UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA INDUSTRIAL - PEI

MESTRADO EM ENGENHARIA INDUSTRIAL

VICENTE BRAGA BARBOSA

UMA NOVA ABORDAGEM NA SELEÇÃO DE VARIÁVEIS PARA ANALISADORES VIRTUAIS VIA REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

PEI
Programa de Pós-Graduação em Engenharia Industrial

SALVADOR
2019

UNIVERSIDADE FEDERAL DA BAHIA

ESCOLA POLITÉCNICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA INSDUSTRIAL

VICENTE BRAGA BARBOSA

UMA NOVA ABORDAGEM NA SELEÇÃO DE VARIÁVEIS PARA ANALISADORES VIRTUAIS VIA REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

SALVADOR
2019

VICENTE BRAGA BARBSOSA

UMA NOVA ABORDAGEM NA SELEÇÃO DE VARIÁVEIS PARA ANALISADORES VIRTUAIS VIA REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Industrial, da Universidade Federal da Bahia, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Industrial.

Orientadora: Karla Patrícia Oliveira Esquerre.

SALVADOR
2019

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas (SIBI/UFBA),
com os dados fornecidos pelo(a) autor(a).

Braga Barbosa, Vicente
UMA NOVA ABORDAGEM NA SELEÇÃO DE VARIÁVEIS PARA ANALISADORES VIRTUAIS VIA REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAS / Vicente Braga Barbosa, Vicente Barbosa. -- Salvador, 2019.
70 f. : il

Orientador: Karla Patricia Santos Oliveira Rodriguez Esquerre.
Dissertação (Mestrado - Engenharia Industrial) -- Universidade Federal da Bahia, PEI, 2019.

1. Analisadores Virtuais. 2. Seleção de Variáveis. 3. Custos de Modelo. 4. PLS. II. Barbosa, Vicente. I. Santos Oliveira Rodriguez Esquerre, Karla Patricia. II. Título.

UMA NOVA ABORDAGEM NA SELEÇÃO DE VARIÁVEIS PARA ANALISADORES VIRTUAIS VIA REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

VICENTE BRAGA BARBOSA

Dissertação submetida ao corpo docente do programa de pós-graduação em Engenharia Industrial da Universidade Federal da Bahia como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Industrial.

Examinada por:

Profª. Drª. Karla Patrícia Santos O. R. Esquerre
Doutora em Engenharia Química, pela Universidade de Campinas, Brasil, 2003

Profª. Drª. Karen Valverde Pontes
Doutora em Engenharia Química, pela Universidade Estadual de Campinas e pela Universidade RWTH-Aachen, Brasil e Alemanha, 2008

Drª. Daniela Domingos
Doutora em Química, pela Universidade Federal da Bahia, Brasil, 2014

Salvador, BA - BRASIL
Dezembro/2019

AGRADECIMENTOS

À Prof.ª Karla, pela inspiração e suporte.

Aos colegas Ana Rosa e Adelmo, pela troca de experiência e apoio, sem os quais essa dissertação não passaria de uma ideia discutida durante um café.

À minha mãe Tânia, ao meu pai José, à minha irmã Joana e os demais membros de minha família, por valorizarem e entenderem a importância da Educação e da Pesquisa.

À minha namorada Emilly, cujo apoio emocional me sustentou durante os momentos de maior necessidade.

Aos amigos de Engenharia Química da UFBA, que me inspiram a ser uma pessoa melhor.

À Braskem, pela cooperação e por tornar possível esse projeto.

À PASB, pelo incentivo à minha qualificação e flexibilidade em entender os momentos em que não pude estar presente.

RESUMO

Analisadores virtuais ocupam uma posição estratégica na indústria petroquímica devido a capacidade destes de estimar variáveis de controle a partir de modelos matemáticos. Entretanto, para garantir uma estimativa, é necessário assegurar a confiança e a disponibilidade dos dados de entrada. Desta forma, há um esforço financeiro para garantir a manutenibilidade dos instrumentos de medição que aferem as variáveis utilizadas com entrada do sistema. O objetivo deste trabalho é propor uma nova abordagem na seleção de variáveis em modelos de Mínimos Quadrados Parciais (PLS) através da introdução de um indicador que avalia o ganho de capacidade preditiva do modelo em função do aumento de custo associado ao se acrescentar determinada variável como entrada. Para isto, as variáveis são hierarquizadas a partir do escore VIP (Importância da Variável na Projeção) e, uma a uma, introduzidas no modelo PLS. O novo indicador mede a razão entre a diferença dos coeficientes de correlação linear (r) dos valores observados e os estimados pelos modelos com e sem a variável adicionada, e a diferença entre o custo padrão associado aos respectivos modelos. Desta forma, quantifica-se a razão entre o ganho de performance e o aumento de custos associados a introdução de uma variável. A nova abordagem é aplicada na seleção de variáveis de um modelo que estima nove pontos de temperatura (ponto inicial de ebulição, 5%, 10%, 30%, 50, 70%, 90%, 95% de vaporizado, e ponto final de ebulição) de uma nafta média, utilizados para avaliar a qualidade desta. Havia disponibilidade de 121 variáveis de processo (VPs), as quais incluem fluxo, temperatura, pressão e nível. O modelo PLS desenvolvido a partir da nova abordagem selecionou 37 das 121 VPs, com um custo total associado (c_T) que representa 34% do c_T do modelo com as 121 variáveis, e 88% do c_T onde a seleção de variáveis é feita por VIP. Os erros quadráticos médios de predição do modelo variam entre 1,184°C e 3,108°C para saídas num intervalo de 90°C a 155°C. Os rs do grupo de validação variam entre 0,875 e 0,932, com exceção da temperatura com 95% de vaporizado, cujo r foi de 0,753. Assim sendo, a seleção de variáveis utilizando a nova abordagem proposta foi capaz de desenvolver um modelo preditivo adequado à aplicação em analisadores virtuais. Dessa forma, os resultados indicam que a inclusão de um traço econômico no processo de seleção de variáveis, que costuma ser quase que puramente estatístico, contribui para procedimentos mais orientados por dados durante o processo de tomada de decisões em ambientes industriais complexos, particularmente da petroquímica.

Palavras-chave: PLS; Seleção de Variáveis; Analisadores Virtuais; Custos de Modelo

ABSTRACT

Virtual analyzers play a strategic role in the petrochemical industry due to its capacity to estimate control variables through mathematical models. Yet, in order to guarantee the model output, it is essential to ensure the availability and reliability of input data, which implies greater financial efforts to guarantee the maintainability of the measuring devices used to generate model input. The aim of this work is to propose a new approach to variable selection for Partial Least Square (PLS) models through the introduction of an indicator which evaluates the gain in predictive power of a model as a function of the increase in its combined costs due to the addition of a process variable as model input. For that, process variables are ranked with VIP (Variable Importance in Projection) scores and, one by one, introduced in different PLS models. The new indicator measures the ratio between the difference in correlation coefficient (r) of the observed and predicted output from the models with and without the input variable, and the difference between the combined standardized costs of the respective models. Thus, it quantifies the ration between gain in performance and increase in costs due to the introduction of an input variable in the model. The proposed approach is used to select input variables for a model that estimates nine temperature points (initial point of distillation 5%, 10%, 30%, 50%, 70%, 90%, 95% vaporized mass, and final point of distillation) of a naphtha, which are used as parameters to determine the naphtha's quality. There was an availability to select data from 121 process variables (PV), which include flow, temperature, pressure and level variables. The PLS model built under the new approach selected 37 out of the 121 PVs, with a combined cost (c_T) that accounts for 34% of the c_T of the combined cost for the model with 121 VPs, and 88% of the combined cost with variable selection through VIP. The root mean square error of prediction for the model varied between 1.184 °C and 3.108 °C for variables whose values ranged from 90°C to 155°C. The r for the validation group varied between 0.875 and 0.932, with the exception of the point with 95% vaporized mass, which displayed an r of 0.753. Thus, the variable selection through the new proposed approach was able to develop an empirical model with adequate predictive power to be used in a virtual analyser. As such, the results suggest that the presence of an economic trait during the process of selecting variables, which tends to be purely statistical, contributes to procedures which are more data oriented in the decision-making process of complex industrial environments, particularly in the petrochemical industry.

Key words: PLS; Variable Selection; Virtual Analysers; Model Costs

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks* / Redes Neurais Artificiais
BIAS	Erro sistemático
CV	*Cross Validation* / Validação Cruzada
KSS	*Kennard-Stone Sampling* / Particionamento por Kennard-Stone
LOF	*Local Outlier Factor* / Fator Local de Outlier
LRC	Logaritmo do fator RC
LV	*Latent Variable* / Variável Latente
MLR	*Multiple Linear Regression* / Regressão Linear Múltipla
PCA	*Principal Components Analysis* / Análise de Componentes Princiapais
PCR	*Principal Componentes Regression* / Regressão por Componentes Principais
PRESS	*Predicted Residual Error Sum of Squares* / Soma dos Erros Residuais Quadráticos de Predição
r	Coeficiente de correlação linear entre valores observados e preditos
RMSEP	*Root Mean Square Error of Prediction* / Erro Quadrático Médio de Predição
PLS	*Partial Least Squares* / Mínimos Quadrados Parciais
PLSR	*Partial Leas Squares Regression* / Regressão por Mínimos Quadrados Parciais
VP	Variável de Processo
SLRC	Soma Logarítmica do fator RC
VIP	*Variable Importance in Projection* / Importância da Variável na Projeção

SUMÁRIO

1. INTRODUÇÃO

A qualidade de diversos produtos da indústria petroquímica é determinada por análises laboratoriais, que podem ser feitas por laboratórios externos ou pertencentes a empresa produtora. Apesar de vantagens no uso de laboratórios para análise, as quais incluem alta exatidão nos resultados, há algumas desvantagens em seu uso para determinação da qualidade de produtos. Entre estas, Silva (2017) destaca o alto tempo de residência da amostra, o que pode levar à uma possível perturbação no sistema no tempo entre a coleta e o resultado laboratorial. Desta forma, o processo de tomada de decisão do operador fica comprometido, impossibilitando um controle do sistema em tempo real. Massa (2017) ressalta a importância dos analisadores em linha, que fornecem uma reposta muito mais rápida em comparação às análises laboratoriais, o que permite um controle maior do sistema. Entretanto, devido à natureza das substâncias utilizadas nas correntes de um processo petroquímico, analisadores em linha precisam de manutenções frequentes, períodos onde pode haver a perda de dados essenciais ao funcionamento do processo. Além disso, analisadores em linha costumam ser equipamentos com custos elevados, tanto de implantação como de manutenção.

Neste contexto, analisadores virtuais ocupam uma posição estratégica na indústria petroquímica devido a sua capacidade utilizar modelos matemáticos para aferir determinas variáveis importantes ao controle de qualidade do processo. Como descrito por Bakhtadze (2004), analisadores virtuais são usados como base de algoritmos nos mais diversos sistemas de controle de processos industriais. Tal controle pode ser feito de duas formas distintas: (1) o sistema apenas monitora as variáveis de processo (VPs), auxiliando as tomadas de decisão do operador; (2) o sistema não somente monitora VPs, mas também indica ações de correção ao operador. A primeira forma de controle é mais comum na literatura, como em trabalhos recentes de Nogueira et al. (2017) e Yadykin *et al.* (2015). Nogueira propõe um modelo que monitora a qualidade de um processo de polimerização, enquanto Yadykin desenvolve um modelo que avalia a estabilidade de um Sistema de Potência, indicando possível risco de uma falha em cascada.

O desenvolvimento de um modelo que represente de forma adequada um processo não é trivial. Há diversas técnicas para construção de modelos preditivos, algumas mais simples, como a Regressão Linear Múltipla (LMR), até outras mais complexas, como Redes Neurais

Artificias (ANN). Em seu trabalho, Nogueira constrói um modelo empírico por redes neurais, enquanto Yadykin mescla modelos fenomenológicos e empíricos. Ambos possuem algo em comum: para garantir a exatidão da resposta do modelo é preciso assegurar que os valores das variáveis usadas como entrada serão confiáveis e estarão disponíveis no momento em que o modelo preditivo precise ser utilizado para evitar uma possível parada no processo. Isso implica em esforços maiores para garantir a manutenibilidade dos dispositivos usados para gerar dados de entrada, o que pode elevar custos. Desta forma, existe um *trade off* entre os custos associados a um modelo de predição e a exatidão de sua resposta, uma característica econômica desconsiderada pela maior parte dos pesquisadores na construção de modelos. Essa consideração pode, entretanto, existir de maneira indireta quando há seleção de variáveis na construção do modelo, uma prática considerada crucial para garantir alta capacidade preditiva do mesmo (Kano e Fujiwara, 2012).

De fato, a natureza dos processos químicos e dos sistemas de engenharia que os monitoram induzem a coleta de centenas, ou até vezes milhares, de variáveis do processo. Entre estas estão incluídas, por exemplo, temperaturas em diferentes pontos do processo, fluxos de vazão, valores de pressão e indicadores de níveis em tanques. Desta forma, existe a necessidade da seleção de variáveis para garantir o sucesso do modelo, já que nem todas contribuem com informação relevante à predição da propriedade desejada, algumas inclusive geram ruído suficiente para prejudicar a capacidade preditiva do modelo (Morais Júnior, 2011). Há diversos métodos de seleção de variável, alguns mais simples, como o método de busca exaustiva, e outros mais rebuscados, como a Importância da Variável na Projeção (VIP). Apesar destes métodos levarem a uma redução implícita do custo associado ao modelo, quando comparado a um modelo em que todas as variáveis são utilizadas, o custo de monitoramento das variáveis de entrada não é, em si, parte do desenvolvimento do mesmo. As técnicas de seleção de variável partem do pressuposto que as variáveis selecionadas estarão disponíveis para serem utilizadas no modelo preditivo a qualquer momento, sem considerar o custo associado à manutenibilidade do equipamento de medição das variáveis selecionadas. A relevância da pesquisa aqui desenvolvida está exatamente em introduzir este fator econômico na seleção de variáveis, transformando um processo puramente estatístico em outro mais abrangente.

O texto desta dissertação se desdobra por mais cinco capítulos, além desta introdução. No próximo capítulo, de Fundamentação Teórica, estão descritos as técnicas e os conceitos

consolidados na literatura e que foram utilizados neste trabalho. No capítulo que se segue, é feita uma revisão crítica do estado da arte da área em que esta pesquisa está inserida. No capítulo de Metodologia, está descrita a forma com que os resultados foram alcançados, e a discussão destes é feita no capítulo seguinte, de Resultados e Discussão. Neste se avalia a eficácia da abordagem de seleção de variáveis proposta neste trabalho. Por fim, são apresentadas as conclusões e sugestões para trabalhos futuros.

2. OBJETIVOS

2.1 Objetivos Geral

Propor e avaliar a eficácia uma nova abordagem na seleção de variáveis para analisadores virtuais via regressão de Mínimos Quadrados Parciais (PLS) através da introdução de um indicador que avalia a razão entre o ganho de capacidade preditiva do modelo e o aumento de custo associados à manutenibilidade dos instrumentos de medição das variáveis selecionadas.

2.2 Objetivos Específicos

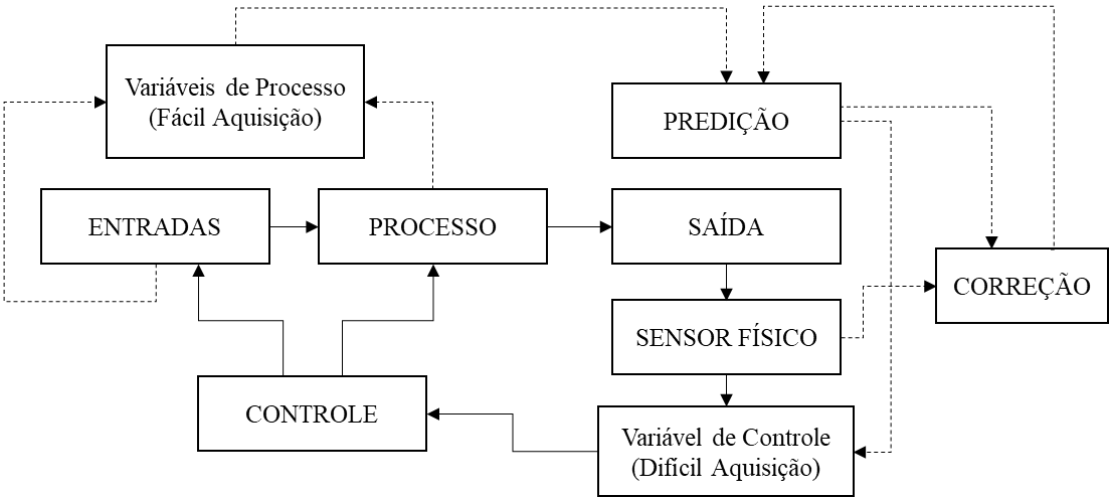
- Definir um fator, chamado de RC, que mede a razão entre o possível ganho de capacidade preditiva e o aumento de custos associados ao modelo;
- Construir um modelo PLS com as variáveis de processo e hierarquizar estas através do escore VIP;
- Selecionar variáveis utilizando o novo fator proposto;
- Modelar nove pontos de temperatura de uma nafta média através de um modelo PLS onde há seleção de variáveis, utilizando a abordagem proposta;
- Testar e validar o modelo construído.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 ANALISADORES VIRTUAIS

Segundo Faccin (2005), o termo *Analisador Virtual* se refere a "*algoritmos capazes de estimar ou inferir variáveis de difícil aquisição de forma contínua"*. A Figura 1 destaca a estrutura básica da aplicação de um analisador virtual no controle de um processo.

Figura 1 – Estrutura básica de aplicação de um analisador virtual



Fonte: Facchin (2005), adaptado

O modelo inferencial utilizado no analisador virtual está representado pelo bloco "Predição" na Figura 1. Neste há uma relação matemática entre as variáveis de processo, que incluem valores de entrada ou medidos ao longo do processo, e a variável de controle, a ser estimada. Nota-se que a variável de controle pode também ser fornecida por um sensor físico, ou por análise laboratorial. De fato, há uma comparação entre o valor estimado pelo modelo inferencial e o fornecido pelo sensor físico no bloco "Correção", onde possíveis ajustes no modelo do analisador virtual podem ser feitos para melhorar a qualidade da resposta estimada.

