realmente apresentar as potencialidades de cada estratégia.

3.3.1 D-Optimal Subset

FERRÉ e RIUS (1997) descrevem o algoritmo de seleção de dados através do critério da D-optimalidade de uma maneira bastante simples. Segundo os autores, as amostras são selecionadas de modo a reduzir a região de confiança $100 \cdot (1-\alpha)\%$ dos parâmetros do modelo, produzindo, dessa forma, estimativas confiáveis.

A seleção das amostras do conjunto de calibração é realizada de modo a maximizar o determinante da matriz de calibração $\boldsymbol{X_{Cal}}$ - $Det(X'_{cal} \cdot X_{cal})$. O procedimento tem início através de uma seleção aleatória de amostras, que são movidas para o conjunto de calibração. A partir daí o algoritmo substitui, de forma iterativa, uma das amostras presentes no conjunto de calibração por uma amostra do conjunto de dados restantes, procurando maximizar $Det(X'_{cal} \cdot X_{cal})$. Uma vez que a substituição de uma amostra inicialmente introduzida por uma externa não acarretar aumento no determinante da matriz $\boldsymbol{X_{Cal}}$ o procedimento é encerrado.

FERRÉ e RIUS (1997) argumentam que para evitar a seleção de um subconjunto sub-ótimo de dados, ou seja, parar em um ótimo local, é necessário repetir-se o procedimento algumas vezes, sendo que os autores concluíram que o número de repetições entre 5 e 10 já é suficiente para a localização do melhor subconjunto de dados para a calibração do modelo.

3.3.2 Seleção Aleatória

Esse é o método mais simples de segregação de dados. As amostras são distribuídas nos subconjuntos de calibração e validação de forma aleatória. Muito utilizado em testes de validação cruzada, sendo que em função de sua natureza estocástica o procedimento deve ser repetido várias vezes para que o resultado obtido possa ser considerado significativo do ponto de vista estatístico.

3.3.3 Seleção y-Rank

Esse procedimento é descrito em um *toolbox* desenvolvido pela FABI (1997) para o desenvolvimento de modelos a serem utilizados em métodos de quimiometria, sendo o código desenvolvido a única fonte de referência disponível.

A seleção de dados é realizada utilizando somente os valores da variável de resposta $\boldsymbol{y}$. Esse procedimento seleciona as amostras a partir dos valores da variável de resposta *y*. Os dados são classificados em ordem crescente em relação ao vetor $\boldsymbol{y}$. A partir de uma razão de separação pré-determina o algoritmo distribui os dados em dois subconjuntos.

Para clarificar a idéia dessa metodologia, consideremos o seguinte exemplo. Um conjunto de dados formados por 12 amostras é apresentado na Tabela 3.2. Será adotado nesse exemplo uma razão de 3/1, ou seja, 75% dos dados serão utilizados para calibração e os 25% restantes para a validação do modelo.

Tabela 3.2: Dados referentes ao exemplo da Seleção y-Rank

Índice	*y*	x_1	x_2
1	6.02E+00	9.00E-01	8.44E-01
2	1.31E+00	-5.38E-01	4.76E-01
3	-2.81E+00	2.14E-01	-6.47E-01
4	-9.99E-01	-2.80E-02	-1.89E-01
5	5.92E+00	7.83E-01	8.71E-01
6	5.22E+00	5.24E-01	8.34E-01
7	-1.07E+00	-8.71E-02	-1.79E-01
8	2.01E+00	-9.63E-01	7.87E-01
9	-3.14E+00	6.43E-01	-8.84E-01
10	-1.69E+00	-1.11E-01	-2.94E-01
11	3.59E+00	2.31E-01	6.26E-01
12	-3.73E+00	5.84E-01	-9.80E-01

A distribuição espacial das amostras contidas na Tabela 3.2 é apresentada na Figura 3.7. Nota-se claramente a presença de 4 regiões de concentração das amostras

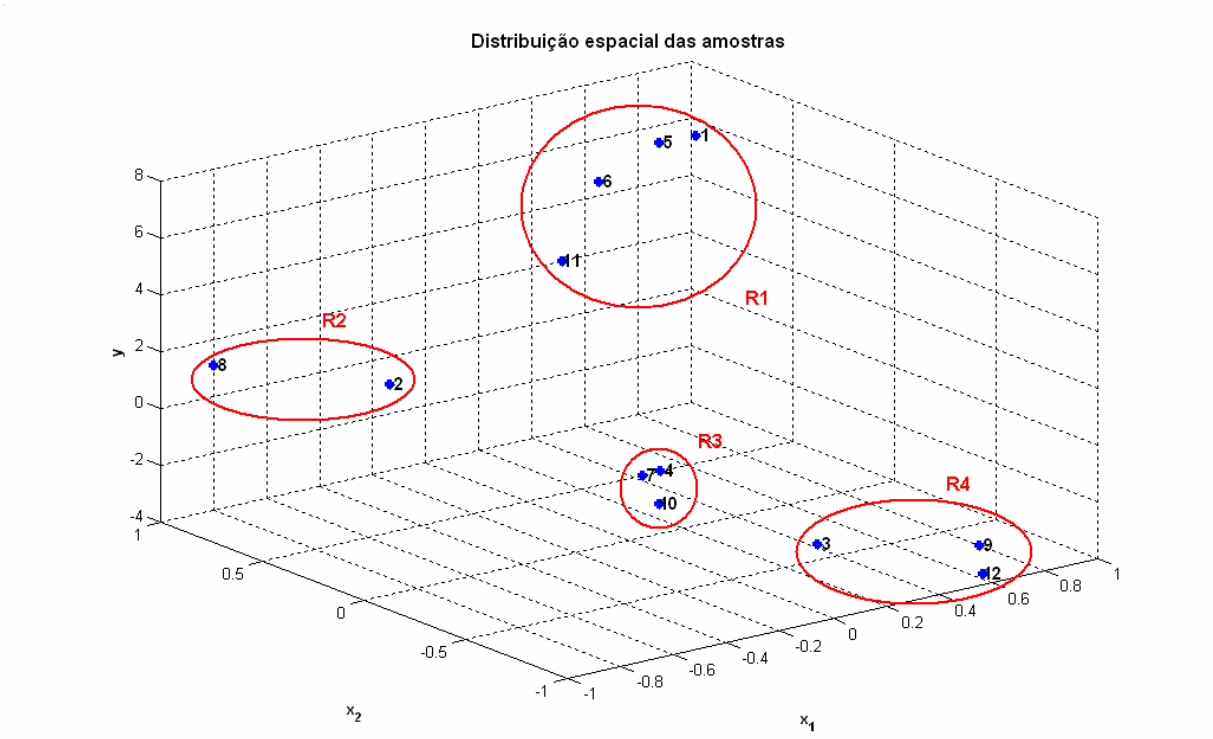


Figura 3.7: Distribuição espacial das amostras utilizadas no exemplo

Primeiramente os dados são ordenados, ou ranqueados, em ordem crescente em relação a variável ***y***. O processo de seleção move os três primeiros dados ordenados (amostras 12, 9 e 3) para o conjunto de calibração, o próximo ponto (amostra 10) é movido para o conjunto de validação. Posteriormente, as amostras 7, 4 e 2 são movidas para o subconjunto de calibração, enquanto que a amostra 8 é transferida para o conjunto de validação. O procedimento se repete até que todos os dados tenham sido transferidos para um dos subconjuntos. A

Tabela 3.3 apresenta os resultados produzidos por essa metodologia para os dados utilizados nesse exemplo.

Tabela 3.3: Dados ordenados em ordem crescente em relação a variável de resposta

Índice	y	x_1	x_2	Calibração	Validação
12	-3.73E+00	5.84E-01	-9.80E-01	✓	
9	-3.14E+00	6.43E-01	-8.84E-01	✓	
3	-2.81E+00	2.14E-01	-6.47E-01	✓	
10	-1.69E+00	-1.11E-01	-2.94E-01		✓
7	-1.07E+00	-8.71E-02	-1.79E-01	✓	
4	-9.99E-01	-2.80E-02	-1.89E-01	✓	
2	1.31E+00	-5.38E-01	4.76E-01	✓	
8	2.01E+00	-9.63E-01	7.87E-01		✓
11	3.59E+00	2.31E-01	6.26E-01	✓	
6	5.22E+00	5.24E-01	8.34E-01	✓	
5	5.92E+00	7.83E-01	8.71E-01	✓	
1	6.02E+00	9.00E-01	8.44E-01		✓

Utilizando a Tabela 3.3 e a Figura 3.7 como guias de análise pode-se notar que o algoritmos foi capaz de selecionar pontos para calibração pertencentes a todas as regiões de operação, e selecionando pontos de validação em regiões distintas, o que, a princípio, garante que o modelo seja testado também em diferentes regiões, o que pode ser verificado através da Figura 3.8.

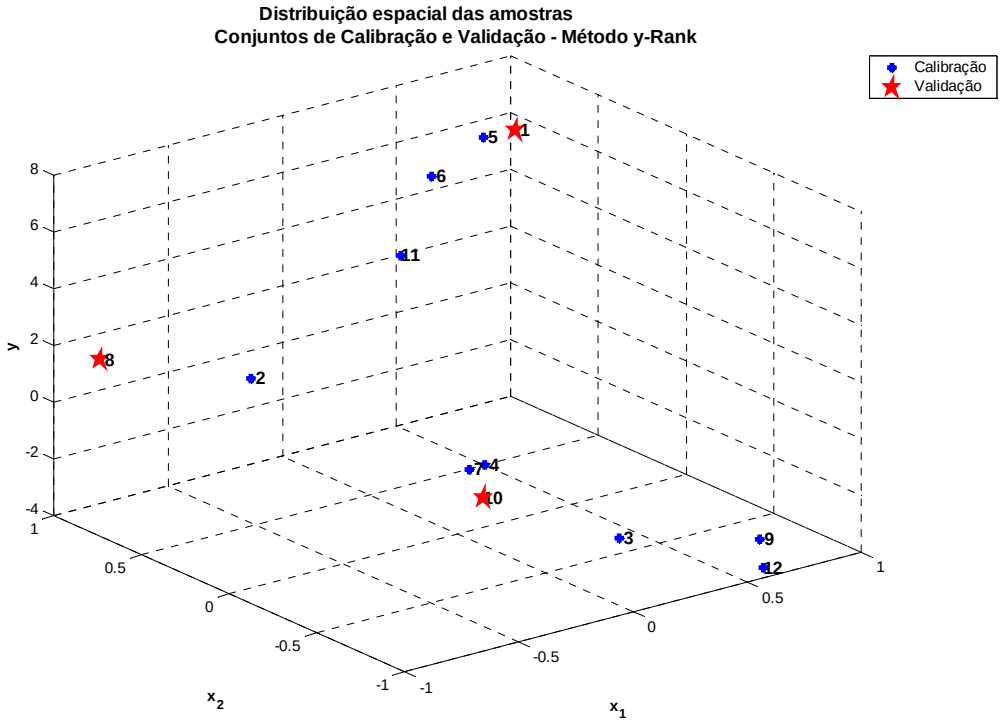


Figura 3.8: Conjuntos de calibração e validação obtidos pelo método y-Rank

O uso da variável de resposta como guia na seleção de dados, permite a escolha de amostras em diferentes regiões de dados. Essas regiões podem, no caso de indústrias de processo, representar diferentes pontos de operação das unidades.

3.3.4 Algoritmo de Kennard e Stone

Dentre os diversos algoritmos desenvolvidos, o de Kennard e Stone é o mais conhecido entre os analistas químicos (DASZYKOWSKI *et al*., 2002). O critério de seleção de amostras é baseado nas distâncias entre elas.

O procedimento se inicia pela determinação da distância de todas as amostras em relação ao valor média das amostras. Seleciona-se então o ponto mais distante, ou o mais próximo conforme determinação do usuário, dela, esse ponto será denominado de s_1.

A seleção de amostras segue, então, um procedimento seqüencial, tendo como base a norma quadrática às amostras já selecionadas. A norma quadrática entre a *i*-ésima e *j*-ésima amostra é definida pela Eq. 3.11.

$$d_{ij}^2 = \left\| x_i - x_j \right\| = \sum_k \left(x_{ik} - x_{jk} \right)^2 \tag{3.11}$$

A segunda amostra selecionada, s_2, é aquela mais distante de s_1. A terceira é a mais distante de s_1 e s_2 e assim por diante.

Os principais passos do algoritmo Kennard e Stone podem ser resumidos da seguinte forma:

1. Selecionar a amostra mais próxima/afastada da média e adicioná-la ao subconjunto de calibração
2. Calcular a distância entre as amostras restantes no conjunto e às já movidas para o subconjunto de calibração
3. Selecionar a amostra mais afastada daquelas já adicionadas ao subconjunto de calibração retornar a 2 até que o número desejado de amostras tenha sido atingido.

A aplicação dessa metodologia ao conjunto de dados utilizados na subseção anterior produziu os conjuntos de calibração e validação apresentados na Tabela 3.4. Os dados estão organizados na ordem em que foram selecionados pelo algoritmo. Como ponto inicial a integrar o conjunto de calibração optou-se pelo ponto mais afastado da média.

Novamente pode-se fazer uma breve análise dos resultados observando a Tabela 3.4 e a Figura 3.9. Apesar de o método ter produzido um conjunto de calibração com pontos em todas as regiões, o conjunto de validação ficou concentrado em apenas duas regiões, o que impede que o modelo já testado em todas as regiões disponíveis.

Tabela 3.4: Conjuntos de calibração e validação produzidos pela metodologia de *Kennard-Stone*

Índice	*y*	x_1	x_2	Calibração	Validação
8	2.01E+00	-9.63E-01	7.87E-01	✓	
12	-3.73E+00	5.84E-01	-9.80E-01	✓	
1	6.02E+00	9.00E-01	8.44E-01	✓	
7	-1.07E+00	-8.71E-02	-1.79E-01	✓	
11	3.59E+00	2.31E-01	6.26E-01	✓	
2	1.31E+00	-5.38E-01	4.76E-01	✓	
3	-2.81E+00	2.14E-01	-6.47E-01	✓	
6	5.22E+00	5.24E-01	8.34E-01	✓	
5	5.92E+00	7.83E-01	8.71E-01	✓	
4	-9.99E-01	-2.80E-02	-1.89E-01		✓
9	-3.14E+00	6.43E-01	-8.84E-01		✓
10	-1.69E+00	-1.11E-01	-2.94E-01		✓

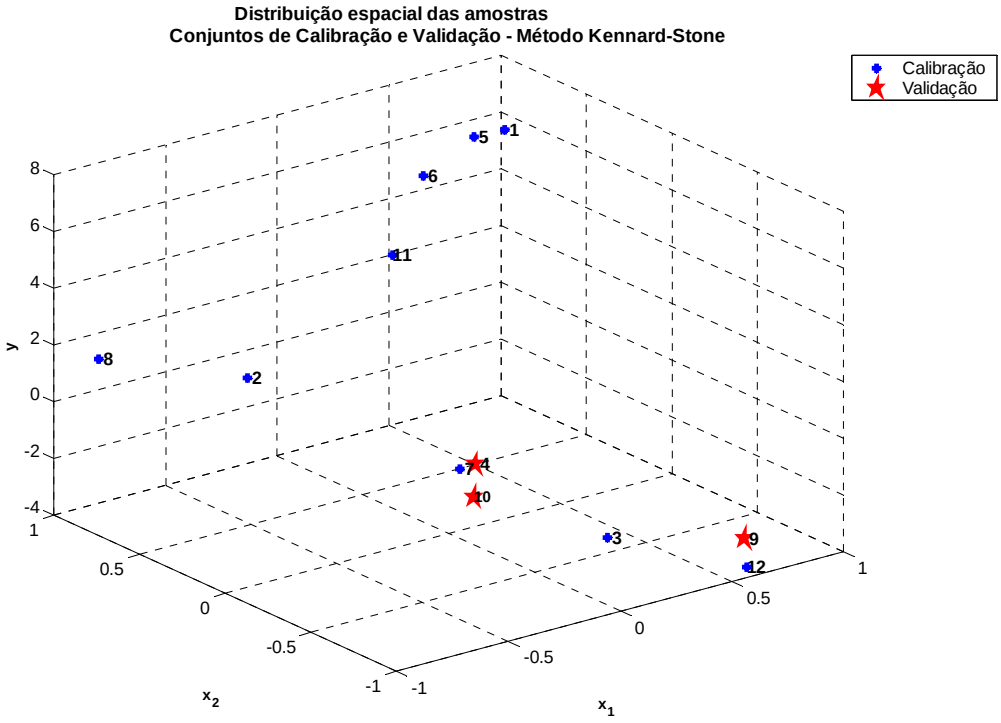


Figura 3.9: Conjuntos de calibração e validação obtidos pelo método Kennard-Stone

3.3.5 Modificação do Algoritmo de Kennard e Stone

O algoritmo proposto por Kennard e Stone procura selecionar as amostras para o conjunto de calibração e validação de modo que essas sejam as mais distantes entre si em relação somente as variáveis explicativas ou de entrada. Essa estratégia é particularmente eficaz quando o modelo a ser calibrado apresenta natureza linear. Em modelos não-lineares, no entanto, utilizar somente as variáveis explicativas pode não ser o procedimento mais adequado, pois uma distância grande entre as entradas pode não garantir o mesmo comportamento nas saídas.

Dessa forma, propomos nesta dissertação incorporar as variáveis de resposta, ou de saída, ao processo de seleção de amostras proposto por Kenard e Stone. A Eq. 3.11 é alterada para que

possa contabilizar também a distância entre as variáveis de resposta do problema a ser estudado.

$$\begin{aligned} d_{ij}^2 &= w_x \cdot \left\| x_i - x_j \right\| + w_y \cdot \left\| y_i - y_j \right\| \\ &= w_x \cdot \sum_k \left(x_{ik} - x_{jk} \right)^2 + w_y \cdot \sum_k \left(y_{ik} - y_{jk} \right)^2 \end{aligned} \tag{3.12}$$

Na Eq. 3.12 os termos w_x e w_y são ponderações, as quais assumem valores entre 0 e 1, submetidas a restrição $w_x + w_y = 1$.

A simples alteração dos valores das ponderações pode levar ao método KS clássico, $w_x = 1$ e $w_y = 0$, ou a um método que considere inteiramente a distância entre as saídas, $w_x = 0$ e $w_y = 1$.

A aplicação da metodologia desenvolvida ao exemplo apresentado anterior produziu os conjuntos de calibração e validação apresentados na Tabela 3.5. Para a execução do algoritmo optou-se por considerar $w_x = 0$ e $w_y = 1$, e novamente o primeiro ponto a ser inserido no conjunto de calibração foi o mais afastado da média. A ordem dos pontos na Tabela 3.5 segue a ordem de seleção produzida pela metodologia.

Tabela 3.5: Conjuntos de calibração e validação produzidos pela metodologia de *Kennard-Stone* Modificado

Índice	*y*	x_1	x_2	Calibração	Validação
1	6.02E+00	9.00E-01	8.44E-01	✓	
12	-3.73E+00	5.84E-01	-9.80E-01	✓	
2	1.31E+00	-5.38E-01	4.76E-01	✓	
7	-1.07E+00	-8.71E-02	-1.79E-01	✓	
11	3.59E+00	2.31E-01	6.26E-01	✓	
3	-2.81E+00	2.14E-01	-6.47E-01	✓	
6	5.22E+00	5.24E-01	8.34E-01	✓	
8	2.01E+00	-9.63E-01	7.87E-01	✓	
10	-1.69E+00	-1.11E-01	-2.94E-01	✓	
4	-9.99E-01	-2.80E-02	-1.89E-01		✓
5	5.92E+00	7.83E-01	8.71E-01		✓
9	-3.14E+00	6.43E-01	-8.84E-01		✓

A Figura 3.10 apresenta a distribuição das amostras entre os conjuntos de calibração e validação obtidos pelo método proposto. Utilizando-se a Tabela 3.5 e a Figura 3.10 para avaliar os subconjuntos resultantes, pode-se notar que o método foi capaz de produzir um conjunto de calibração contendo amostras em todas as regiões. O conjunto de validação é formado por amostras em três regiões distintas, garantindo, em teoria, um modelo com melhores características.

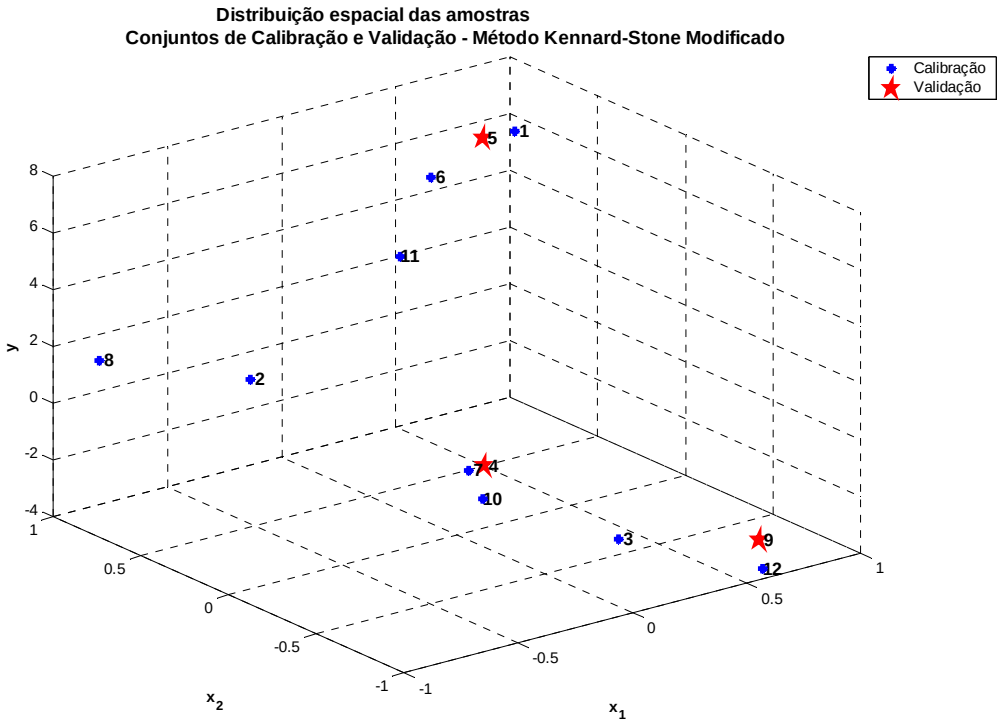


Figura 3.10: Conjuntos de calibração e validação obtidos pelo método Kennard-Stone Modificado

3.4 Estudo Comparativo e Proposição de Novas Técnicas de Seleção de Variáveis

Os procedimentos tradicionais de seleção de variáveis utilizam um procedimento de validação cruzada onde os subconjuntos de calibração e validação são gerados aleatoriamente. Em função da característica estocástica desse procedimento, é necessário que sejam avaliados um grande número de subconjuntos em cada etapa, fazendo que o resultado produzido, representado pelo parâmetro de performance escolhido, apresente significância estatítica.

O algoritmo SRMP, discutido na seção 3.2.2, e os métodos heurísticos como os algoritmos genéticos utilizam essa estratégia para a seleção de variáveis. Em função da necessidade de repetição esses algoritmos acabam por demandar um elevado esforço computacional, que pode ser crítico quando o vetor de variáveis candidatas a integrarem o modelo for grande.

Essa seção apresenta novas abordagens para os procedimentos de seleção de variáveis através da combinação das técnicas de seleção de seleção de dados apresentados na seção anterior e mecanismos de seleção de variáveis clássicos, como os métodos seqüenciais e o algoritmo genético.

As novas abordagens são comparadas com a clássica geração aleatória de subconjuntos por meio de três diferentes estudos de caso, sendo dois deles lineares e um terceiro não-linear. Ao final dessa seção é realizada uma análise comparativa das alternativas avaliadas quanto a suas acuracidades, capacidade de selecionar as variáveis verdadeiras, esforço computacional,

medido através do tempo necessário para avaliação das técnicas e da qualidade dos modelos obtidos, utilizando medidas de erro médio quadrático.

3.4.1 Proposição de Novas Técnicas de Seleção de Variáveis

Os algoritmos de seleção/segregação de dados apresentados na seção anterior tem a capacidade de gerar conjuntos de calibração e validação de tal forma que os modelos obtidos com a utilização desses subconjuntos sejam otimizados em termos de qualidade do modelo (FLÅTEN e WALMSLEY, 2004).

A combinação dessas técnicas com mecanismos de seleção de variáveis é apresentada pioneiramente nesse trabalho, gerando três novas alternativas ao procedimento clássico de validação cruzada para a determinação das variáveis que devem ser consideradas em modelos do tipo MLR, PCR e PLS.

As novas alternativas propostas nesse trabalho são sumarizadas na Tabela 3.6, onde a primeira parte da denominação (FS e GA) referem-se ao método se seleção utilizado, Seleção por Adição e Algoritmos Genéticos respectivamente, e a segunda parte (_02, _03 e _04) indica qual técnica de construção dos conjuntos de calibração e validação foi empregada.

Na mesma Tabela 3.6, está contida a forma classicamente utilizada no procedimento de validação cruzada utilizando a geração aleatória dos subconjuntos de calibração/validação, denotadas pela extensão _01 ao lado do método utilizado.

Tabela 3.6: Combinação de técnicas abordadas para a seleção de variáveis

Denominação	Estratégia de Seleção de Variáveis	Estratégia de Divisão dos Dados
FS_01	Forward Selection	Aleatória
FS_02	Forward Selection	y-Rank
FS_03	Forward Selection	KS Clássico
FS_04	Forward Selection	KS Modificado
GA_01	Algoritmos Genéticos	Aleatória
GA_02	Algoritmos Genéticos	y-Rank
GA_03	Algoritmos Genéticos	KS Clássico
GA_04	Algoritmos Genéticos	KS Modificado

Obs.: Para fins de visualização as alternativas propostas nesse trabalho são destacadas com preenchimento cinza.

3.4.2 Estudos de Caso

A fim de avaliar a potencialidade das estratégias desenvolvidas, utilizou-se 3 casos sintéticos, sendo dois deles lineares e o outro um caso não-linear. Todos os códigos referentes aos algoritmos apresentados foram desenvolvidos em ambiente MATLAB®, Versão 5.3. Para a execução dos estudos de caso utilizou-se um computador com as características contidas na Tabela 3.7. A utilização de conjuntos de dados simulados foi adotada em função das prerrogativas propostas por BAUMANN (2003), onde afirma que a avaliação de estratégias

de seleção de variáveis é melhor conduzida nesses casos simulados, uma vez que a solução exata é conhecida *a priori*.

Tabela 3.7: Características principais do computador utilizado na geração e avaliação de resultados

Processador	Intel Pentium 4 (Northwood)
Velocidade	1.6 GHz
Memória RAM Disponível	192 Mb (128 + 64)
Tipo de Memória	SDRAM

Os critérios utilizados para comparar as diferentes estratégias propostas foram a capacidade de selecionar as variáveis verdadeiras, o tempo computacional utilizado e o desempenho dos modelos em relação a sua capacidade preditiva, medido através da $PRESS_{Int}$ que é o erro médio quadrático medido no conjunto de validação criado pelo técnica de segregação de dados de cada alternativa.

Para os casos Linear 02 e Não-linear foi ainda calculado o valor da PRESS em um conjunto independente de dados com o objetivo de explorar a capacidade preditiva do modelo em amostras externas de dados. Esse procedimento não foi realizado para o Caso Linear 01, pois os dados foram extraídos da literatura (FINKLER, 2003), os quais foram apresentados em forma de tabela, não sendo fornecidas as equações utilizadas na geração dos mesmos.

O parâmetro de performance adotados para a seleção de variáveis foi a $PRESS_{Int}$, que nada mais é que a PRESS calculada no conjunto de validação gerado pela técnica de segregação de dados de cada alternativa, sendo que o modelo que apresentar o menor valor desse índice é considerado o melhor. Nos casos onde se utilizou a geração aleatória dos subconjuntos de dados, identificados pela extensão _01 na Tabela 3.6, em cada etapa do algoritmo de seleção de variáveis foram realizadas 100 computações e o valor médio da $PRESS_{Int}$ resultante dessas repetições foi considerado com característico do modelo.