Denn (1986) destaca que a relação matemática representada pelo bloco "Predição" na Figura 1 pode ser determinada através de três modelos diferentes:

- Modelo Fenomenológico: a relação é derivada a partir da utilização de teorias fundamentais e princípios básicos da natureza, agregando princípios de conservação de massa, energia e quantidade de movimento, além de outras leis fundamentais da física e da química;

- Modelo Empíricos: a relação é produto da observação direta de experimentos ou de dados históricos;
- Modelos Análogos: a relação é descrita através de equações de um sistema análogo, onde as variáveis em um modelo são análogas às variáveis no sistema de referência.

Ressalta-se que a capacidade de extrapolação de modelos fenomenológicos é muito superior aos demais. Entretanto, em processos altamente complexos, como os da indústria petroquímica, a construção de modelos fenomenológicos, mesmo que simplificados, demanda tempo e conhecimento que nem sempre estão disponíveis. Geralmente, há a necessidade da utilização de parâmetros que não estão bem determinados na literatura, ou que são específicos às condições do processo. Desta forma, modelos empíricos são favorecidos na construção de modelos inferenciais para analisadores virtuais na indústria petroquímica (Facchin, 2005). Apesar de não possuírem capacidade extrapolativa comparada à modelos fenomenológicos, modelos puramente empíricos (denominados de *caixa preta*), resultantes de métodos puramente matemáticos e estatísticos, permitem a modelagem do comportamento dos dados e a estimação da variável de difícil aquisição sem exigir a modelagem do processo em si, o que favorece a sua utilização.

3.2 PRÉ-PROCESSAMENTO DE DADOS E DETECÇÃO DE OUTLIERS

Ferreira *et al.* (1999) destaca a importância de uma etapa que precede a calibração multivariada e que visa, entre outras coisas, remover interferentes ou informações superpostas. Massa (2017) ressalta que a ausência de manipulações matemáticas com o objetivo de reduzir variações aleatórias ou sistemáticas, as quais não costumam possuir qualquer relação com o problema estudado, tendem a gerar resultados insatisfatórios. Entre estas manipulações matemáticas, Silva (2017) evidencia a ampla utilização da centralização (subtração dos valores de cada vetor variável pelo valor médio desse vetor ao longo de todas as observações) e do escalonamento (divisão do resultado da centralização em cada vetor variável pelo desvio padrão do vetor em todas as observações) das varáveis de entrada, cujo objetivo é tornar as médias de tais variáveis nulas e os respectivos desvios-padrão unitários. Estas técnicas são amplamente aplicadas em dados onde há coleta de variáveis com diferença de unidade de medição (pressão e temperatura, por exemplo), o que dificulta a comparação destas. A centralização e o escalonamento permitem a conversão dos dados em valores que podem ser diretamente

comparados (Walach, Filzmoser e Hron, 2018), além do escalonamento ser um dos pré-requisitos para a aplicação da PCA e do PLS (Facchin, 2005). A Equação (1) descreve o escalonamento por variância (*variance scaling*), também conhecido como padronização, onde z_{ij} é o valor escalonado da observação i na variável j, cuja média é $\overline{x_j}$ e desvio-padrão é s_j.

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j} \tag{1}$$

A correta detecção e remoção de *outliers* é essencial à análise de dados. Segundo Andersen e Bro (2010), esta se torna ainda mais importante quando há aplicação de um método de seleção de variáveis, como proposto neste trabalho, já que a maior parte destes métodos se embasa na avaliação de diferenças mínimas na qualidade do modelo, o que os tornam ainda mais sensíveis a presença de *outliers* do que a própria calibração do modelo. De fato, os autores sugerem que a detecção e remoção de *outliers* seja feita antes da seleção de variáveis, com os dados originais, para evitar que a seleção em si afete a detecção de *outliers*. Entretanto, é importante salientar que, em casos onde haja presença de uma variável completamente irrelevante ao contexto do modelo, a sua remoção é sugerida antes de detecção de *outliers* para evitar que pontos normais sejam identificados como anômalos.

Li *et al.* (2016) detalham que diversos métodos de detecção de *outliers* foram desenvolvidos nas últimas décadas devido ao efeito significativo que a presença destes exerce sobre a qualidade da calibração de modelos. Estes incluem desde métodos de agrupamento (*clustering*), das mais diversas formas (*fuzzy, K-means, C-means, Mountain Clustering*), utilizados por Morais Junior (2011) e Wille *et al.* (2018), até a análise de resíduos-Q e distância T^2 de Hotelling, presentes em Ferreira (1999), Silva (2017) e Massa (2017). Apesar destes métodos se mostrarem eficientes nos casos em que foram aplicados, Li *et al.* (2016) argumentam que, na presença de múltiplos *outliers*, efeitos de mascaramento (*masking*) e inundação (*swamping*) podem torná-los ineficientes em outros casos, particularmente em modelos PLS. Para evitar tais efeitos, os autores propuseram um método de detecção de *outliers* chamado de "diagnóstico do modelo" (*model diagnostics*), resumido na subseção a seguir. Uma descrição mais detalhada sobre este método e suas aplicações pode ser encontrada em Li *et al.* (2016) e Breunig *et al.* (2000).

3.2.1 Detecção de *outliers* através do "diagnóstico do modelo"

Li *et al.* (2016) partem do pressuposto que uma observação deve ser considerada anômala numa calibração de modelo por PLS se, na palavra dos autores, "tal não se ajusta ao modelo construído com as demais observações" e, consequentemente, "se comporta diferente da massa de dados". Esse comportamento anômalo pode ser reconhecido através dos escores de cada observação nas variáveis latentes (também conhecidas como componentes), que são combinações lineares das variáveis utilizadas para a construção do modelo PLS. Para isso, é necessário que o modelo seja construído com todos os indivíduos da massa de dados, isto é, não há seleção de amostras para um grupo de validação. A maior parte das observações terá seus escores, para cada componente, distribuídos num certo intervalo, enquanto que os *outliers* terão escores significativamente diferente dos demais.

No método de diagnóstico do modelo, o critério de identificação de *outliers* é embasado no valor numérico do Fator Local de Outlier (LOF, *Local Outlier Factor*), introduzido primeiramente por Breunig *et al.* (2000). O LOF é baseado num conceito de densidade local, e quantifica, de fato, o desvio local de uma dada observação em relação aos k vizinhos mais próximos. As Equações (2) e (3) definem matematicamente o valor numérico do LOF para um determinado k, onde $N_k(A)$ é o conjunto dos k vizinhos mais próximos de A.

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|} \tag{2}$$

$$lrd_k(A) = \left(\frac{\sum_{B \in N_k(A)} rd_k(A,B)}{|N_k(A)|} \right)^{-1} \tag{3}$$

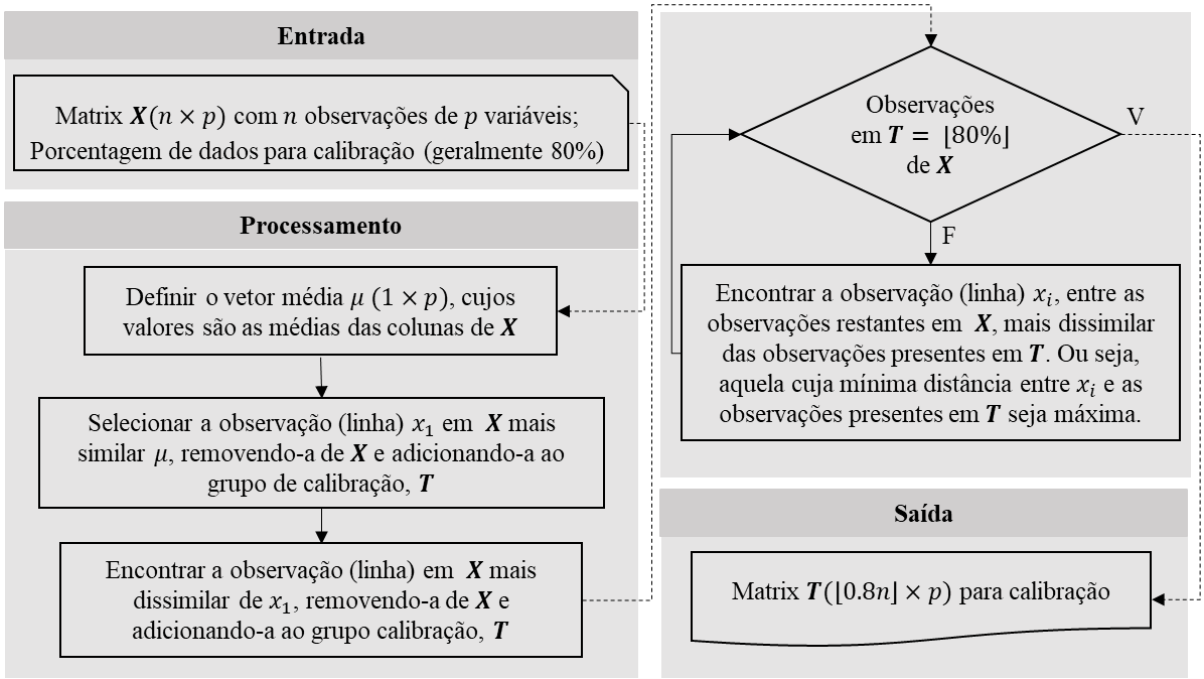
A função $rd_k(A,B)$ na Equação (3) é conhecida como *reachability-distance* (distância de alcance, na tradução literal). Seu valor é dado como o máximo entre a distância de um ponto A ao seu k vizinho mais próximo e a distância entre o ponto B e seu k vizinho mais próximo. O inverso da média das distâncias de alcance entre A e todos os seus k vizinhos é utilizada para calcular a $lrd_k(A)$, conhecida como *local reachability density* (densidade de alcance local). O $lrd_k(A)$ é um indicativo da distância entre dois pontos dentro de uma certa vizinhança. Como é um valor de densidade, quanto menor seu valor, maior a distância entre vizinhos. O $LOF_k(A)$ compara, de acordo com Breunig (2003), a densidade de alcance local de um ponto A com seus

k vizinhos. Caso $lrd_k(A)$ seja muito menor que a densidade de seus vizinhos ($LOF_k(A) \gg 1$), o ponto está posicionado em uma região esparsa e deverá ser considerado um *outlier*. No método proposto por Li *et al*. (2016), observações que possuam um LOF maior que a média dos valores de LOF mais três vezes o desvio padrão dos valores de LOF, sem a presença dos supostos *outliers*, devem ser consideradas anômalas.

3.3 SELEÇÃO DE AMOSTRAS PARA CALIBRAÇÃO/VALIDAÇÃO DO MODELO

A necessidade de selecionar indivíduos que estejam distribuídos uniformemente no conjunto de pontos possíveis durante o planejamento (*design*) de experimentos é uma discussão antiga na literatura. No final da década de 60, R.W. Kennard e L.A. Stone publicaram um artigo intitulado *Computer Aided Design of Experiments* (Planejamento de Experimentos auxiliados por Computador) cujo objetivo era auxiliar no planejamento de experimentos para a construção de superfícies de respostas (Kennard e Stone, 1969). Inicialmente, o algoritmo desenvolvido por Kennard e Stone foi chamado de "mapeamento uniforme", mas devido ao amplo uso na seleção de amostras para calibração de modelos inferenciais, particularmente na Quimiometria, acabou ficando conhecido como KSS, do inglês Kennard-Stone *sampling* (Ramiro-Lopez *et al.*, 2014). O algoritmo KSS é detalhado na Figura 2.

Figura 2 - Algoritmo KSS



Na Figura 2, o conceito de similaridade entre duas observações p e q é baseado na distância geométrica entre estas, definida pela Equação (4).

$$d(p,q) = \sqrt{\sum_{i=1}^{p} (x_{p\times i} - x_{q\times i})^2} \quad (4)$$

Diversos outros algoritmos de partição de amostras para calibração foram desenvolvidos desde a concepção do KSS. Alguns funcionam como um ajuste do KSS, tal como algoritmo SPXY, desenvolvido por Galvão (2005) e aplicado por Silva (2017). Neste, a distância geométrica de um par de pontos não é baseada nos valores correspondentes da matriz de entrada $\boldsymbol{X}$, mas sim pelos valores correspondentes da matriz de variáveis resposta, $\boldsymbol{Y}$. Outros algoritmos mais complexos utilizam técnicas probabilísticas de agrupamento, como o *Fuzzy c-means*, amplamente utilizado nas análises de espectroscopia por NIR e no mapeamento de solos (de Gruijter, McBratney e Taylor, 2010). Entretanto, é importante salientar que tais algoritmos mais complexos não irão, necessariamente, afetar a performance do modelo desenvolvido. De fato, segundo Ramiro-Lopez *et al.* (2014), o tipo de algoritmo de partição só influencia a capacidade preditiva do modelo se o número de amostras para calibração for relativamente pequeno. Caso contrário, o tipo de algoritmo de partição não é crítico à performance do modelo.

3.4 REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS – PLSR

O algoritmo conhecido como Mínimos Quadrados Parciais (PLS, *Partial Least Squares*) foi desenvolvido durante as décadas de 60 e 70 por Herman Wold como solução de problemas na área de Econometria (Akarachantachote *et al.*, 2014). Algumas décadas depois, nos anos 1980, foi adaptada por Svante Wold, filho de Herman Wold, e Harald Martens, e utilizada como método de regressão na área de Quimiometria, após sua aplicação ter sido inicialmente proposta por Kowalski em 1982 (Geladi e Kowalski, 1986). Desde então, é uma das técnicas mais utilizadas como alternativa para regressões múltiplas simples, cuja performance é altamente afetada pela presença de colinearidade entre as variáveis preditoras, sendo aplicada principalmente nas áreas de bioinformática, *machine learning* e Quimiometria (Akarachantachote *et al.*, 2014). O algoritmo da PLSR usa informação contida em ambas as matrizes dos dados de entrada, **X**, e dos dados de saída, **Y**, durante a calibração, de forma a

explicar a variabilidade em ambas as matrizes, e não somente em **X**, como ocorre em outras técnicas de regressão. Dessa forma, a calibração reduz o impacto de possíveis variabilidades em **X** que não são relevantes a **Y** (Romía e Bernàdez, 2009).

Na regressão por PLS, a variável a ser estimada é uma combinação linear de variáveis latentes (LVs), que por sua vez são combinações lineares das variáveis de entrada (variáveis de fácil aquisição no caso da indústria petroquímica). É importante destacar que a regressão por PLS busca um conjunto de componentes, ou LVs, no processo de decomposição simultânea de $\boldsymbol{X}$ e $\boldsymbol{Y}$ que maximiza a explicação da covariância entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ (Abdi, 2010). Como o número de LVs é significantemente menor que o número de variáveis de entrada, a PLSR é também interessante numa possível redução de dimensionalidade dos dados, além de fornecer possíveis fatores que podem ser utilizados na construção de cartas de controle (Kourti e MacGregor, 1995). Por fim, Morellato (2010) argumenta que a PLSR consegue modelar regressões com saídas múltiplas sem ser afetada por multicolinearidade, além de produzir fatores com alto poder de predição, dado que estes possuem elevadas covariâncias com a variável resposta.

3.4.1 Descrição Matemática da PLSR

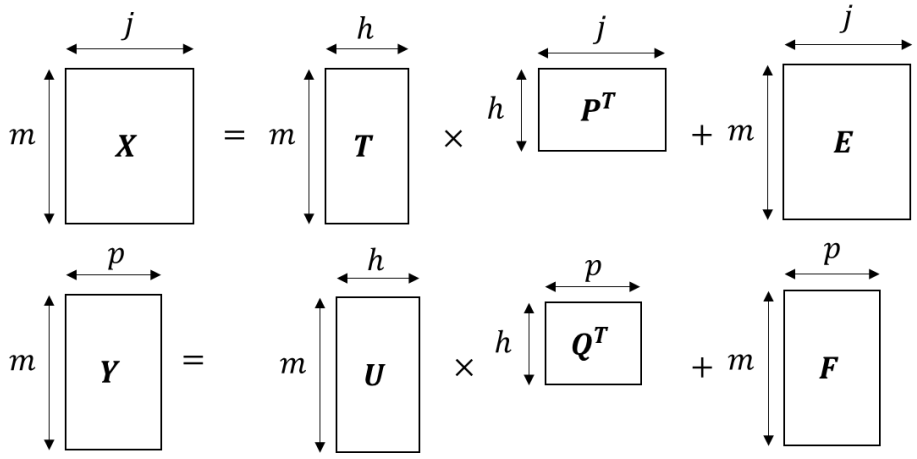
Dadas m observações, j variáveis de entrada, p variáveis de saída, após centralização e escalonamento das matrizes $\boldsymbol{X}(m \times j)$ e $\boldsymbol{Y}(m \times p)$, estas devem ser decompostas em uma soma de h variáveis latentes, de acordo com as Equações (5) e (6) (Romía and Bernàdez, 2009):

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} = \sum_{i=1}^{h} t_i p_i^T + \boldsymbol{E} \tag{5}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} = \sum_{i=1}^{h} u_i q_i^T + \boldsymbol{F} \tag{6}$$

Em (5) e (6), as matrizes $\boldsymbol{T}(m \times h)$ e $\boldsymbol{U}(m \times h)$ são as matrizes de escores de $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente, assim como as matrizes $\boldsymbol{P}^T(h \times k)$ e $\boldsymbol{Q}^T(h \times p)$ são as matrizes de cargas fatoriais de $\boldsymbol{X}$ e $\boldsymbol{Y}$. $\boldsymbol{E}$ e $\boldsymbol{F}$ são, respectivamente, os resíduos de $\boldsymbol{X}$ e $\boldsymbol{Y}$. As matrizes $\boldsymbol{Q}$ e $\boldsymbol{P}$ são formadas por colunas ortogonais, isto é, $\boldsymbol{Q}\boldsymbol{Q}^{\boldsymbol{T}} = \boldsymbol{I}$ e $\boldsymbol{P}\boldsymbol{P}^{\boldsymbol{T}} = \boldsymbol{I}$, onde $\boldsymbol{I}$ é a matriz identidade. A decomposição das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$, descrita pelas equações (5) e (6), é ilustrada na Figura 3.

Figura 3 – Decomposição das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ no algoritmo do PLS



Fonte: Brereton, 2007 (Adaptado)

O modelo PLS é obtido através do mapeamento da relação linear entre $\boldsymbol{U}$ e $\boldsymbol{T}$, de acordo com a equação (7) (Brereton, 2007).

$$\boldsymbol{U}(m \times h) = \boldsymbol{T}(m \times h) \times \boldsymbol{B}(h \times h) \tag{7}$$

Em (7), a matriz $\boldsymbol{B}$ é definida como a pseudo-inversa da matriz $\boldsymbol{T}$. Dessa forma, a Equação (6) pode ser rescrita, usando a Equação (7), gerando a Equação (8).

$$\boldsymbol{Y} = \boldsymbol{T} \times \boldsymbol{B} \times \boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \tag{8}$$

A predição de uma matriz resposta ($\widehat{\boldsymbol{Y}}$) a partir de uma nova matriz de entrada ($\boldsymbol{X}^*$) é então dada pela Equação (9).

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}^* \times \boldsymbol{P} \times \boldsymbol{B} \times \boldsymbol{Q}^{\boldsymbol{T}} \tag{9}$$

Para utilizar a Equação (9), é preciso obter as matrizes $\boldsymbol{P}$ (cargas fatoriais de $\boldsymbol{X}$), $\boldsymbol{B}$ (pseudeo-inversa de $\boldsymbol{T}$) e $\boldsymbol{Q}^{\boldsymbol{T}}$ (transposta da matriz de cargas fatoriais de $\boldsymbol{Y}$). A princípio, qualquer conjunto de vetores (variáveis latentes) ortogonais pode formar as colunas da matriz $\boldsymbol{T}$ e da matriz $\boldsymbol{Q}$. Porém, o algoritmo do PLS visa extrair a máxima covariância entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. Dessa forma, o objetivo do algoritmo do PLS é obter um primeiro conjunto de vetores, $\boldsymbol{t}$ e $\boldsymbol{u}$, onde $\boldsymbol{t} = \boldsymbol{Xp}$ e $\boldsymbol{u} = \boldsymbol{Yq}$, de forma que $\boldsymbol{tt}^{\boldsymbol{T}} = 1$, $\boldsymbol{uu}^{\boldsymbol{T}} = 1$ e $\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{u}$ é máximo. Após o vetor que representa a primeira variável latente é encontrado, ele é subtraído de ambas $\boldsymbol{X}$ e $\boldsymbol{Y}$, e o

procedimento é repetido até que X se torne uma matriz nula. (Abdi, 2010). Esse algoritmo de determinação das variáveis latentes é conhecido como Mínimos Quadrados Parciais Iterativo Não Linear, ou simplesmente NIPALS (*Non-linear Iterative Partial Least Squares*). Uma descrição mais detalhada sobre o NIPALS pode ser encontrada em Wold, Sjostrom e Eriksson (2001).

3.4.2 O número adequado de Variáveis Latentes (LVs)

Abdi (2010) afirma que a capacidade preditiva de modelos PLS nem sempre melhora com aumento do número de variáveis latentes. De fato, a qualidade da predição de um modelo PLS tende a aumentar com o número de LVs somente até certo ponto, quando a adição de uma outra LV causa uma piora significativa. Nestes casos, ocorre um fenômeno descrito como *overfitting*, ou sobreajuste, onde o modelo construído é capaz de predizer com exatidão somente as observações do conjunto de calibração, não sendo capaz de manter uma capacidade preditiva adequada para novas observações. Desta forma, é necessário testar a significância preditiva de cada LV adicionada ao modelo PLS, e não mais adicionar componentes quando a inclusão destas não é mais significativa à predição de novas observações.

Wold, Sjostrom e Eriksson (2001) recomendam que validação cruzada (CV) seja utilizada para determinação do PRESS (Soma dos Erros Residuais Quadráticos de Predição, *Predicted Residual Error Sum of Squares*). Na CV, o conjunto de dados é dividido em grupos (entre cinco e dez), e modelos PLS são calibrados sempre deixando um dos grupos de fora. O PRESS é então calculado sempre considerado o grupo que não foi utilizado na calibração. Como exemplo, considere um conjunto de dados que foi dividido em 5 grupos, A, B, C, D e E. O primeiro modelo PLS é calibrado utilizando os dados dos grupos A, B, C, e D, e uma parcela do PRESS é calculada utilizando o grupo E. Isto feito, um outro modelo PLS é então calibrado utilizando os dados dos grupos B, C, D e E, e outra parcela do PRESS é calculada utilizando o grupo A. Segue-se até que todos os grupos tenham ficado de fora da calibração uma vez, e soma-se as parcelas do PRESS de cada grupo. Em seu modo sequencial, a validação cruzada é executada considerando uma variável latente por vez. Desta forma, para uma dada LV a, define-se $PRESS_a$ como a soma dos quadrados das diferenças entre os elementos de $\widehat{\boldsymbol{Y}}_{\boldsymbol{a}}$ (matriz de valores preditos utilizando a LVs) e os correspondentes em $\boldsymbol{Y}$ (matriz dos valores de referência). A equação que define o PRESS está presente na Tabela 1. Segundo Wold, Sjostrom e Eriksson

(2001), variáveis latentes são significativas somente quando a razão θ dada pela Equação (10) é menor que 0,9 para pelo menos uma das variáveis da matriz de saída $\boldsymbol{Y}$. Na Equação (10), SS_{a-1} é a soma quadrática dos elementos de $\widehat{\boldsymbol{Y}}_{\boldsymbol{a-1}}$ corrigidos pela média, ou seja, a soma dos quadrados das diferenças entre os elementos de $\widehat{\boldsymbol{Y}}_{\boldsymbol{a-1}}$ (matriz de valores preditos utilizando $a-1$ LVs) e a média destes. Nota-se que a primeira LV deve ser sempre considerada, já que SS_0 não é definida.

$$\theta = \frac{PRESS_a}{SS_{a-1}} \tag{10}$$

Apesar do método de escolha do número de variáveis latentes descrita por Wold, Sjostrom e Eriksson (2001) ser amplamente utilizado, Abdi (2010) argumenta que um método mais robusto pode ser aplicado. Neste, para cada componente a, calcula-se a razão entre o $PRESS_a$ e a soma residual quadrática, RESS (do inglês *Residual Sum of Squares*), do modelo desenvolvido com $a-1$ variáveis latentes. A equação que define o RESS está presente na Tabela 1. Esta razão é denominada Q_a^2, matematicamente definida pela Equação (11), onde $RESS_0 = p \times (m-1)$, dado p o número de variáveis de saída, e m o número de observações. Uma variável latente a deve ser mantida caso a razão Q_a^2 seja maior do que 0,05 para $m \leq 100$ ou maior do que zero pra $m > 100$.

$$Q_a^2 = 1 - \frac{PRESS_a}{RESS_{a-1}} \tag{11}$$

3.5 AVALIAÇÃO MODELOS DE CALBRIÇAO MULTIVARIADA

Massa (2017) destaca que, após a calibração de modelos PLS, é necessário utilizar índices de desempenho, conhecido como figuras de mérito, para avaliar a capacidade preditiva do modelo criado. Abdi (2010) argumenta que, uma vez determinado o número de variáveis latentes, essa avaliação deve ser feita através da similaridade entre as matrizes $\widehat{\boldsymbol{Y}}_{\boldsymbol{a}}$ e $\boldsymbol{Y}$, que pode ser medida de diversas maneiras. As figuras de mérito mais utilizadas na avaliação de modelos PLS são descritas na Tabela 1, que inclui alguns conceitos já discutidos em seções anteriores.

Ressalta-se que as figuras de mérito descritas na Tabela 1 não devem ser usadas de forma isolada. A primeira delas, o BIAS, mede um erro sistemático. Dessa forma, um valor de BIAS muito próximo de zero não garante que o modelo possui capacidade preditiva adequada,

apenas indica que, em termos absolutos, há uma equidade entre pontos subestimados pelo modelo (resíduos negativos) e pontos sobrestimados (resíduos positivos), o que é necessário, mas não suficiente. Além disso, as figuras de mérito que lidam com a soma quadrática de resíduos (PRESS, RESS, RMSEC e RMSEP) tendem a ser bastante sensíveis à presença de *outliers*, além da difícil interpretação do valor, devido a presença de dimensão (a mesma da variável de saída no caso do RMSEC e RMSEP, e o quadrado desta no caso do PRESS e RESS).

Tabela 1 - Figuras de Mérito para Avaliação de Modelos PLS

Símbolo	Parâmetros	Fórmula	Intervalo e Valor desejado
BIAS	$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$: predição para valor de referência, $\boldsymbol{y}_{\boldsymbol{i}}$; $\boldsymbol{N}$: número de amostras do conjunto, podendo ser de calibração ou de validação	$\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{N}$	BIAS $\in R$ Próximo de 0
PRESS	$\widehat{\boldsymbol{y}}_{\boldsymbol{CV,i}}$: valor previsto na etapa de validação cruzada; $\boldsymbol{y}_{\boldsymbol{i}}$: o valor de referência para a amostra *i*	$\frac{\sum_{i=1}^{N}\left(\hat{y}_{CV,i} - y_i\right)^2}{N}$	PRESS ≥ 0 Próximo de 0
RESS	$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, $\boldsymbol{y}_{\boldsymbol{i}}$, $\boldsymbol{N}$	$\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}$	RESS ≥ 0 Próximo de 0
RMSEC	$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, $\boldsymbol{y}_{\boldsymbol{i}}$, $\boldsymbol{N}_{\boldsymbol{cal}}$: número de amostras de calibração; $\boldsymbol{A}$: número de variáveis latentes	$\sqrt{\frac{\sum_{i=1}^{N_{cal}}(\hat{y}_i - y_i)^2}{(N_{cal} - A - 1)}}$	RMSEC ≥ 0 Próximo de 0
RMSEP	$\widehat{\boldsymbol{y}}_{\boldsymbol{CV,i}}$, $\boldsymbol{y}_{\boldsymbol{i}}$, $\boldsymbol{N}$.	$\sqrt{\frac{\sum_{i=1}^{N}\left(\hat{y}_{CV,i} - y_i\right)^2}{N}}$	RMSEP ≥ 0 próximo de 0
$\mathbf{R^2}$	$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, $\boldsymbol{y}_{\boldsymbol{i}}$, $\boldsymbol{N}$	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$	$R^2 \leq 1$ Próximo de 1
$\boldsymbol{r}$	$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, $\boldsymbol{y}_{\boldsymbol{i}}$, $\boldsymbol{N}$	$\frac{\sum_{i=1}^{N}(y_i - y_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}}_i)^2}}$	$-1 \leq r < 1$ Mais próximo de 1

Fonte: Abdi (2010), Massa (2017)

O coeficiente de determinação, R^2, e o coeficiente de correlação linear, r, são as figuras de mérito mais amplamente utilizadas na literatura, apesar de existirem claras limitações em figuras baseadas na correlação entre o valor predito e o valor de referência (Ritter e Carpena, 2013). Uma das maiores limitações do coeficiente de determinação é o fato deste indicador apontar valores próximos à 1 mesmo quando há diferença significativa de magnitude entre o valor predito e o valor de referência, desde que a distribuição de $\widehat{\boldsymbol{y}_i}$ em função de $\boldsymbol{y_i}$ seja linear (Legates e McCabe, 1999). Isto é, considerando $\boldsymbol{y_i} = A\widehat{\boldsymbol{y}_i} + B$, um modelo com alta capacidade preditiva deve ter, além de R^2 próximo de 1, coeficiente angular (A) mais próximo possível de 1, e um coeficiente linear (B) mais próximo possível de zero. Entretanto, um modelo com A e B significativamente diferentes de 1 e 0 ainda pode apresentar R^2 alto desde que A e B sejam relativamente constantes ao longo da distribuição. Esta limitação é também encontrada no coeficiente de correlação linear, apesar deste ser bastante utilizado, juntamente como RESS, para avaliação da qualidade de modelos PLS (Abdi, 2010).

Desta forma, além dos indicadores descritos na Tabela 1, é essencial que se verifique o comportamento da distribuição dos resíduos apresentados pelos conjuntos de calibração e de validação externa para avaliar a qualidade preditiva do modelo desenvolvido. Espera-se que a distribuição dos resíduos seja homoscedástica, apesar do algoritmo de regressão por PLS não assumir qualquer tipo de distribuição dos resíduos (Morelatto, 2010). Além disso, o resíduo percentual deve ficar dentro de uma faixa adequada ao processo, como relatado por Massa (2017), que cita como exemplo um resíduo percentual de no máximo 2% em 95% dos pontos de validação num analisador virtual que estima o teor de um contaminante na saída de um reator do tipo *trickle bed*. Por fim, porém não menos importante, é necessário avaliar a significância estatística dos modelos através do p-valor dos coeficientes de regressão do modelo final.

3.6 MÉTODOS DE SELEÇÃO DE VARIÁVEL

Nesta seção estão descritas as metodologias consolidadas na Literatura e utilizadas na seleção de variáveis durante o desenvolvimento de modelos inferenciais. Entre estas, inclui-se o método de busca exaustiva, mais antigo e com alto custo computacional, e alguns de seus derivados, tais como os métodos sequenciais e algoritmos genéticos (Facchin, 2005). Além destas, descreve-se também um método específico à modelos PLS, denominado de VIP, cujo

valor numérico será utilizado indiretamente na metodologia de seleção proposta por este trabalho, e métodos de regularização, amplamente utilizados na literatura atual (Clark, 2013).

3.6.1 Método de busca exaustiva

O método de busca exaustiva (conhecido em inglês como *All Possible Regressions*), como o próprio nome indica, analisa todas as possibilidades de combinação dentro de um conjunto de variáveis preditoras (Facchin, 2005). Modelos são então gerados para cada combinação e avaliados de acordo com uma figura de mérito, que pode ser uma das descritas na Tabela 1. Supondo que há J variáveis disponíveis para regressão, e considerando que um termo de intercessão seja incluído em todos os modelos, haverá então 2^J modelos a serem construídos. Como exemplo, para $J = 10$, é necessário a construção de um pouco mais de mil modelos. Caso dobre-se a disponibilidade de variáveis preditoras, para $J = 20$, será exigido a construção de mais de um milhão de modelos.

Montgomery, Peck e Vining (2012) alertam para o alto custo computacional que o método de busca exaustiva exige, já que há um crescimento exponencial no número de modelos a serem desenvolvidos em função do número de variáveis preditoras. Entretanto, o aumento da capacidade de processamento junto com o advento de códigos eficientes em computadores modernos permitiu que o método de busca exaustiva fosse mais amplamente utilizado. Ainda assim, em casos onde há um elevado número de variáveis preditoras (como na indústria petroquímica), o custo computacional ainda é relevante e o método tende a não ser utilizado.

3.6.2 Métodos Sequenciais

Com o objetivo de aliviar o custo computacional do método de busca exaustiva, outros métodos de seleção de variáveis foram desenvolvidos de forma a analisar apenas um subconjunto das possibilidades de combinação das variáveis preditoras. Entre estes, destacam-se os métodos sequenciais, onde se avaliam os efeitos da adição ou remoção de uma das variáveis por vez (Facchin, 2005). Montgomery, Peck e Vining (2012) classificam os métodos sequenciais em três categorias:

(i) passo à frente (*forward selection*), onde variáveis são selecionadas por vez;

(ii) passo atrás (*backward elimination*), onde variáveis são removidas por vez;

(iii) seleção *stepwise*, onde variáveis são adicionadas por vez e há teste de redundância para possível eliminação de umas variáveis já selecionadas.

Cada uma das três categorias acima é brevemente descrita abaixo. Uma explanação mais detalhada sobre estas categorias pode ser encontrada em Montgomery, Peck e Vining (2012).

Em (i), parte-se do pressuposto que é possível construir um modelo sem nenhuma variável preditora, contendo apenas um coeficiente linear de interseção. Constrói-se então J modelos com apenas uma variável, e seleciona-se aquela que compõe o modelo com melhor capacidade preditiva (determinada por uma figura de mérito). Testa-se então se a variável selecionada, x_1, causou melhora significativa na capacidade de predição do modelo em comparação ao modelo com o coeficiente linear de interseção apenas. Caso positivo, x_1 é mantida e constroem-se então modelos com duas variáveis de predição, sendo uma destas necessariamente x_1. Verifica-se entre os modelos construídos qual apresenta melhor capacidade preditiva e se esta é significativamente diferente do modelo construído somente com x_1, determinado a partir de um teste de hipóteses. Caso positivo, seleciona-se então a segunda variável, x_2. Repete-se esse procedimento até que não haja mais melhora significativa do modelo, ou até que todas as variáveis disponíveis tenham sido selecionadas.

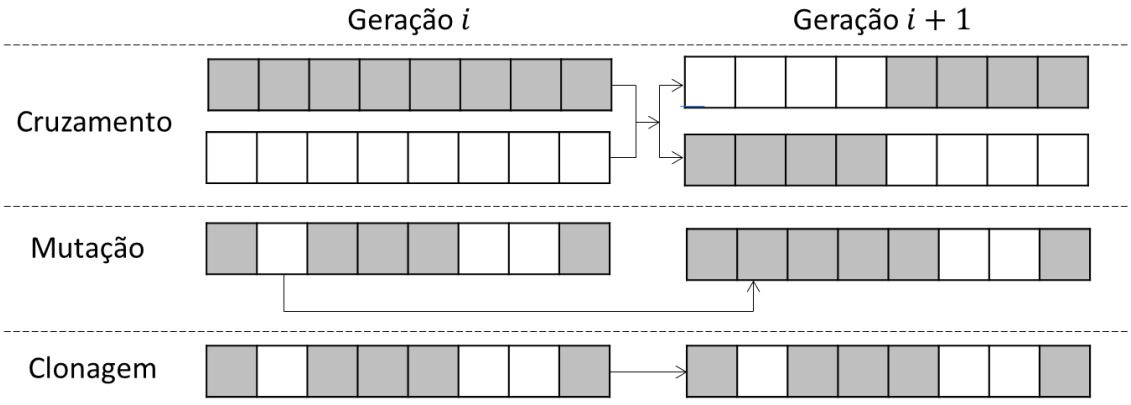
Em (ii), parte-se de um modelo construído com todas as J variáveis preditoras disponíveis. A partir daí, são desenvolvidos J modelos com $J-1$ variáveis preditoras, num cenário onde cada uma das variáveis disponíveis tenha sido removida de um dos J modelos. Verifica-se então, entre os modelos construídos, qual não apresentou piora significativa na capacidade preditiva em comparação ao modelo com todas as J variáveis disponíveis, e remove-se a variável que não está presente naquele. O procedimento é repetido até que todos os modelos construídos após redução do número de variáveis apresentem piora significativa na capacidade de predição, ou até que todas as variáveis tenham sidos eliminadas.