Como pode ser observado na Tabela 3.6, são utilizadas duas populares metodologias de seleção de variáveis, a Seleção por Adição e o Algoritmo Genético. A escolha de uma estratégia da Seleção por Adição como representante das classes seqüenciais de seleção de variáveis é fundamentada no fato de se possuir o conhecimento prévio de que o número de variáveis verdadeiras é inferior ao número de variáveis candidatas.

Enquanto que a implementação do método FS seguiu um algoritmo clássico, no desenvolvimento do GA foi inserida uma modificação representada pela limitação do número máximo de variáveis a serem selecionadas.

Essa limitação objetiva forçar o mecanismo de busca a selecionar a melhor combinação possível com um número fixo de variáveis. Os parâmetros do algoritmo genético utilizados nesse trabalho são apresentados na Tabela 3.8.

Tabela 3.8: Parâmetros adicionais necessários para o Algoritmo Gerado

Parâmetro	Valor Adotado	Parâmetro	Valor Adotado
População inicial	100	Taxa de Mutação	20%
Número de Genes	Variáveis candidatas	Taxa de indivíduos na elite	40%
Número de Geração	40	Reinicializazões a cada	5 gerações
Taxa de Reprodução	80%	Indivíduos preservados	10

Caso Linear 01

Esse caso foi o exemplo utilizado por FINKLER (2003) para comparar o desempenho do algoritmo SRMP frente ao algoritmo SROV. O conjunto de dados foi construído a partir da geração de dois vetores independentes t_1 e t_2. Esses vetores são constituídos por valores normalmente distribuídos entre 0 e 1. A variável de resposta y foi obtida através da soma desses dois vetores.

As variáveis x_1, x_2, x_3 e x_4 foram criadas através de combinações lineares dos vetores t_1 e t_2, sendo essas as variáveis que devem compor o subconjunto verdadeiro de variáveis que deverá ser selecionada pelos algoritmos. A matriz X que contém as variáveis candidatas foi obtida pela inclusão dos vetores de x_1, x_2, x_3 e x_4 e por mais 6 vetores, x_5, x_6, x_7, x_8, x_9 e x_{10}, constituídos por valores aleatoriamente distribuídos entre 0 e 1. A esses dados foi adicionado um ruído com magnitude de 0.1%. Os dados utilizados são apresentados na Tabela 3.9.

Tabela 3.9: Dados referentes ao Estudo de Caso 01

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1.260	0.394	0.651	0.481	0.275	0.067	0.235	0.179	0.701	0.690	1.634
0.714	0.263	0.393	0.288	0.439	0.676	0.369	0.627	0.680	0.499	0.911
0.910	0.442	0.562	0.409	0.406	0.274	0.807	0.167	0.985	0.553	1.106
0.796	0.274	0.427	0.314	0.227	0.402	0.138	0.638	0.579	0.595	1.023
1.190	0.441	0.655	0.481	0.351	0.670	0.061	0.962	0.108	0.506	1.513
0.764	0.457	0.520	0.377	0.158	0.816	0.186	0.803	0.873	0.627	0.889
0.483	0.193	0.274	0.201	0.832	0.811	0.679	0.717	0.982	0.676	0.606
0.537	0.313	0.361	0.262	0.269	0.074	0.989	0.308	0.857	0.299	0.627
1.387	0.544	0.781	0.573	0.065	0.888	0.466	0.691	0.367	0.855	1.745
0.587	0.237	0.334	0.245	0.625	0.514	0.574	0.862	0.240	0.053	0.737
0.735	0.127	0.321	0.240	0.564	0.586	0.254	0.210	0.763	0.002	1.005
0.691	0.242	0.372	0.275	0.877	0.917	0.515	0.569	0.414	0.444	0.887
1.147	0.526	0.689	0.504	0.275	0.084	0.962	0.903	0.622	0.499	1.406
0.739	0.200	0.364	0.270	0.747	0.949	0.476	0.913	0.676	0.446	0.972
0.531	0.220	0.305	0.224	0.900	0.427	0.647	0.493	0.714	0.101	0.662
0.977	0.380	0.549	0.403	0.802	0.179	0.030	0.623	0.352	0.610	1.234
0.403	0.049	0.165	0.125	0.247	0.991	0.005	0.530	0.491	0.951	0.559
0.551	0.110	0.249	0.186	0.337	0.539	0.914	0.280	0.979	0.336	0.746
1.217	0.478	0.686	0.503	0.833	0.862	0.744	0.670	0.142	0.271	1.533
0.280	0.039	0.116	0.088	0.468	0.055	0.681	0.694	0.017	0.023	0.385
0.986	0.490	0.615	0.448	0.632	0.817	0.640	0.151	0.694	0.497	1.194
0.538	0.312	0.361	0.262	0.315	0.886	0.668	0.036	0.014	0.560	0.630
0.902	0.189	0.415	0.308	0.296	0.792	0.330	0.136	0.293	0.217	1.216
0.646	0.151	0.304	0.227	0.070	0.937	0.912	0.708	0.196	0.853	0.861
0.724	0.322	0.429	0.314	0.273	0.907	0.091	0.020	0.235	0.373	0.895

No caso de modelos PLS é necessária a determinação do número de variáveis latentes a serem utilizadas, como na geração do conjunto verdadeiro de variáveis foram utilizados 2 vetores independentes (***t₁*** e ***t₂***), esse é o número de variáveis latentes que dever ser considerado. O parâmetro referente ao número máximo de variáveis a serem selecionada para as alternativas que utilizam algoritmo genético adotado foi igual a 4.

Caso Linear 02

Esse caso é similar ao Caso Linear 01, só que ao invés de se utilizar somente dois vetores independentes foram utilizados 4 – ***t₁***, ***t₂***, ***t₃*** e ***t₄*** – sendo que esses eram constituídos por 50 amostras distribuídas aleatoriamente entre –1 e 1. O conjunto verdadeiro de variáveis explicativas foi gerado através da multiplicação desses vetores, agrupados em uma matriz ***T*** por uma matriz β, apresentada na Eq. 3.13, que foi gerada de forma aleatória.

$$\beta = \begin{bmatrix} -7 & 2 & 5 & 10 & -4 & -9 \\ -3 & 4 & 1 & -5 & 9 & -1 \\ -7 & 2 & -6 & 9 & -5 & -6 \\ 10 & 9 & -8 & 10 & 1 & 3 \end{bmatrix} \tag{3.13}$$

$$X_B = \begin{bmatrix} t_1 & t_2 & t_3 & t_4 \end{bmatrix} \cdot \beta \tag{3.14}$$

Dessa forma, o conjunto verdadeiro de variáveis é constituído por 6 vetores, referidos como ***x₁***, ***x₂***, ***x₃***, ***x₄***, ***x₅*** e ***x₆***. Essas variáveis são armazenadas em uma matriz $\boldsymbol{X_B}$. A variável de resposta *y* foi gerada através da multiplicação da matriz $\boldsymbol{X_B}$, por um vetor ***b***, apresentado na Eq. 3.15.

$$b = \begin{bmatrix} -3 & 3 & -5 & 2 & -8 & 4 \end{bmatrix} \tag{3.15}$$

$$y = X_B \cdot b^T \tag{3.16}$$

Por fim, a matriz de variáveis candidatas é construída pela adição de 4 vetores – ***x₇***, ***x₈***, ***x₉*** e ***x₁₀*** – formados por valores uniformemente distribuídos entre –1 e 1, a matriz $\boldsymbol{X_B}$. Os dados são escalonados e, com a finalidade de simular ruídos de processo, às variáveis de entrada é adicionado ruído com desvio padrão igual a 0.05.

O número de variáveis latentes adotado para esse caso foi 4 e o número máximo de variáveis a ser considerado nas alternativas que fazem uso de algoritmos genéticos foi igual a 6.

Caso Não-linear 01

Esse estudo de caso tem por objetivo avaliar a aplicabilidade das alternativas desenvolvidas em selecionar as variáveis quando a variável de resposta apresenta uma relação não-linear com as variáveis explicativas.

Inicialmente é gerada uma matriz composta por 10 vetores contendo 100 amostras aleatórias uniformemente distribuídas com um desvio padrão igual a 1. Esses vetores são armazenados em uma matriz ***X***. Adicionalmente criou-se uma matriz de ruído com 10 vetores e 100 amostras uniformemente distribuídas com desvio padrão igual a 0.25. A variável de resposta ***y*** é obtida através da Eq. 3.17.

$$y = \frac{\exp(x_5)}{(\exp(x_7) + eps)} \cdot \text{sen}(x_{10}) \cdot \cos(x_4)^2 \tag{3.17}$$

Para o procedimento de seleção de variáveis, somou-se a matriz ***X*** com a matriz de ruídos e esses sinais foram então utilizados na geração do sinal de saída e constituem-se no conjunto de variáveis que terão que ser selecionados. Como mencionado anteriormente, durante a obtenção de um modelo do tipo PLS é necessário também se determinar o número de variáveis latentes.

A determinação do número ótimo de variáveis latentes foi realizada através da análise da variância explicada cumulativa (VEC) da variável de saída em função do número de variáveis latentes utilizadas. O procedimento é realizado através da criação de um modelo PLS com o número de variáveis latentes iguais ao número de variáveis candidatas. Analisa-se então a contribuição de cada variável latente no incremento da VEC. O número ótimo de variáveis latentes é selecionado verificando-se quando a adição de uma nova variável latente não acarreta melhora significativa nesse parâmetro. Esse procedimento foi utilizado por FINKLER (2003) em todos estudos de caso executados.

Tabela 3.10: Variância explica cumulativa para as variáveis de entrada e saída em função do número de variáveis latentes consideradas

Número de Variáveis Latentes	VEC Entradas (%)	VEC Saída (%)
1	11.86	54.30
2	22.80	65.97
3	31.76	68.95
4	42.70	72.08
5	50.45	73.95
6	60.68	75.32
7	70.01	75.73
8	79.62	76.24
9	89.72	76.45
10	100.00	76.45

Analisando-se a Tabela 3.10, nota-se que a partir da sexta variável latente não há aumento significativo na variância explicada pelo modelo, logo para os procedimentos de seleção de variáveis para esse caso foram utilizadas 6 variáveis latentes.

3.4.3 Metodologia

Cada uma das alternativas propostas na Tabela 3.6 utiliza uma seqüência de operações na sua execução. A seguir é realizada de forma breve uma descrição das peculiaridades apresentada por cada uma delas.

As metodologias baseadas na geração aleatória de subconjuntos de validação/calibração (FS_01 e GA_01) são avaliadas em função da PRESS média de 100 computações, ou seja, em cada etapa do método seqüencial são avaliados 100 modelos para cada uma das combinações das variáveis candidatas a integrarem os modelos. Da mesma forma, para os algoritmos genéticos, para cada gene gerado, a PRESS associada á ele é a média de 100 conjuntos de calibração e validação avaliados.

As alternativas baseadas na seleção de dados através do método *y-Rank* realiza o ordenamento e seleção dos dados para o conjunto de calibração e validação apenas uma vez para cada etapa do algoritmo seqüencial e uma avaliação para gene gerado no algoritmo genético.

Para as alternativas baseadas nos métodos *Kennard-Stone* o procedimento de segregação de dados nos conjuntos de calibração e validação também é realizada uma única vez em cada etapa dos algoritmos seqüenciais e uma única vez por gene na utilização dos algoritmos genéticos. Para os casos que utilizam a metodologia de *Kennard-Stone* Modificado foi utilizada apenas as distâncias entre as saídas ($w_y = 1$ e $w_x = 0$) para comparar a sua eficiência com o método *y-Rank* que também só considera as variáveis de resposta na segregação dos dados.

3.4.4 Resultados e Análises

Os resultados obtidos através das alternativas propostas na Tabela 3.6 serão aqui descritos e analisados. Inicialmente serão comparados os resultados obtidos entre as alternativas que utilizam a mesma estratégia de seleção de variáveis, ou seja, serão comparadas os resultados obtidos pelas técnicas Seleção por Adição FS_01, FS_02, FS_03 e FS_04 entre si e os obtidos pelas técnicas de Algoritmos Genéticos GA_01, GA_02, GA_03 e GA_04. Num segundo momento será apresentada uma comparação entre as duas classes de técnicas de seleção de variáveis entre si – Seleção por Adição e Algoritmos Genéticos.

Propostas baseadas em método seqüencial

Os resultados obtidos pelas diferentes variações da metodologia FS para o primeiro caso linear são apresentados na Tabela 3.11. Os resultados contidos nessa tabela apresentam a alternativa obtida através da utilização do método de segregação de dados baseado no método *y-Rank* como sendo o único capaz de selecionar o verdadeiro conjunto de variáveis. Os métodos baseados nos métodos *Kennard-Stone*, tanto o clássico (FS_03), quanto o modificado (FS_04) apresentaram os piores resultados, não sendo capazes de selecionar as variáveis verdadeiras.

O método tradicional, ou seja, a utilização de segregação aleatória de dados (FS_01) foi capaz de selecionar duas variáveis verdadeiras, apresentando uma capacidade preditiva adequada,

mas o tempo computacional requerido por ele foi muito superior ao utilizado pelo FS_02, sendo que esse último foi o que obteve a melhor capacidade preditiva entre as alternativas estudadas.

Tabela 3.11: Resultados obtidos pelas alternativas *Stepwise* para o primeiro caso

	Oráculo	FS_01	FS_02	FS_03	FS_04
Tempo (s)		7	0.14	1.2	1.4
Variáveis Selecionadas	[1 2 3 4]	[1 4]	[1 2 3 4]	[5 6]	[7]
$PRESS_{Int}$	1.6E-6	3E-6	1.6E-6	1.7E0	1.6E0

Os resultados obtidos pelas alternativas de Seleção por Adição para o segundo caso – Caso Linear 02 – são apresentados na Tabela 3.12.

Tabela 3.12: Resultados obtidos pelas alternativas *Stepwise* para o segundo caso

	Oráculo	FS_01	FS_02	FS_03	FS_04
Tempo (s)		21	0.2	5	20
Variáveis Selecionadas	[1 2 3 4 5 6]	[1 2 3 4 5 6]	[1 2 3 4 5]	[10]	[1 8 7]
$PRESS_{Int}$	2.86E-01	2.86E-01	2.89E-01	1.06E+00	3.95E-01
$PRESS_{Ext}$	4.85E-01	4.85E-01	5.04E-01	1.40E+00	7.12E-01

A coluna “Oráculo” presente nas tabelas de resultados se referem-se ao modelo obtido com o verdadeiro conjunto de variáveis.

Os resultados contidos na Tabela 3.12 indicam que o algoritmo FS_02 novamente apresentou bons resultados, sendo que apesar de não ter selecionado todas as verdadeiras variáveis, ele não incluiu nenhuma variável espúria, tendo como efeito um bom poder preditivo. Outra característica interessante da alternativa FS_02 é o baixo tempo requerido, sendo no mínimo 25 vezes mais rápido que a segunda melhor alternativa em termos de velocidade – FS_03.

Novamente as metodologias derivadas dos métodos *Kennard-Stone* (FS_03 e FS_04) apresentaram os piores resultados em termos de seleção de variáveis, tendo efeito diretamente na capacidade preditiva do modelo gerado. A alternativa clássica de seleção de variáveis (FS_01) apesar de ter sido a única a ter conseguido selecionar todo o verdadeiro conjunto de variáveis, essa alternativa demandou mais de 100 vezes o tempo requerido pela técnica FS_02.

Para o caso não-linear deve-se antes fazer uma ressalva. Os resultados serão uma combinação entre a capacidade de o modelo não-linear escolhido (QPLS) em representar os dados e da capacidade das alternativas de identificar variáveis que possibilitem ao modelo alcançar um ajuste adequado.

Os resultados para o caso não-linear são apresentados na Tabela 3.13. Novamente o Oráculo representa o modelo construído com todas as verdadeiras variáveis. Nota-se novamente a incapacidade dos modelos baseados no método *Kennard-Stone* de identificar as variáveis corretas. A alternativa FS_03 selecionou somente uma variável, a qual não faz parte do

conjunto das verdadeiras, o FS_04 selecionou três variáveis, sendo que apenas uma faz parte do conjunto verdadeiro, mas em contra partida demandou o mais alto esforço computacional entre as quatro alternativas aqui comparadas.

A seleção com segregação aleatória apresentou, para esse caso, resultados semelhantes as apresentadas pela FS_03, selecionando apenas uma variável, a qual também não faz parte do conjunto verdadeiro. Novamente a alternativa FS_02 se mostrou a mais eficiente entre as avaliadas. Essa técnica foi capaz de identificar três variáveis verdadeiras e adicionou ao conjunto apenas uma variável espúria, o que acabou por produzir um modelo com boa capacidade preditiva.

Fato interessante é que o modelo produzido pela técnica FS_02 foi capaz de produzir valor de $PRESS_{Ext}$ inferior ao modelo Oráculo. Esse fato pode estar relacionado ao grau de ajuste do modelo Oráculo, ou seja, na etapa de calibração os parâmetros foram determinados de modo a representar os dados, porém em um conjunto externo de amostras o modelo produziu resultados piores.

Tabela 3.13: Resultados obtidos pelas alternativas *Stepwise* para o terceiro caso

	Oráculo	FS_01	FS_02	FS_03	FS_04
Tempo (s)		21	4.6	49	171
Variáveis Selecionadas	[4 5 7 10]	[3]	[4 7 8 10]	[6]	[2 4 5]
$PRESS_{Int}$	8.71E+00	1.46E+01	1.16E+01	1.47E+01	1.18E+01
$PRESS_{Ext}$	3.13E+00	3.75E+00	2.48E+00	3.52E+00	5.38E+00

Propostas baseadas em Algoritmos Genéticos

Os resultados obtidos pela combinação das diferentes alternativas de segregação de dados com a técnica de seleção baseada em algoritmos genéticos são analisados nessa seção. Similarmente ao que foi realizado para as alternativas baseados na técnica *Stepwise*, serão analisados os resultados para cada um dos casos estudados.

Os resultados obtidos para o Caso Linear 01 são apresentados na Tabela 3.14. Para esse caso, as alternativas GA_02, GA_03 e GA_04 apresentaram resultados similares em termos de capacidade preditiva. As três alternativas selecionaram somente variáveis pertencentes ao grupo das verdadeiras, no entanto a alternativa GA_02 produziu um modelo mais enxuto, somente duas variáveis, enquanto que as outras duas produziram modelos com três variáveis.

Analisando-se o esforço computacional, a técnica GA_02 foi a mais eficiente, sendo 10 vezes mais rápida que a GA_03 e 20 vezes mais rápida que a GA_04. A combinação de geração aleatória de subconjuntos de calibração e validação em combinação com algoritmos genéticos – GA_01 – foi a que produziu os piores resultados para esse caso, consumindo o maior tempo, selecionando apenas duas variáveis, sendo que uma delas não pertence ao grupo das verdadeiras. O resultado disso pode ser visto no valor da $PRESS_{Int}$ resultante, três ordens de grandeza superiores aos demais.

Tabela 3.14: Resultados obtidos pelas alternativas GA para o primeiro caso

	Oráculo	GA_01	GA_02	GA_03	GA_04
Tempo (s)		983	11	145	244
Variáveis Selecionadas	[1 2 3 4]	[1 6]	[1 4]	[1 2 3]	[1 2 4]
$PRESS_{Int}$	1.6E-6	1.5E-2	1.7E-5	2.8E-5	1.7E-5

Tabela 3.15: Resultados obtidos pelas alternativas GA para o segundo caso

	Oráculo	GA_01	GA_02	GA_03	GA_04
Tempo (s)		1165	13	1004	1780
Variáveis Selecionadas	[1 2 3 4 5 6]	[3 5 6 10]	[1 3 5 6]	[2 3 4 6 7 10]	[1 3 5 6 8]
$PRESS_{Int}$	1.97E-03	4.33E-01	1.97E-03	1.52E-02	2.21E-03
$PRESS_{Ext}$	9.27E-03	3.04E-01	1.18E-02	6.95E-02	1.29E-02

Analisando-se a Tabela 3.15, a qual contém os resultados obtidos para o segundo caso linear utilizando-se as alternativas baseadas em algoritmos genéticos, pode-se ver que para esse caso a alternativa GA_02 foi a única capaz de selec ionar somente variáveis verdadeiras. Os demais adicionaram variáveis espúrias, o que acabou por refletir em um decréscimo na capacidade preditiva do modelo.

Do ponto de vista computacional, novamente a alternativa GA_02 se mostrou a mais eficiente, sendo quase 100 vezes mais rápida que a segunda mais veloz. Comparando-se as demais técnicas, vê-se que a distribuição aleatória dos dados – GA_01 – selecionou três variáveis verdadeiras e uma espúria, o GA_04 foi capaz de selecionar 4 variáveis verdadeiras e também uma espúria. O GA_03 selecionou 6 variáveis, sendo que 4 fazem parte do conjunto verdadeiro e duas são espúrias.

Para o caso não-linear as mesmas ressalvas feitas para a metodologia *Stepwise* são válidas. Os resultados produzidos por cada uma das alternativas apresentados na Tabela 3.16.

Tabela 3.16: Resultados obtidos pelas alternativas GA para o terceiro caso

	Oráculo	GA_01	GA_02	GA_03	GA_04
Tempo (s)		8.6244E+04	8.86E+02	1.2915E+04	2.2121E+04
Variáveis Selecionadas	[4 5 7 10]	[4 5 8 10]	[4 7 8 10]	[3 5 910]	[1 9]
$PRESS_{Int}$	9.90E+01	1.07E+02	1.35E+02	9.78E+01	1.55E+02
$PRESS_{Ext}$	5.69E+01	3.76E+01	2.35E+01	7.01E+01	1.11E+02

No caso não-linear, novamente tem-se a alternativa GA_02 como a que apresentou melhor performance em relação ao tempo computacional e capacidade preditiva. Fato interessante é que os modelos gerados através das alternativas GA_01 e GA_02 produziram modelos com capacidade preditiva superior ao próprio Oráculo. Isso pode ser resultado de o modelo Oráculo apresentar um elevado grau de sobreajuste, sendo essa observação embasada pelo valor de $PRESS_{Int}$, uma vez que os modelos que apresentaram os mais baixos valores desse parâmetros, resultaram nos modelos com pior capacidade preditiva.

Em termos de seleção do conjunto correto de variáveis, as alternativas GA_01 e GA_02 foram as que apresentaram a melhor performance, selecionando cada uma 4 variáveis, sendo três verdadeiras e uma espúria. A grande diferença entre essas duas alternativas está no esforço computacional, sendo a GA_02 cerca de 100 vezes mais eficiente.

Os algoritmos fundamentados na técnica *Kennard-Stone* (GA_03 e GA_04) foram os que apresentaram os piores resultados, em termos de acuracidade, aqui referente a capacidade de identificar as variáveis corretas, capacidade preditiva e esforço computacional.

Outro fato que merece ressalva é que o conjunto selecionado pelos métodos baseados na técnica *y-Rank* foram idênticos para as duas alternativas testadas, algoritmos genéticos (GA_02) e *Forward Selection* (FS_02), o que, de alguma forma, aponta para coerência da alternativa, pois foi capaz de identificar as mesmas variáveis com duas abordagens completamente distintas.

Discussão comparativa entre Algoritmos Genéticos e Métodos Seqüenciais

Na literatura encontra-se argumentações de que os métodos derivados da metodologia *Stepwise*, ou seqüenciais, tendem a selecionar um número menor de variáveis, pois tendem a ficar presos no primeiro ótimo local da função objetivo escolhida como parâmetro de performance (BAUMANN, 2003).

Métodos baseados em algoritmos de busca aleatória não sofrem esse problema, uma vez que são capazes de explorar uma região muito mais ampla. De fato HAN e YANG (2004) enunciam que apesar de os métodos baseados em algoritmos genéticos não garantirem a convergência para o ótimo global, leia-se aqui o verdadeiro, ou melhor, subconjunto de variáveis, a probabilidade da resposta produzida pelo mesmo estar próxima desse ponto é alta.

Outra desvantagem dos métodos seqüenciais, *Stepwise*, é o fato de não poderem avaliar a correlação existente entre as variáveis, sendo essas avaliadas independentemente (HAN e YANG, 2004). Como exemplo, as variáveis x_i e x_j podem não ser importantes separadamente, mas quando conjugadas são capazes de fornecer informação útil ao modelo.

Os resultados produzidos no decorrer desse estudo não foram conclusivos sobre os aspectos das dimensões dos modelos obtidos nos diferentes casos, isso é decorrência direta da existência de variáveis puramente aleatórias, não possuindo nenhuma informação útil para a construção do modelo. Conclusões sobre as dimensões dos modelos gerados por uma ou outra técnica devem ser obtidas a partir de um conjunto muito maior de variáveis candidatas e que possuam, em menor ou maior grau, informação para o modelo.

Em termos de esforço computacional é óbvio que os métodos da família *Stepwise* são mais econômicos, uma vez que o número de modelos testados é muito inferior que os desenvolvidos através dos algoritmos genéticos.

Nesse ponto é válido ressaltar o cuidado necessário na determinação dos parâmetros utilizados nos algoritmos genéticos. Uma escolha de inadequada do número de gerações,

número de indivíduos iniciais e taxas de reprodução, mutação e clonagem podem resultar em um número de avaliações superior ao resultante do método de busca exaustiva.

De fato, esse efeito foi verificado na estruturação dos resultados fornecidos pelas alternativas que utilizam o algoritmo genético nos estudos de caso aqui apresentados. Uma análise do número de modelos gerados nas alternativas que utilizam GA resultou em um número maior a todas as combinações de variáveis possíveis, que para os casos estudados seria de 1024. Porém, apresar dessa má parametrização dos algoritmos genéticos, pode-se concluir que a utilização da estratégia de criação de conjuntos de calibração e validação baseadas na técnica *y-Rank* foi a que resultou, em geral, em modelos com capacidade preditiva superior.

Em relação as combinações propostas na Tabela 3.6, pode-se considerar que a utilização da estratégia de geração de conjuntos de calibração e validação baseadas na técnica *y*-Rank (FS_02 e GA_02) foram as que, em geral, apresentaram o melhor compromisso entre acuracidade, esforço computacional e capacidade preditiva, sendo essas duas consideradas as melhores alternativas entre as 8 aqui testadas e comparadas.

Apesar de as metodologias FS_02 e GA_02 terem conseguido na maioria dos casos estudados determinar as variáveis verdadeiras para cada modelo, é válido sempre considerar a observação feita por BAUMANN (2003), de que os algoritmos de busca não são capazes de diferenciar correlação causal da casual, sendo que eles buscam simplesmente minimizar ou maximizar a função objetivo.

Por conta disso, as variáveis pré-selecionadas devem ser consideradas somente como uma sugestão preliminar, sendo a decisão final de quais variáveis devem ser utilizados ficar a cargo de um especialista no sistema que está sendo modelado.

Capítulo 4

Analisadores Virtuais para Colunas de Destilação

O controle de composição das correntes que deixam uma coluna de destilação procura garantir a qualidade dos produtos efluentes. Classicamente se utilizam duas metodologias de controle de composição em uma coluna, o controle através de temperaturas e o controle diretamente da composição através de analisadores.

Nesse capítulo, inicialmente, será apresentado brevemente as duas estratégias citadas anteriormente, enunciando suas estruturas, características dinâmicas, vantagens e desvantagens. Essa parte inicial está fundamentada no livro “Distillation Operation” de KISTER (1989).

Posteriormente será apresentada a aplicação de técnicas PLS para a elaboração de analisadores virtuais para colunas de destilação, bem como um estudo de caso baseado em simulações de uma unidade depropanizadora, comparando-se as diferentes estratégias testadas durante a realização desse trabalho.

4.1 Controle de Composição

O objetivo do controle de composição vai além de garantir a pureza adequada do produto. A malha de controle de composição manipula uma corrente como o refluxo, o refervimento ou o destilado. (KISTER, 1989). Um controle instável de composição acaba por introduzir perturbações em todo o sistema.

Existem duas principais metodologias para o controle de composição, a primeira utiliza a forte correlação existente entre a temperatura e a composição em cada estágio de separação e é denominada de controle de temperatura. A segunda utiliza diretamente resultados oriundos de analisadores para efetuar o controle da unidade.