Em (iii), há uma mescla entre o passo à frente e o passo atrás. Inicialmente, aplica-se o *forward selection* para seleção da primeira variável. A partir da seleção da segunda variável, entretanto, aplica-se o *backward elimination* e verifica-se a possibilidade da variável adicionada ser redundante à capacidade preditiva do modelo. Isto é, após a inclusão de uma variável x_n num modelo com n variáveis, constroem-se n modelos com $n-1$ variáveis e, através do passo atrás, verifica-se a possibilidade de exclusão de uma das variáveis previamente selecionadas.

3.6.3 Algoritmos Genéticos

Algoritmos genéticos são métodos de seleção heurísticos inspirados na teoria de evolução natural de Charles Darwin. Por se basear num método estocástico, não há garantia que a seleção de variáveis determinada pelo algoritmo seja ótima, porém há altas chances desta seleção se encontrar próxima ao ótimo global. A inicialização destes tipos de algoritmos se dá por meio da criação aleatória de uma população de subconjuntos que contém k das j variáveis disponíveis. Os subconjuntos são denominados *cromossomos*, enquanto que as variáveis são chamadas de *gene*. Haverá uma seleção dos *genes* da população original, onde a probabilidade de sobrevivência deste está atrelada a resposta de um cromossomo a uma função custo (Facchin, 2005), que é geralmente uma das figuras de mérito para avaliação de modelos descritas na Tabela 1. Cromossomos com resposta favoráveis possuem maior probabilidade de seleção, transmitindo seus *genes* (variáveis) para a próxima geração, que será obtida através de operadores genéticos entre os cromossomos selecionados, os quais incluem cruzamento, mutação e clonagem, descritos na Figura 4.

Figura 4 - Mecanismos utilizados nos Algoritmos Genéticos



As probabilidades de ocorrência dos mecanismos descritos na Figura 4 são parâmetros de inicialização do algoritmo. Estes mecanismos levarão a novas gerações até um critério de parada, que pode ser um parâmetro de performance ou similaridade entre os cromossomos gerados. Uma descrição mais sucinta do algoritmo pode ser encontrada em Han e Yang (2004).

3.6.4 Importância da Variável na Projeção (VIP, *Variable Influence for the Projection*)

Durante a regressão de um modelo PLS, uma variável x_k pode ser importante na modelagem da matriz de saída $\boldsymbol{Y}$, o que faz com que x_k possua um alto coeficiente de regressão

no modelo PLS. De outra maneira, uma variável x_p pode ser importante na descrição da matriz de entrada $\boldsymbol{X}$, o que faz com que x_p possua uma alta carga fatorial. Wold, Sjostrom e Eriksson (2001) relatam que o VIP foi desenvolvido como um índice capaz de resumir a importância de uma variável em ambas $\boldsymbol{X}$ e $\boldsymbol{Y}$, o qual foi chamado de VIP *(Variable Importance for the Projection*, ou, em português, Importância da Variável na Projeção). Este índice é definido, para cada variável, como a raiz da soma ponderada dos quadrados dos pesos do modelo PLS, detalhado matematicamente na Equação (12).

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 . SSY_f . J}{SSY_{total} . F}} \tag{12}$$

em que:

- w_{jf} são os pesos da variável j na componente f do modelo PLS;
- SSY_f é a soma dos quadrados da variância explicada por cada componente;
- J é o número de variáveis;
- SSY_{total} é a soma dos quadrados da variância explicada total; e
- F é o número total de componentes.

Os pesos w_{jf} do modelo PLS refletem, matematicamente, a covariância entre as variáveis de entrada e de saída, e a sua inclusão no cálculo do VIP permite, segundo Andersen e Bro (2010), que o índice pondere não somente na descrição adequada, pelo modelo, das variáveis de entrada, mas também no quão importante a informação contida em cada variável *j* é essencial à predição das variáveis de saída. Wang *et al.* (2015) cita a regra do "maior que um" como o critério amplamente utilizado para seleção de variáveis por VIP. Nesta, somente as variáveis com escores VIP maiores que a unidade são considerados significantes ao modelo, enquanto que as demais devem ser descartadas.

Entretanto, Akarachantachote *et al.* (2014) destaca que a diversidade da estrutura dos dados utilizados em modelos PLS implica que o limite de uma unidade para a seleção de variáveis não é sempre a forma mais adequada de seleção. De fato, Wold, Johansson e Cocchi

(1993) aconselham que a estrutura do modelo seja sempre apreciada antes da inclusão ou remoção de variáveis. Particularmente em modelos PLS, não é aconselhável apenas remover todas as variáveis com escore VIP menor que a unidade (Andersen e Bro, 2010). Como alternativa, as variáveis com os escores mais baixos devem ser removidas primeiro e, caso haja aumento da capacidade preditiva do modelo, as demais variáveis com menor VIP devem ser removidas continuamente até que não se observe mais melhora significativa no modelo.

3.6.5 Métodos de regularização

Clark (2013) relata que o resíduo associado à resposta de um modelo pode ser dividido em três partes: uma associada a variância do valor de referência, uma associada à diferença entre a média dos valores estimados e a média dos valores de referência (conhecida como enviesamento, ou *bias*) e uma última parte que é inerente a qualquer sistema, e, portanto, inevitável. Num modelo ideal, nível baixos de *bias* e de variância são desejáveis. Porém, em modelos onde há grande número de variáveis, observa-se um aumento expressivo do erro associado à variância em detrimento da redução do erro relacionado ao *bias*. Uma redução no número de variáveis (através de seleção) causará uma redução do erro relacionado a variância, porém é possível que haja aumento do erro relacionado ao *bias*. Desta forma, há um ponto ótimo entre o ganho de erro relacionado ao *bias* e a perda de erro relacionada variância. Métodos de regularização, tais como o *LASSO*, o *RIDGE* e o *Elastic Net* são fundamentos nesta premissa, e um detalhamento maior de cada um destes pode ser encontrado em Clark (2013).

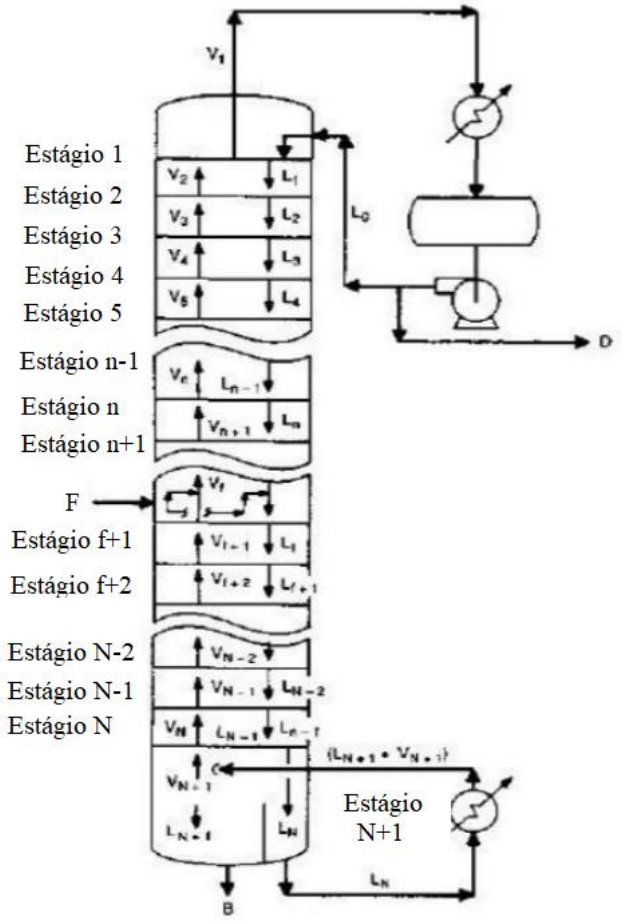
3.7 O PROCESSO DE DESTILAÇÃO FRACIONADA DA NAFTA

A destilação fracionada é um processo amplamente utilizado na indústria petroquímica. Neste, há sucessivas vaporizações e condensações do material introduzido numa coluna com o objetivo de separar os componentes mais voláteis (que tendem ao topo) dos menos voláteis (que tendem ao fundo). A nafta é um dos produtos da destilação fracionada do petróleo cru, e a sua posterior destilação ou craqueamento é uma das etapas do processo de produção dos gases eteno e do propeno, derivados do petróleo largamente utilizados na indústria petroquímica (Antunes, 2007). A nafta também pode ser utilizada para produção dos BTX (benzeno, tolueno e xileno), uma fração líquida com altíssimo valor agregado.

Como descrito em Kister (1990) e representado na Figura 5, a destilação resume-se em alimentar a coluna num estágio conhecido como *feed* (F, ou alimentação) e permitir o contato

entre o vapor (V) que ascende do estágio (ou prato) inferior e o líquido (L) que escoa do prato superior. Desta forma, o fluxo de vapor irá para o estágio superior mais rico nos componentes mais voláteis, enquanto que o líquido escoará ao prato inferior mais rico nos componentes menos voláteis. O vapor percorrerá a coluna e será coletado no topo, enquanto que o líquido será conduzido ao fundo da coluna pela ação da gravidade. É possível também retirar produtos em estágios intermediários entre o fundo e o topo, se este for de interesse comercial. Caso haja coleta de produtos em estágios intermediários, o processo é denominado de destilação fracionada.

Figura 5 - Coluna de Destilação



O número de estágios requerido para alcançar a separação desejada depende de inúmeros fatores, que incluem a eficiência da coluna e a composição do *feed*. Para uma coluna de destilação com um número de pratos já definido, a composição do *feed* é essencial para alcançar a separação desejada e retirar os produtos pretendidos. No caso da nafta, é comum que

sua composição seja indiretamente analisada através da determinação do ponto inicial de destilação (P-000), e das temperaturas onde se observa 5% (P-005), 10% (P-010), 30% (P-030), 50% (P-050), 70% (P-070), 90% (P-090), 95% (P-095) de nafta vaporizada, além do ponto final de destilação (P-100) (Bezerra, 2005). Existe um intervalo para cada um destes nove pontos de temperatura que irá determinar se a nafta a ser destila possui qualidade adequada ou não para gerar os produtos desejados. Entretanto, é preciso que coletar uma amostra do *feed* da coluna e, através de análise laboratorial, definir os nove pontos de temperatura, processos que demandam tempo e elevam os custos de operação.

4. REVISÃO DE LITERATURA

4.1 ANALISADORES VIRTUAIS E MODELOS PLS

A problemática da medição de variáveis essenciais ao controle de processos químicos (como a composição de um destilado) é discutida na literatura há décadas. Tham *et al.* (1990) relata que estudos na área de controle das variáveis de saída de processos industriais geram publicações desde a década de 1970. Ainda assim, os autores citam dificuldades em detectar de forma eficiente perturbações em processos químicos que causam um descontrole da saída. Essa dificuldade é gerada, entre outros fatores, por limitações na medição de certas variáveis de controle, que vão desde restrições na coleta de amostras até uma demora significativa no tempo de resposta, a depender da técnica aplicada. Estudos de Soderstrom (1980) e Parrish & Brosilow (1985) levaram ao desenvolvimento de algoritmos especiais para o controle dessas variáveis de difícil aquisição. Soderstrom formulou situações de controle onde a entrada era manipulada nos intervalos de amostragem da variável de saída, com resultados limitados a processos de primeira ordem e sem possibilidade de comparação com sistema mais complexos. Parrish & Brosilow também utilizaram uma técnica baseada na reconstrução dos efeitos de um distúrbio manipulado, porém eles desenvolveram parâmetros de controle através de regras de afinação heurísticas por meio dos resultados obtidos.

Além da abordagem utilizada por Soderstrom e Parrish & Brosliow, uma outra solução encontrada para a problemática do controle de variáveis de difícil aquisição, e relatada por Tham *et al.* (1990), seria o uso de informações contidas em outras variáveis de fácil medição, como temperatura e pressão, para estimar a variável de controle. Numa situação ideal, o funcionamento da planta seria completamente observável por estas variáveis secundárias, e técnicas como o Filtro de Kalman e o Filtro de Kalman Estendido poderiam ser utilizadas para modelar o estado de funcionamento da planta e, consequentemente, a variável de controle desejada. Entretanto, essas técnicas limitam-se a processos onde é possível determinar o estado de funcionamento da planta através de variáveis secundárias, o que nem sempre é o caso, como relatado no estudo por Tham *et al.* (1990), onde os autores sugerem dois estimadores adaptativos (há atualização a partir de novos dados) que utilizam variáveis de rápida amostragem para inferir sobre variáveis de saída que estão sujeitas a altos tempo de resposta,

com objetivo de tornar o controle da composição de topo de uma coluna de destilação mais eficiente.

Os estudos de Tham *et al.* (1990) ilustram uma das soluções encontradas para a problemática de medição de variáveis de difícil aquisição, onde haveria uma substituição de um sensor físico por um não físico, também chamado de *soft sensor* ou "analisador virtual". Tais sensores são baseados em modelos inferenciais construídos a partir de dados históricos de variáveis facilmente medidas, onde se estabelece uma relação matemática (linear ou não) entre estas e a variável de controle. Apesar de utilizar as medições de variáveis secundárias, os autores não utilizam uma relação matemática puramente empírica para estimar diretamente a variável de saída, já que a inferência desta se dá indiretamente através do que os autores chamam de *estimadores adaptativos*. Já havia, entretanto, publicações onde modelos do tipo caixa-preta haviam sido empregados com esse propósito, a exemplo de modelos de regressão por PLS na área de Quimiometria citados por Geladi e Kowalski (1986). A publicação de 1986 de Geladi e Kwalski tinha como objetivo, segundo os autores:

> *"... [fornecer] um tutorial no método de regressão por mínimos quadrados parciais. Pontos fracos em outros métodos de regressão são apresentados e a regressão é desenvolvida de forma a se apresentar como solução para tais limitações. Um algoritmo para um PLS preditivo e outras dicas práticas para sua utilização são fornecidas".* (Geladi e Kowalski, 1986, tradução nossa)[1]

De fato, o tutorial fornecido por Geladi e Kowalski é uma das principais fontes para pesquisadores que utilizam regressão por PLS, tendo sido citado mais de 4860 vezes, de acordo com a plataforma *Science Direct*. Matematicamente descrito na seção 2.4.1 deste trabalho, o algoritmo conhecido como Mínimos Quadrados Parciais (PLS, da sigla em inglês *Partial Least Squares*) foi desenvolvido durante as décadas de 1960 e 1970 por Herman Wold como solução de problemas na área de Econometria, tendo sido adaptado por Svante Wold (filho de Herman Wold) e Harald Martens na década seguinte, onde passou a ser largamente utilizado como

[1] *"A tutorial on the partial least-squares (PLS) regression method is provided. Weak points in some other regression methods are outlined and PLS is developed as a remedy for those weaknesses. An algorithm for a predictive PLS and some practical hints for its use are given."*

método de regressão na área de Quimiometria (Geladi e Kowalski, 1986). Desde então, é uma das técnicas mais utilizadas como alternativa para regressões múltiplas simples, cuja performance é altamente afetada pela presença de colinearidade entre as variáveis preditoras, sendo aplicada principalmente também nas áreas de bioinformática, *machine learning* controle de processos (Akarachantachote *et al.*, 2014).

Além da regressão múltipla simples, diversos outros métodos de regressão para analisadores virtuais com modelos do tipo caixa-preta foram desenvolvidos e/ou consolidados na literatura desde a criação do PLS na década de 1960. Massa (2017) cita algumas destas técnicas em seu estudo, onde a mesma destaca três destes métodos: a Regressão por Componentes Principais (PCR), a Análise Canônica Independente (ICA) e a Identidicação subespacial (SSI). Há diversos relatos na literatura onde analisadores virtuais foram construídos através destes métodos, além da Regressão Linear Múltipla (MLR) e da Regressão por Mínimos Quadrados Parciais (PLSR). A Tabela 2 ilustra a quantidade de publicações na qual um dos objetivos foi o desenvolvimento de analisadores virtuais usando os métodos citados, nos últimos cinco anos, em inglês, onde houve um número igual ou superior a cinco citações e que estão disponíveis na plataforma *Science Direct*.

Tabela 2 – Número de publicações na plataforma Science Direct relacionadas ao desenvolvimento de analisadores virtuais (por método)

Método	**Trabalhos publicados (em inglês) com mínimo de cinco citações**
MLR	47
PCR	183
PLSR	187
ICA	68
SSI	39

Fonte: Plataforma Science Direct (2019)

Dentre os métodos descritos na Tabela 2, destaca-se o uso da PLSR e da PCR. Como observado por Massa (2017), estes são capazes de lidar com situações onde há alta correlação entre variáveis de entrada, o que afeta significativamente a capacidade preditiva de modelos que assumem independência entre as variáveis preditoras, tal como o MLR. Entretanto, estudos de Maitira e Yan (2008) indicam que, entre o PLSR e o PCR, o primeiro apresenta melhor

desempenho quando existe alta correlação entre uma das variáveis de entrada e a variável a ser estimada, situação comum em modelos da indústria petroquímica. Além disso, os autores citam os principais propósitos de ambas as técnicas.

> *"PCA e PLS têm dois propósitos na análise de regressões. Primeiro, ambas técnicas são utilizadas para converter um conjunto de variáveis altamente correlacionadas num conjunto de variáveis independentes. Segundo, ambas técnicas são utilizadas para redução de variáveis. Quando a variável dependente para uma regressão é especificada, a regressão por PLS é mais eficiente do que a por PCA para essa redução devido a natureza supervisionada do algoritmo"* (Maitira e Yan, 2008, tradução nossa)[2]

Não há somente melhor desempenho do modelo por rgressão PLS quando comparada a PCR em casos onde se especifica a variável de saída. Pesquisas de Abdi (2010) atentam para o fato de que na PCR os componentes ortogonais que serão utilizados na regressão da variável de saída são desenvolvidos de forma a explicar máxima variabilidade da matriz de entrada $\boldsymbol{X}$. Dessa forma, não há garantia que tais componentes serão relevantes na predição de $\boldsymbol{Y}$. A PLSR, entretanto, desenvolve componentes ortogonais de forma a explicar a máxima covariância entre $\boldsymbol{X}$ e $\boldsymbol{Y}$, o que tende a melhorar a capacidade preditiva de modelos via regressão PLS em comparação aos modelos desenvolvidos via PCR.

4.2 SELEÇÃO DE VARIÁVEIS E VIP

Apesar de significativo, o tutorial de Geladi e Kowalski (1986) não discute alguns aspectos relevantes ao desenvolvimento de modelos via regressão por PLS, tais como a detecção de outliers, o tratamento de dados faltantes, testes F e t, além da seleção de variáveis. Morais Júnior (2011) destaca que a seleção adequada de variáveis que serão utilizadas como entrada num modelo inferencial é determinante para o sucesso do mesmo. Uma seleção inadequada pode comprometer as características preditivas de um modelo e reduzir significativamente sua capacidade inferencial. De fato, Yun *et al.* (2019a) destaca que em

[2] *"PCA and PLS serve two purposes in regression analysis. First, both techniques are used to convert a set of highly correlated variables to a set of independent variables by using linear transformations. Second, both of the techniques are used for variable reductions. When a dependent variable for a regression is specified, the PLS technique is more efficient than the PCA technique for dimension reduction due to the supervised nature of its algorithm."*

modelos preditivos com grande número de variáveis, apenas umas parte destas deve estar relacionada com a propriedade de interesse, e métodos de seleção de variável partem do pressuposto que uma escolha adequada de um número menor de variáveis pode não somente melhorar a capacidade preditiva do modelo, mas também tornar a calibração mais confiável, além de facilitar a interpretação dos resultados, particularmente em modelos PLS.

> *"A seleção de variáveis, que é um passo crucial na calibração multivariada, pode ser utilizada por três razões: (1) fornecer variáveis com menor tempo de resposta e com maior relação de custo-benefício, devido à redução de dimensionalidade; (2) melhorar a predição da performance das variáveis selecionadas; (3) fornecer um melhor entendimento e interpretação dos modelos obtidos"* (Yun *et al.*, 2019b, tradução nossa)

Nota-se que apesar de afirmar que a seleção de variáveis leva a uma melhor relação de custo-benefício no modelo, devido à redução de dimensionalidade, Yun *et al.* (2919a) não propõe nenhuma forma de análise dos custos associados durante a seleção de variáveis em si. A melhora na relação do custo-benefício se dá de maneira implícita, já que a redução de variáveis ocasionaria numa respectiva redução de custos associados. De fato, os métodos de seleção descritos na seção 2.6 possuem abordagem puramente estatística, incluindo o VIP, que foi desenvolvido especificamente para modelos de regressão PLS.

De fato, estudos de Wold, Sjostrom e Eriksson (2001) sobre modelos de regressão PLS concluíram que uma variável de entrada deveria ser considerada importante em duas ocasiões diferentes: 1) a variável é importante à modelagem da matriz $\boldsymbol{Y}$, o que geraria altos coeficientes de regressão; 2) a variável é importante à modelagem da matriz $\boldsymbol{X}$, o que geraria altas cargas fatoriais. É este o propósito do escore VIP, que foi desenvolvido de forma a avaliar a importância de uma variável em ambas as situações, como demonstra a equação (12). Desde seu desenvolvimento, o VIP é uma das técnicas de seleção de variáveis mais aplicadas em modelos de regressão por PLS, como detalha Andersen e Bro (2010).

4.3 MODELOS PLS EM PROCESSOS INDUSTRIAIS

Pesquisas recentes publicadas na literatura demonstram que a regressão por Mínimos Quadrados Parciais tem tido um papel relevante no desenvolvimento de modelos preditivos que serão utilizados em ambientes industriais. Na Tabela 3 estão descritos os autores, os objetivos, os métodos utilizados, as variáveis analisadas e os principais resultados obtidos dos produtos

gerados por quatro destas pesquisas. Os estudos de Massa (2017) e de Silva (2017) se desenvolveram a partir de dados gerados na mesma planta industrial que gerou os dados analisados neste trabalho. A utilização da mesma técnica de regressão em estudos com objetivos distintos que utilizam dados de seções diferentes da mesma planta industrial evidenciam a relevância do desenvolvimento de modelos via regressão PLS. Além disso, os estudos de Massa (2017) sugerem, em suas considerações finais, a necessidade de analisar a seleção de variáveis sob a luz de um indicador que seja capaz de avaliar o ganho de performance em função do aumento de custos associados durante a seleção de variáveis, questão central tratada neste trabalho. As demais publicações na Tabela 3, de Harrou, Nounoue e Madakyaru (2015) e Hug et al. (2015), evidenciam a quantidade de citações que pesquisas com essa temática tem tido na literatura atual. O número de citações segundo a plataforma *Science Direct* para estas publicações é de 25 e 39, respectivamente. Apesar da grande quantidade de citações, ambas pesquisas não consideram o custo associado a medição das variáveis de entrada no desenvolvimento dos modelos PLS, uma ausência também notada nos trabalhos de Massa (2017) e Silva (2017).

Estudos de Kano e Ogawa (2010) indicam que, dentro da indústria petroquímica, o processo de destilação é onde se encontra a maior aplicação de *soft sensors*, já que a medição exata de composição do produto é raramente utilizada como uma variável de controle devido à dificuldade de medição da mesma em tempo real. De fato, a maioria dos analisadores em linha (que são cromatógrafos a gás ou espectrômetros no infravermelho próximo, NIR) apresentam atraso significativo na reposta de composição e/ou altíssimos custos de investimento e manutenção, o que induz ao desenvolvimento de analisadores virtuais. Não há, entretanto, nenhum relato da análise da manutenibilidade dos equipamentos que fornecem dados para estes *soft sensors* nos estudos de Kano e Ogawa.

Tabela 3 – Aplicações recentes de modelos PLS a processos industriais

Autor(es) (Ano)	Objetivos	Variáveis avaliadas	Métodos	Principais resultados obtidos
Massa (2017)	Desenvolver um analisador virtual para estimar o teor de contaminante MAPD na saída de um reator trickle bed.	Pressão, vazão, temperatura, teor de MAPD na carga fresca.	PLS, VIP	Modelos PLS-VIP fornecem uma estimativa confiável do teor de MAPD na saída; há alto potencial de uso dos analisadores virtuais para estimar o teor de MAPD.
Silva (2017)	Transferência de calibração de um espectrômetro NIR de bancada para um NIR ultracompleto; desenvolver um simulador de octanagem da gasolina automotiva.	espectros de componentes puros da gasolina e misturas diesel/biodiesel.	PLS	Simulações foram capazes de fornecer valores preditos confiáveis; há potencial para utilização do modelo construído em instrumentos portáteis como mais uma ferramenta de otimização.
Harrou, Nounoue e Madakyaru (2015) *39 citações*	Combinar as vantagens da univariada EWMA (média-móvel exponencialmente ponderada) com o multivariado PLS para melhorar a um modelo de detecção de falhas.	Temperaturas, pressão e fluxos de uma coluna de destilação.	PLS	Simulações demonstraram que o modelo é mais eficaz na detecção de falhas quando comparado ao modelo PLS tradicional, particularmente quando há presença de falhas de pequena magnitude.
Hug *et al.* (2015) *25 citações*	Identificar compostos que covariam com a atividade mutagênica em amostras de ambientes complexos.	Características químicas de efluentes de uma planta de tratamento de águas residuais de plantas industriais.	PLS	A regressão por PLS conseguiu identificar quatro compostos que apresentam covariância com a atividade mutagênica, entre eles um não identificado anteriormente, tornando possível separa-los dos menos relevantes toxicologicamente.

5. METODOLOGIA

Um fluxograma da metodologia desenvolvida e aplicada neste trabalho é apresentado na Figura 7. Uma descrição mais detalhada de cada etapa da metodologia é feita abaixo.

5.1 AQUISIÇÃO E ORGANIZAÇÃO DE DADOS

A coluna de destilação onde a nafta será introduzida faz parte de uma planta de produção de propileno de uma grande indústria brasileira. Como em qualquer outra indústria moderna, há diversos sensores ao longo da planta que monitoram variáveis sensíveis ao controle da mesma. Entre estes estão equipamentos que monitoram fluxos de vazão, indicadores de nível, pressão e temperatura, além de outros equipamentos que monitoram variáveis sensíveis ao processo, tais como o teor de um contaminante, por exemplo. Estes sensores possuem monitoramento real e dados históricos das variáveis de processo sendo monitoradas estão disponíveis para construção do modelo.

Os dados utilizados na construção dos modelos presentes neste trabalho representam o período cronológico de um ano, começando em 1 de janeiro de 2018, e terminando em 31 de dezembro de 2018. A periodicidade das análises laboratoriais que determina a qualidade da nafta é diária, onde um relatório é emitido com os nove pontos de temperatura em função da percentagem de nafta vaporizada (de P-000 a P-100). Estes dados foram organizados em uma matriz de saída, denominada $\boldsymbol{Y}$, com 9 colunas e 365 linhas. Dentre os sensores disponíveis na planta, os 129 que aferem medidas de vazão, nível de equipamento, pressão e temperatura antes da saída da nafta da coluna de destilação foram indicados por operadores da planta como candidatos para construção do modelo. Estas foram divididas em variáveis do tipo FC (fluxo) do tipo LC (nível), tipo PC (pressão) e do tipo TC (temperatura). Assim sendo, organizou-se uma matriz de entrada $\boldsymbol{X}$, com 129 colunas e 365 linhas.

Por questões de sigilo, o *tag* que identifica a variável no sistema foi substituído pela letra V, seguido por três números. Desta forma, cada variável recebeu a identificação VNNN, onde N é um algarismo entre 0 e 9. As variáveis também foram agrupadas de acordo com o tipo. O primeiro grupo representa as variáveis de fluxo. Assim sendo, a V001 é variável de fluxo que aparece primeiramente no processo, enquanto que a V002 é a segunda variável de fluxo, e assim por diante. A Tabela 4 detalha o intervalo de cada tipo de variável, bem como a unidade em que esta é medida.

Tabela 4 - Identificação das Variáveis de Entrada

Intervalo	Tipo (unidade de medição)
V001 a V027	Fluxo (ton/h ou kg/h)
V028 a V041	Pressão (kgf/cm^2)
V042 a V079	Nível (m)
V080 a V129	Temperatura ($°C$)

5.2 ANÁLISE EXPLORATÓRIA E DETECÇÃO DE *OUTLIERS*

Uma extensa análise exploratória das 129 variáveis de entrada e 9 variáveis de saída, que incluem a inspeção das estatísticas básicas de cada uma (média, desvio, quartis, entre outras), além da análise dos gráficos das séries temporais, foi feita para entender o comportamento das variáveis que a serem utilizadas no modelo. Devido a diferença de dimensão entre as variáveis, estas foram escalonadas segundo a técnica descrita na seção 2.2 deste trabalho. Isto feito, o método de diagnóstico de modelo proposto por Li *et al.* (2016) e descrito na subseção 2.2.1 foi executada para eliminar pontos considerados aberrantes. Por fim, uma matriz de correlação linear, com os respectivos coeficientes de correlação linear e p-valores, foi produzida para analisar, de forma preliminar, de quais formas as variáveis estudadas se relacionam entre si.

5.3 MODELO COMPLETO E HIERARQUIZAÇÃO DE VARIÁVIS

Os dados limpos, sem a presença de outliers, foram particionados em subconjuntos de calibração (80%) e validação (20%) através do algoritmo KSS, detalhado na seção 2.3. Um único modelo PLS foi construído com a presença de todas as variáveis disponíveis (tanto de entrada, como de saída). É importante notar que a regressão por PLS é capaz de modelar e analisar diferentes variáveis de saída concomitantemente, ao invés de gerar um modelo diferente para cada variável resposta.

Uma análise da capacidade preditiva do modelo sem seleção de variáveis é realizada através das figuras de mérito deste e dos resíduos gerados. Isto feito, os escores VIP são calculados através das equações descritas na seção 2.6.3. Estes escores serão então utilizados para hierarquizar as variáveis de entrada.

5.4 CUSTO TOTAL DO MODELO

A nova abordagem para seleção de variáveis proposta por esse trabalho introduz um traço econômico no algoritmo de seleção. Para tanto, é preciso primeiro definir o custo total associado a um modelo inferencial ao longo de t anos. Este custo é definido pela Equação (13)

$$C_t = C_s(J) + \sum_{j=1}^{J} \left(C_{rj}\tilde{k}_j\right) \tag{13}$$

Onde $C_s(J)$ representa a soma dos custos fixos associados aos equipamentos de medição de cada variável, que é uma função do número J de variáveis; C_{rj} representa o custo de substituição do equipamento utilizado na medição da variável j; e $\tilde{k}_j$ representa o número de falhas desse equipamento ao longo de t anos. Assume-se que o número de falhas de um certo equipamento segue uma distribuição de Poisson, comumente utilizada para descrever processos de contagem associados a falhas em industrias (Ross, 2014). A Equação (14) descreve a função densidade de probabilidade da distribuição de Poisson, onde k é o número de eventos em certo intervalo de tempo, λ é a taxa de falha do equipamento nesse intervalo e e o número de Euler.

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{14}$$

O número de falhas de um certo equipamento ao longo de t anos, $\tilde{k}_j$ na Equação (13), é então determinado com a mediana da distribuição descrita pela Equação (14). Isto é, o valor de k de forma que $P(0 \leq X \leq k) = 0{,}5$. Dados de manutenção e dos fornecedores dos equipamentos utilizados na medição das variáveis de entradas descritas na Tabela 4 foram utilizados para estimar as respectivas taxas de falha λ de cada um dos sensores. Estas foram então utilizadas na Equação (14) para determinar o número mediano de falhas de cada equipamento ao longo de um período de $t = 5$ anos. Estes valores foram utilizados na Equação (13), juntamente com o custo de substituição de cada equipamento, para determinar o custo total do modelo;

5.5 OTIMIZAÇÃO POR FATOR RC

Para um número J de variáveis disponíveis como entrada, existem J diferentes modelos PLS em potencial. O primeiro modelo a ser construído deve incluir todas as J variáveis e ter h variáveis latentes. Os escores VIP para todas as variáveis são determinados e as variáveis são

hierarquizadas por ordem crescente de VIP. O algoritmo de seleção começa então calculando o custo total máximo, onde todas as J variáveis de entrada (com seus respectivos custos de substituição e taxas de falhas dos equipamentos) são inseridos na Equação (13). A variável com o menor escore no VIP é removida e um novo modelo PLS é construído, com um novo custo total associado. O valor de h pode ser alterado ou não, desde que os mesmos critérios utilizados na decisão de h sejam mantidos. Este procedimento se repete até que haja somente uma variável restante. Dessa forma, para cada modelo construído com $n = J, J-1, J-2, \ldots$ número de variáveis, existem figuras de mérito (que medem a capacidade preditiva do modelo) e um custo associado ao modelo. O Fator RC é então definido pela Equação (15)

$$RC_n = \frac{r_n - r_{n-1}}{c_n - c_{n-1}}, \quad 2 \leq n \leq J \tag{15}$$

em que

- r é o coeficiente de correlação entre o valor observado da referência (y) e o valor predito pelo modelo PLS utilizando n variáveis, ($\hat{y}$);
- c_n é o custo padrão associado ao modelo PLS com n variáveis, sendo o padrão o custo total máximo, onde $n = J$.

O custo padronizado associado ao modelo, c_n, cresce em função do número n de variáveis utilizadas no modelo. Apesar do mesmo ser esperado para o coeficiente de correlação, um crescimento só será verificado se a nova variável inserida aumentar a capacidade preditiva do modelo, o que não é sempre garantido. Desta forma, um valor negativo de fator RC indica uma piora na capacidade inferencial do modelo. Segundo a Equação (15), quanto maior for o valor de RC, maior será o ganho de performance em função do aumento de custo.