4.1.1 Controle de Temperatura

KISTER (1989) argumenta que o controle de temperatura talvez seja a maneira mais popular para controlar a composição dos produtos em uma coluna de destilação. O controle de temperatura substitui o controle de composição, utilizando a correlação existente entre a temperatura e a composição.

Uma variação na temperatura de controle está associada a uma variação correspondente na concentração do componente chave na corrente do produto. Assim, por exemplo, uma elevação nas temperaturas na seção de topo da coluna representa a elevação na composição do componente chave pesado nessa seção.

O controle de temperatura é largamente utilizado em função de utilizar meios fáceis e baratos para o controle de composição. A instrumentação utilizada possui alta confiabilidade, baixa necessidade de manutenção e que apresenta um atraso dinâmico muito pequeno, quando comparado a analisadores de composição.

As principais desvantagens do controle de temperatura, segundo KISTER (1989), é de que a temperatura de controle pode não se correlacionar bem com a composição do produto ou ela pode apresentar baixa sensibilidade em relação a variações na composição do produto.

Os critérios da correta seleção da temperatura de controle, bem como da estratégia a ser utilizada, isso é controle através de temperaturas ou diferenças de temperaturas, bem como o efeito de outros distúrbios são apresentados de forma bastante didática em KISTER (1989).

4.1.2 Controle através de Analisadores

A grande vantagem dos analisadores de composição é o fato de medirem diretamente a composição nas correntes. O meio mais usual de análise em linha é a cromatografia gasosa.

Essa estratégia, no entanto, sofre das desvantagens de utilizar instrumentação cujo custo de aquisição, instalação e manutenção são geralmente elevados. Além dos aspectos financeiros desfavoráveis, o controle de composição através de analisadores sofre de um grande atraso dinâmico, a instrumentação demanda elevados tempo e custo de manutenção e pode sofrer interferência de contaminantes presentes na corrente de produto.

O atraso dinâmico de um sistema desses é resultado da soma de diversos fatores, sendo principalmente representados por atrasos de processo, geralmente decorrentes do tempo de residência no vaso de refluxo, atrasos devido a transferência das amostras do ponto de coleta até os analisadores, esse se mostra particularmente grande quando o analisador é posicionado em uma área externa a coluna em função de questões de segurança, e o tempo de transferência da amostra da entrada do analisador até o sensor. Além desses, se o analisador for compartilhado com outras correntes, deve-se considerar o tempo de ciclo das análises.

Esse atraso, segundo KISTER (1989), é normalmente da ordem de 10 a 20 minutos, no entanto não é incomum encontrar tempos de 30 minutos ou maiores. Esse atraso dinâmico

acaba por se refletir como um tempo morto no sistema de controle, já que é necessário aguardar o processamento de uma nova amostra para executar uma nova ação de controle. Além disso, a resposta do analisador geralmente é suavizada para prevenir ações de controle bruscas ou resultado de análises espúrias. Essa suavização impõe mais um atraso ao sistema.

Os analisadores em linha são muito mais dispendiosos que os medidores de temperatura. Sua utilização fica restrita a sistema de larga escala, onde mesmo uma pequena melhoria representa um grande retorno ou em sistemas onde há uma grande diferença entre os valores dos produtos.

Colunas de alta pureza, que realizam separações de produtos similares ou que apresentam problemas na implementação de um controle de temperatura são as candidatas a utilizar o controle de composição. Cabe ainda salientar, que em alguns casos, tais como a separação de propano de propeno, a diferença de temperatura entre topo e fundo é tão baixa a ponto de não ter a sensibilidade necessária para acompanhar a composição através do controle de temperatura. Nestes casos, os analisadores são fundamentais para permitir o acompanhamento da qualidade do produto final.

4.1.3 Analisadores Virtuais para Colunas de Destilação

As metodologias de controle de composição através do controle de temperatura utilizam apenas uma temperatura, ou uma diferença de temperaturas, para controlar a composição de cada componente chave que se deseja monitorar.

A limitação a apenas uma temperatura está relacionada ao comportamento de um perfil de temperaturas ao longo de uma coluna. Em geral, as temperaturas em uma coluna de destilação são variáveis altamente correlacionadas, ou seja, frente a uma mudança qualquer, por exemplo, aumento na vazão de destilado, mantendo-se todas as demais variáveis constantes, causará um aumento na temperatura de todos os pratos da coluna (KRESTA *et al*., 1993).

O desenvolvimento de modelos com múltiplas temperaturas através de regressão multivariáveis simples, como MLR, pode levar a erros numéricos e instabilidades, tornando o modelo inutilizável para fins práticos.

Como apresentado no capítulo 2, técnicas de redução de dimensionalidade como PCR/PCA e PLS são ferramentas estatísticas ótimas para operarem com dados correlacionados, o que aponta para uma possibilidade de serem utilizadas para a estimação de propriedades em processos químicos, os quais geralmente apresentam variáveis fortemente correlacionadas.

KRESTA *et al.* (1993) enuncia que as primeiras aplicações da metodologia PLS para a construção de estimadores com finalidade de controle foram desenvolvidas por MEJDELL e SKOGESTAD (1990) e KRESTA *et al*. (1990 a,b).

Nesses estudos preliminares foram utilizados modelos PLS e PCR estacionários, ou seja, sem incorporar nos modelos efeitos dinâmicos. KANO *et al*. (2000) enunciaram que os resultados obtidos pelos autores apontaram para uma boa eficiência dos modelos desenvolvidos, sendo esses quase tão bons quantos os obtidos através da técnica de Filtro de Kalman dinâmico.

O sucesso dos analisadores baseados em PCR/PCA e PLS estacionários está vinculado a natureza preditiva dos mesmos, ou seja, a associação direta das temperaturas dos pratos em um instante *k* com a composição do componente chave nesse mesmo instante.

KANO *et al.* (2000), baseados nos resultados obtidos por MEJDELL e SKOGESTAD (1990, 1993), concluíram que a utilização de outras variáveis de processo diferentes das temperaturas, tais como vazão de refluxo e carga térmica no refervedor, necessitam da inserção de características dinâmicas aos modelos, pois como essas variáveis geralmente são as manipuladas em sistemas de controle, elas não podem afetar os valores de composição sem um atraso.

A incorporação de dinâmica em técnicas PCA, PCR e PLS é feita considerando como variáveis de entrada não somente as medidas no instante atual, mas também medidas passadas, dessa forma aumentando as dimensões da matriz de entradas $\boldsymbol{X}$.

A incorporação de dados passados resulta em uma matriz com colunas fortemente correlacionas, mas isso não vem a ser um problema para as técnicas com redução de dimensionalidade, pois elas são capazes de lidar facilmente com essa estrutura de dados.

A utilização de técnicas PLS para o desenvolvimento de analisadores virtuais tem sido largamente explorada na literatura, pode-se citar como exemplo KOMULAIMEN *et al.* (2004), CHEN e LIU (2001), KANO *et al.* (2000), KRESTA *at al.* (1994), entre outros.

4.2 Estudo de Caso – Torre Depropanizadora

A utilização de simulações estacionárias e/ou dinâmicas para a realização de um estudo comparativo entre diferentes metodologias tem como principal motivação o fato de se ter controle total sobre o sistema e, acima de tudo, sobre as análises, uma vez que todos os distúrbios são conhecidos, o que não acontece em dados provenientes de plantas industriais.

Para a comparação das diversas implementações realizadas durante o desenvolvimento desse trabalho será utilizada a simulação dinâmica de uma Torre Depropanizadora de uma central de matérias primas de um pólo petroquímico, realizadas no *software* Aspen Dynamics® Versão 12.1.

Inicialmente será apresentada uma breve descrição do sistema simulado, caracterizando a unidade quanto ao número de alimentações, configuração da torre e estruturas de controle presentes na mesma. Posteriormente será apresentada a metodologia

utilizada para a criação de dados que representem o comportamento típico de uma unidade industrial com características semelhantes ao sistema sob estudo.

Os conjuntos de dados gerados serão discutidos quanto as perturbações utilizadas na sua concepção e o efeito das mesmas sobre a concentração do principal componente da corrente de topo da torre, o 1,3 Butadieno. Todos os dados apresentados estarão em sua forma escalonada, visto que a simulação estacionária utilizada para a criação da simulação dinâmica contém dados operacionais da unidade industrial real, os quais não podem ser apresentados em função de acordo de sigilo.

4.2.1 Descrição do Processo

A torre depropanizadora é uma coluna de destilação multicomponente cujo objetivo é produzir uma corrente rica no corte C3, ou seja, uma corrente constituída principalmente por compostos alifáticos com três carbonos em sua cadeia. Essa corrente sofrerá posterior processamento para produzir o propeno e o propano utilizado nas indústrias de segunda geração.

A torre simulada é composta por 42 estágios ideais de separação, incluindo o condensador total e o refervedor, a alimentação da torre é composta por três correntes oriundas do fundo de outras colunas de destilação. Essas correntes são distribuídas através de duas alimentações independentes. A primeira corrente é composta pela mistura da corrente de fundo de uma torre deetanizadora e o fundo de uma torre de reciclo de corte C3. Por essas duas correntes serem constituídas majoritariamente por compostos leves, elas são misturadas e alimentadas no 20° prato. A numeração dos pratos se inicia pelo topo, sendo que o prato de topo recebe a numeração 1, e segue em ordem crescente até o prato de fundo, prato 40.

A corrente restante, originada do fundo de uma torre de retificação de GLP, é composta por uma mistura com peso molecular médio maior e é introduzida sobre o 32° prato da depropanizadora. Em termos quantitativos, a carga é composta principalmente pelas correntes de fundo da torre de retificação de GLP e da Deetanizadora, enquanto que a torre de reciclo corresponde a menos de 3% da carga total da torre.

Em termos de estruturas de controle, a torre possui além do controle de inventário, realizado através do ajuste das vazões de destilado e de fundo, um controle de temperatura que utiliza a indicação do prato 27 para manipular a carga térmica alimentada ao refervedor da unidade. A vazão de refluxo tem seu *set point* determinado manualmente, não sendo afetado por nenhum controle adicional. A pressão é assumida se manter constante, o que também é verificado na unidade industrial. O diagrama esquemático da coluna é apresentado na Figura 4.1.

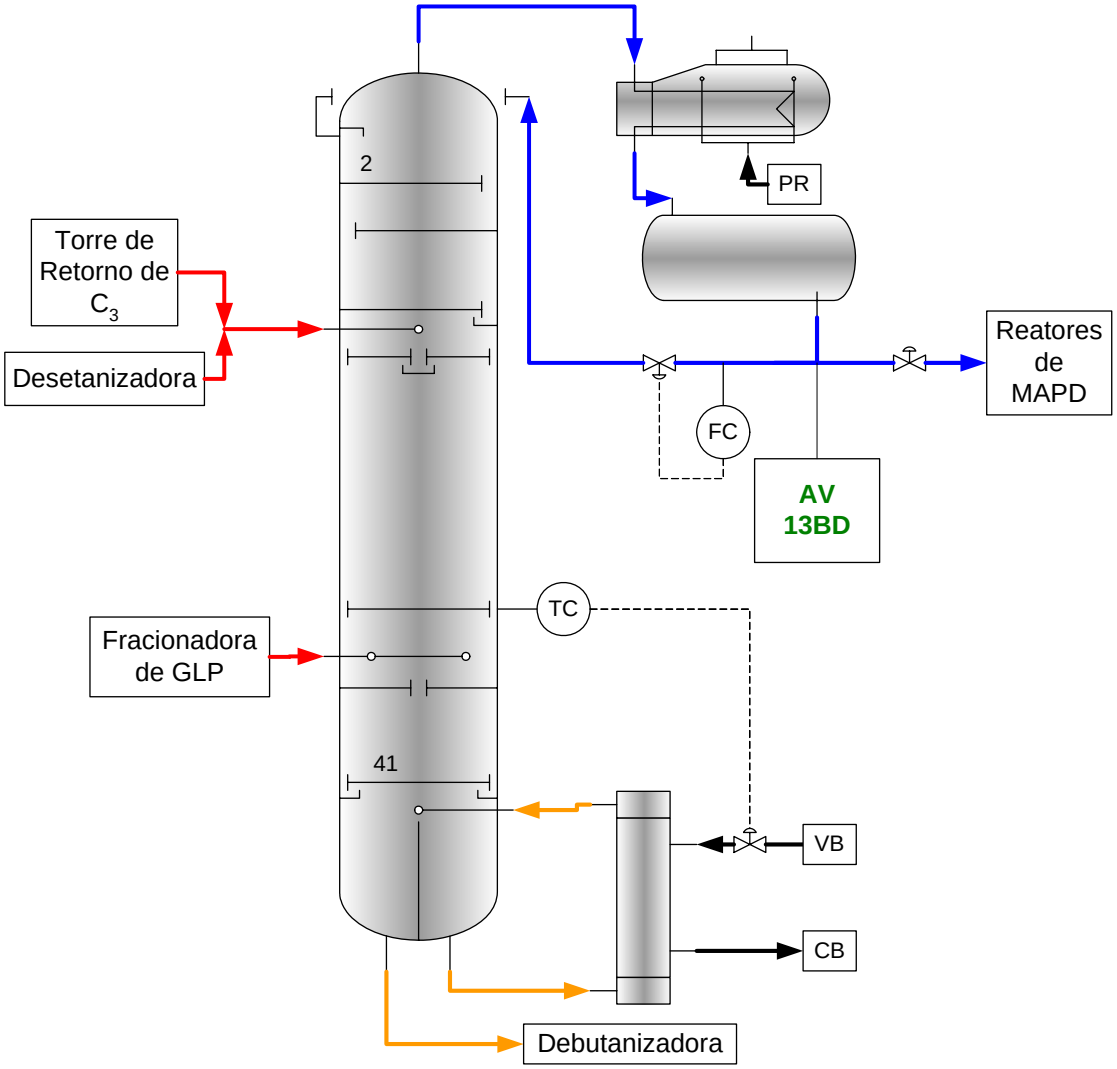


Figura 4.1: Esquema simplificado da Torre Depropanizadora

4.2.2 Simulações da Unidade

As perturbações realizadas nas diferentes simulações seguiram a metodologia apresentada por KANO *et al.* (2000). Visando simular a operação típica da unidade foram realizadas perturbações na vazão de refluxo, principal variável manipulada pelos operadores, e nas vazões e composições das correntes de alimentação, sendo essas as principais fontes de distúrbio da unidade. As variações nas composições de alimentação se restringiram as duas principais alimentações, uma vez que a corrente efluente da torre de retorno de C3 representa menos de 3% da carga total e análises realizadas demonstraram que sua composição apresenta variabilidade próxima à zero.

As perturbações foram desenvolvidas com base em um período de testes na unidade industrial, onde foram acompanhadas além das variáveis de processo, as composições das principais correntes afluentes e efluentes da torre depropanizadora. Os limites máximos e mínimos adotados para as simulações dinâmicas são apresentados de forma percentual ou em função do desvio padrão típico da variável (σ) na Tabela 4.1.

Tabela 4.1: Caracterização das perturbações utilizadas nas simulações

Variável	Magnitude Máxima da Perturbação
Vazão de Fundo Desetanizadora	±5%
Vazão de Fundo Retificadora de GLP	±10%
Vazão de Fundo Reciclo de C3	±10%
Vazão de Refluxo Depropanizadora	±10%
Fração molar de Butadieno Fundo da Retificadora de GLP	±2σ
Fração molar de Butadieno Fundo da Desetanizadora	±2σ

As simulações foram desenvolvidas através da criação de seqüências de perturbações aleatórias tendo como limites os valores apresentados na Tabela 4.1. As variações nas vazões que compõem a carga da unidade foram realizadas de forma conjunta, novamente reproduzindo o comportamento típico da unidade industrial.

4.2.3 Conjuntos de Dados

Para realizar o estudo comparativo foram utilizados três conjuntos independentes de dados, sendo que cada um corresponde a uma corrida da simulação dinâmica com duração aproximada de 34 horas. A principal diferença entre os conjuntos de dados é o grau de contaminação da corrente de destilado por 1,3 Butadieno, sendo um conjunto caracterizado por um elevado grau de contaminação, um segundo conjunto com contaminação intermediária e um terceiro com baixo índice de butadieno na corrente de topo.

O primeiro conjunto, alta contaminação, será utilizado para a calibração do modelo, ou seja, a obtenção dos parâmetros do modelo, o segundo e o terceiro, contaminação e intermediária e baixa respectivamente, serão utilizados para testar a capacidade preditiva do modelo desenvolvido no conjunto de calibração e serão denominados de conjunto de validação e teste.

O tempo de amostragem utilizado em todos os casos de simulação foi igual a 3 minutos e, com o intuito de simular um analisador em linha, para o conjunto utilizado para a calibração dos modelos adotou-se um tempo de amostragem de 15 minutos.

As composições resultantes desses três casos são apresentadas na Figura 4.2 , sendo que os dados foram normalizados em relação ao conjunto de menor índice de contaminação. Nota-se que a relação entre os níveis de contaminação entre o conjunto de elevada contaminação e baixo nível de contaminação é superior a 25 vezes.

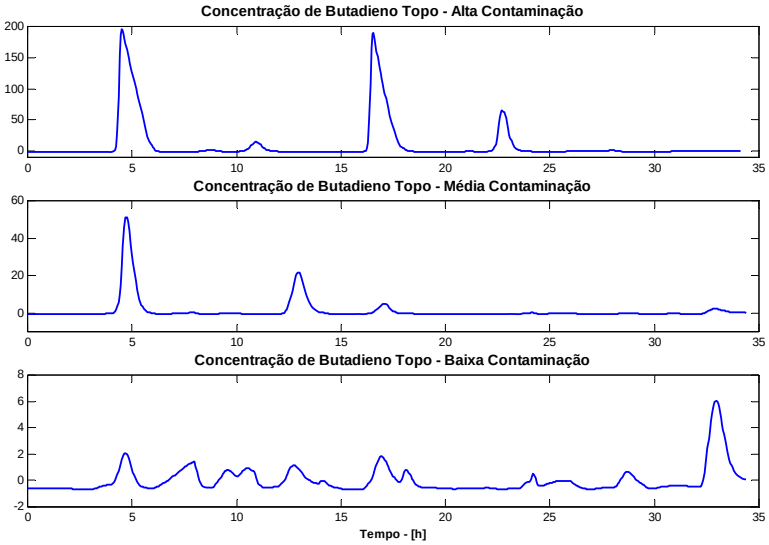


Figura 4.2: Composição de 1,3 butadieno no topo da Depropanizadora

As principais variáveis utilizadas nas simulações dinâmicas são apresentadas nas figuras a seguir. Os dados apresentados estão normalizados em função do estado estacionário inicial, uma vez que todas as simulações partiram de um mesmo estado estacionário.

O processo de normalização foi utilizado em função de um acordo de sigilo, pois as variáveis de processo da simulação refletem os valores e comportamento daquelas presentes na unidade industrial. Todas as variáveis graficadas nas figuras são apresentadas em função de variações em relação ao estado estacionário inicial. Para cada uma dessas varáveis, o parâmetro BIAS contido nos gráficos corresponde ao valor do estado estacionário inicial da variável.

Para a temperatura do Prato 01, por exemplo, o valor de BIAS é igual ao valor da temperatura desse prato no estado estacionário inicial ($T_{1,0}$) e a linha apresentada na Figura 4.3 referente ao Prato 01 é obtida através da Eq. 4.1, t um instante de tempo qualquer.

$$\begin{aligned} \Delta T_{1,t} &= T_{1,t} - BIAS \\ BIAS &= T_{1,0} \end{aligned} \tag{4.1}$$

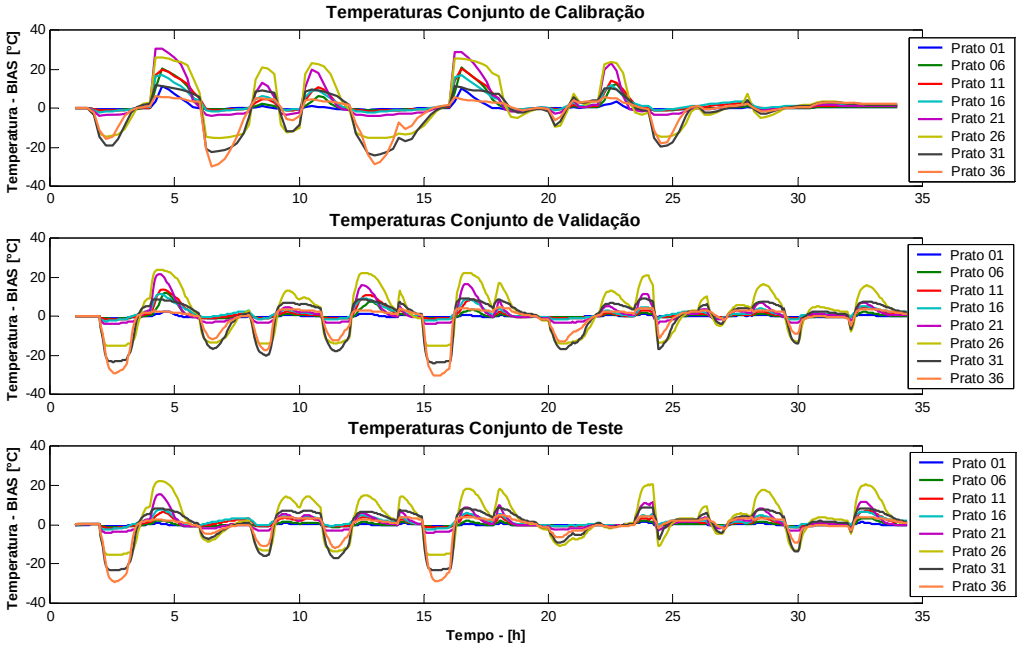


Figura 4.3: Perfis das variações de temperatura para os conjuntos de Calibração, Validação e Teste

Na Figura 4.3 é apresentado somente parte do perfil de variações de temperatura da coluna. O conjunto completo de dados utilizados nos estudos desse capítulo podem ser obtidos a partir de contato com o autor. Esses dados não foram incluídos em função do grande volume que demandariam.

De forma a reproduzir os principais distúrbios da unidade, vazões e composições de 1,3 Butadieno nas alimentações, adotaram-se perturbações do tipo degrau, ilustradas nas Figura 4.4 e Figura 4.5 respectivamente, novamente os dados contidos nas figuras estão normalizados em função do estado estacionário inicial da variável graficada.

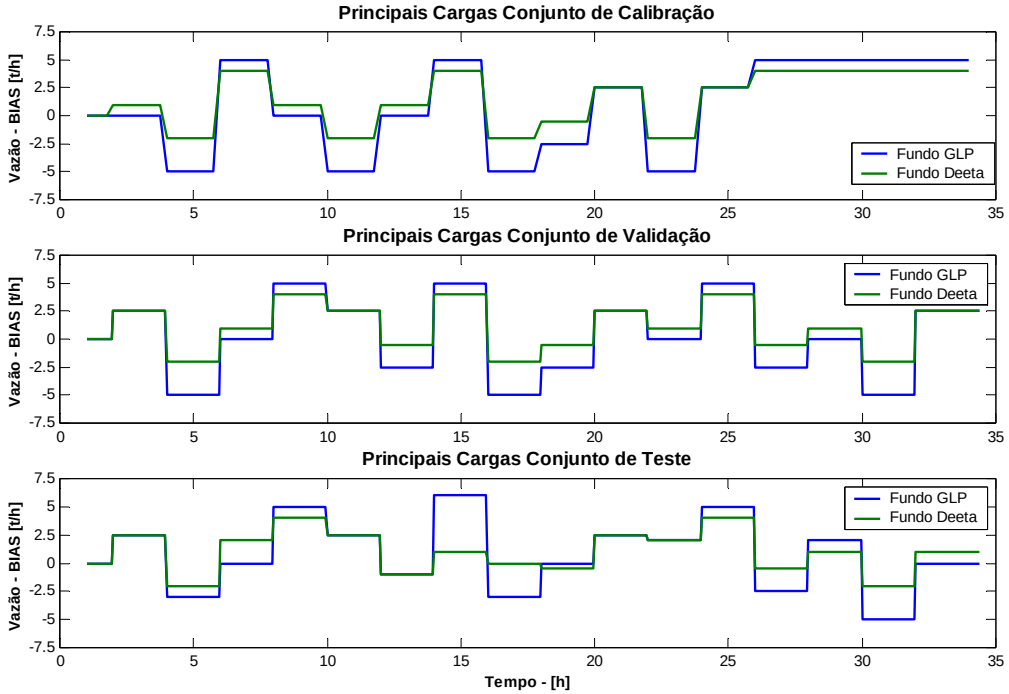


Figura 4.4: Conjuntos de perturbações adotados nas simulações para as duas principais cargas da unidade, corrente de Fundo da Torre Retificadora de GLP e corrente de Fundo da Torre Deetanizadora

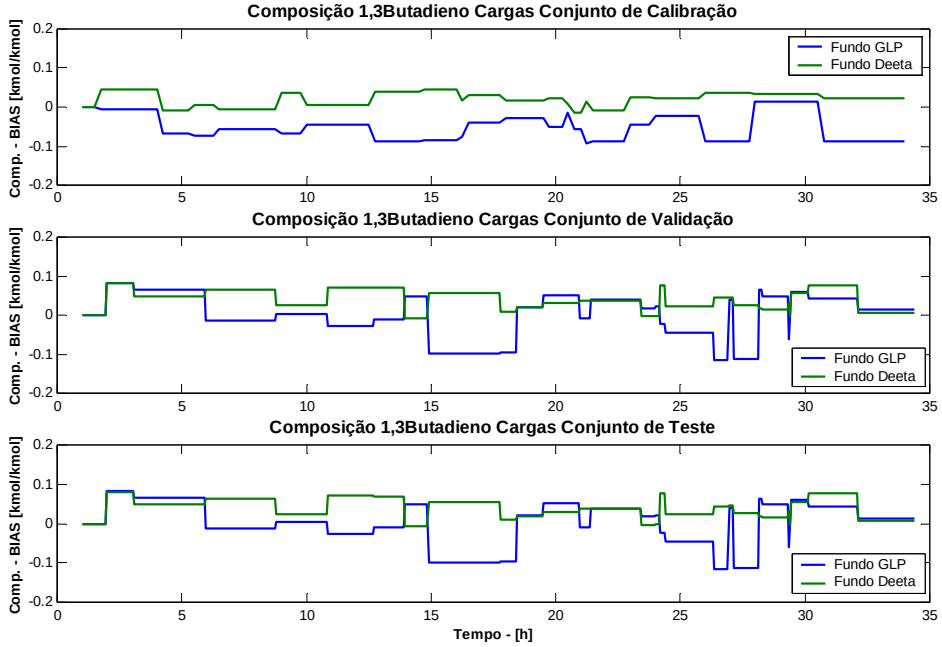


Figura 4.5: Conjuntos de perturbações para as composições de 1,3 Butadieno para as duas principais cargas da unidade.

Nota-se que para o conjunto de Validação e Teste foi utilizada a mesma seqüência de perturbações na composição das cargas, porém os valores de vazões de alimentação foram diferentes.

As variações sobre a principal variável manipulada, a vazão de refluxo, utilizadas durante as simulações são apresentadas através da Figura 4.6. A relação Refluxo/Carga resultante para as três simulações está contida na Figura 4.7.