Como alguns equipamentos de medição das variáveis de entrada (tais como sensores de temperatura) possuem valores de substituição ou de taxa de falha muito pequenos, a diferença entre c_n e c_{n-1} pode ser ínfima, o que induz a valores muito altos de RC segundo a Equação (15). Este fato torna difícil a comparação entre os diferentes valores de RC para os J modelos PLS construídos. De forma a facilitar essa comparação, uma escala logarítmica pode ser utilizada. Entretanto, deve-se considerar a possibilidade de valores negativos de RC (impossibilitando a utilização de um logaritmo) e de valores muito próximos a zero, que resultariam em logaritmos extremamente negativos. De forma a manter a interpretação do fator

RC (quanto mais positivo, maior o aumento na capacidade preditiva do modelo, e quanto mais negativo, maior a piora na capacidade preditiva do modelo) e facilitar a interpretação numérica em caso onde a diferença entre c_n e c_{n-1} é ínfima, define-se o fator logaritmo de RC (LRC) de acordo com função definida pela Equação (16).

$$\begin{aligned} RC_n \geq 0, \quad & LRC_n = \ln(RC_n + 1) \\ RC_n < 0, \quad & LRC_n = \ln(1 - RC_n) \times \frac{|RC_n|}{RC_n} \end{aligned} \tag{16}$$

A Equação (16) cria uma escala logarítmica dos valores de RC sem remover a natureza positiva/negativa dos fatores, e suas respectivas interpretações. É importante salientar que o pico do fator LRC implica no melhor *trade off* entre aumento de custo e ganho de performance. O algoritmo poderia então selecionar o número de variáveis que maximiza o fator LRC. Porém, nota-se que esse pico sinaliza que a variável adicionada acarretou num maior ganho de performance em função do aumento de custo em relação as variáveis presentes no modelo anterior. Entretanto, as magnitudes de r e c_n devem ser analisadas antes da seleção de variáveis. Para $LRC > 0$ e $r > 0{,}85$, cria-se uma zona de aceitação, onde estão presentes os modelos que atendem a estas restrições. Um dos três critérios a seguir deverá ser aplicado para escolha do modelo que contém as variáveis selecionadas:

(i) modelo com menor custo associado, sugerido em casos onde haja diferença considerável entre os custos dos modelos na zona de aceitação, e as capacidades preditivas destes modelos sejam similares;

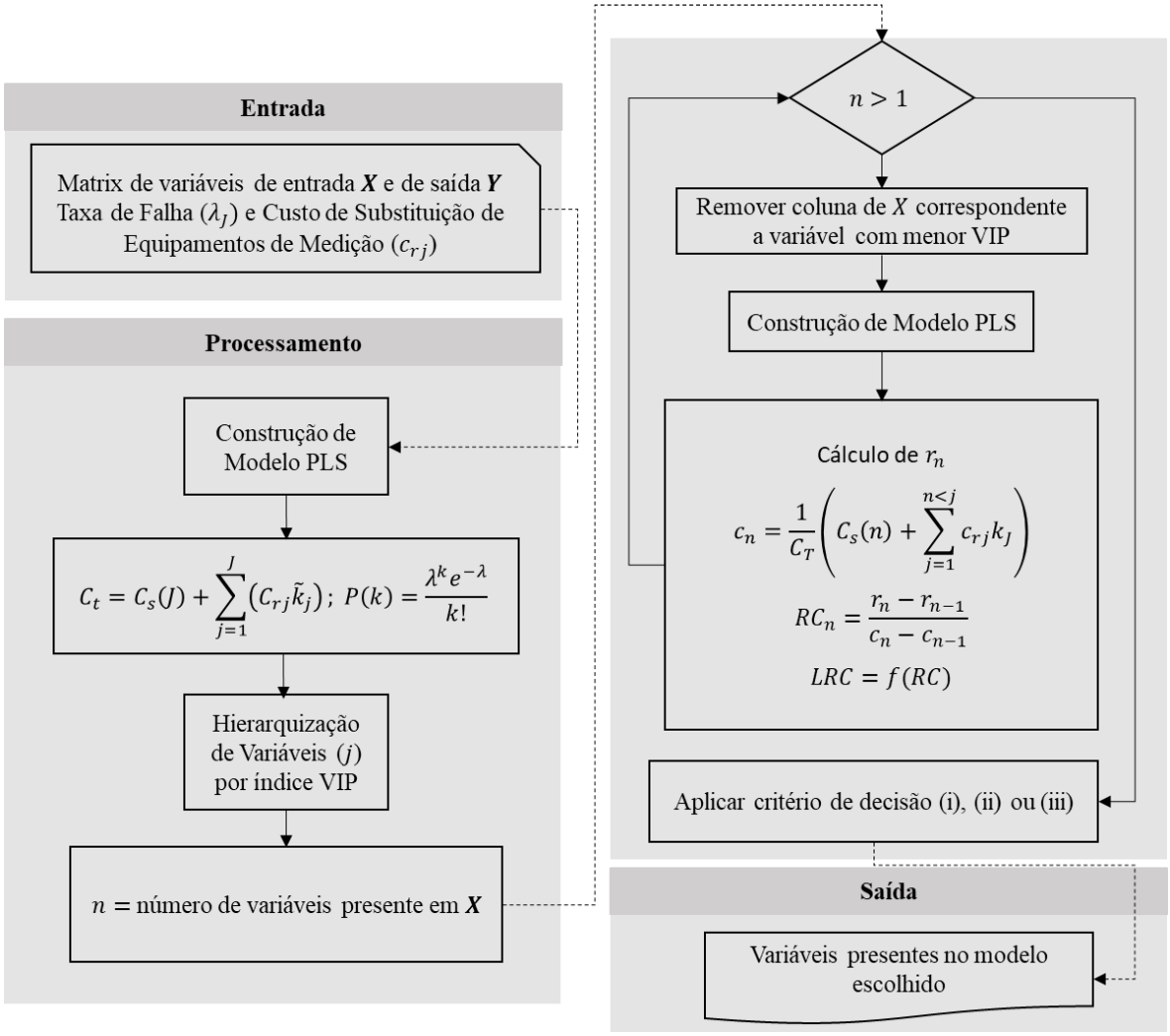
(ii) modelo com maior coeficiente de correlação, sugerido em casos onde não haja diferença considerável entre os custos dos modelos da zona de aceitação, e as capacidades preditivas sejam significativamente diferentes;

(iii) modelo com maior fator LRC, sugerido em casos onde não haja diferença considerável entre os custos dos modelos da zona de aceitação, e as capacidades preditivas destes modelos sejam similares.

Em todas as três situações descritas acima, o termo *custos consideráveis* é subjetivo ao tamanho, capacidade econômica e planejamento financeiro da instituição que pretende aplicar

a abordagem proposta neste trabalho. A Figura 6 descreve o proposto algoritmo para seleção de variáveis.

Figura 6 – Novo Algoritmo para Seleção de Variáveis



É importante salientar que cada variável de saída irá gerar um fator LRC diferente para cada modelo gerado com n variáveis de entrada. Nota-se que a inclusão de uma variável de entrada pode causar aumento significativo da capacidade preditiva de uma das variáveis de saída, mas reduzir a capacidade do modelo em predizer uma outra variável resposta. De forma a analisar o efeito global, isto é, em todas as nove variáveis de saída, o parâmetro a ser utilizado nos critérios (i), (ii) e (iii) para escolha do melhor modelo é a *soma* dos fatores LRC, conhecido como SLRC e detalhado na Equação (17).

$$SLRC_n = \sum_{a=1}^{p} LRC_{n,a} \quad (17)$$

Onde $LRC_{n,a}$ é o fator LRC da variável de saída a num modelo com n variáveis de entrada; e p é o numero de variáveis de saída, nove no caso estudado. De forma análoga, é preciso analisar a capacidade preditiva global de um modelo com n varáveis de entrada em estimar as nove variáveis de saída. Para isto, a média harmônica das correlações entre os valores observados e os preditos para as saídas ($\overline{r_h}$), calculadas com o conjunto de validação, é determinada. A média harmônica, descrita na Equação (18), é considerada mais conservadora do que a média aritmética e é amplamente utilizada para analisar medidas de tendência central (Xu, 2009), já que esta é muito mais sensível a variabilidade entre os valores do que a média aritmética.

$$\overline{r_h} = \frac{9}{\sum_{j=1}^{9} \frac{1}{r_j}} \quad (18)$$

Após a definição do fator $SLRC$ para cada um dos J modelos e da média harmônica das correlações entre os valores observados e estimados pelo modelo nas nove variáveis de saída, $\overline{r_h}$, a construção de um gráfico de $SLRC$ versus $\overline{r_h}$ ilustra a relação entre a capacidade preditiva do modelo e o ganho de performance em função do aumento de custo associado a adição de uma variável. O critério utilizado para seleção do modelo que possui as variáveis selecionadas é o (i), que considera o modelo com menor custo associado.

5.6 TESTE E VALIDAÇÃO DO MODELO COM SELEÇÃO DE VARIÁVEIS

As figuras de mérito RMSEP e r, descritas na Tabela 1, serão utilizadas para avaliar a capacidade preditiva do modelo construído com seleção de variáveis. Ressalta-se que os modelos construídos na seção 4.5 utilizam as observações do mesmo grupo de calibração (80% dos dados) utilizado no desenvolvimento do modelo completo na seção 4.3. De forma análoga, o r é calculado com as observações do correspondente grupo de validação.

Um gráfico com pontos cujas coordenadas são os valores observados na saída (obtidos por análise laboratorial) e os estimados pelo modelo é construído para cada uma das nove saídas, com o objetivo de ilustrar a performance do modelo na estimação dos pontos de

temperatura da nafta. Retas de suporte de diferentes cores com 0% (azul), 1% (verde), 2% (amarela) e 5% (vermelha) de erro percentual auxiliam na análise da performance do modelo, que será implementado no sistema de controle do processo.

5.7 SOFTWARE R-STUDIO

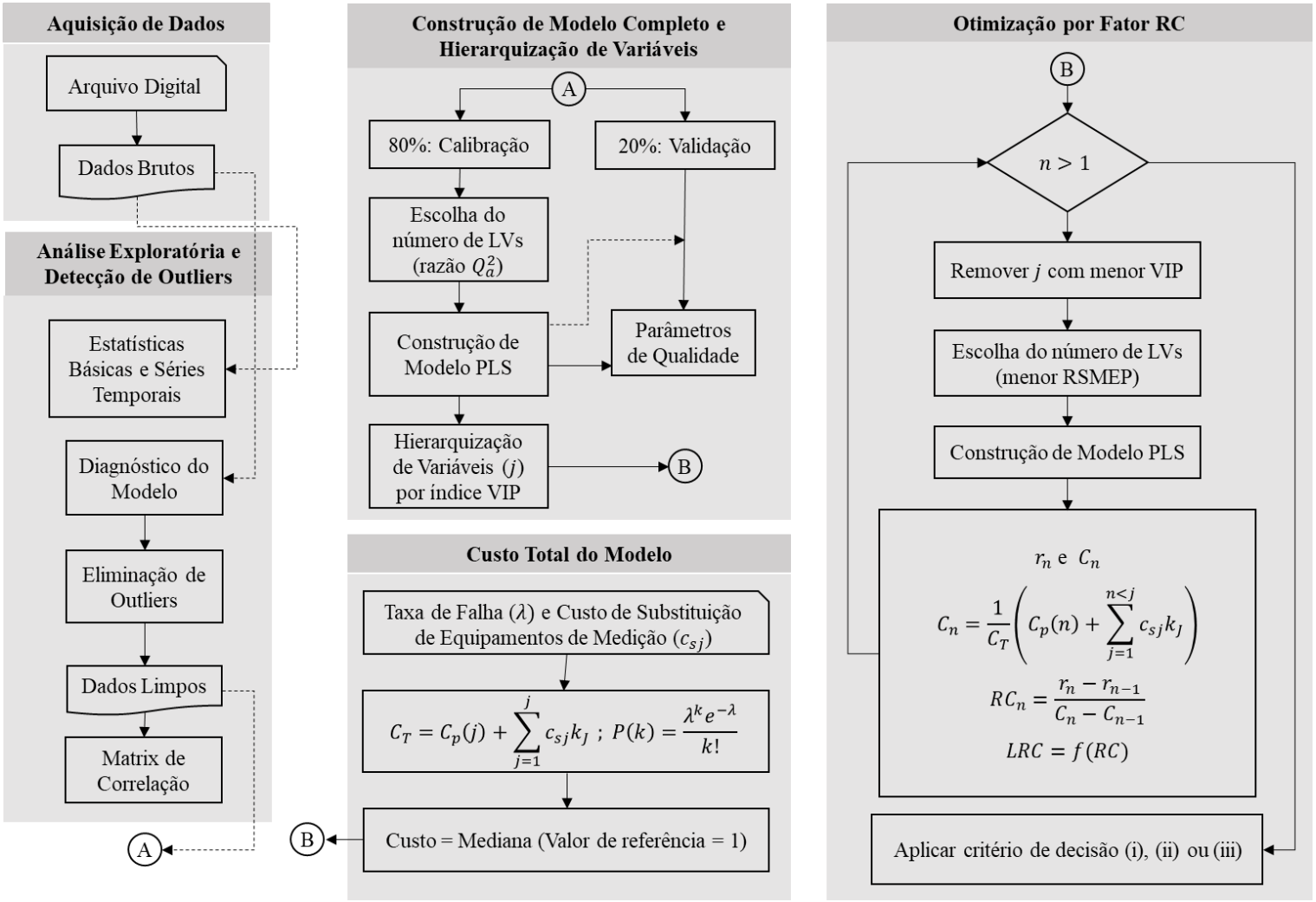
Todas as ações necessárias à produção deste trabalho, que incluem desde a organização de dados em matrizes e análises estatísticas/gráficas, até a construção de modelos matemáticos, foram realizadas por meio do IDE (Ambiente de Desenvolvimento Integrado, da sigla em inglês *Integrated Development Environment*) do RStudio®. O IDE do RStudio contempla uma plataforma moderna onde é possível importar dados, realizar atividades de programação e gerar imagens em um único ambiente. Este ambiente opera através da linguagem de programação *R*, uma linguagem aberta e em crescente uso na comunidade científica.

A Tabela 5 descreve e referencia os pacotes de funções que foram utilizados em pelo menos uma das etapas metodológicas presentes neste trabalho:

Tabela 5 - Pacotes de Função do R Utilizados na Metodologia

Nome	**Referência**
plotrix	Lemon, J. (2006)
plsdepot	Gaston Sanchez (2012)
pls	Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2019)
mixOmics	Kim-Anh Le, Florian Rhart, Ignacio Gonzalez, Sebastien Dejean (2017)
prospectr	Antoine Stevens and Leornardo Ramirez-Lopez (2013)
reshape2	Hadley Wickham (2007)
ggplot2	H. Wickham (2016)
dplyr	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019)
ggthemes	Jeffrey B. Arnold (2019)
ggExtra	Dean Attali and Christopher Baker (2018)
readxl	Hadley Wickham and Jennifer Bryan (2019)
magrittr	Stefan Milton Bache and Hadley Wickham (2014)
tibble	Kirill Müller and Hadley Wickham (2019)
hydroGOF	Mauricio Zambrano-Bigiarini (2017)
Rlof	Yingsong Hu, Wayne Murray, Yin Shan and Australia (2015)
gghighlight	Hiroaki Yutani (2018)

Figura 7 – Fluxograma da Metodologia de Otimização por Fator RC



6. RESULTADOS E DISCUSSÃO

6.1 PRÉ-PROCESSAMENTO

O pré-processamento dos dados identificou que, dentre as 129 variáveis preditoras disponíveis para o desenvolvimento do modelo, havia três pares de variáveis do tipo LC (nível), cujos valores medidos eram idênticos em cada par, ou seja, diferentes equipamentos foram utilizados para monitorar a mesma variável no mesmo intervalo de tempo. Verificou-se que tais variáveis são críticas ao processo, e é comum que o monitoramento dessas seja feito por mais de um equipamento, de modo a diminuir a chance de indisponibilidade de medição. Entretanto, a inclusão de variáveis preditoras idênticas em modelos inferenciais não aumenta a capacidade de predição do modelo, e, de forma a construir modelos mais simples, a inclusão de ambas não é recomendada (Ferreira *et al.*, 1999). Desta forma, preservou-se apenas uma das variáveis de cada par para o desenvolvimento do modelo. Também foram identificados dois trios de variáveis do tipo PI (indicador de pressão), cujos valores medidos eram idênticos entre si ou possuíam uma diferença dentro do incremento digital. De forma análoga, manteve-se apenas umas das variáveis em cada trio. Assim sendo, o pré-processamento dos dados eliminou 8 das 129 variáveis disponíveis para construção do modelo.

6.2 ANÁLISE EXPLORATÓRIA INICIAL

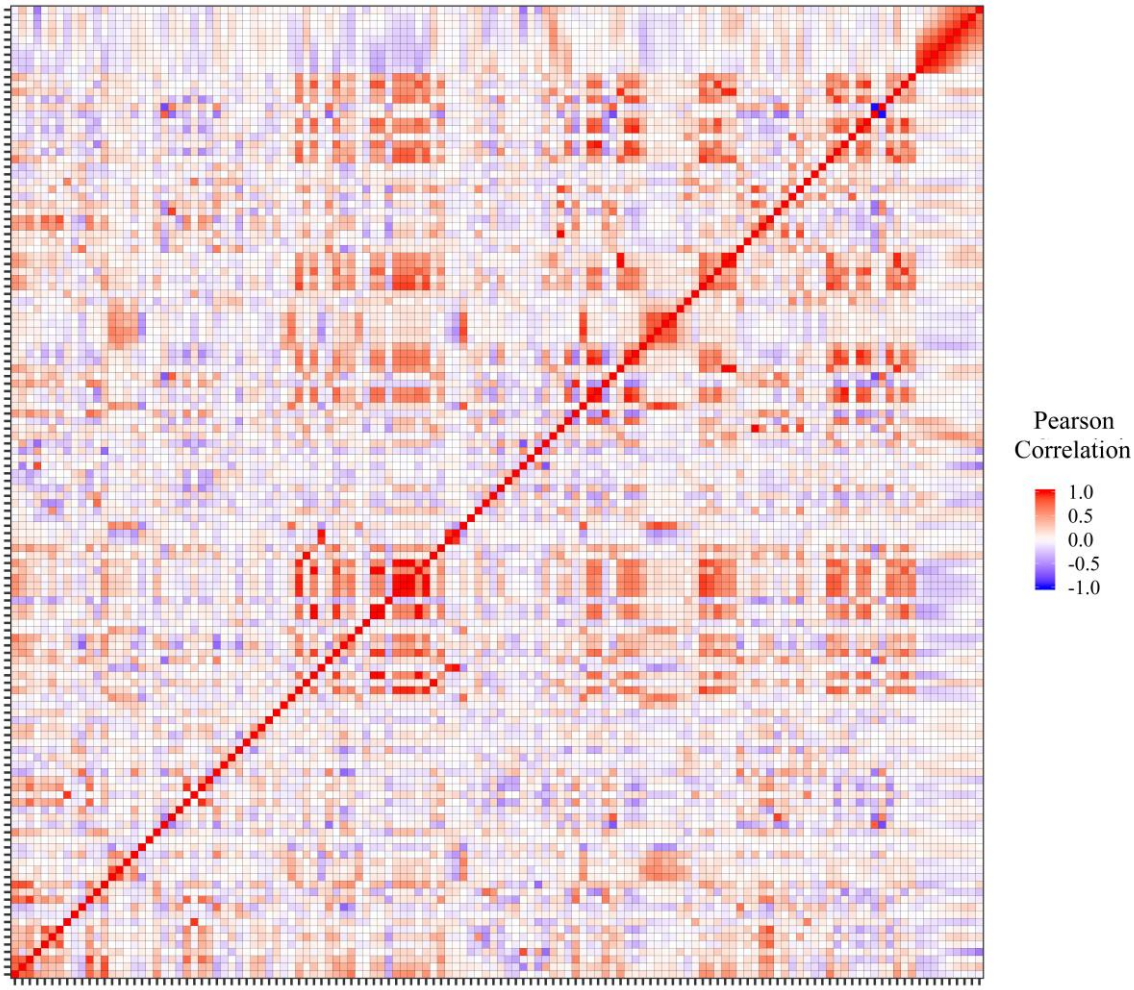
6.2.1 Correlação Entre Variáveis de Processo

As correlações entre as 121 variáveis de processo e as 9 variáveis resposta de temperatura estão representadas na Figura 8. Devido ao grande número variáveis, optou-se por representar as correlações por meio de pixels presentes numa imagem. Na Figura 8, pixels mais escuros representam valores absolutos de correlação mais altos (próximo a unidade), enquanto pixels mais claros representam valores absolutos menores (próximo a zero). Pixels de coloração branca indicam que não há correlação linear entre as variáveis. Correlações positivas possuem coloração vermelha, enquanto que correlações negativas são representadas pela cor azul.