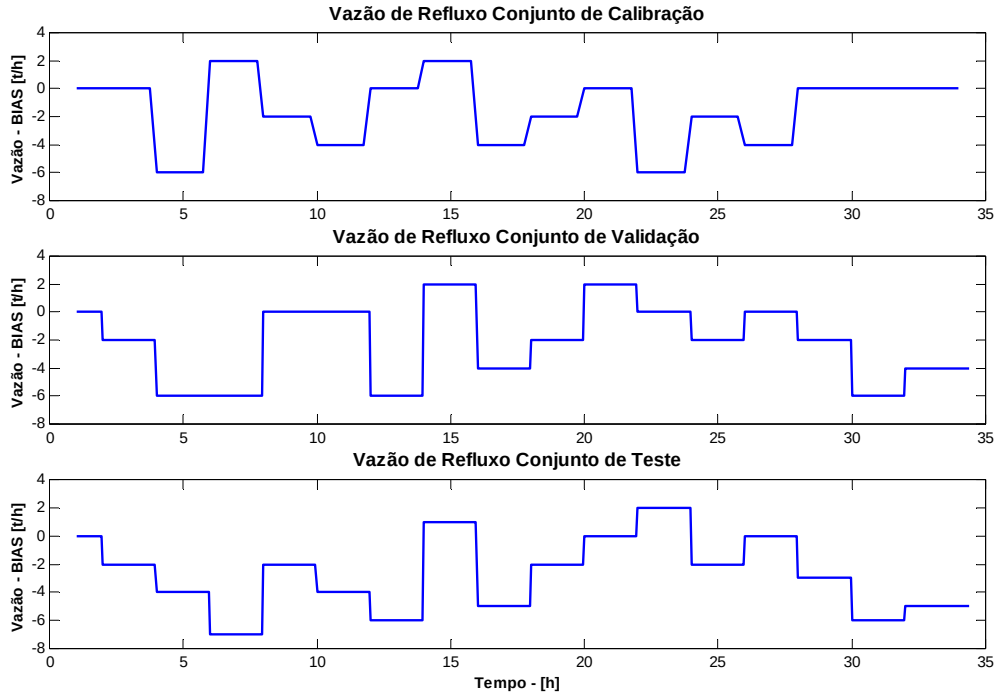


Figura 4.6: Perturbações adotadas nas simulações para a vazão de refluxo.

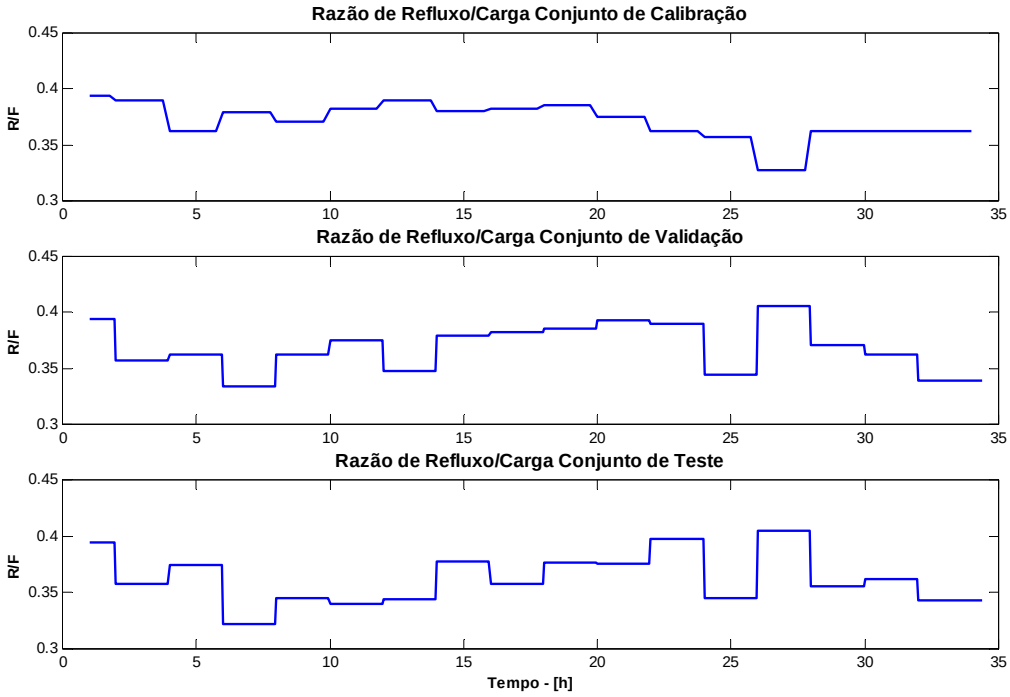


Figura 4.7: Razões de refluxo/carga para os conjuntos de Calibração, Validação e Teste

A Figura 4.8 apresenta as ações de controle efetuadas pela malha de controle de temperatura sobre a carga térmica no refervedor da coluna para os três conjuntos de dados analisados.

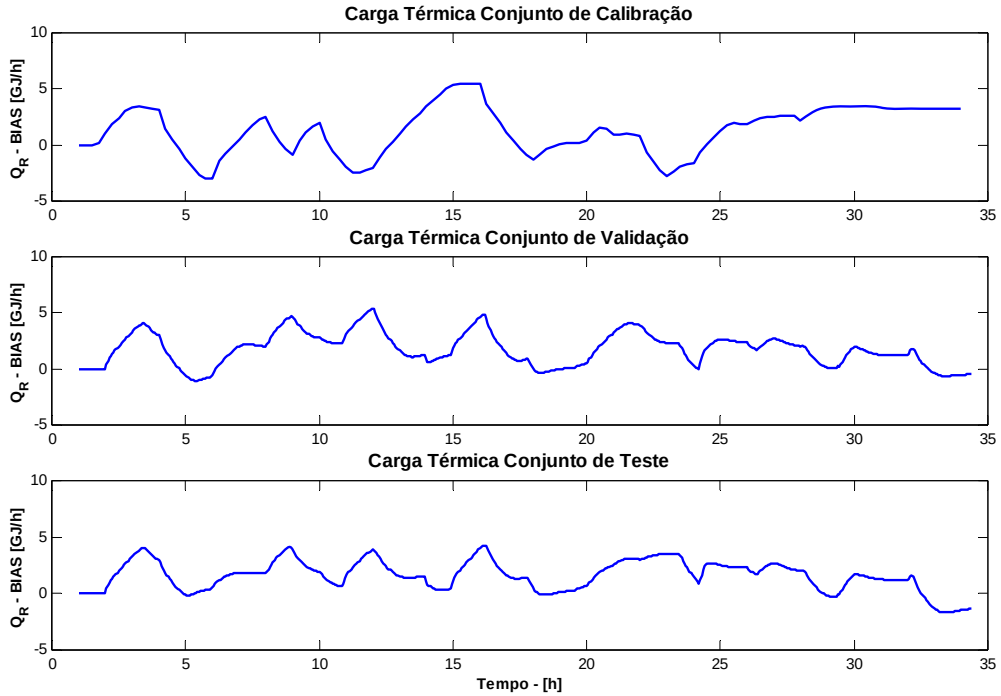


Figura 4.8: Cargas térmicas para os conjuntos de Calibração, Validação e Teste.

As razões carga térmica/carga da unidade resultantes são apresentadas na Figura 4.9.

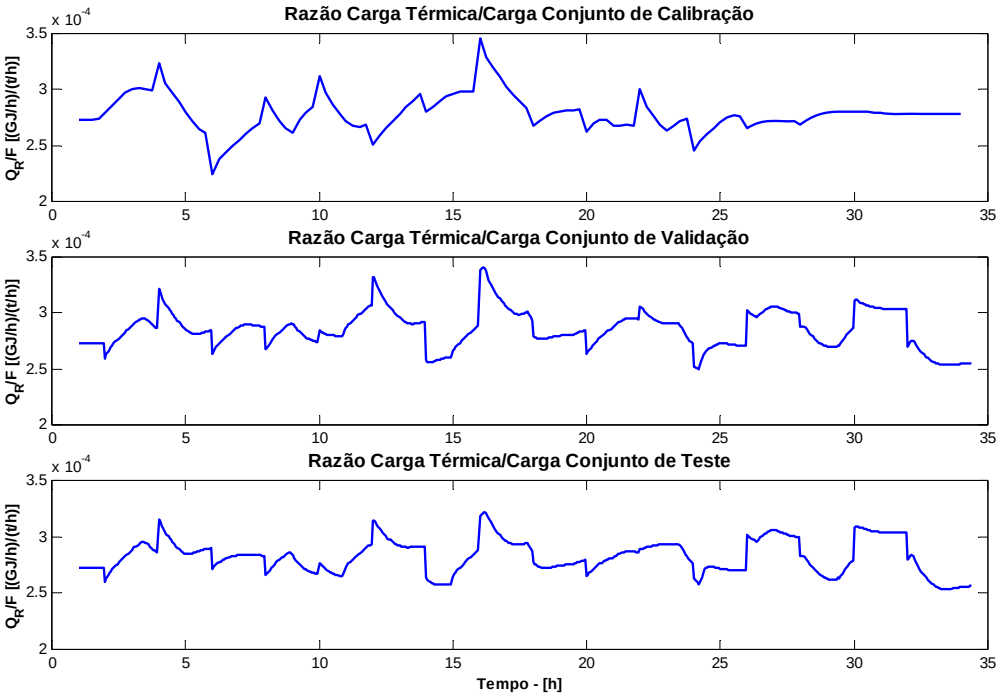


Figura 4.9: Razões carga térmica/carga da unidade para os conjunto de Calibração, Validação e Teste

4.3 Alternativas Estudadas

A combinação das diferentes estratégias de seleção de variáveis apresentadas no Capítulo 3 com as diversas variações das técnicas PLS existentes resulta em um amplo universo de alternativas a serem exploradas. Visando realizar uma análise mais cuidadosa nesse trabalho serão analisadas as duas melhores técnicas de seleção de variáveis apresentadas no capítulo 3, GA_02 (algoritmos genéticos com segregação de dados através da técnica *y-Rank*) e FS_02 (Adição Seqüencial com segregação de dados através da técnica *y-Rank*)

O objetivo é avaliar como essas técnicas se comportam quando todas as variáveis pertencentes ao conjunto de variáveis candidatas possuem informação útil, uma vez que nos casos apresentados no capítulo anterior apenas parte das variáveis continha informação, sendo que o restante não passava de oscilações aleatórias.

Em relação à estrutura de modelos serão avaliados modelos de PLS Lineares, PLS Quadráticos e a combinação de PLS Linear com a composição logarítmica, apresentado em KRESTA *et. al* (1994). As metodologias avaliadas são apresentadas na Tabela 4.2.

Poderia-se ainda combinar a técnica QPLS com a compensação logarítmica da composição, a qual poderia, em teoria, ser capaz de lidar com não linearidades ainda mais complexas. Essa alternativa, no entanto, não é abordada nesse trabalho.

Tabela 4.2: Descrição dos casos avaliados

Caso	Descrição
LT(pc)	PLS Linear com o perfil de temperatura completo (40 temperaturas)
LogLT(pc)	PLS Linear com o perfil de temperatura completo e composição logarítmica
LT(fs)	PLS Linear com Seleção de Temperaturas pelo método FS-Rank
LT(ga)	PLS Linear com Seleção de Temperaturas pelo método GA-Rank
QT(pc)	PLS Quadrático o perfil de temperatura completo (40 temperaturas)
QT(fs)	PLS Quadrático com Seleção de Temperaturas pelo método FS-Rank
QT(ga)	PLS Quadrático com Seleção de Temperaturas pelo método GA-Rank
LogLT(fs)	PLS Linear com Seleção de Temperaturas pelo método FS-Rank e composição logarítmica
LogLT(ga)	PLS Linear com Seleção de Temperaturas pelo método GA-Rank e composição logarítmica

KANO *et al.* (2000) enunciaram que a utilização de outras variáveis de processo além das temperaturas dos pratos, tais como a vazão de refluxo e carga térmica, melhoram a qualidade da estimação. Visando avaliar esse efeito foram propostas três variantes dos casos apresentadas na Tabela 4.3.

Tabela 4.3: Descrição dos subcasos avaliados

Subcaso	Descrição
a	Somente temperaturas
b	Temperaturas e razões de refluxo/carga e carga térmica/carga
c	Temperaturas e vazão de refluxo e carga térmica

4.4 Análise dos Resultados

A fim de comparar os diferentes modelos desenvolvidos será utilizado como critério o Erro Médio Quadrático (MSE) de cada um dos modelos nos conjuntos de calibração, validação e teste. Para a construção dos modelos foi utilizado o conjunto de dados contendo altos índices de contaminação. O conjunto que apresentou nível de contaminação intermediário foi utilizado para a validação do modelo e o de baixa contaminação como conjunto adicional de teste.

4.4.1 Seleção de Variáveis

No capítulo anterior as estratégias de seleção de variáveis foram testadas em conjuntos de dados onde se sabia da existência de variáveis espúrias, sendo que a metodologia concebida através da combinação do método de segregação de dados baseado no ordenamento crescente do conjunto de variáveis de resposta e dos algoritmos de seleção seqüencial e algoritmos genéticos se mostraram os métodos mais eficientes na determinação do melhor subconjunto de variáveis explicativas.

Visando avaliar essas duas técnicas em um conjunto de variáveis de processo, onde todas as variáveis contém um certo nível de informação, essas duas estratégias foram utilizadas para selecionar as variáveis para a construção do analisador virtual para a unidade aqui estudada. Para efeito de seleção foram consideradas somente as temperaturas como variáveis candidatas, sendo as variáveis de processo relacionadas a hidráulica da coluna e ao sistema de aquecimento analisadas através dos três subcasos propostos na Tabela 4.3.

A presença de *fs* e *ga* no campo entre parênteses dos modelos significa que para esse caso adotou-se a respectiva estratégia para a seleção de temperaturas que deveriam integrar o modelo.

Os subconjuntos de variáveis selecionados para cada um dos casos são apresentados na Tabela 4.4. Os números representam os pratos da coluna, sendo que o prato de topo o número 1 o prato de fundo o de número 40.

Tabela 4.4: Variáveis selecionadas por cada uma das estratégias

Caso	Pratos selecionados
LT(fs)	[1 2 3]
LT(ga)	[1 2 5 7 21 27]
QT(fs)	[1 2 3]
QT(ga)	[1 6 8 16 32 40]

Através da análise da Tabela 4.4 nota-se que o método de seleção seqüencial, casos LT(fs) e QT(fs), selecionou exatamente o mesmo subconjunto de variáveis para modelos do tipo PLS linear e PLS quadrático. O algoritmo genético, por sua vez, produziu conjunto de variáveis bastante díspares para as duas metodologias de PLS abordadas nesse trabalho.

A avaliação de qual estratégia foi capaz de produzir o modelo mais adequado será realizada através da comparação do MSE produzido por cada um dos modelos nos subconjuntos utilizados na calibração, validação e teste dos modelos.

4.4.2 Avaliação da Capacidade Preditiva

Os modelos desenvolvidos, compreendendo todos os casos e subcasos, totalizaram um total de 27 alternativas. Os gráficos de barras dos MSE's para cada um dos modelos desenvolvidos estão apresentados na Figura 4.10.

Através da análise dos gráficos contidos na Figura 4.10 pode-se observar que os modelos lineares com compensação logarítmica da composição (logLT(pc), logLT(fs) e logLT(ga)) foram capazes de produzir os mais baixos valores para os parâmetros MSE para os conjuntos de calibração e teste, enquanto que para o conjunto de validação sua performance foi inferior a todos os demais .

Os modelos não lineares (QT(pc), QT(fs) e QT(ga)) apresentaram performance superior aos seus equivalentes lineares (LT(pc), LT(fs) e LT(ga)) nos três conjuntos de dados estudados. Se comparados as variantes lineares com compensação logarítmica os modelos não lineares apresentam performance similar, sendo melhores para alguns casos e piores para outros.

Outro fator a ser analisado é a dimensão dos modelos gerados. Os modelos construídos com todas as temperaturas disponíveis (LT(pc), logLT(pc) e QT(pc)) apresentaram elevada performance preditiva, especialmente o linear com composição logarítmica e o não linear. O puramente linear é que resultou nos maiores valores de MSE entre esses três modelos aqui comparados.

Em relação aos modelos desenvolvidos com menos temperaturas (LT(fs), LT(ga), QT(fs), QT(ga), logLT(fs) e logLT(ga)) nota-se que os modelos puramente lineares (LT(fs) e LT(ga)) apresentaram performance preditiva inferior, sobretudo o modelo LT(fs) que teve suas temperaturas selecionadas através do algoritmo de seleção seqüencial combinado com o método Rank. Os modelos construídos através da seleção de temperaturas pelos algoritmos genéticos apresentaram melhores resultados que os obtidos pelo método de adição seqüencial.

Dentre os modelos não lineares desenvolvidos (QT(pc), QT(fs) e QT(ga)) pode-se notar um pequeno aumento de capacidade preditiva do modelo construído com todas as temperaturas, sobretudo quando comparado ao modelo QT(fs). No entanto, quando se compara os modelos QT(pc) e QT(ga) vê-se que o acréscimo da capacidade preditiva é pequeno e que o algoritmo genético conseguiu selecionar um conjunto de temperaturas capaz de produzir um modelo muito mais simples e com poder de predição equivalente ao modelo com todas as temperaturas.

Os modelos desenvolvidos através da aplicação da compensação logarítmica aos modelos LT(fs) e LT(ga) resultaram em modelos compactos e com alta capacidade preditiva,

sobretudo o modelo logLT(ga) que é resultado da aplicação da compensação logarítmica de composição e a seleção de variáveis através dos algoritmos genéticos.

Em relação a utilização de outras variáveis de processo além das temperaturas – subcasos *b* e *c* – para o problema analisado aqui, a inclusão dessas variáveis em geral não trouxe melhora significativa aos modelos, sendo que para esse caso os resultados apontam para conclusão diferente da obtida por KANO *et al.* (2000) onde os autores concluíram que a inclusão dessas melhorava significativamente os modelos.

Mas cabe salientar, que na maioria das colunas industriais, não se dispõe de um número suficiente de temperaturas, ou mesmo que as temperaturas não possuem a sensibilidade necessária para quantificar o grau de separação. Nestas situações é fundamental que se utilize as medições e vazão e carga térmica.

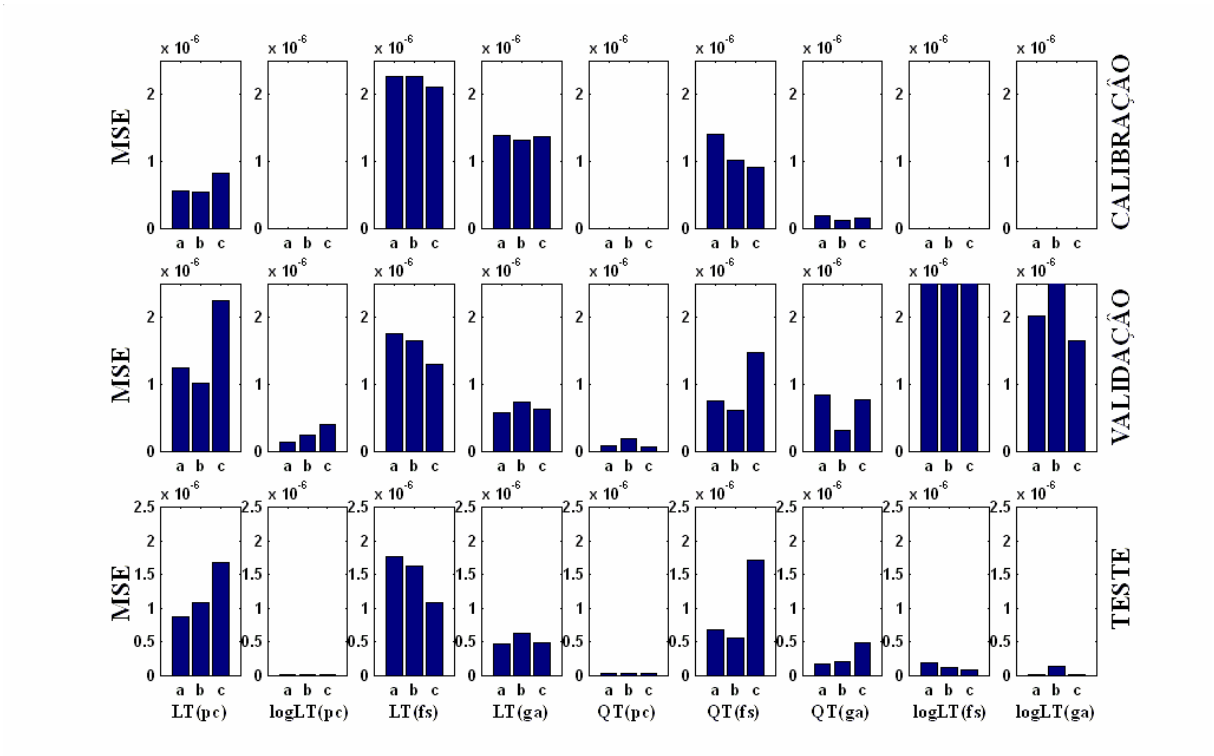


Figura 4.10: MSE obtidos para os modelos desenvolvidos

As predições retornadas pelos modelos $LT(ga)_a$, $QT(ga)_a$ e $logLT(GA)_a$ para o conjunto de calibração são apresentadas na Figura 4.11.

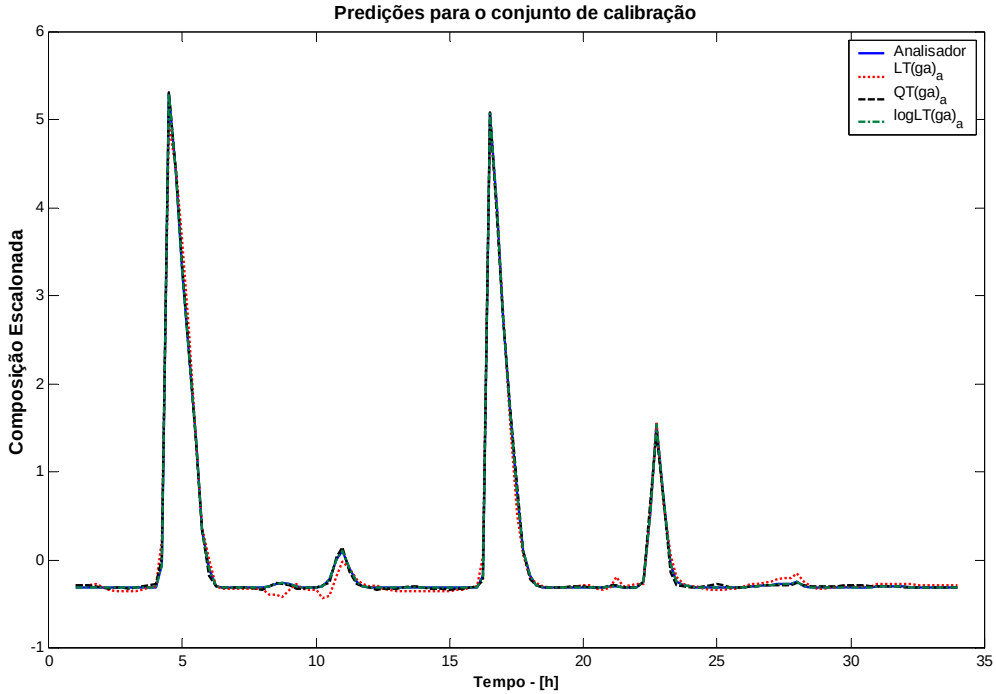


Figura 4.11: Inferências produzidas pelos modelos para o conjunto de calibração

Nota-se que os três modelos apresentaram resultados compatíveis, se adequando ao longo de todo o conjunto de dados. O modelo não linear ($QT(ga)_a$) e o com compensação logarítmica de composição ($logLT(ga)_a$) conseguiram reproduzir quase que perfeitamente os dados de calibração, enquanto que o modelo puramente linear ($LT(ga)_a$) apresentou pequenos desvios, principalmente nas regiões próximas a ausência de contaminação.

A performance de predição desses três modelos em conjuntos de dados complementares, diferentes dos utilizados na calibração, é ilustrada através da Figura 4.12 e da Figura 4.13. Para o conjunto de média contaminação, representado na Figura 4.12, o modelo não linear ($QT(ga)_a$) obteve o melhor desempenho, conseguindo reproduzir os picos de elevada contaminação e adaptando-se bem aos pontos de contaminação mais baixos.

O modelo com compensação logarítmica apesar de conseguir reproduzir com exatidão os pontos de baixo índice de contaminação resultou em desvios nos picos, produzindo valores maiores e menores que os realmente existentes. O modelo linear ($LT(ga)_a$) apesar de conseguir reproduzir os picos de contaminação elevada, produziu os resultados menos precisos para o restante da faixa.

Para o conjunto de baixa contaminação, representado na Figura 4.13, o modelo com melhor performance foi o linear com compensação logarítmica ($logLT(ga)_a$), sendo esse o modelo que melhor conseguiu acompanhar a tendência da concentração de 1,3 Butadieno na corrente. Os demais modelos ($LT(ga)_a$ e $QT(ga)_a$) acabaram por produzir resultados que oscilavam em torno da variável a ser monitorada, degradando a qualidade da inferência.

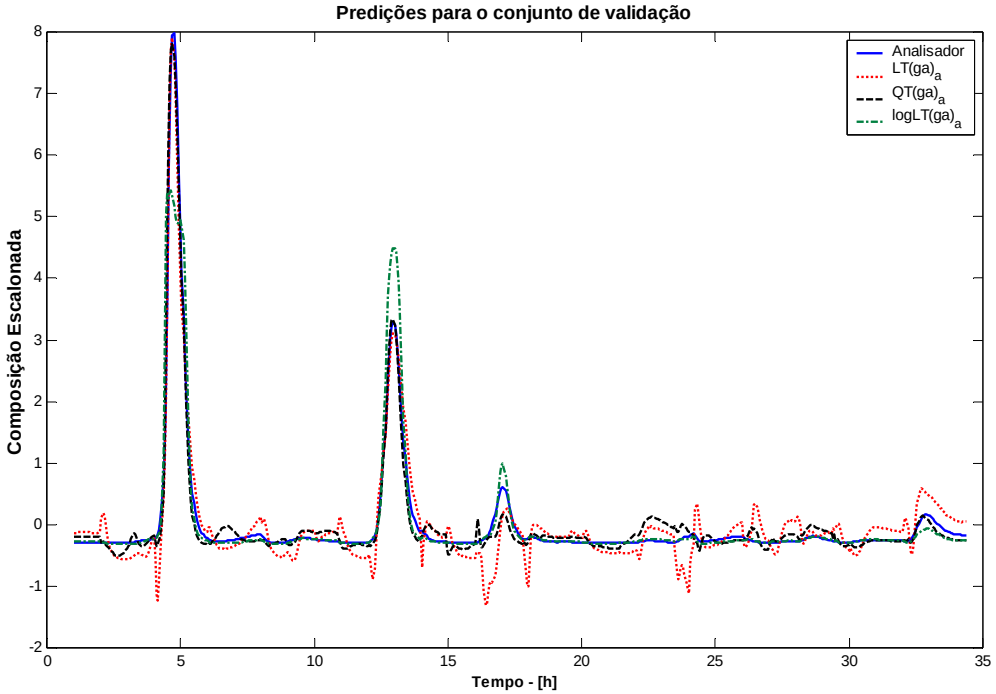


Figura 4.12: Inferências produzidas pelos modelos para o conjunto de validação

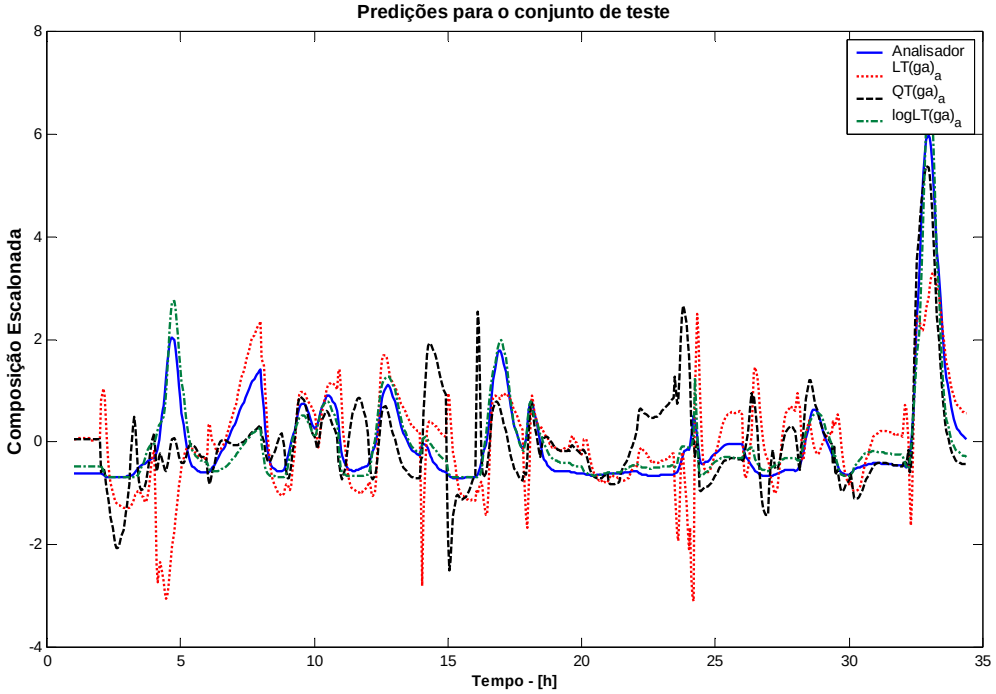


Figura 4.13: Inferências produzidas pelos melhores modelos para o conjunto de teste

4.4.3 Análise dos Resultados

Analisando-se os modelos através de sua capacidade preditiva, Figura 4.10, nota-se que os modelos desenvolvidos com o auxílio da seleção de variáveis baseadas em técnicas *Stepwise*, casos LT(fs) e QT(fs), obtiveram as piores performances em termos de predição e

ajuste dos dados. Esse resultado reforça as observações feitas por HAN e YANG (2004), de que as técnicas *Stepwise* por terem a tendência de gerar modelos mais enxutos não são capazes de fornecer o melhor modelo.

Os modelos gerados com o auxílio da seleção de variáveis baseadas em algoritmos genéticos, casos LT(ga) e QT(ga), forneceram resultados melhores do que aqueles obtidos a partir das técnicas *Stepwise*. Quando comparados com os modelos gerados com todas as variáveis das colunas, pode-se observar que os últimos apresentaram resultados em geral melhores que os obtidos através das técnicas de seleção de variáveis.

Comparando-se as dimensões dos modelos gerados através das técnicas de seleção de variáveis com os modelos gerados com todas as temperaturas (LT(pc) e QT(pc)) pode-se notar que os modelos LT(fs) e QT(fs), obtidos através da técnica *Stepwise* de seleção de variáveis, tendem a ser espartanos demais, resultando em modelos com baixo poder de ajuste e predição.

Os modelos LT(ga) e QT(ga), no entanto, produziram resultados satisfatórios, fornecendo modelos com um bom compromisso entre dimensão, cada um deles possui 6 temperaturas, e poder de ajuste. Certamente que quando comparados com os modelos LT(pc) e QT(pc) apresentam resultados inferiores, mas o grau de melhora fornecido por esses últimos, formados por todas as 40 temperaturas da torre, pode não se justificar quando da aplicação industrial, uma vez que os primeiros demandariam uma manutenção mais criteriosa em um menor número de sensores e sofreriam muito mais com os erros de medição e calibração.

Em relação a tipologia dos modelos, ou seja PLS (LT(pc), LT(fs) e LT(ga)), QPLS (QT(pc), QT(fs) e QT(ga)) e PLS mais compensação logarítmica das composições (logLT(pc), logLT(fs) e logLT(ga)), pode-se observar que os modelos puramente lineares resultaram em modelos com os piores resultados, principalmente os modelos com menos variáveis.

Os modelos não lineares do tipo QPLS e PLS com compensação logarítmica das composições, em geral, apresentaram resultados compatíveis, sendo que em alguns casos os QPLS apresentaram melhores resultados e nos demais os com composição logarítmica geraram os melhores resultados.

Analisando-se os resultados em relação à adição de outras variáveis de processo, subcasos b e c, não foi possível se determinar um efeito conclusivo. Para alguns modelos a utilização de razões entre as correntes trouxe benefício, em outros a utilização diretamente das vazões produziu resultados superiores, porém em outras situações a utilização somente de temperaturas foi suficiente para produzir modelos satisfatórios.

Em uma comparação final, pode-se dizer que os modelos que apresentaram os melhores resultados foram os $LT(ga)_a$, $QT(ga)_a$ e $logLT(ga)_a$, sendo seus resultados apresentados para os três casos simulados através da Figura 4.11, Figura 4.12 e Figura 4.13. O

conjunto de calibração utilizado foi a de alta contaminação, mas os resultados produzidos pelos três modelos para o conjunto de média contaminação foram equivalentes.

Para o conjunto de baixa contaminação, o modelo que mais representou o valor da composição do 1,3 Butadieno foi o modelo $\log LT(ga)_a$, sendo que os demais oscilavam em torno do real valor da variável.

A princípio qualquer um dos três modelos apresentados seria capaz de ser utilizado para a aplicação em um analisador virtual, pois os resultados produzidos são satisfatórios. No entanto considerando-se além da performance o critério da capacidade extrapolativa do modelo, teria-se que optar entre os modelos $LT(ga)_a$ e $\log LT(ga)_a$, pois esses são baseados em modelos lineares, o que garante que os ganhos de cada umas das variáveis apresentam comportamento monótono, enquanto que o modelo $QT(ga)_a$ possui um funcional não linear – uma parábola que relaciona as variáveis latentes $\boldsymbol{u}$ e $\boldsymbol{t}$ – e que pode levar a condição de inversão de ganho em uma região diversa daquela de calibração.

Capítulo 5

Aplicação Industrial

No capítulo anterior foi apresentada a utilização das metodologias desenvolvidas nessa dissertação em uma simulação dinâmica de uma coluna depropanizadora, ambiente onde se detém o controle de todas as variáveis e distúrbios que possam afetar a unidade.

No presente capítulo, a unidade real que deu origem a simulação é utilizada para a descrição das etapas envolvidas na implementação industrial de um analisador virtual, que para a unidade em questão tem por objetivo inferir a composição de 1,3 Butadieno (13BD) na corrente de destilado de uma Coluna Depropanizadora de uma central petroquímica.

Inicialmente será apresentada uma descrição do sistema onde a torre objeto desse estudo está localizada, exaltando a importância da unidade dentro do contexto global da planta e a importância do conhecimento da concentração do 13BD, principal contaminante da corrente de topo da torre depropanizadora.

Posteriormente serão descritas as etapas desenvolvidas para a obtenção de um modelo capaz de representar o comportamento da variável em questão. Essas etapas envolvem desde o reconhecimento da unidade, mapeamento de instrumentação disponível, levantamento de problemas operacionais e utilização de ferramentas para a modificação do paradigma operacional utilizado na unidade.

Por fim será descrito o modelo desenvolvido, ressaltando o ambiente para construção, as limitações e potencialidades do mesmo e os resultados obtidos pelo modelo selecionado para a construção do Analisador Virtual.

5.1 Descrição da Unidade

5.1.1 Sistema de Tratamento de Compostos Olefínicos

A obtenção dos petroquímicos olefínicos básicos, tais como o eteno e o propeno, a partir de nafta ou GLP é realizada através do craqueamento térmico desses compostos em presença de vapor d’água em fornos de pirólise. Nesses fornos a matéria prima, geralmente composta

por cadeias compostas por 5 a 9 átomos de carbono, é convertida em cadeias menores sob a ação o intenso calor.

A corrente que deixa o forno é composta por uma ampla gama de produtos que vão desde gases leves, como o hidrogênio, até compostos aromáticos e gasolina de pirólise. A separação e purificação de cada um desses produtos é realizada através da utilização de inúmeros sistemas de destilação, reatores de purificação, unidades de compressão e separação, entre outros processos.

A Copesul possui hoje duas unidades industriais para a produção dos mais diversos petroquímicos, desde eteno e propeno, produtos essenciais para as indústrias de segunda geração, até solventes aromáticos e combustíveis, como a gasolina, GLP e Diesel. As duas plantas possuem tecnologias um pouco diferentes. A Planta Um, primeira unidade a partir em 1982, adota a configuração *tail end*, enquanto que a Planta Dois, cuja partida se deu em 2000, adota a configuração *front end*. A torre 13T04 está inserida na área de tratamento de olefinas da Planta Um da Copesul, e por esse motivo é que serão dados maiores detalhes sobre ela. A Figura 5.1 apresenta um diagrama de blocos simplificado da configuração *tail end* presente na Planta 1 da Copesul.

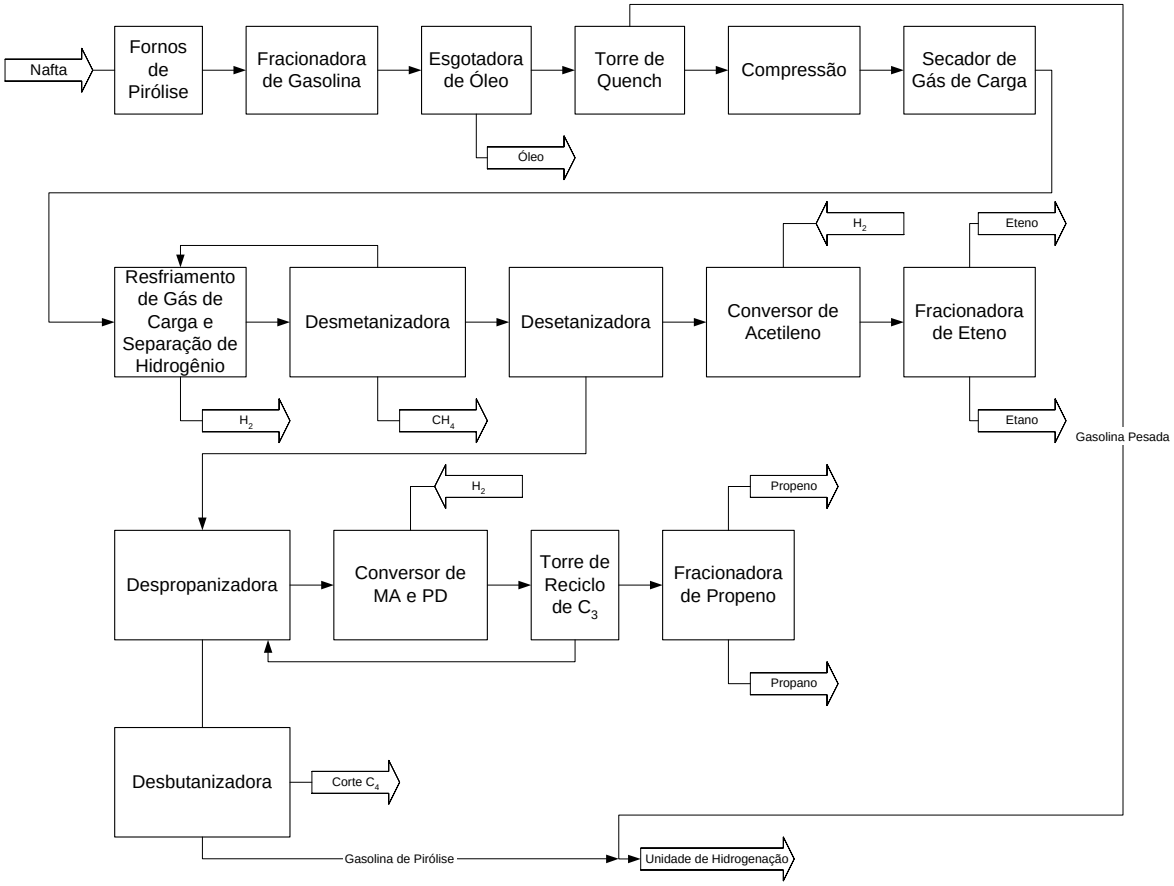


Figura 5.1: Fluxograma esquemático de uma central petroquímica com configuração *tail end*

O início da cadeia de produção dos petroquímicos produzidos na Copesul ocorre nos fornos de pirólise. Nesses equipamentos a nafta petroquímica, constituída por cadeias compostas por 5 a 9 átomos de carbono, é quebrada em cadeias menores na presença de vapor

d'água. A reação ocorre no interior dos tubos dos fornos, e produzem desde gases leves, como o hidrogênio e o metano, até compostos mais pesados, como a gasolina de pirólise.

Para a separação e purificação dessa mistura, são utilizados processos de destilação e conversão catalítica. A primeira etapa é separação dos compostos pesados, representados principalmente pela gasolina e óleo. Esses compostos são removidos nas unidades de fracionamento de gasolina, esgotamento de óleo e na torre de *quench*. Tanto a gasolina, quanto o óleo são produtos mais pesados e, portanto são retirados pelo fundo dessas unidades, sob a forma de líquidos. O produto de topo é uma mistura de gases leves que é comprimida e resfriada, para a sua liquefação, e então encaminhada as demais unidades de processo.

A corrente de topo, já liquefeita, é alimentada a torre desmetanizadora, onde são removidos pelo topo o metano e o hidrogênio, os compostos mais leves dessa corrente, e pelo fundo sai uma corrente constituída por uma mistura de compostos olefínicos com mais de 2 átomos de carbono em sua estrutura. Essa mistura é encaminhada a torre desetanizadora, onde é divida em duas outras correntes. No topo dessa unidade é removida uma corrente composta predominantemente por cadeias com 2 átomos de carbono, como o eteno, etano e outros contaminantes como o acetileno.

A remoção desses contaminantes é realizada através da hidrogenação desses compostos. Essa reação é realizada em reatores de leito fixo. A corrente que deixa os reatores passa por um sistema de secagem e é direcionada para a torre de separação de eteno/eteno. A corrente de topo dessa torre é constituída principalmente por eteno, enquanto que a corrente de fundo é composta em sua maioria por etano.

A corrente de fundo da desetanizadora é composta por uma mistura de compostos com mais de 3 átomos de carbono em sua estrutura. Essa corrente é alimentada a torre depropanizadora, que separa essa corrente em duas outras, uma rica em compostos com 3 átomos de carbono, que é removida pelo topo, e outra rica em compostos mais pesados, retirada pelo fundo.

A corrente de topo contém propeno, propano, metilacetileno, propadieno e outros contaminantes. Como o metilacetileno e o propadieno são compostos que interferem nos processos das indústrias de segunda geração, eles são convertidos a propeno em uma série de reatores de leito fixo. A corrente efluente desses reatores é uma mistura de propeno, propano e compostos denominados de *green oil*. Os *green oil* são subprodutos das reações de hidrogenação que ocorrem nesses reatores. A mistura é separada na torre de retorno de C_3. Nessa unidade são geradas três correntes, a de topo, composta principalmente por hidrogênio não convertido e metano, uma corrente lateral, composta por uma mistura de propeno/propano, e outra de fundo composta principalmente por *green oil* que é encaminhada novamente a depropanizadora para recuperar traços de propeno que possam ter ficado nessa corrente.

A corrente de fundo da depropanizadora é enviada a torre debutanizadora para a separação de compostos com 4 átomos de carbono e compostos com mais de 4 átomos de

carbono. A corrente de topo é constituída pelo corte C_4 e é encaminhada para tancagem. A corrente de fundo, denominada de gasolina de pirólise é enviada para a unidade de hidrogenação para a saturação de compostos olefínicos insaturados de alto peso molecular antes de serem enviados a unidade de aromáticos.

5.1.2 Unidade de Tratamento do Corte C_3

A torre depropanizadora é uma das quatro torres de destilação presentes na unidade de tratamento do corte C_3. O sistema, além de quatro torres de destilação, é composto por uma bateria de reatores de hidrogenação. O fluxograma contido na Figura 5.2 representa de forma esquemática a unidade em questão

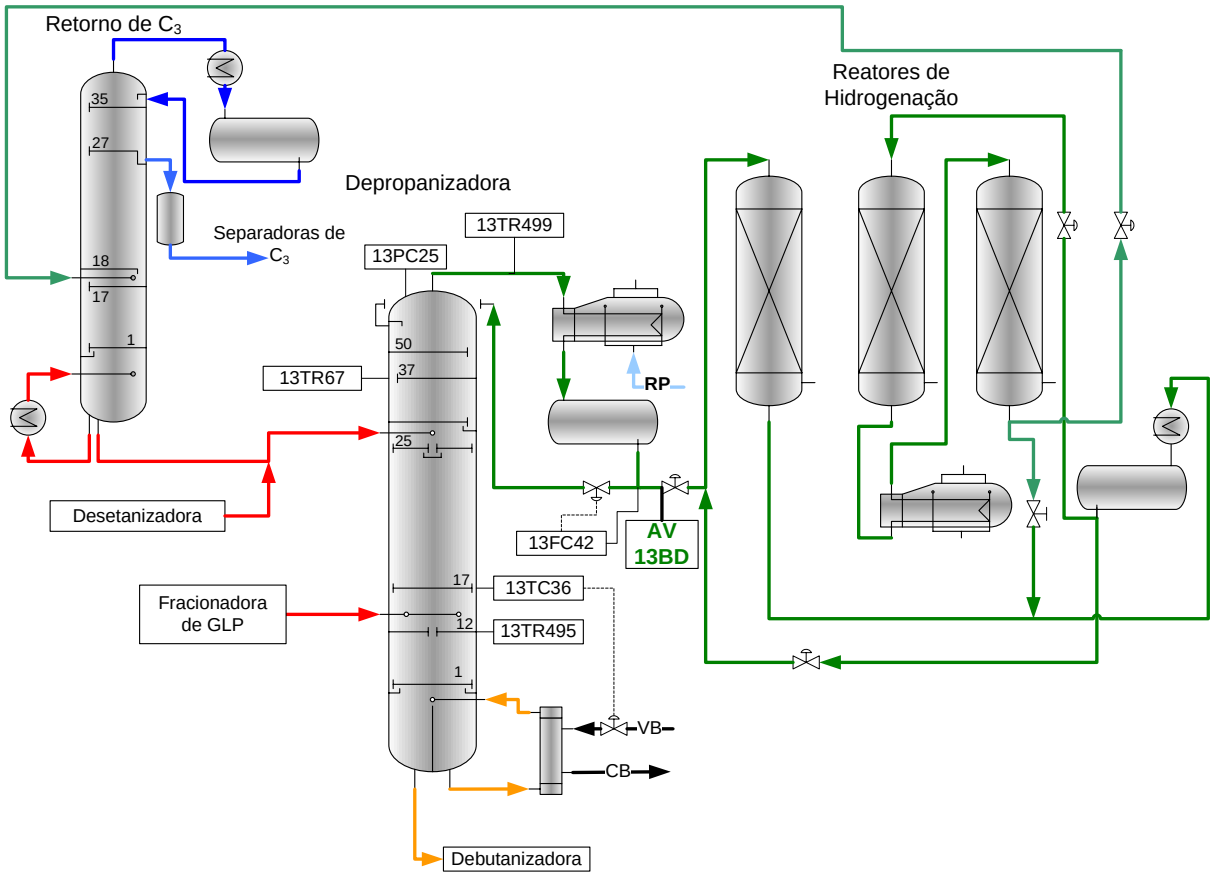


Figura 5.2: Fluxograma simplificado da Unidade de Tratamento de Corte C_3

A torre de depropanizadora, como pode ser visto na Figura 5.2, está localizada no centro da unidade, sendo responsável pela segregação das correntes que alimentarão os reatores de hidrogenação e a torre debutanizadora.

A carga da torre depropanizadora é composta por três correntes, sendo duas oriundas da corrente de fundo de unidades a montante dessa unidade e a outra é uma corrente de reciclo, sendo que as duas primeiras respondem por aproximadamente 97% da carga da depropanizadora.

Em termos construtivos, a torre depropanizadora é constituída por 50 pratos, possui um refervedor alimentado com vapor de baixa pressão. O condensador é do tipo *kettle*, tendo como fluído térmico propeno líquido.

A alimentação da torre é feita em dois pratos, sendo uma corrente introduzida no 25° prato a partir do fundo e outra no 12° segundo prato. A corrente alimentada no prato 25 é composta pela mistura das correntes de fundo da torre deetanizadora e da torre de reciclo de C_3. A outra fração é oriunda do fundo da torre de retificação de GLP, sendo que essa divisão foi adotada em função da natureza química das correntes.

A torre depropanizadora se caracteriza pelo baixo nível de instrumentação disponível, possuindo medidores em 3 pratos, além de um situado na linha de topo e outro na linha de fundo. O controle da unidade se caracteriza pela existência de uma malha que relaciona a temperatura do prato 17 à vazão de vapor alimentada ao refervedor.

O controle de inventário é realizado através do nível no vaso de refluxo e no fundo da torre. A vazão de refluxo tem seu *set point* fixado pelos operadores, não sendo alterada por nenhuma malha de controle.

5.2 Desenvolvimento do Analisador

A necessidade do conhecimento da composição de 1,3 Butadieno na corrente de topo da depropanizadora está vinculada ao fato de que esse composto compete com o propadieno e com o metilacetileno nos reatores de hidrogenação subseqüentes a depropanizadora. A reação do 1,3 Butadieno com o hidrogênio é muito mais eficiente do que com os outros dois compostos. Assim, a presença de um elevado índice desse contaminante na corrente de alimentação dos reatores, acarreta a passagem de metilacetileno e propadieno para as etapas subseqüentes do processo, acabando por chegar as indústrias de segunda geração, onde eles funcionam como venenos nos processos industriais.

Na corrente de topo da unidade há um analisador em linha, porém o seu uso para a monitoração de 1,3 Butadieno foi interrompido em função de constantes problemas com a coluna devido a interferência de um contaminante desconhecido.

A manutenção de baixos índices de 1,3 Butadieno é conseguida através da utilização de elevadas vazões de refluxo, que garantem a especificação da corrente. Apesar de essa medida garantir a pureza da corrente de topo ela é onerosa, pois o condensador utiliza propeno líquido como fluído de troca térmica, cuja obtenção é extremamente dispendiosa.

O não funcionamento do analisador em linha e a inexistência de rotinas de análises para a torre depropanizadora forçaram o desenvolvimento de uma metodologia para a criação do analisador virtual para essa unidade, baseada em simulações estacionárias e dinâmicas e testes na unidade industrial.

5.3 Simulações da Unidade

A realização de testes em unidades industriais não pode ser feita a esmo, pois geralmente esses testes implicam na geração de produtos fora de especificação, desenvolvimento de um plano especial de análise e mudança na rotina de operação normal, resultando em custos e *stress* operacional.

Nesse sentido, deve-se desenvolver um plano de perturbações capaz de agregar o maior número de informações possíveis, com o menor número de perturbações necessárias. Visando a geração desse plano de análise foi desenvolvida uma metodologia fundamentada em simulações estacionárias, interpolações e simulações dinâmicas.

5.3.1 Simulações Estacionárias

As simulações estacionárias foram realizadas no simulador Aspen Plus®, versão 12.1 a partir de um arquivo fornecido pelo Time de Processo da Copesul. O arquivo inicial sofreu alterações com a finalidade de reter processos relacionados somente com a unidade em questão, uma vez que o arquivo inicial continha uma simulação utilizada para o balanço material de toda unidade de olefinas.

Após as devidas modificações, foram realizadas diversas simulações, utilizando dados estacionários obtidos da unidade industrial para ajustar e validar a simulações, fazendo com que a mesma reproduzisse o comportamento das variáveis de processo encontradas na unidade real. Uma vez que as variáveis de processo obtidas na simulação se mostraram em concordância com as disponíveis na planta para as mesmas condições foi desenvolvida uma análise de sensibilidade da composição de 1,3 Butadieno na corrente de topo em função de alterações realizadas nas duas variáveis manipuladas disponíveis: a vazão de refluxo e a temperatura do prato de controle.

Utilizando-se combinações dessas duas variáveis foram realizadas um total de 42 simulações estacionárias, cobrindo, dessa forma, todas as condições normalmente utilizadas na unidade.

A partir dos resultados obtidos através das simulações estacionárias foi possível criar um “mapa” da concentração de 1,3 Butadieno em função dos valores das variáveis manipuladas, presente na Figura 5.3.

Mais detalhes do desenvolvimento dessas simulações e seu impacto no planejamento dos testes industriais podem ser encontrados em KOHMANN (2004), o qual se trata de um relatório de estágio supervisionado, cujo foco principal era a elaboração das simulações estacionárias e dinâmicas para essa unidade.

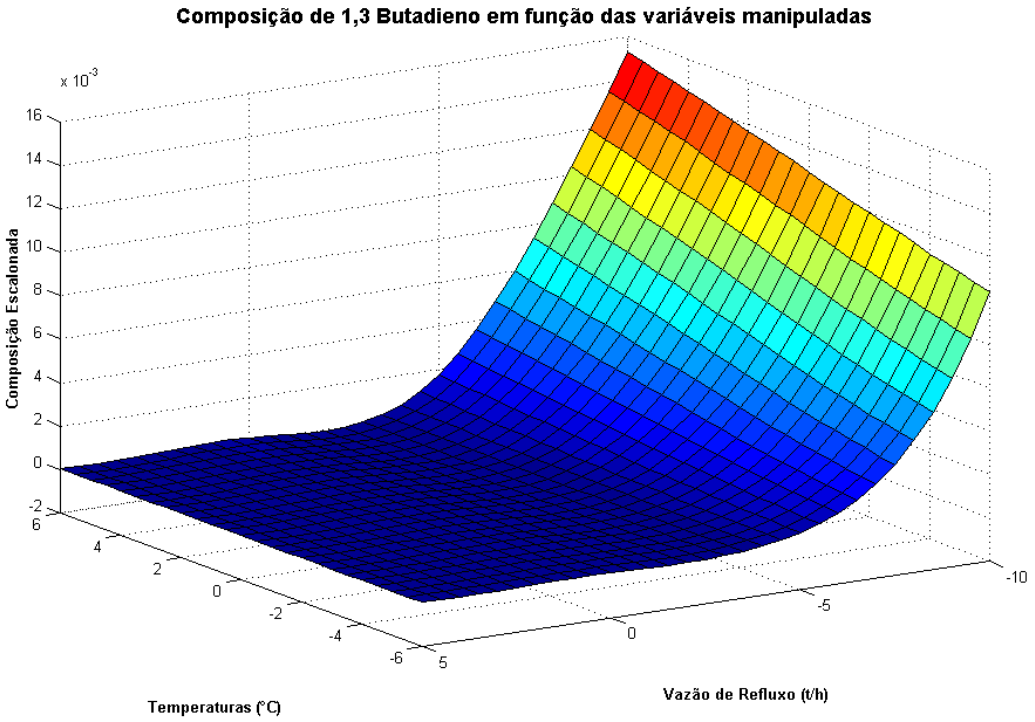


Figura 5.3: Mapa de sensibilidade da composição de 1,3 Butadieno em função das variáveis manipuladas

A partir da análise dos resultados foi possível identificar que a unidade opera em um estado de superespecificação, uma vez que a operação normal é realizada em pontos próximos a origem do gráfico.

Vê-se claramente que essa região é um platô, sendo possível se trabalhar em vazões de refluxo menores sem comprometer a qualidade da corrente de topo da depropanizadora. Esse resultado secundário foi de extrema importância para o convencimento do pessoal de operação, uma vez que informalmente esse comportamento já era conhecido, mas não havia nenhum estudo mais profundo sobre esse assunto e, por segurança os operadores operavam em vazões maiores de refluxo.

Os resultados estacionários foram utilizados para o planejamento das perturbações realizadas na unidade industrial, focando principalmente nas regiões próximas a contaminação, regiões onde as perturbações poderiam produzir informação útil.

5.3.2 Simulações Dinâmicas

As simulações estacionárias forneceram informações de como a concentração de 1,3 Butadieno varia em função das duas principais variáveis manipuladas em sucessivos estados estacionários.

As simulações dinâmicas permitem que além de se obter o valor em estado estacionário se consiga visualizar o comportamento dinâmico de variáveis frente a uma perturbação, bem como obter uma idéia do tempo de resposta da variável.

As simulações dinâmicas foram desenvolvidas em ambiente Aspen Dynamics® versão 12.1 a partir da simulação estacionária. Além do controle de inventário, adicionado automaticamente pelo programa, foi introduzido um controle de temperatura similar ao encontrado na unidade industrial. O controlador foi ajustado através do método da síntese direta.

O conjunto de perturbações efetuadas em estado estacionário foi reproduzido no ambiente dinâmico através de um *task* (conjunto de perturbações) dinâmico, onde foram variados a vazão de refluxo e o *set point* da malha de controle de temperatura.