Na Figura 8 é possível perceber que não há evidência substancial da existência de correlação linear entre as variáveis resposta e as variáveis de processo, devido à escassez de pixels mais escuros à direita da figura e no topo desta. De fato, apenas três variáveis de processo possuem correlação linear moderada (com valor entre 0,5 e 0,7) com pelo menos uma das variáveis resposta. Entre estas VPs, uma mede temperatura (V073), uma afere o valor de

pressão (V069) e uma quantifica um valor de fluxo (V004). Como detalhado nas seções seguintes, estas são as variáveis que possuem maior influência (escores VIP elevados) nos modelos PLS.

Figura 8 – Matrix (pixels) de Correlação



Apesar de não haver fortes evidências de correlação linear com as variáveis preditoras, a região na extrema direita superior da Figura 8 indica que há correlação linear entre as nove variáveis resposta. De fato, pontos de temperatura com porcentagem vaporizada próximos, tais como P-005 e P-010, cuja correlação é de 0,978, são mais fortemente correlacionados do que pontos mais distantes, como P-005 e P-100, cuja correlação é de 0,115. A Tabela exibe tal tendência, que é fisicamente esperada, já que há maior semelhança de composição, e, portanto, de comportamento, entre amostras com porcentagem de vaporizado similar. À exceção de três ocasiões, indicadas com um asterisco, os p-valores das correlações na Tabela são menores que 0,05. Assim, à 95% de confiança, as demais correlações são estatisticamente significantes.

Tabela 2 – Correlação Entre Variáveis Reposta

	P-000	**P-005**	**P-010**	**P-030**	**P-050**	**P-070**	**P-090**	**P-095**	**P-100**
P-000	1,000	0,893	0,851	0,698	0,466	0,268	0,057*	0,001*	-0,082*
P-005		1,000	0,978	0,898	0,717	0,531	0,287	0,170	0,115
P-010			1,000	0,945	0,801	0,635	0,392	0,257	0,197
P-030				1,000	0,943	0,832	0,610	0,448	0,389
P-050					1,000	0,963	0,798	0,626	0,564
P-070						1,000	0,909	0,757	0,668
P-090							1,000	0,923	0,722
P-095								1,000	0,648
P-100									1,000

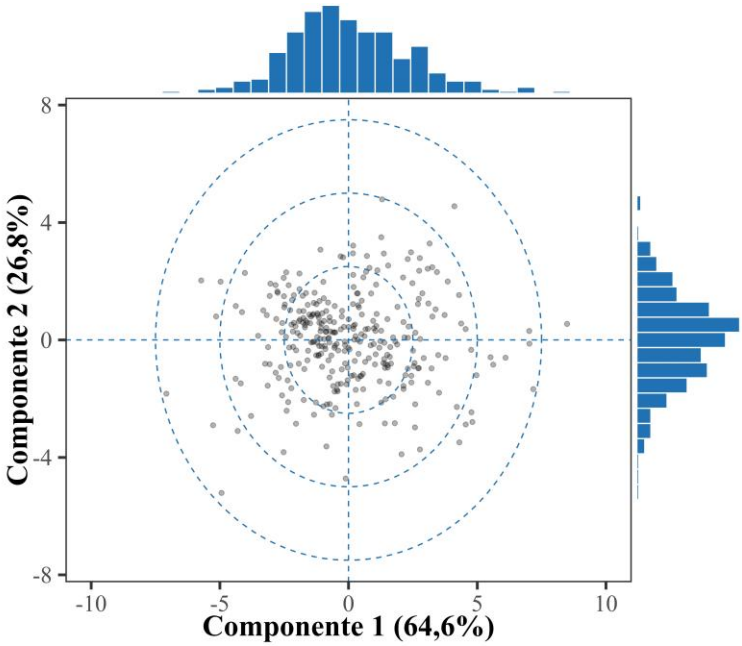
* *p*-valor acima de 0,05

Os valores numéricos da Tabela embasam a decisão de utilizar um modelo PLS único para representar as nove variáveis de saída. Segundo Wold, Sjostrom e Eriksson (2001), quando há correlação entre as variáveis da matriz $\boldsymbol{Y}$, estas devem ser analisadas conjuntamente, e não de forma separada. O autor alerta, porém, para a possibilidade da presença de grupos bem definidos na matriz $\boldsymbol{Y}$. Nesta situação, é necessário o desenvolvimento de modelos PLS distintos para cada grupo de forma a manter a capacidade preditiva de novas observações.

De forma a avaliar a possibilidade de agrupamento, os escores das duas primeiras componentes de uma PCA feita com a matriz $\boldsymbol{X}$ são apresentados na Figura 9. Graficamente, não há evidência que justifique a presença de grupos bem definidos. Desta forma, o desenvolvimento de um único modelo PLS é adequada.

A Figura 8 também indica a presença de correlações moderadas entre as variáveis do tipo FC (fluxo) e do tipo LC (nível) com as demais VPs, porém a grande maioria dos pixels posicionados na esquerda da figura é de coloração mais clara, havendo predominância da cor branca. Desta forma, as variáveis do tipo FC e do tipo LC tendem a não serem correlacionadas com as demais. Por outro lado, há maior presença de pixels mais escuros nas faixas das variáveis que medem pressão e temperatura, localizadas mais ao centro e à direita da Figura 8. Tal fato é esperado, já que há uma relação física entre pressão e temperatura, particularmente se as medições são feitas em pontos onde há pequena distância espacial. São esses pixels de coloração mais escura que explicam a necessidade de construir um modelo que leve em consideração a colinearidade das variáveis preditoras, tal como o PLS.

Figura 9 – Escores da duas primeiras PCs – Matriz ***Y***

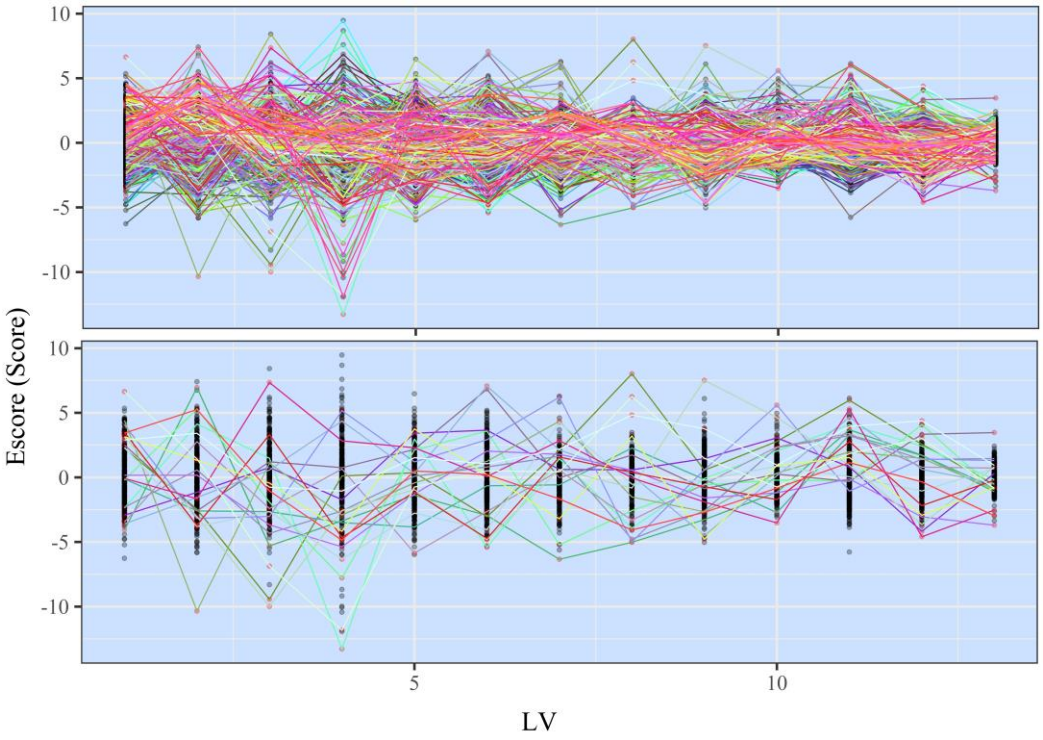


6.2.2 Identificação de *Outliers*

A Figura 10 ilustra os escores de cada observação em cada umas das 13 variáveis latentes num modelo PLS desenvolvido com as 121 variáveis de processo e calibrado com todas as observações. Na parte superior da Figura 10, o escore de cada observação numa LV é conectado ao escore na LV seguinte. Segundo a técnica do diagnóstico do modelo, proposta por Li *et al.* (2016) e descrita na seção 2.2.1, as observações que possuírem escores significativamente diferente das demais serão consideradas anômalas. Na parte inferior da Figura 10, a linha conectando os escores das LVs é mantida somente nas observações que foram identificadas como *outliers*.

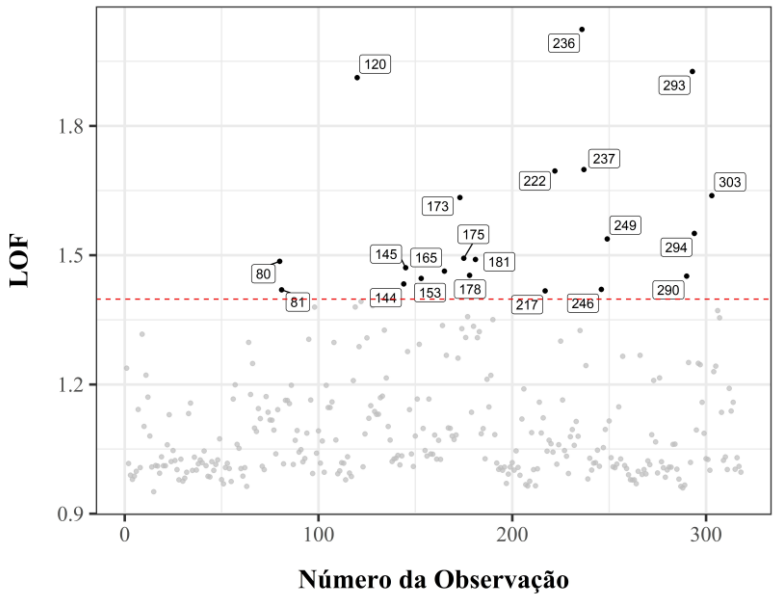
Na Figura 10 é possível notar que algumas observações possuem escore visivelmente diferente da distribuição de escores numa dada LV. Como exemplo, há um ponto na LV de número 2 cujo valor do escore é claramente menor do que o observado nos demais pontos. Entretanto, é importante notar que o critério de identificação de uma observação anômala engloba os escores de todas as LVs, e não somente de uma LV específica. De fato, é possível observar na parte inferior da Figura 10 que as variáveis latentes de número 1, 3,4, 5 e 11 possuem pontos com escores visivelmente diferentes da demais observações em cada LV, mas que não foram identificados como *outliers*, já que a distinção é feita através do Fator Local de Outlier.

Figura 10 - Identificação de Outliers por Diagnóstico do modelo



A Figura 11 ilustra o Fator Local de Outlier, LOF, de cada observação e destaca aquelas cujo valor numérico ficou acima da média dos valores de LOF mais três vezes o desvio padrão, desconsiderando os valores de LOF dos *outliers*. Após 7 iterações, o valor limite do LOF foi determinado como 1,398. Das 319 observações, 21 obtiveram LOF acima de 1,398 e, portanto, foram identificadas como anômalas, o que representa cerca de 6,6% do total de dados.

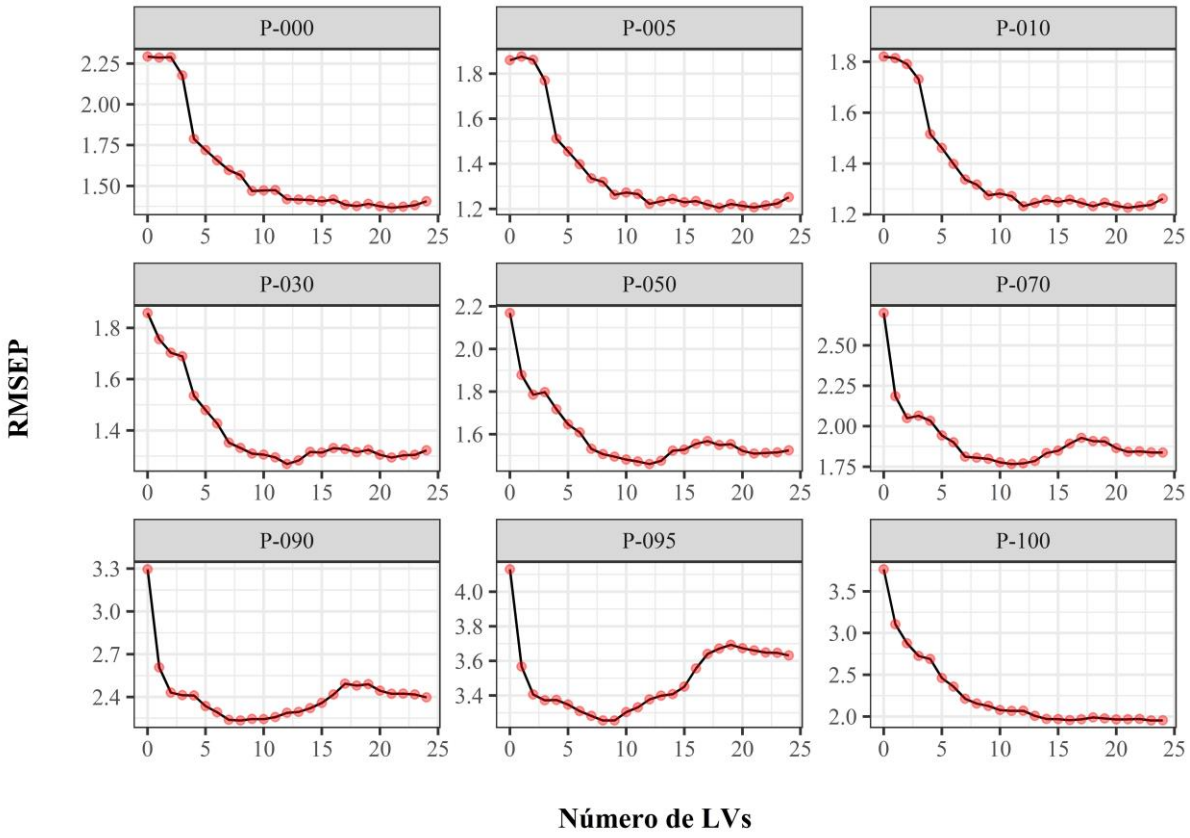
Figura 11 - Fator Local de Outlier, LOF



6.3 MODELO PLS SEM SELEÇÃO DE VARIÁVEIS

A Figura 12 detalha o Erro Quadrático Médio de Predição (RMSEP) para cada variável de saída em função do número de variáveis latentes (LVs) utilizada no modelo.

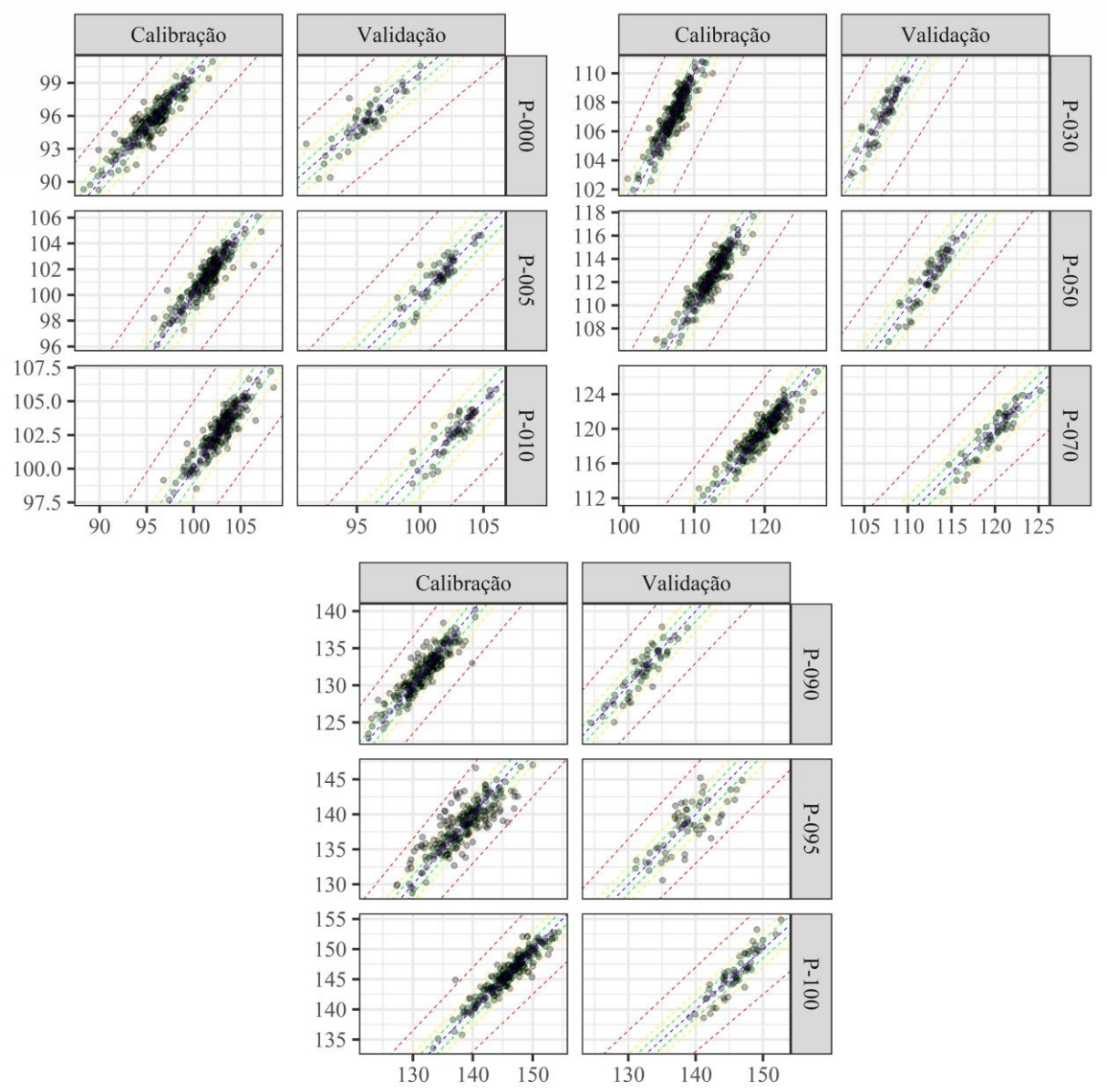
Figura 12 - RMSEP em função do Número de LVs



Na Figura 12 destaca-se que o número de LVs que minimiza o RMSEP não é único entre as variáveis. De fato, algumas destas apresentam aumento expressivo do RMSEP em função do aumento do número LVs, como é o caso da P-095. Este tipo de comportamento é relatado por Clark (2013), que argumenta que em modelos onde há aumento de complexidade (neste caso, o número de LVs) existe a possibilidade de um erro maior em função do aumento da variância introduzida pelas novas variáveis, apesar de uma possível redução no *bias*. O critério de seleção do número ótimo de LVs, descrito na seção 2.4.2, levou a um número de 24 componentes. Entretanto, segundo a Figura 12, este número de LVs eleva o RMSEP, quando comparado com um número menor, de cinco das nove variáveis resposta. Este aumento é graficamente mais expressivo nas variáveis P-090 e P-095. Desta forma, este número de componentes pode comprometer a qualidade da calibração destas variáveis, provocando um

aumento significativo no desvio destas. Entretanto, como visto na Figura 13, que representa a relação entre os valores observados de temperatura e os preditos pelo modelo, este número de componentes não gerou um modelo com capacidade preditiva pobre e, como o modelo PLS completo tem como função, segundo a metodologia deste trabalho, apenas hierarquizar as variáveis de entrada, seu comportamento inicial é apenas um indicativo do modelo final, cujo número de componentes ainda será decidido.