Os resultados obtidos, em termos de composição de 1,3 Butadieno, diferiram dos obtidos através das simulações estacionárias, o que está em acordo com as conclusões de CONZ (2005), onde através de um estudo comparativo realizado, observou-se a discrepância dos resultados obtidos através de simulações estacionárias e dinâmicas, mesmo sendo os simuladores pertencentes a uma mesma desenvolvedora, no caso a Aspen Tech.

Os resultados apesar de não caracterizar quantitativamente os valores de contaminação na corrente de topo, foram úteis, pois foram capazes de capturar a tendência da contaminação, ilustrando a característica de elevação brusca ao se atingir uma determinada vazão de refluxo. Esse comportamento é representado na Figura 5.4, onde os dados estão devidamente escalonados.

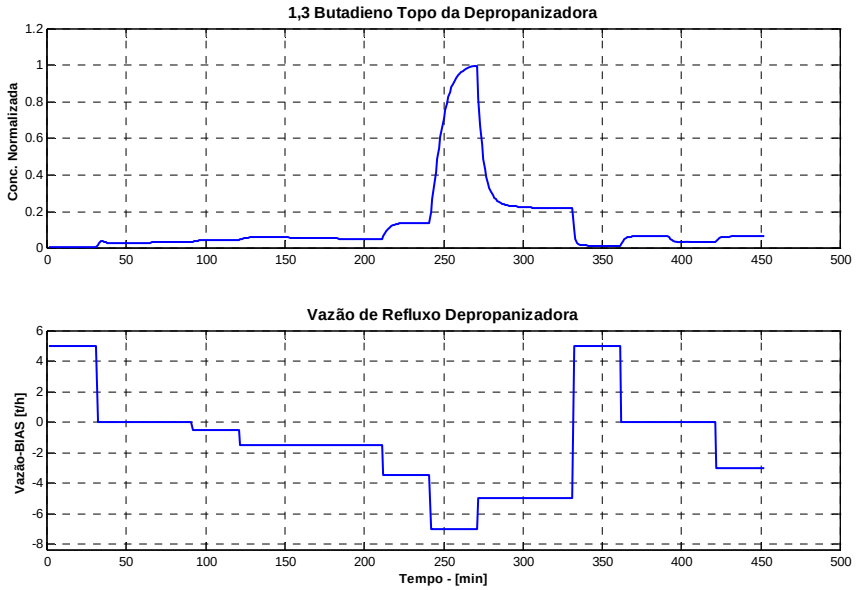


Figura 5.4: Comportamento dinâmico da concentração de 1,3 Butadieno frente a diferentes variações na vazão de refluxo

Como pode ser observado na Figura 5.4, há um súbito acréscimo na concentração de 1,3 Butadieno a partir de um determinado valor de vazão de refluxo. Essa elevação abrupta também é verificada na unidade industrial.

5.4 Testes Industriais

A idéia inicial desse trabalho era utilizar os dados gerados nas simulações estacionárias e dinâmicas como ferramenta para gerar os dados disponíveis para a construção do modelo do analisador virtual, uma vez que a unidade em questão não apresenta nenhuma rotina de análises, tão pouco um histórico desse tipo de registro. Em função de os dados gerados pelo simulador não apresentarem concordância total com os dados de processo, principalmente nos valores de concentração de 1,3 Butadieno na corrente de topo, optou-se por uma abordagem tradicional, com a realização de testes em planta, combinados com análises laboratoriais com a finalidade de se obter dados para a calibração do modelo.

5.4.1 Reconhecimento da Unidade

O primeiro passo para a realização de qualquer teste em planta é a verificação das condições atuais da unidade, passando por uma análise da instrumentação presente e das rotinas operacionais adotadas pelo pessoal de operação.

Instrumentação

Na verificação da instrumentação disponível na unidade depropanizadora deparou-se com os primeiros problemas. Em termos de medidores de temperatura verificou-se a existência de apenas 5 sensores, sendo 1 na linha de topo, 1 na linha de fundo e somente 3 ao longo de toda coluna.

Além disso, dos 3 medidores instalados na coluna, um está localizado em um dos pratos de carga, sofrendo forte influência das oscilações na corrente de alimentação, o segundo é utilizado para o controle de carga térmica da unidade e o último estava desativado, por apresentar uma compensação através da pressão de topo, a muito desatualizada. A primeira modificação sugerida e prontamente atendida foi a reativação desse medidor, através da desvinculação da compensação da pressão e disponibilizar essa leitura no sistema PIMS (*Plant Information Management System*) da unidade.

Outro problema verificado na unidade está atrelado a presença de oscilações constantes na temperatura do prato de controle. Esse comportamento pode ser verificado através da Figura 5.5 para um período de 12 horas de operação. A origem dessas perturbações é a configuração do sistema de controle de nível de condensado no refervedor.

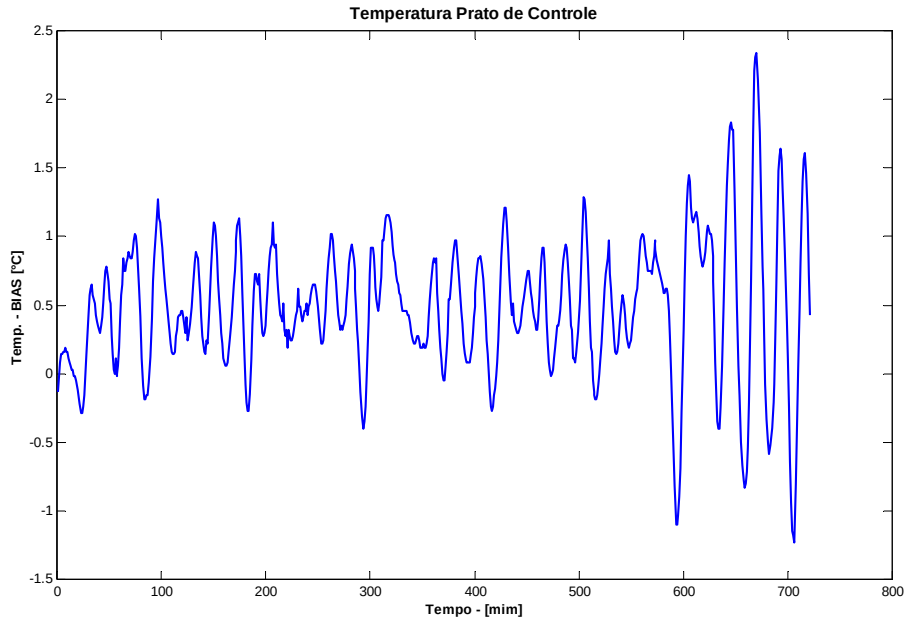


Figura 5.5: Oscilações na temperatura do prato de controle para um período de 12 horas de operação

Essas oscilações acabam por influenciar no sistema de refervimento, prejudicando o desempenho do sistema de controle. A Figura 5.6 apresenta os reflexos das oscilações da temperatura sobre o *set point* de vazão de vapor para o refervedor.

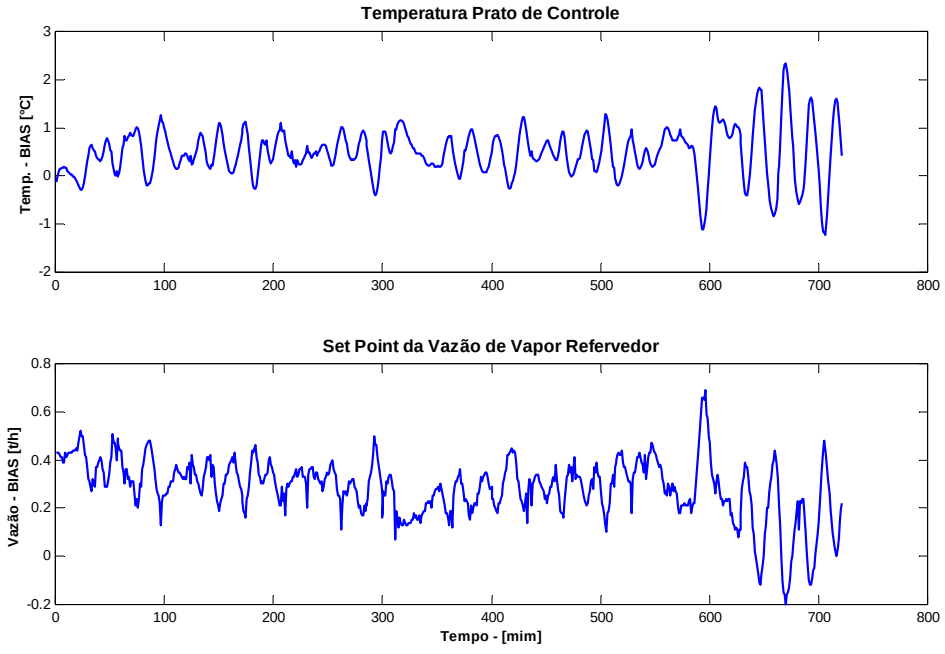


Figura 5.6: Efeitos das oscilações na temperatura do prato de controle sobre a vazão de vapor do refervedor

As oscilações nessa variável são acarretadas pelo sistema de refervimento presente na unidade. Esse sistema utiliza um termo sifão alimentado por vapor de baixa pressão dessuperaquecido. O calor necessário para o aquecimento da corrente proveniente da torre é conseguido pela condensação desse vapor.

Para garantir que ocorra condensação do vapor no interior do refervedor é necessário que se mantenha um nível de condensado, e é nesse ponto que surge o problema da oscilação. O nível de condensado é "controlado" através de purgador na linha de condensado e uma válvula de *by pass*, a qual é ajustada manualmente pelos operadores.

O sistema funciona pelo bloqueio da linha de condensado, por ação do purgador, no momento em que esse detecta a presença de vapor vivo na linha de condensado. O aparecimento de vapor na corrente de condensado é acompanhado por um decréscimo na carga térmica fornecida à coluna, uma vez que o vapor deixa de condensar no interior do refervedor, pois o único calor transferido é o sensível, o que faz com que a temperatura do prato de controle caia.

Imediatamente o controlador retorna uma ação de controle, aumentando a vazão de vapor para o refervedor, com o objetivo de aumentar a temperatura do prato de controle. Como a linha de condensado está bloqueada pelo purgador, o nível de condensado no refervedor é refeito e a condensação volta a ocorrer, aumentando a quantidade de calor fornecido à torre, dando origem os ciclos de temperatura observados na Figura 5.5 e na Figura 5.6.

Durante a fase de avaliação da unidade foi proposta uma melhoria no sistema de controle de nível do vaso de refervimento, no entanto essa modificação só pode ser realizada durante paradas de manutenção.

Os demais instrumentos disponíveis na unidade, compostos principalmente por medidores de vazão e pressão, apresentavam leituras estáveis e com nível de ruído aceitável. Em relação à pressão, verificou-se que esta apresentava variância quase nula, sendo que suas leituras foram consideras como constantes e, portanto, não contribuindo com informação útil para o desenvolvimento do analisador. A Tabela 5.1 apresenta os TAGs e a descrição dos instrumentos presentes na unidade de tratamento do corte C3 envolvidos no desenvolvimento do analisar virtual.

Tabela 5.1: Instrumentos utilizados no desenvolvimento do analisador

TAG	Descrição
13TR499	Temperatura corrente de Topo Depropanizadora
13TR67	Temperatura do Prato 37 da Depropanizadora
13FC42	Vazão de Refluxo Depropanizadora
12FC12	Vazão da Corrente de Fundo da Fracionadora de GLP
13FC64	Vazão da Corrente de Fundo da Desetanizadora
13FC38	Vazão da Corrente de Fundo da Torre de Retorno de C_3
R/F	Razão Refluxo/Carga

Obs.: a carga da Torre Depropanizadora é composta pelas correntes 12FC12, 13FC64 e 13FC38

As razões de refluxo/carga e carga térmica/carga são variáveis que caracterizam o real ponto de operação da unidade, pois as principais variáveis manipuladas são normalizadas pela carga da unidade, incluindo assim, os efeitos das oscilações na alimentação da unidade (LUYBEN, 1992).

Práticas Operacionais

Além de verificar a disponibilidade e o estado da instrumentação disponível, durante o período de reconhecimento da unidade estudaram-se também as principais práticas operacionais adotadas pelo pessoal responsável pela unidade.

Os engenheiros da unidade relataram que a torre opera em um elevado índice de superespecificação da corrente de topo, sendo isso resultado dos elevados valores de refluxo adotados na unidade. O mais interessante desse fato é de que apesar de os operadores terem consciência dessa situação, os mesmos recusavam-se a reduzir o valor dessa variável. A justificativa por eles fornecida era de que como eles não tinham indicação direta nenhuma sobre a concentração de 1,3 Butadieno na corrente de topo, optava-se por operar em uma zona "confortável", onde qualquer distúrbio que pudesse influenciar na concentração do contaminante seria eliminado pelos altos valores da vazão de refluxo.

A única indicação indireta de contaminação que os operadores tinham era a elevação de temperatura na corrente de topo (13TR499). O fato agravante nessa estratégia é de que o intervalo de temperatura que separa a região de contaminação e limpeza é extremamente pequeno, em torno de 0,5 a 1°C. Esse comportamento pode ser visualizado através da Figura 5.7, obtida no período de testes industriais realizados no desenvolvimento desse estudo.

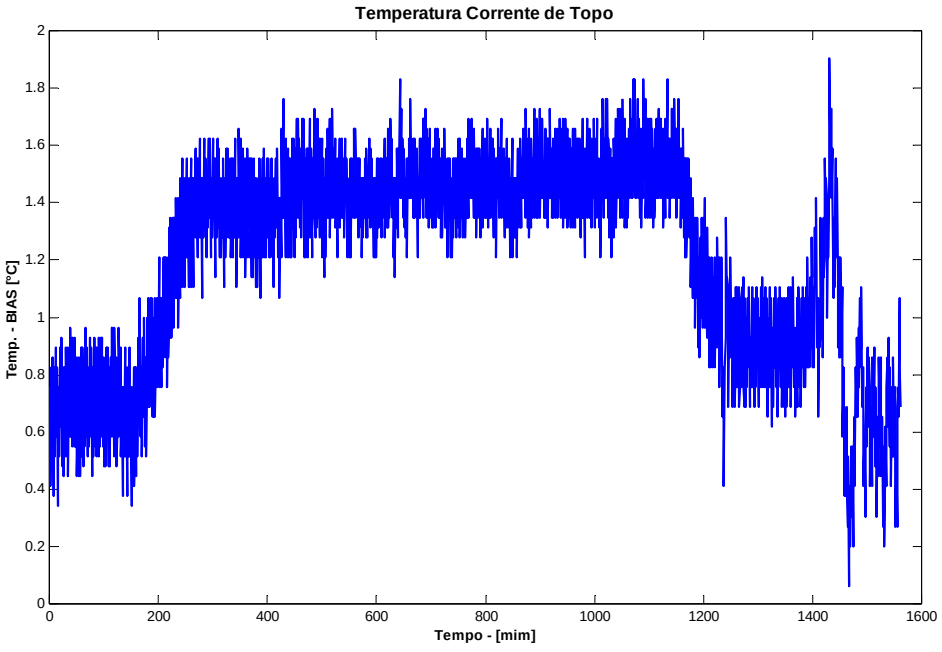


Figura 5.7: Gráfico de tendência da temperatura da corrente de topo durante o período de testes industriais

Um dos benefícios colaterais decorrentes dos testes na unidade foi a reativação do medidor 13TR67, o qual, como mencionado anteriormente, estava fora de operação. Esse medidor mostrou-se muito mais sensível à contaminação por 1,3 Butadieno do que a temperatura de topo, sendo que o intervalo de temperatura que separa as regiões de contaminação e limpeza é em torno de 10°C, como pode ser observado na Figura 5.8.

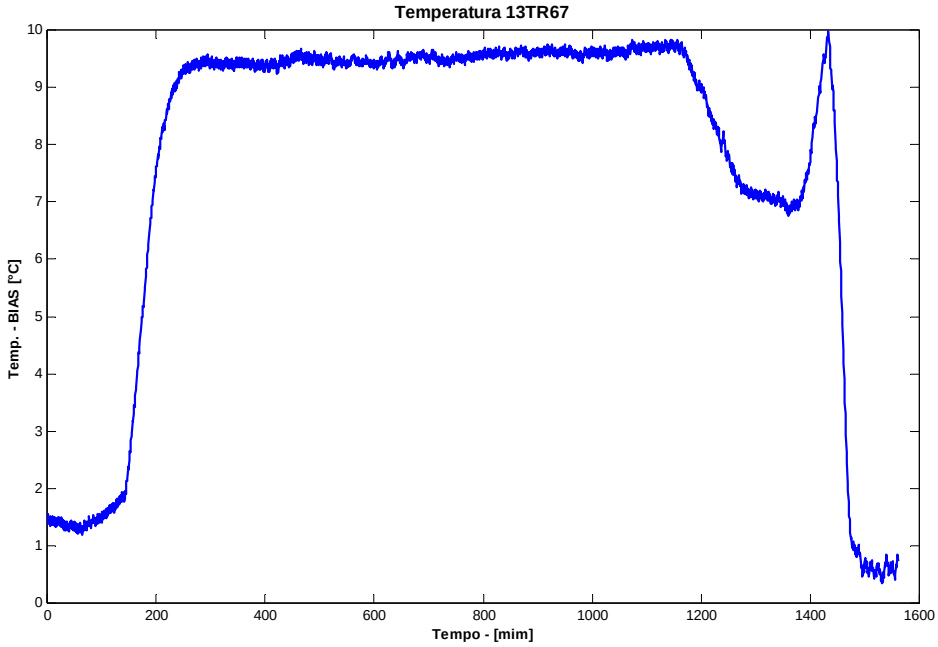


Figura 5.8: Gráfico de tendência da temperatura do 13TR67 durante o período de testes industriais

A adoção, por parte do pessoal de operação, dessa temperatura (13TR67) como indicativo de contaminação da corrente de topo possibilitou a redução na vazão de refluxo da torre, o que terminou por representar um grande ganho econômico, energético e operacional, pois como o condensador utiliza propeno líquido com fluído de troca, a redução na vazão de refluxo, diminui a demanda desse oneroso insumo e garante maior flexibilidade operacional de toda a unidade, sobretudo nos meses de verão, onde a disponibilidade desse passa a ser crítica.

Planejamento e Execução

Testes industriais requerem um planejamento refinado, para que não se tornem algo ainda mais crítico para a operação da unidade. Normalmente eles demandam um plano de análises especiais, dedicação apurada por parte dos operadores e geração de produtos fora de especificação.

Na unidade em questão, esses fatos foram agravados em função da inexistência de qualquer rotina de análise, sendo necessário o retreinamento do pessoal de laboratório na caracterização das correntes a serem analisadas.

Com o objetivo de minimizar o impacto dos testes sobre a unidade, as simulações estacionárias e dinâmicas foram utilizadas como base para o planejamento de um *step test* capaz de produzir a maior quantidade de informação com o mínimo de distúrbio na unidade.

Inicialmente pretendia-se perturbar as duas principais variáveis manipuladas da unidade, a temperatura do prato de controle, responsável pela carga térmica no refervedor, e a vazão de refluxo. Após análise do comportamento dessas variáveis foi observado que as perturbações na temperatura de controle não trariam muita informação para a construção do analisador,

pois em função de seu comportamento oscilatório, não foi possível a detecção de um valor estacionário, ou quase estacionário.

O teste resumiu-se então a variações no *set point* da vazão de refluxo, sendo que uma seqüência tipicamente utilizada é apresentada na Figura 5.9. As variações possuíam um intervalo mínimo de 3 horas entre cada coleta de amostra, tido, pelo pessoal da operação, como tempo necessário para as temperaturas da torre atingirem o novo valor estacionário.

Os efeitos dessas perturbações sobre a composição de 1,3 Butadieno e sobre as temperaturas disponíveis podem ser verificados através da análise da Figura 5.9. Nota-se claramente que a composição de 1,3 Butadieno possui uma tendência muito similar às apresentadas pelas temperaturas e são fortemente influenciadas pela vazão de refluxo.

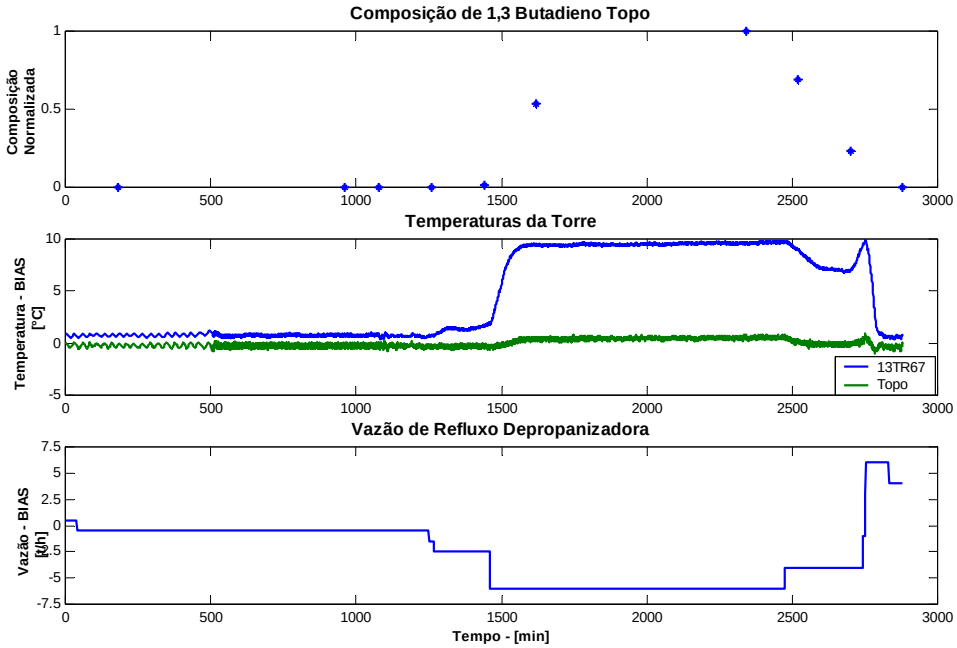


Figura 5.9: Composição de 1,3 Butadieno na corrente de topo de temperaturas da depropanizadora em função da vazão de refluxo.

Nota-se que na parte final do teste houve um súbito aumento nas temperaturas, sem que houvesse uma redução no refluxo. Esse fato conduziu a uma breve investigação para definir as causas dessa elevação. Análises revelaram um súbito aumento na carga da torre, decorrente de oscilações em unidades a montante da depropanizadora.

Através da análise da relação refluxo/carga (RF), apresentada na Figura 5.10, pode-se notar claramente a correlação existente dessa nova variável com as temperaturas, sendo que o comportamento das temperaturas é praticamente uma imagem especular da relação RF. Em função disso, verificou-se a importância da inclusão dessa variável no desenvolvimento do analisador.

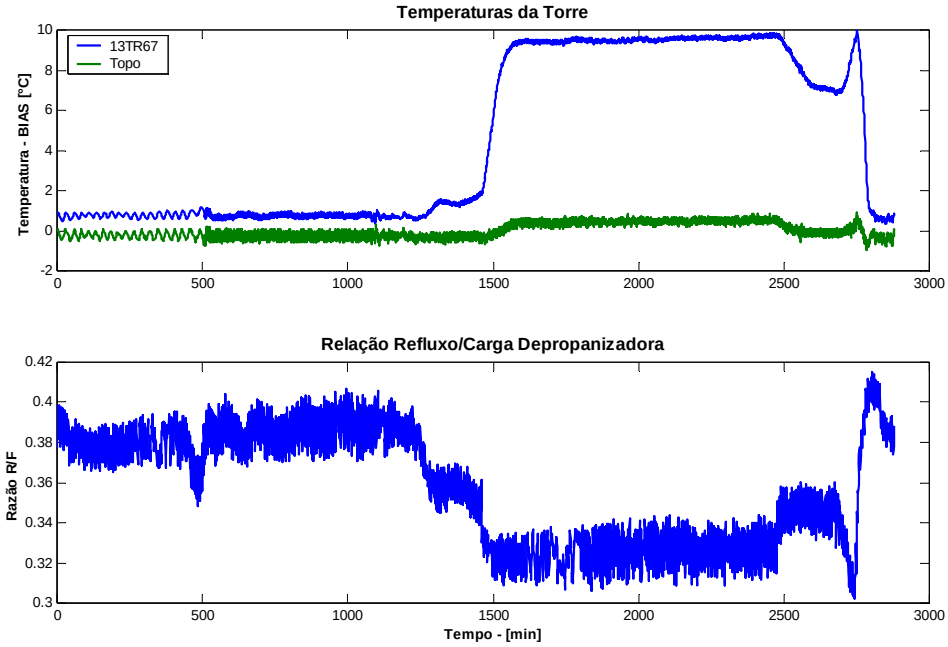


Figura 5.10: Gráficos de tendência das temperaturas da depropanizadora e relação vazão de refluxo/carga da unidade.

5.5 Calibração do Modelo

O analisador a ser utilizado industrialmente foi desenvolvido em ambiente Aspen IQ®, uma ferramenta desenvolvida pela Aspen Tech para a elaboração de analisadores virtuais. Um detalhamento dessa ferramenta pode ser encontrado em CONZ (2005).

A ferramenta disponibiliza modelos baseados em técnicas PLS lineares, PLS não lineares utilizando lógica Fuzzy e modelos híbridos de PLS linear e redes neuronais. O processo de seleção de variáveis utilizado pelo programa é baseado em uma combinação de algoritmos genéticos e PLS Linear, podendo se optar pela execução de validação cruzada para se determinar o número de variáveis latentes a serem consideradas.

A única estratégia de atualização de modelos disponível é através da soma de um BIAS, tendo seu valor determinado pela diferença entre o valor obtido através de análise e a valor predito pelo modelo. Esse valor é mantido até o momento em que uma nova análise esteja disponível no sistema.

Em função das condições disponíveis para a elaboração do analisador virtual optou-se pela adoção de um modelo PLS linear, uma vez a quantidade de amostras disponíveis para a calibração do modelo é reduzido, e os modelos não lineares disponibilizados pela ferramenta sabidamente demandam um elevado número de amostras para a sua calibração.

Ao longo do desenvolvimento desse trabalho foram desenvolvidos diversos modelos, sendo eles avaliados e modificados em testes realizados na unidade industrial. Ao final da etapa de calibração/desenvolvimento foram selecionados dois modelos, os quais apresentaram

os melhores resultados nos testes, sendo que esses modelos serão comparados na etapa de validação, para a escolha do modelo final a ser utilizado no analisador virtual. As características desses modelos são apresentadas na Tabela 5.2.

Tabela 5.2: Características dos modelos finais

	Modelo 1	Modelo 2
Tipo	PLS Linear	PLS Linear
Variáveis Integrantes	13TR499, 13TR67 e 13FC42	13TR499, 13TR67, 13FC42 e R/F

A calibração do modelo foi realizada com base nos pontos obtidos durante o período de testes, uma vez que essa unidade não possui nenhuma rotina de análises periódicas e os analisadores em linha para a avaliação de 1,3 Butadieno há muito não funcionam. Os resultados produzidos pelos modelos candidatos a analisadores são apresentados na Figura 5.11, sendo os valores da concentração do contaminante normalizados em função do limite de especificação.

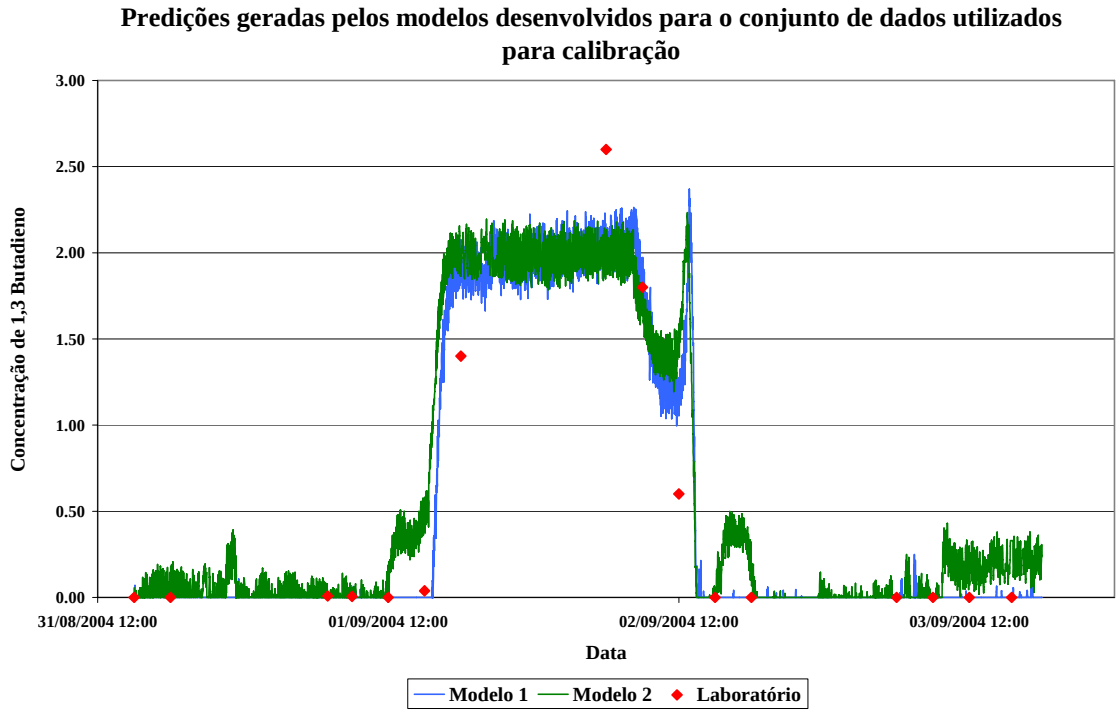


Figura 5.11: Predições produzidas pelos modelos candidatos a analisadores virtuais para o período de calibração dos modelos

Pode-se notar que o analisador foi capaz de reproduzir o comportamento da contaminação em termos de tendência, não sendo capaz de reproduzir o valor numérico do maior valor de contaminação. Isso não chega a ser um problema, pois esse valor está muito acima do valor limite para a contaminação da corrente de topo por 1,3 Butadieno, que é de cerca de 38% desse valor máximo.

Essa incapacidade de representar o valor de máxima contaminação está associada a inúmeros fatores, tais como, o baixo número de amostras utilizadas na calibração do modelo,

num total de 16,a saturação das variáveis secundárias, distúrbios não medidos na composição das cargas que alimentam a unidade, entre outros.

Os resultados obtidos, no entanto, foram considerados satisfatórios pelos responsáveis pela operação da unidade, uma vez que anteriormente não se tinha nenhuma indicação dos valores de contaminação da unidade e que dentro das condições operacionais possíveis para a unidade o modelo foi capaz de capturar o comportamento da composição de 1,3 Butadieno na corrente de topo.

5.6 Validação e Implantação do Modelo

Antes de se colocar o modelo a disposição do pessoal de operação é necessário validar o analisador desenvolvido, ou seja, avaliar os resultados produzidos por ele em um conjunto de dados diferente dos utilizados para a sua calibração.

Essa etapa é feita de maneira análoga à etapa de calibração, ou seja, o resultado predito pelo analisador é comparado com o obtido através de análises laboratoriais da corrente de topo da depropanizadora.

As predições produzidas pelos dois modelos para um período de avaliação pode ser verificada através da Figura 5.12, os valores de contaminação estão novamente normalizados em relação a especificação do 1,3 Butadieno. Os resultados fornecidos pelos modelos para o conjunto de validação mostram que ambos os modelos, na maioria das situações, conseguiram reproduzir o comportamento da composição de 1,3 Butadieno. Em geral, eles foram capazes de representar a tendência de contaminação da unidade, produzindo resultados conservativos, ou seja, a composição fornecida pelos modelos é superior a real.

O fato de os resultados produzidos serem superiores às reais concentrações não chega a ser um problema, pois por se tratar de um contaminante, as ações tomadas pelos operadores serão no sentindo de limpar a torre, provavelmente através do aumento da vazão de refluxo, o que levará a torre para uma condição de superespecificação.

Analisando-se a Figura 5.12 nota-se que para uma amostra as predições fornecidas pelo modelo foram muito diferentes do resultado da análise, ponto de mais alta contaminação. As possíveis razões para justificar essa inabilidade dos modelos em preverem tal ponto são: o pequeno número de amostras utilizadas na calibração do modelo, sendo que poucos apresentavam valores elevados de contaminação, e a saturação das variáveis secundárias, principalmente as temperaturas, impedindo que o analisador pudesse representar essa contaminação.

Outra possível causa bastante provável, seja uma falha no procedimento analítico, uma vez que as variáveis secundárias não apresentaram qualquer sinal da presença do contaminante.

Os modelos, no entanto, foram capazes de representar a elevação da contaminação acima dos índices de especificação, o que, em uma situação normal de operação, sofreria uma interferência por parte dos operadores, com uma elevação na vazão de refluxo da depropanizadora, movendo a unidade para uma situação de limpeza.

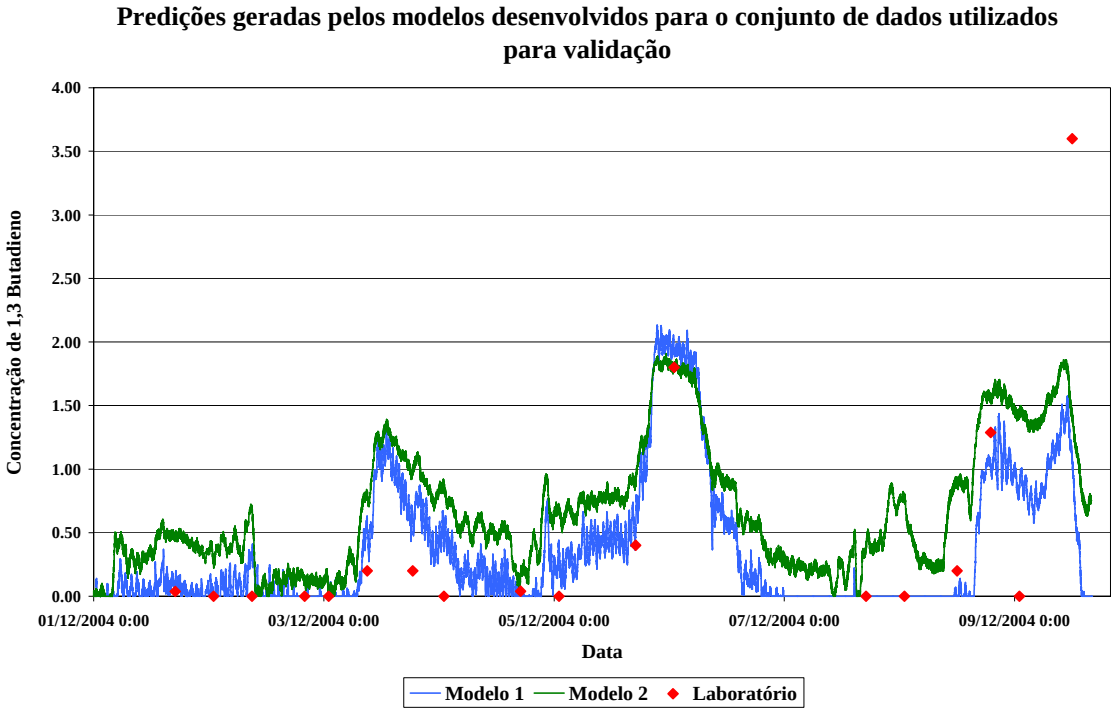


Figura 5.12: Predições gerados pelos modelos para o primeiro conjunto de validação

Os modelos desenvolvidos estão em fase final de avaliação e o escolhido será utilizado pelos responsáveis da unidade para o acompanhamento da contaminação da corrente de topo por 1,3 Butadieno e otimização econômica da unidade, através da redução na vazão de refluxo.

5.7 Ferramenta Off-Line

Os modelos desenvolvidos, apesar de terem sido desenvolvidos como simuladores podem também ser utilizados para análise do impacto de cada uma das variáveis sobre a concentração de 1,3 Butadieno na corrente de topo da Depropanizadora.

Para suprir essa necessidade, foi desenvolvida, em ambiente Excel, uma ferramenta que permita avaliar o impacto de cada uma das variáveis sobre as condições de contaminação da torre. Esse analisador virtual *off-line* permite que as variáveis sejam inseridas manualmente ou seja adquiridas automaticamente pelo sistema PIMS instalado na unidade industrial. A interface dessa ferramenta está contida na Figura 5.13.

Além da predição de contaminação, resultado principal dos modelos, é apresentada também a relação refluxo/carga para a unidade. Esse parâmetro tem por objetivo familiarizar

os operadores a essa grandeza, a qual possui relação intrínseca com o grau de contaminação da corrente de topo.

A interface tem com padrão a aquisição automática das variáveis de processo, porém se o usuário desejar o efeito da alteração de uma das variáveis sobre as predições do modelo, basta deselecionar o campo de aquisição automática da variável em questão e executar o cálculo.

Os resultados e variáveis de processo utilizados no cálculo também podem ser exportados para uma planilha, facilitando uma análise posterior, através do botão exportar. Uma vez selecionado o modelo final, os campos referentes a Modelo 1 e Modelo 2 serão substituído pela resposta do modelo selecionado.

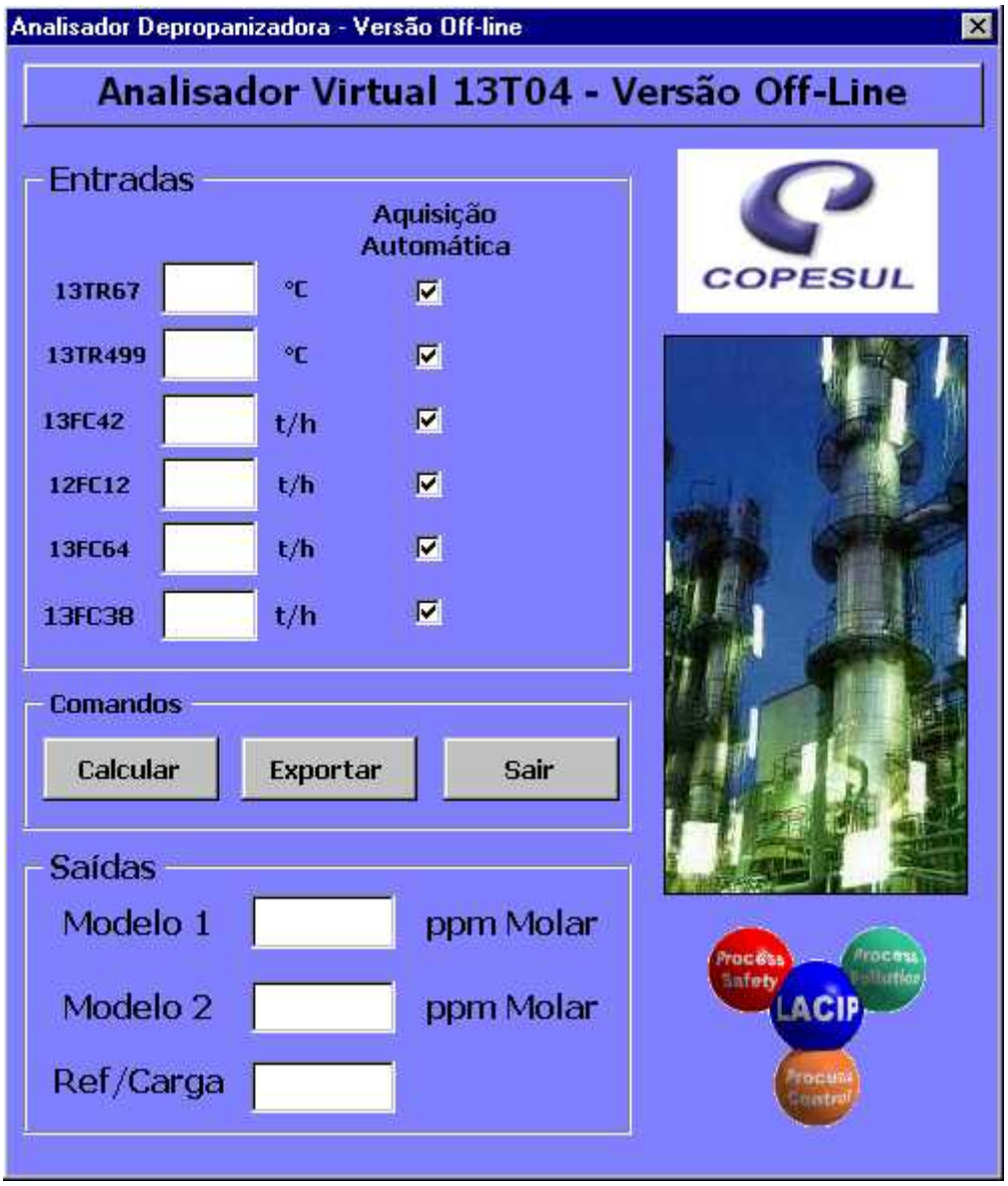


Figura 5.13: Interface para a ferramenta *off-line* desenvolvida

Capítulo 6

Conclusão

6.1 Considerações Finais

Os analisadores virtuais são um dos mais recentes desenvolvimentos dentro dos diversos campos que constituem o universo do controle de processos moderno. Sendo assim, são um terreno fértil para a consolidação e desenvolvimento de metodologias de modelagem e atualização de modelos com o objetivo de inferência de propriedades.

Os resultados obtidos durante o desenvolvimento desse trabalho representam uma contribuição para o desenvolvimento de analisadores virtuais, sobretudo aqueles que utilizam modelos baseados em técnicas de análise multivariável, como PCA, PCR e PLS. A aplicação de técnicas de seleção de variáveis, comumente empregadas na área da quimiometria, foi utilizada de forma pioneira para a seleção das temperaturas que devem fazer parte de um analisador virtual dessa natureza.

As contribuições resultantes desse trabalho abrangem o campo de seleção de variáveis em modelos baseados em análises multivariáveis, comparação entre modelos PLS lineares, não lineares (PLS Quadrático) e lineares com compensação logarítmica de composição para o desenvolvimento de analisadores virtuais para colunas de destilação e uma metodologia para o planejamento de perturbações para testes industriais.

Sucintamente as principais contribuições e conclusões obtidas durante o desenvolvimento desse trabalho foram:

- Seleção de variáveis: a combinação de estratégias de seleção de dados, normalmente empregadas na etapa de concepção de modelos, com metodologias clássicas de seleção de variáveis se mostrou como uma alternativa satisfatória na identificação das variáveis que devem compor um modelo baseado em análise multivariável. Essa afirmação é fundamentada nos resultados obtidos no Capítulo 3, onde a combinação da metodologia *y-Rank* com as técnicas de

seleção de variáveis do tipo *Forward Selection* e Algoritmos Genéticos foi capaz de identificar as verdadeiras variáveis nos estudos de caso propostos. Afora os problemas verificados nas alternativas baseadas em Algoritmos Genéticos, essa técnica foi capaz de identificar as variáveis que resultaram em modelos com capacidade preditiva superior. Ainda no campo da seleção de variáveis foi possível obsrvar, fundamentado nos modelos gerados no Capítulo 4, resultados encontrados na literatura de que as metodologias de seleção de variáveis da família *Stepwise* tendem a produzir modelos mais enxutos, parando no primeiro mínimo local. Algoritmos Genéticos, no entanto, por serem métodos heurísticos tem menor probabilidade de ficarem presos em mínimos locais, o que leva a seleção de um maior conjunto de variáveis podendo produzir um melhor modelo.

- Analisadores Virtuais: a comparação sistemática de diferentes alternativas de modelos PLS, lineares e não lineares, mostrou que os modelos não lineares, baseados em PLS Quadráticos, apresentaram performance superior aos PLS lineares, se mostrando uma alternativa simples e eficaz de contornar as não linearidades presentes em processos industriais. A compensação logarítmica da composição do componente de interesse em combinação com a técnica PLS linear também se mostrou uma alternativa extremamente poderosa, pois consegue atenuar as não lineares da variável de resposta e mantém a elevada capacidade extrapolativa, típica dos modelos lineares.

- Planejamento de Perturbações: a utilização de simulações estacionárias e dinâmicas para o planejamento de perturbações a serem realizadas em unidades industriais se mostrou uma ferramenta poderosa para a identificação de variáveis sensíveis e regiões ricas em informação, racionalizando, dessa forma, o número de perturbações a serem realizadas na unidade industrial, o que resulta em redução dos custos do testes. As simulações também se mostraram como uma forma de convencimento do pessoal da operação sobre práticas operacionais conservativas, que resultam em perda de performance da unidade.

- Aplicação Industrial: a aplicação industrial, embora necessite ainda de ajustes finais, foi capaz de mostrar as potencialidades das técnicas PLS para a construção de analisadores virtuais. O analisador virtual desenvolvido foi capaz de apontar tendências de contaminação, permitindo aos operadores uma intervenção antes que a contaminação acarrete problemas de especificação de outras correntes. Além do próprio analisador, única indicação disponível para os operadores sobre o estado de contaminação da corrente de topo da depropanizadora, surgiram desse trabalho outras melhorias, sendo a mais significativa a redução na vazão de refluxo da unidade, o que terminou por gerar grande economia energética em todo o sistema. Pode-se citar ainda como melhoria a identificação e sugestão de melhoria do sistema de refervimento da torre estudada, que quando implementada poderá resultar em economia de vapor e considerável melhoria do sistema de controle da torre.

Outra importante conclusão, adversa a qualquer resultado desse trabalho, resultante do convívio com o pessoal de operação durante os teste industriais, é a importância do convencimento e participação desses na realização de qualquer modificação de práticas operacionais. A participação do corpo operacional é fundamental para a mudança de qualquer paradigma operacional existente.

6.2 Sugestões para trabalhos futuros

Como mencionada anteriormente, a área de analisadores virtuais é relativamente nova e possui um potencial enorme para futuros desenvolvimentos, tanto em relação aos campos de aplicação quanto a desenvolvimentos de modelos e alternativas de atualização dos mesmos.

A exploração de outras alternativas de modelos PLS não lineares, tais como PLS combinados com redes neurais e PLS combinado com lógica Fuzzy, parece ser o próximo passo natural na exploração de analisadores virtuais baseados em técnicas de análise multivariável.

A utilização de modelos do tipo PLS dinâmicos também se mostra com um caminho promissor para o desenvolvimento de modelos, não só para uso em analisadores virtuais, mas também como modelos de processo para outras aplicações na área de controle, como identificação de modelos para controle avançado.

Em relação aos métodos de seleção de variáveis, sugere-se a utilização de métodos mais modernos de otimização heurística, destacando-se os métodos da Colônia de Formigas e Enxame de Partículas. Ainda na linha da seleção de variáveis fica a sugestão da elaboração de uma função capaz de predizer o número de modelos avaliados pelas técnicas de algoritmos genéticos em função do número de gerações, tamanho da população inicial e em função das taxas de mutação, reprodução e clonagem, evitando assim a avaliação exagerada de modelos.

A utilização de técnicas de adaptação de modelos, como Mínimos Quadrados Recursivos e Filtro de Kalman, em combinação a modelos do tipo PLS e a elaboração de uma ferramenta de desenvolvimento de analisadores virtuais baseados em metodologias de análise multivariável em combinação a essas técnicas de atualização é uma sugestão prática, já que as ferramentas disponíveis comercialmente não dispõem de tais procedimentos.

Referências Bibliográficas

AKAIKE, H., **A NEW LOOK AT THE STATISTICAL MODEL IDENTIFICATION**, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, V.19, 716-723, 1974

BAFFI, G., MARTIN, E. B., MORRIS, A. J., **NON-LINEAR PROJECTION TO LATENT STRUCTURES REVISETED: THE QUADRATIC PLS ALGORITHM,** COMPUTERS & CHEMICAL ENGINEERING, V.23 , 395-411, 1999A,

BAFFI, G., MARTIN, E. B., MORRIS, A. J., **NON-LINEAR PROJECTION TO LATENT STRUCTURES REVISETED: THE NEURAL NETWORK PLS ALGORITHM**, COMPUTERS & CHEMICAL ENGINEERING, V.23, 1293-1307, 1999B.

BARATTI, R., BERTUCCO, A., DA ROLD, A., MORBIDELLI, M., **DEVELOPMENT OF A COMPOSITION ESTIMATOR FOR BINARY DISTILLATION COLUMNS. APPLICATION TO A PILOT PLANT,** CHEMICAL ENGINEERING SCIENCE, V.50(10), 1541-1550, 1995.

BARROS, A. S., RUTLEDGE, D. N., **GENETIC ALGORIOTHM APPLIED TO SELECTION OF PRINCIPAL COMPONENTS,** CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.40, 65-81, 1998

BLANCO, M., COELLO, J., ITURRIAGA, H., MASPOCH, S., PAGES, J., **NIR CALIBRATION IN NON-LINEAR SYSTEMS: DIFFERENT PLS APPROACHES AND ARTIFICIAL NEURAL NETWORKS,** CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.50, 75–82, 2000.

BAUMANN, K, **CROSS-VALIDATION AS THE OBJETIVE FUNCTION FOR VARIABLE-SELECTION TECHNIQUES**, TRENDS IN ANALYTICAL CHEMISTRY, V.22(6), 395-406, 2003

BISCAIA JR., E. C., VIEIRA, R. C., **MÉTODOS HEURÍSTICOS DE OTIMIZAÇÃO,** CURSO DA ESCOLA PILOTO PEQ-COPPE, 2004

BROADHURST, D., GOODACRE, R., JONES, A., ROWLAND, J. J., KELL, D. B., **GENETIC ALGORITHM AS A METHOD FOR VARIABLE SELECTION IN MULTIPLE LINEAR REGRESSION AND PARTIAL LEAST SQUARES REGRESSION WITH APPLICATIONS TO PYROLYSIS MASS SPECTORMETRY,** ANALYTICA CHIMICA ACTA, V.348, 71-81, 1997

BROWNE, M. W., **CROSS- VALIDATION METHODS**, JOURNAL OF MATHEMATICAL PSYCHOLOGY, V.44, 108-132, 2000

BROWN, R. G., HWANG, P. Y. C., **INTRODUCTION TO RANDOM SIGNALS AND APPLIED KALMAN FILTERING**, SECOND EDITION, JONH WILEY AND SONS, INC, 1992

CHEN, J., LIU, K., **ON-LINE BATCH PROCESS MONITORING USING DYNAMIC PCA AND DYNAMIC PLS MODELS,** CHEMICAL ENGINEERING SCIENCE, V.57, 63-75, 2002

CHIANG, L H., RUSSEL, E. L., **FAULT DETECTION AND DIAGNOSIS IN INDUSTRIAL SYSTEMS**, SPRINGER-VERLAG LONDON LIMITED, 2001

COGWILL, M. C., HARVEY, R. J., **A GENETIC ALGORITHM APPROACH TO CLUSTER ANALYSIS,** COMPUTERS AND MATHEMATICS, V.37, 99-108, 1999

CONZ, V., **DESENVOLVIMENTO DE ANALISADORES VIRTUAIS APLICADOS A COLUNAS DE DESTILAÇÃO INDUSTRIAIS,** DISSERTAÇÃO DE MESTRADO, PPGEQ/UFRGS, 2005

DASZYKOWSKI, M., WALCZAK, B., MASSART, D. L., **REPRESENTAIVE SUBSET SELECTION,** ANALYTICA CHYMICA ACTA, V.468, 91-103, 2002

DAYAL, B. S., MacGREGOR, J. F., **RECURSIVE EXPONENTIALLY WEIGHTED PLS AND ITS APPLICATIONS TO ADAPTIVE CONTROL AND PREDICTION**, JOURNAL OF PROCESS CONTROL, V.7(3), 169-179,1997

DENN, M. M., **PROCESS MODELING**, LONGMAN INC., NEW YORK, 1986

DING, Q., SMALL, G. W., ARNOLD M. A., **EVALUATION OF NONLINEAR MODEL BUILDING STRATEGIES FOR THE DETERMINATION OF GLUCOSE IN BIOLOGICAL MATRICES BY NEAR-INFRARED SPECTROSCOPY,** ANALYTICA CHIMICA ACTA, V.384(3), 333-343,1999.

FABI, VUB, **THE STANDARD TOOLBOX FOR MATLAB**, 1997
HTTP://WWW.VUB.AC.BE/FABI/

FERRÉ, J., RIUS, F. X., **CONSTRUCTING D-OPTIMAL DESIGNS FROM A LIST OF CANDIDATE SAMPLES**, TRENDS IN ANALYTICAL CHEMESTRY, V.16 (2), 70 - 73,1997.

FINKLER, T. F., **DESENVOLVIMENTO DE UMA FERRAMENTA PARA OBTENÇÃO DE MODELOS EMPÍRICOS** – DISSERTAÇÃO DE MESTRADO, PPGEQ/UFRGS, 2003

FLÅTEN, G. H., WALMSLEY, A. D., **A DESIGN OF EXPERIMENT APPROACH INCORPORATING LAYERD DESIGNS FOR CHOOSING THE RIGHT CALIBRATION MODEL**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.73, 55-66, 2004

FORTUNA L., GRAZIANI S., XIBILIA M. G., **SOFT SENSORS FOR PRODUCT QUALITY MONITORING IN DEBUTANIZER DISTILLATION COLUMNS**, CONTROL ENGINEERING PRACTICE, V.13(4), 499-508, 2005.

GELADI, P, KOWALSKI, B. R., **PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL**, ANALYTICA CHIMICA ACTA V.185, 1-17, 1986

GELB, A, **APPLIED OPTIMAL ESTIMATION**, MIT PRESS, CAMBRIDGE, MA, 1974.

GESTAL, M., GÓMEZ-CARRACEDO, M. P., ANDRADE, J.M., DORADO, J., FERNÁNDEZ, E., PRADA, D., PAZOS, A., **CLASSIFICATION OF APPLE BEVERAGES USING ARTIFICIAL NEURAL NETWORKS WITH PREVIOUS VARIABLE SELECTION**, ANALYTICA CHIMICA ACTA, V.524, 225–234,2004

GRANGER, C. W. J., **STRATEGIES FOR MODELLING NONLINEAR TIME-SERIES RELATIONSHIPS**, THE ECONOMIC RECORD, V.69, 233-238, 1993

HAN, S. H.,YANG, H., **SCREENING IMPORTANT DESIGN VARIABLES FOR BUILDING A USABILITY MODEL: GENETIC ALGORITHM-BASED PARTIAL LEAST SQUARES APPROACH,** INDUSTRIAL ERGONOMICS, V.33, 159-171, 2004

HÖSKULDSSON, A., **DIMENSION OF LINEAR MODELS**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.32, 37-55, 1996

HYÖTYNIEMI, H., **MULTIVARIATE REGRESSION – TECHNIQUES AND TOOLS**, HELSINKI UNIVERSITY OF TECHLOGY, CONTROL ENGINEERING LABORATORY, 2001
http://saato014.hut.fi/hyotyniemi/publications/01_report125.htm

KAMOHARA, H., TAKINAMI, A., TAKEDA, M., KANO, M., HASEBE, S., HASHIMOTO, I., **IMPROVEMENT OF DISTILLATION COMPOSITION CONTROL BY USING PREDICTIVE INFERENTIAL CONTROL TECHNIQUE**, JOURNAL OF CHEMICAL ENGINEERING OF JAPAN, V.34(8),1026-1032, 2001.

KANO, M., MIYAZAKI, K., HASEBE, S., HASHIMOTO, I., **INFERENTIAL CONTROL SYSTEM OF DISTILLATION COMPOSITION USING DYNAMIC PARTIAL LEAST SQUARES REGRESSION**, JOURNAL OF PROCESS CONTROL, V.10, 157-166, 2000

KISTER, H. Z., **DISTILLATION OPERATION**, MCGRAW-HILL, CAPÍTULO 18, 545-576, 1990

KOHMANN, C. M., **RELATÓRIO DE ESTÁGIO SUPERVISIONADO**, DEQUI/UFRGS, 2004

KOMULAINEN, T, SOURANDER M., JÄMSÄ-JOUNELA, S., **AN ONLINE APPLICATION OF DYNAMIC PLS TO A DEAROMATIZATION PROCESS**, COMPUTERS AND CHEMICAL ENGINEERING, V.28(12), 2611-2619, 2004.

KOURTI, T., MacGREGOR, J. F., **PROCESS ANALYSIS, MONITORING AND DIAGNOSIS, USING MULTIVARIATE PROJECTION METHODS,** CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.28, 3-21, 1995.

KOWALSKI B., GERLACH R., WOLD H., **CHEMICAL SYSTEMS UNDER INDIRECT OBSERVATION**, in SYSTEMS UNDER INDIRECT OBSERVATION (K. JORESKOG AND H. WOLD (EDS.)), 1982, 191-209, NORTH-HOLLAND,AMSTERDAM.

KRUGER U., CHEN Q., SANDOZ D. J., MCFARLANE R. C., **EXTENDED PLS APPROACH FOR ENHANCED CONDITION MONITORING OF INDUSTRIAL PROCESSES,** AICHE JOURNAL, V.47(9), 2076-2091,2001.

KRESTA, J., MARLIN, T., MacGREGOR, J., **CHOOSING INFERIENTIAL VARIABLES FOR MULTICOMPONENT DISTILLATION USING PROJECTION TO LATENT STRUCTURES**, CSCHE ANNUAL MEETING, HALIFAX, NOVA ESCÓCIA, JULHO 1990A

KRESTA, J., MARLIN, T., MacGREGOR, J., **CHOOSING INFERIENTIAL VARIABLES FOR MULTICOMPONENT DISTILLATION USING PROJECTION TO LATENT STRUCTURES**, AICHE ANNUAL MEETING, HALIFAX, CHIGAGO, ILINNOIS, NOVEMBRO 1990B

KRESTA, J. V., MARLIN, T. E., MacGREGOR, F. J., **DEVELOPMENT OF INFERENTIAL PROCESS MODELS USING PLS**, CUMPUTERS CHEM. ENGNG, V.18(7), 597-611, 1994

LEARDI, R., SEASHOLTZ, M. B., PELL, R. J., **VARIABLE SELECTION FOR MULTIVARIATE CALIBRATION USING A GENETIC ALGORITHM: PREDICTION OF ADDITIVE CONCENTRATIONS IN POLYMER FILMS FROM FOURIER TRANSFORM-INFRARED SPECTRAL DATA,** ANALYTICA CHIMICA ACTA, V.461, 189–200, 2002.

LI, B., MARTIN, E. B., MORRIS, A. J., **BOX-TIDWELL TRANSFORMATION BASED PARTIAL LEAST SQUARES REGRESSION**, COMPUTERS AND CHEMICAL ENGINEERING, V.25, 1219-1233, 2001

LI, B., MORRIS, J., MARTIN, E. B., **MODEL SELECTION FOR PARTIAL LEAST SQUARES REGRESSION,** CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.64, 79-89, 2002

LUYBEN, W. L., **PRACTICAL DISTILATION CONTROL**, Van NOSTRAND REINHOLD, NOVA IORQUE, 1992

MacGREGOR, J. F., KOURTL. T., **STATISTICAL PROCESS CONTROL OF MULTIVARIATE PROCESSES,** CONTROL ENGINNERING PRACTICE, V.3(3), 403-414, 1995.

McSHANE, M. J., CAMERON, B. D., COTE, G. L., MOTAMEDI, M., SPIELGEMAN, C. H., **A NOVEL PEAK-HOPPING STEPWISE FEATURE SELECTION METHOD WITH APPLICATIONS TO RAMAN SPECTROSCOPY**, ANALYTICA CHIMICA ACTA, V.388, 251-264, 1999

MEJDELL, T., SKOGESTAD, S., **COMPOSITION CONTROL OF DISTILLATION COLUMNS USING MULTIPLE TEMPERATURE MEASURENTS**, AICHE ANNUAL MEETING. CHIGAGO, ILINNOIS, NOVEMBRO 1990

MOHR, T., **DESENVOLVIMENTO DE UM ANALISADOR VIRTUAL PARA UMA PLANTA DE POLIETILENO DE ALTA DENSIDADE**, DISSERTAÇÃO DE MESTRADO – PPGEQ – UFRGS, 2004

MONTGOMERY, D. C., PECK, E. A., **INTRODUCTION TO LINEAR REGRESSION ANALYSIS,** JOHN WILEY & SONS, NOVA IORQUE, 1982

NISHIMO, T., NAGAMACHI, M., TSUCHIYA, T., MATSUBARA, Y., COOPER, D., **A GENETICS-BASED APPROACH TO AUTOMATED DESIGN BASED ON KANSEI ENGINEERING,** PROCEEDINGS OF THE THIRD PAN-PACIFIC CONFERENCE ON OCCUPATIONAL ERGONOMICS, SEOUL, KOREA., 162-166, 1994

OISIOVICI, R. M., CRUZ, S. L., **STATE ESTIMATION OF BATCH DISTILLATION COLUMNS USING AN EXTENDED KALMAN FILTER,** CHEMICAL ENGINEERING SCIENCE, V.55(20), 4667-4680, 2000.

QI, M, ZHANG, G. P., **AN INVESTIGATION OF MODEL SELECTION CRITERIA FOR NEURAL NETWORK TIME SERIES FORECASTING**, EUROPEAN JOURNAL OF OPERATIONAL RESEARCH V.132, 666-680, 2001

QIN, S. J., MCAVOY, T. J., **NONLINEAR PLS MODELING USING NEURAL NETWORKS**, COMPUTERS CHEMICAL ENGINEERING, V.16(4), 379-391, 1992

van der BERG, F., **INTRODUCTION TO MATLAB + MATHEMATICAL ASPECTS OF BILINEAR FACTOR MODELS (PCA AND PLS),** THE ROYAL VETERINARY AND AGRICULTURAL UNIVERSITY, KVL. HTTP://WWW.MODELS.KVL.DK/COURSES/INTROMATLAB/INDEX.ASP

THAM, M. T., MONTAGUE, G. A., MORRIS A. J., LANT P. A., **SOFT-SENSORS FOR PROCESS ESTIMATION AND INFERENTIAL CONTROL,** JOURNAL OF PROCESS CONTROL, V.1, 3-14, 1991.

TSUCHIYA, T. MAEDA, T., MATSUBARA, Y., NAGAMACHI, M., **A FUZZY RULE INDUCTION METHOD USING GENETIC ALGORITHM**, INTERNATIONAL JOURNAL OF INDUSTRIAL ERGONOMICS, V.18, 135-145, 1996

WASSERMANN, G. S., SUDJIANTO, A., **ALL SUBSETS REGRESSION USING GENETIC SEARCH ALGORITHM**, COMPUTERS & INDUSTRIAL ENGINEERING, V.27, 489-492, 1994

WELCH, G., BISHOP, G., **AN INTRODUCTION TO KALMAN FILTER**, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL, 2000

WILSON, D. J. H., IRWIN, G. W. E LIGHTBODY, G., **NONLINEAR PLS MODELING USING RADIAL BASIS FUNCTIONS**, AMERICAN CONTROL CONFERENCE, ALBUQUERQUE, NOVO MÉXICO, 4 A 6 DE JUNHO DE 1997

WOLD, S., MARTENS H., RUSSWURM H., **FOOD RESEARCH AND DATA ANALYSIS**, APPLIED SCIENCE PUBLISHERS, LONDON, 1983.

WOLD, S., RUHE, A., WOLD, H., DUNN W., **THE COLLINEARITY PROBLEM IN LINEAR REGRESSION. THE PARTIAL LEAST SQUARES (PLS) APPROACH TO GENERALIZED INVERSES**, SIAM JOURNAL.OF SCIENTIFIC AND STATISTICAL COMPUTING., V.5, 735-743, 1984.

WOLD, S., KETTANEH-WOLD, N., SKAGERBERG, B., **NONLINEAR PLS MODELING**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.7, 53-65, 1989.

YIN, C., LIUA, X., GUOB, W., LINA, T., WANGA, X., WANGA, L., **PREDICTION AND APPLICATION IN QPSR FOR AQUEOUS SOLUBILITY OF SULFUR-CONTAINING AROMATIC ESTERS USING GA-BASED MLR WITH QUANTUM DESCRIPTORS**, WATER RESEARCH, V.36, 2975-2982, 2002

ZAMPROGNA, E., BAROLO, M., SEBORG D. B, **OPTIMAL SELECTION OF SOFT SENSOR INPUTS FOR BATCH DISTILLATION COLUMNS USING PRINCIPAL COMPONENT ANALYSIS,** JOURNAL OF PROCESS CONTROL, V.15, 39–52, 2005.

Apêndice 1

Rotinas Desenvolvidas

Esse apêndice contém algumas das rotinas desenvolvidas e utilizadas no desenvolvimento desse trabalho. Serão apresentadas as rotinas desenvolvidas para os modelos PLS Lineares, PLS não Lineares, Seleção de Variáveis e Seleção de Dados.

As implementações das diferentes estratégias de PLS foram realizadas de forma a apresentarem uma estruturação de entrada e saída de dados bastante similar, sendo que a partir da estratégia linear serão descritos os parâmetros utilizados pelas funções responsáveis pela etapa de modelagem.

1.1 PLS Linear

Os códigos referentes a implementação da estratégia de PLS linear foram adquiridos em um *Toolbox* desenvolvido por van der BERG (2001) e pode ser obtido através do seguinte endereço: : http://www.models.kvl.dk/source/MBToolbox/.

Apesar de o núcleo do PLS linear ter sido obtido do referido *Toolbox* foi necessário o desenvolvimento de uma interface para que se obtivesse um padrão de saídas. Essa interface é apresentada a seguir, servindo também para a descrição dos principais parâmetros contidos nas funções. Os parâmetros são listados e descritos na tab, sendo esses comuns a todas as funções desenvolvidas.

Para a utilização das funções é necessário que se informe a matriz $\boldsymbol{X}$, contendo os dados referentes as variáveis explicativas, a matriz $\boldsymbol{y}$, contendo os dados referentes as variáveis de resposta, e o número de fatores, ou variáveis latentes, a serem considerados, representado pelo parâmetro *nf*.

```
function [yp,modelo,stat] = pls(x,y,nf);
%Dimesões do problema
[lx,cx] = size(x);
[ly,cy] = size(y);
%escalonamento dos dados
[xa,mx,sx] = autosc(x);
[ya,my,sy] = autosc(y);
%criação do modelo
[T,P,W,U,Q,B,ssq,Ro,Rv,Lo,Lv,iter] = mypls(xa,ya,nf);
%predição das variáveis de resposta pelo modelo
Yps = xa*W*inv(P'*W)*Q';
for i=1:cy
   yp(:,i) = Yps(:,i)*sy(i)+my(i);
end
%criação do modelo de forma explícita
Bpls = (W*inv(P'*W)*Q'); %regressores do modelo PLS
Bd = Bpls*sy./sx'; %regressores para modelo explícito
Bo = -mx*Bd+my; %BIAS do modelo direto
%Estruturas de saída
modelo.Bdir = Bd; modelo.Bo = Bo;
modelo.u = U; modelo.q = Q; modelo.b= B; modelo.ymed = my; modelo.ystd = sy;
modelo.t = T; modelo.p = P; modelo.w = W; modelo.xmed = mx; modelo.xstd = sx;
stat.ssq = ssq; stat.ro = Ro; stat.rv = Rv; stat.iter = iter;
stat.lo = Lo; stat.lv = Lv;
```

As saídas geradas pela função totalizam três parâmetros sendo duas estruturas (modelo e stat) contendo informações sobre o modelo e parâmetros estatísticos do mesmo, e uma matriz (***yp***)formada pelas predições produzidas pelo modelo para a o conjunto de calibração. Os parâmetros internos de cada estrutura são apresentados na Tabela 1.1.

Tabela 1.1: Parâmetros utilizados nas funções implementadas

Parâmetro	Descrição
Modelo.Bdir	Vetor de regressores que relacionam diretamente as variáveis explicativas e de resposta
Modelo.Bo	BIAS fixo do modelo
Modelo.u	Matriz de projeções das variáveis de resposta (output scores)
Modelo.q	Matriz com os vetores de projeção das variáveis de resposta (output weights)
Modelo.b	Matriz de regressores internos
Modelo.ymed	Vetor contendo as médias das variáveis de resposta
Modelo.ystd	Vetor contendo os desvios padrões das variáveis de resposta
Modelo.t	Matriz de projeção das variáveis explicativas (input scores)
Modelo.p	Matriz de projeção das variáveis explicativas (input loadings)
Modelo.w	Matriz contendo os pesos das variáveis explicativas (input weights)
Modelo.xmed	Vetor contendo as médias das variáveis explicativas
Modelo.xstd	Vetor contendo os desvios padrões das variáveis explicativas
Stat.ssq	Matriz contendo a variância acumulada por variável latentes para as variáveis explicativas e de resposta
Stat.ro	Matriz de resíduos das variáveis explicativas
Stat.rv	Matriz de resíduos das variáveis de resposta
Stat.iter	Vetor contendo o número de iterações necessárias para a convergência em cada fator
Stat.lo	Matriz contendo as Leverages para as variáveis explicativas
Stat.lv	Matriz contendo as Leverages para as variáveis de resposta

Leverage é uma medida estatística que indica qual a influência de uma amostra sobre os parâmetros de um modelo, sendo que quanto maior esse valor, mais a amostra influencia na determinação dos parâmetros do modelo.

As demais estratégias de PLS implementadas foram simples transcrições de algoritmos encontrados na literatura, em função disso seus códigos não serão apresentados.

1.2 Interfaces de Acesso

Como as implementações seguiram uma estrutura modular, isso é, todas as funções necesitam das mesmas entradas e fornecem as mesmas saídas, forma desenvolvidas interfaces para que o acesso fosse facilitado. A primiera interface é utilizada para a calibração/construção dos modelos, enquanto que a segunda é responsável pela predição/simulação do modelos já construídos.

1.2.1 Interface de Calibração

O quadro abaixo contém função responsável pela calibração dos modelos, possibilitando o acesso a todas as funções desenvolvidas a partir de um só local.

```
function [yp,modelo,stat] = fpls(fun,x,y,nf);
%Função genérica para modelos PLS
%Entradas
%   fun - string correspondente ao modelo a ser utilizado
%       PLS - PLS linear
%       QPLS - PLS Quadrático com atualização dos vetores pesos w
%       BTPLS - Box-Tidwell PLS com atualização dos vetores pesos w
%       RBFPLS - PLS combinado com Rede Neuronal do tipo RBF com atualização dos vetores
pesos w
%   x - matriz contendo as variáveis explicativas - [nXm]
%   y - matriz contendo as variáveis de resposta - [nXk]
%   nf - número de variáveis latentes a ser consideradas - escalar
%Saídas
%   yp - predição gerada pelo modelo
%   modelo - estrutura contendo as informações do modelo gerado
%   stat - estrutura contendo as informações de análise do modelo gerado
%Dimesões do problema
[lx,cx] = size(x); [ly,cy] = size(y);
if nf > cx
   s =['Número de fatores inválido. O número de fatores deve ser menor ou igual a
'(num2str(cx))] error(s)
end
if lx ~= ly
    s = ['ERRO: número de amostras em x (' int2str(lx) ') e y (' int2str(ly) ') devem
ser igauis.']; error(s)
end
[yp,modelo,stat] = feval(fun,x,y,nf);
```

Para a sua utilização, além das matrizes contendo os dados referentes as variáveis explicativas e de resposta, ***X*** e ***y***, e o número de fatores, *nf*, deve ser informado também a tipologia do modelo PLS a ser calibrado. As palavras chave para um das técnicas são apresentadas na Tabela 1.2.

Tabela 1.2: Palavres chave para a utilização da interface de calibração de modelo

Palavra Chave	Modelo Associado
PLS	PLS Linear
QPLS	PLS Quadrático
BTPLS	PLS baseado em transformações Box-Tidwell
RBFPLS	PLS com mapeamento interno realizado através de rede neural com função de ativação de base radial
CSPLS	PLS com mapeamento interno realizado através de rede neural com função de ativação do tipo sigmóide centrada na origem

1.2.2 Interface de Simulação

A interface de simulação tem por objetivo produzir predições para as variáveis de saída tendo como base os modelos desenvolvidos com a utilização da Interface de calibração. O quadro abaixo contém a estrutura dessa função;

```
function yp = fplsfit(fun,x,modelo);
%Entradas
%   fun - string correspondente ao modelo a ser utilizado
%       PLS - PLS linear
%       QPLS - PLS Quadrático com atualização dos vetores pesos w
%       BTPLS - Box-Tidwell PLS com atualização dos vetores pesos w
%       RBFPLS - PLS combinado com Rede Neuronal do tipo RBF com atualização dos vetores
pesos w
%   x - matriz contendo as variáveis explicativas - [nXm]
%   modelo – modelo PLS a ser simulado
fun = [fun 'f'];
yp = feval(fun,x,modelo);
```

A função acima é responsável por acionar cada uma das funções utilizadas para a simulação/predição de novas amostras. Os parâmetros necessários para a sua utilização são a categoria do modelo, conforme Tabela 1.2, a nova matriz de variáveis explicativas $\boldsymbol{X}$ e o modelo correspondente a metodologia adotada.

1.3 Validação Cruzada

Para a realização do procedimento de validação cruzada, procedimento necessário para a determinação da real dimensão do sistema, foi desenvolvida uma função. Essa função utiliza a estratégia de divisão dos dados em múltiplos blocos de calibração e validação.

Os parâmetros de entrada para essa função são o tipo de modelo a ser utilizado, com a utilização das palavras chaves da Tabela 1.2, as matrizes $\boldsymbol{X}$ e $\boldsymbol{y}$, as matrizes contendo os dados referentes as variáveis explicativas e de resposta, e o número de blocos de calibração e validação a serem utilizados, o parâmetro *ng*; O procedimento LOO pode ser facilmente utilizado, fazendo *ng* igaul ao número de amostras na matriz $\boldsymbol{X}$.

```
function [ip,best,sm_err] = fplscv(fun,x,y,ng);
%Escalonamento dos Dados
[xa,mx,sx] = autosc(x); [ya,my,sy] = autosc(y);
%Dimensões do problema
[lx,cx] = size(x); [ly,cy] = size(y);
%Número máximo de dimensões a serem consideradas
max_dim = 15;
if cx > max_dim
   nf_max = max_dim;
else
   nf_max = cx;
end
err_gp = [];
for ngr = 1:ng
   %Criação dos grupos de validação e calibração
   vi=(ngr:ng:lx);
   xv=xa(vi,:);
   yv=ya(vi,:);
   xc=xa;yc=ya;
   xc(vi,:)=[];
   yc(vi,:)=[];
   %Variâncias em X e Y
   varinx=sum(xc.^2); variny=sum(yc.^2);
   %Calibração do modelo
   [yp_c,modelo,stat] = fpls(fun,xc,yc,nf_max);
   %Avaliação do modelo - predição no conjunto de validação
      yp_v = zeros(length(yv),nf_max);
   for nf = 1:nf_max
      yp_v(:,nf) = fplsfitcv(fun,xv,modelo,nf);
   end
   yv_m = yv*ones(1,nf_max);
   [lyv,cyv] = size(yv);
   if lyv>1;
      pred_err = sum((yv_m-yp_v).^2);
   else
      pred_err = (yv_m-yp_v).^2;
   end;
   err_gp = [err_gp;pred_err];
end
sm_err = sum(err_gp); [best,ip] = min(sm_err);
```

O número ótimo de fatores é determinado através da minimização do erro quadrático. A função retorna o número ótimo de fatores, o parâmetro de performance referente a esse fator e o vetor contendo o erro quadrático da variável de saída para cada um dos fatores.

1.4 Seleção de Variáveis

Para os procedimentos de seleção de variáveis foram criadas duas rotinas principais, uma para a seleção via algoritmos genéticos e outra para a seleção por adição (*Forward Selection*). Foram implementadas funções específicas para a geração de cada uma das quatro alternativas de segregação de dados nos conjuntos de calibração e validação, de forma que elas pudessem ser comparadas. A seguir são apresentados os códigos desenvolvidos para a seleção de variáveis e elaboração dos conjuntos de calibração e validação.

1.4.1 Algoritmos Genéticos

Implementações de algoritmos genéticos para a seleção de variáveis podem ser obtidas facilmente na internet. O primeiro código analisado foi obtido em http://www.models.kvl.dk/source/GAPLS/. A implementação desse *Toolbox* utiliza apenas os operadores de mutação e *cross-over*.

Através da literatura, foram encontrados novos operadores, como a clonagem e a reinicilização periódica. Para utilizar esses operadores, que aumentam a performance dos algoritmos genéticos, foi implementada uma função própria de algoritmos genéticos, contendo os operadores de *cross-over*, mutação, clonagem e reinicilização periódica, utilizando codificação binária. Além da inclusão desses dois novos operadores, foi introduzido um mecanismo de limitação do número de variáveis, forçando o algoritmo a procurar o melhor conjunto de variáveis, quando se impõem uma restrição na dimensão do modelo.

```
function [press_min,vb,avalia] = agb(mod,sub,x,y,nf)
%mod: pls,qpls,btpls,cspls,rbfpls
%sub: randga, ranksga, ksga, ksmga
rand('state',sum(100*clock));
nind = 100;    % numero de individuos iniciais
ngens = 10;    % numero de genes
ngera = 40;    % numero de geracoes
trep = 90;     % taxa de reproducao
tmut = 10;     % taxa de mutacao
telit = 40;    % taxa de individuos da elite
nmaxvar = 6;   % numero maximo de variaveis
freset = 1;    % flag do reset da populacao
nreset = 5;    % numero de geracoes para resetar a população
nindarca   = 10;   % numero de individuos q vive apos chacina
gerain = 1;
% geracao da populacao inicial
if freset
   nciclos = ceil(ngera/nreset);
   ngera = nreset;
end
n_gerain = 0; n_gera = 0; n_gera2 = 0;
for k = 1:nciclos
   if gerain
      for i = (nindarca+1):nind
         n_gerain = n_gerain+1;
         id = randb(nmaxvar,ngens);
         nz = find(id);
         if isempty(nz)
            pos = randunifd(1,ngens);
            id(pos) = 1;
         end
         ind(i,:) = id;
         apt(i) = feval(sub,mod,x,y,ind(i,:),nf);
      end
      gerain = 0;
   else
      for i = (nindarca+1):nind
         n_gera = n_gera +1;
         id = randb(nmaxvar,ngens);
         nz = find(id);
         if isempty(nz)
            pos = randunifd(1,ngens);
            id(pos) = 1;
         end
         ind(i,:) = id;
         apt(i) = feval(sub,mod,x,y,ind(i,:),nf);
      end
   end
   for ger = 1:ngera
      [apt,indc] = sort(apt);
      apt = apt(end:-1:1);
      ind = ind(indc(end:-1:1),:);
      for i=1:ceil(nind*tmut/100)% gera os individuos mutados
         indmut = ceil(rand(1)*nind);
         gact = find(ind(indmut,:));
         if sum(gact) == 0
            ind(indmut,ceil(rand(1)*ngens)) = 1;
            gact = find(ind(indmut,:));
         end
         gnact = find(ind(indmut,:) == 0);
         igeac = ceil(rand(1)*length(gact));
         igenac = ceil(rand(1)*length(gnact));
         indnew(i,:) = ind(indmut,:);
         indnew(i,gact(igeac)) = 0;
         indnew(i,gnact(igenac)) = 1;
      end
```

```
nacind = i;
      for i=1:ceil(nind*trep/200)% gera os individuos reproduzidos
         indnew(nacind + i,:) = [ind(2*i-1,1:ceil(ngens/2)), ...
               ind(2*i,(ceil(ngens/2)+1):ngens)];
         nuns = sum(indnew(nacind + i,:));
         while nuns > nmaxvar
            dife = nuns - nmaxvar;
            gzera = ceil(rand(1,dife)*ngens);
            indnew(nacind + i,gzera) = 0;
            nuns = sum(indnew(nacind + i,:));
         end
      end
nacind = i + nacind;
% gera os individuos da elite (Sobreviventes a reinicilização)
      nineli = ceil(nind*telit/100);
      indnew((1:nineli)+nacind,:) = ind(1:nineli,:);
      nacind = nacind + nineli;
      % reatribuicao das variáveis
      ind = indnew; apt = []; nind = nacind;
      % calculo das novas aptidoes
      for i = 1:nind
         n_gera2 = n_gera2+1;
         apt(i) = feval(sub,mod,x,y,ind(i,:),nf);
      end
   end
end
[ymax,idmax] = max(apt);
indmax = ind(idmax,:);
press_min = inv(ymax);
vb = find(indmax);
avalia = [n_gerain n_gera n_gera2];
function nn = randb(nmaxvar,ngens)
np = ceil(rand(1,nmaxvar)*ngens);
nn = zeros(1,ngens); nn(np) = 1;
```

A utilização da função necessita que sejam informados o tipo de modelo a ser utilizado, conforme Tabela 1.2, a estratégia de geração de conjuntos de calibração/validação, seguindo as definições da Tabela 1.3, as matrizes de dados, ***X*** e ***y***, e o número de fatores a ser considerado, *nf*. O parâmetro de performance adotado para a seleção de variáveis e a PRESS.

A função retorna o conjunto de variáveis selecionadas, o valor da PRESS obtida pelo modelo construído com as variáveis selecionadas e o número de modelos avaliados

Tabela 1.3: Funções para geração de conjuntos de calibração e validação para a seleção de variáveis através de algoritmos genético

Palavra Chave	Descrição
randga	Seleção aleatória de dados
ranksga	Seleção de dados através do método y-Rank
ksga	Seleção de dados através do método de Kennard-Stone
ksmga	Seleção de dados através do método de Kennard-Stone Modificado

1.4.2 Seleção por Adição

O código referente a estratégia de seleção de variáveis através da adição sucessiva é apresentado no quadro abaixo.

```
function [vb, press] = vsf(mod,sub,x,y,nf);
%mod: pls,qpls,btpls,cspls,rbfpls
%sub: randss, ranks, ks, ksm
%Dimensões do problema
[lx,cx] = size(x); [ly,cy] = size(y);
flag_01 = 0; %Flag para interromper a adição de variáveis: 0 - para; 1 - continua
adicionando
%Inicialização das variáveis do problema
%Criação do vetor de identificação das variáveis
vnb = (1:cx); vb = [];
press = []; vt = []; pr = [];
%Modelo com 1 variável
for vr = 1:cx
   vt = vnb(vr);
   pr1 = feval(sub,mod,x,y,vt,nf);
   pr = [pr;pr1];
end
[pm,ind] = min(pr); %Valor e localização do menor valor de PRESS
press = [press; pm];
%Adicao da variavel ao bloco basico
vb = [vb vnb(ind)]; vnb(ind) = [];
%Modelos com 2 ou mais variaveis
for j = 2:cx
   [lnb, cnb] = size(vnb);
   press_md = [];
   for i = 1:cnb
      vt = [vb vnb(i)];
      [lvt, cvt] = size(vt);
      pr = feval(sub,mod,x,y,vt,nf);
      press_md = [press_md;pr];
   end
   [pm,ind] = min(press_md); %Valor e localização do menor valor de PRESS
   press = [press;pm];
   if press(j) > press(j-1) & flag_01 == 0
      break
   else
      vb = [vb vnb(ind)];
      vnb(ind) = [];
   end
end
```

A utilização da função necessita que sejam informados o tipo de modelo a ser utilizado, conforme Tabela 1.2, a estratégia de geração de conjuntos de calibração/validação, seguindo as definições da Tabela 1.4, as matrizes de dados, ***X*** e ***y***, e o número de fatores a ser considerado, *nf*. O parâmetro de performance adotado para a seleção de variáveis e a PRESS.

A função retorna o conjunto de variáveis selecionadas, o valor da PRESS obtida pelo modelo construído com as variáveis selecionadas e o número de modelos avaliados

Tabela 1.4: Funções para geração de conjuntos de calibração e validação para a seleção de variáveis através da adição sucessiva

Palavra Chave	Descrição
Randss	Seleção aleatória de dados
Ranks	Seleção de dados através do método y-Rank
KS	Seleção de dados através do método de Kennard-Stone
KSM	Seleção de dados através do método de Kennard-Stone Modificado