Figura 13 - Predito vs. Observado - Modelo sem seleção de variáveis



A Figura 13 indica que o modelo gerado com todas as variáveis já possui uma capacidade preditiva adequada. As linhas de suporte indicam o erro percentual, sendo a de cor

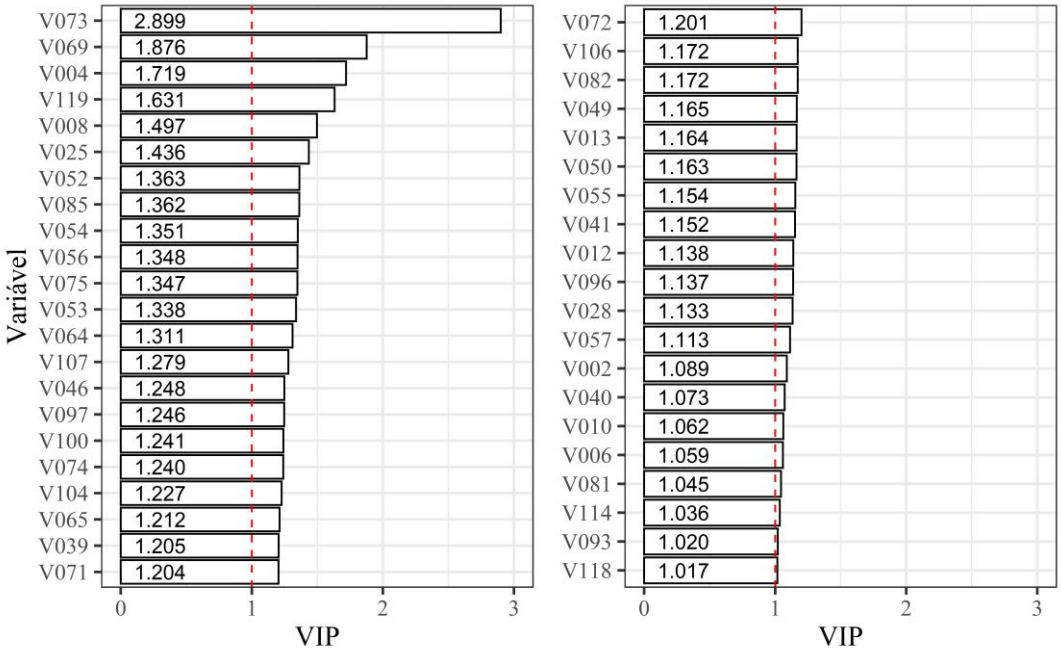
vermelha 5% de erro, a amarela 2%, a verde 1% e a azul 0%. É possível perceber que, com exceção da P-095, a grande maioria dos pontos está localizado ou entre as retas de suporte amarela ou entre as retas de suporte verde, além de estarem bem distribuídos ao redor da reta de suporte azul. Uma outra evidência da boa capacidade preditiva dos modelos está nos coeficientes de correlação linear entre os valores observados e os valores preditos do grupo de validação, os quais variam entre 0,877 (P-000) e 0,935 (P-070), à exceção da P-095, que possui r de 0,764, bastante inferior aos demais. É possível que o número alto de componentes utilizados no modelo PLS tenha afetado a capacidade preditiva da P-095. Entretanto, como dito anteriormente, esta primeira etapa tem como objetivo hierarquizar as variáveis que serão selecionadas no modelo final e, desde que a capacidade preditiva do modelo sem seleção de variáveis não indique completa inviabilidade de predição de uma das variáveis respostas, a inferioridade da capacidade preditiva de P-095 não compromete a metodologia deste trabalho.

As cargas fatoriais das variáveis de entrada estão representadas, para cada uma das 24 variáveis latentes, na Figura 22 do Apêndice A. É importante salientar que as variáveis com maior carga nas primeiras componentes são as que explicam o maior comportamento da massa de dados, enquanto que as variáveis com maior carga fatorial nas últimas componentes tendem a explicar mais o ruído presente nos dados do que sua variabilidade (Abdi, 2010).

6.4 HIERARQUIZAÇÃO DE VARIÁVEIS

A Tabela 5 no Apêndice B detalha os escores VIP das 121 variáveis, enquanto que a Figura 14 apresenta as variáveis com VIP maior que a unidade. É importante notar que, de acordo com Figura 14, as cinco variáveis com maior escore VIP representam medições de vazão (V004 e V008), de pressão (V073 e V069) e de temperatura (V119), variáveis que costumam ser críticas à processos de destilação. Naturalmente, pela própria definição do escore VIP dada pela Equação (12), estas variáveis também são as que possuem maior carga fatorial nas primeiras LVs do modelo PLS, como evidenciado na Figura 22 do Apêndice A. Nota-se também na Figura 14 que a primeira variável do tipo LC (nível) com escore VIP maior que a unidade é a V046, com VIP de 1,248, apenas a décima-quinta em ordem de importância. Ou seja, a hierarquização das variáveis por VIP é representativa do que se espera fisicamente do processo estudado, onde variáveis de vazão, pressão e temperatura são consideradas mais importantes. Na Figura 15 estão detalhadas as variáveis com VIP menor que a unidade.

Figura 14 - Variáveis com VIP maior que 1

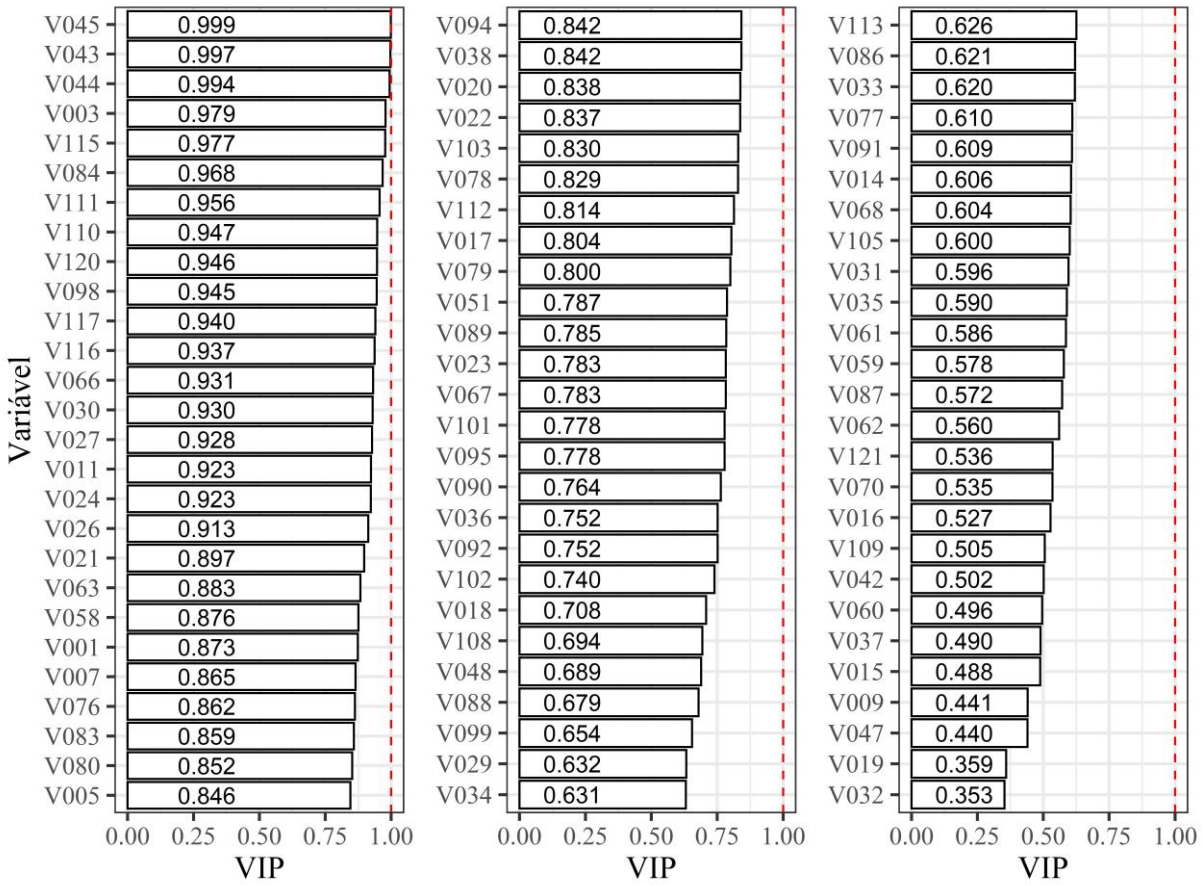


Analogamente às variáveis com escore VIP mais elevado, nota-se, a partir da Figura 15 e da Figura 22, que as variáveis com menor VIP possuem menor carga fatorial nas primeiras LVs do modelo completo. Salienta-se também que algumas variáveis com posições numéricas próximas (que estão fisicamente em regiões adjuntas na planta), tais como V043, V044 e V045, possuem escores VIP semelhantes. Isto significa que tais variáveis carregam informações com nível similar de importância sobre a variável resposta. Este fato ocorre com mais frequência na Figura 15 do que Figura 14, onde somente as variáveis V054 e V056 possuem escores VIP similares. Desta forma, há evidência de que quando há informação repetida em mais de uma variável, a tendência é que estas não sejam consideradas importantes pela técnica do VIP, ou seja, possuem escores menores que a unidade.

Na Figura 15 estão representadas 79 das 121 variáveis utilizadas na construção do modelo completo. Segundo o "critério de menor que a unidade" descrito por Wang et al. (2015), estas variáveis não devem ser selecionadas. Desta forma, o modelo proposto pelo método de seleção do VIP selecionaria 42 variáveis de entrada para o modelo final, descritas na Figura 14. Este método de seleção não considera o *tradeoff* entre a capacidade de predição do modelo e os custos a ele associados. Dessa forma, é possível que, entre as 42 variáveis com escore VIP acima da unidade, existam variáveis cuja importância estatística para o modelo não seja

suficiente para explicar o alto custo associado ao equipamento que faz a medição daquela variável. Dessa forma, o VIP não é ideal para a seleção de variáveis em si. Entretanto, segundo as evidências presentes na Figura 14 e na Figura 15, e discutidas nesta seção, o escore VIP é adequado como índice para hierarquização de variáveis.

Figura 15 - Variáveis com VIP menor que 1



6.5 CÁLCULO DE FATOR LRC/SLRC E SELEÇÃO DE VARIÁVEIS

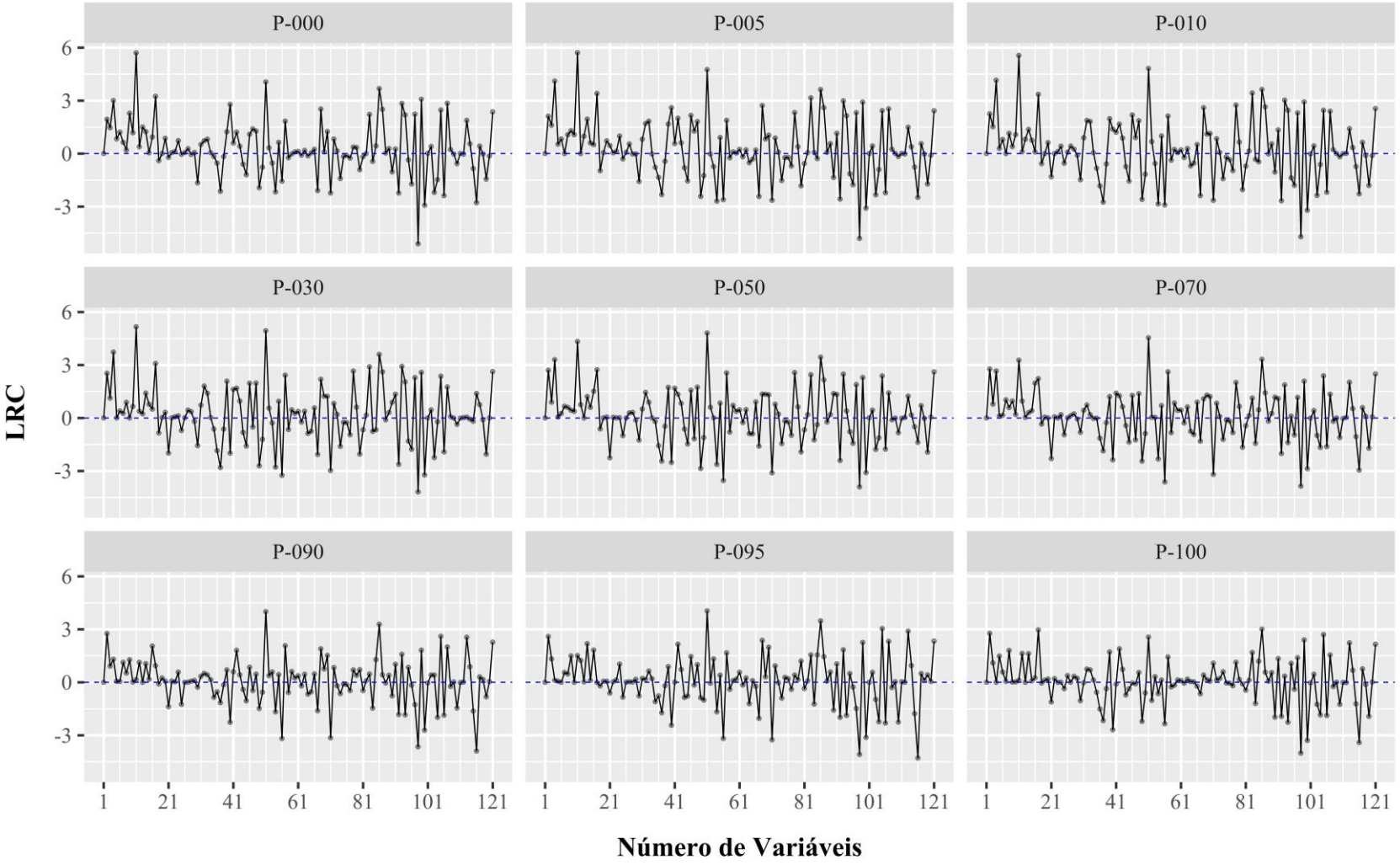
A Figura 16 detalha o valor numérico do fator LRC para cada uma das nove variáveis de saída em função do número de VPs utilizadas no modelo PLS. É importante notar que nos casos onde $LRC_n < 0$ para uma determinada variável de saída, o modelo com n variáveis preditoras possui capacidade menor de estimar a saída quando comparado ao modelo que possui $n - 1$ entradas, enquanto que em casos onde $LRC_n > 0$, a capacidade preditiva do modelo com n entradas é mais alta quando comparada ao modelo com $n - 1$ variáveis preditoras. Observa-se na Figura 16 que os pontos mais a direita dos gráficos estão acima da reta de suporte azul

(onde não há ganho e nem perda na capacidade preditiva, $LRC_n = 0$), o que indica que a adição de variáveis de entrada causa melhora visível na performance dos modelos PLS quando estes possuem um número reduzido de variáveis de entrada, o que é esperado. Observa-se também que, de forma geral, a adição das variáveis de entrada impacta a capacidade preditiva das variáveis de saída em magnitudes diferentes (os valores de LRC_n para diferentes saídas não são uniformes), apesar do efeito negativo ou positivo tender a ser o mesmo.

Ainda assim, observa-se casos em que a adição de certa variável de entrada produz efeitos opostos nas diferentes saídas. Como exemplo, a adição da vigésima primeira variável de maior VIP (V039, segundo a Figura 14), causa uma piora na estimativa de todas a variáveis de saída, com exceção da P-005, onde há uma pequena melhora (RC de 0,581 e LRC de 0,457). Um comportamento análogo é verificado na adição da septuagésima primeira variável de maior VIP (V038), que induz um LRC negativo em todas as variáveis de saída, à exceção da P-100, com um LRC de 1,084. De fato, esse comportamento desigual do efeito da adição de uma entrada nas diferentes saídas é verificado em 42 das 120 adições de VPs, o que representa um percentual de 35%. Destes, em 17 situações o efeito foi diferente em apenas uma das nove variáveis de saída, e entre estes 17, em nove situações o LRC possui magnitude inferior a 0,1, o que indica que a queda ou melhora de performance na saída que se comportou de maneira desigual às demais foi ínfima. Desta forma, a tendência geral é que a adição de uma certa variável de entrada cause efeitos similares (de piora ou de melhora) nas nove variáveis de saída, apesar da diferença em magnitude. Tal comportamento pode ser verificado não somente na Figura 16, como também na Tabela 8 do Apêndice C.

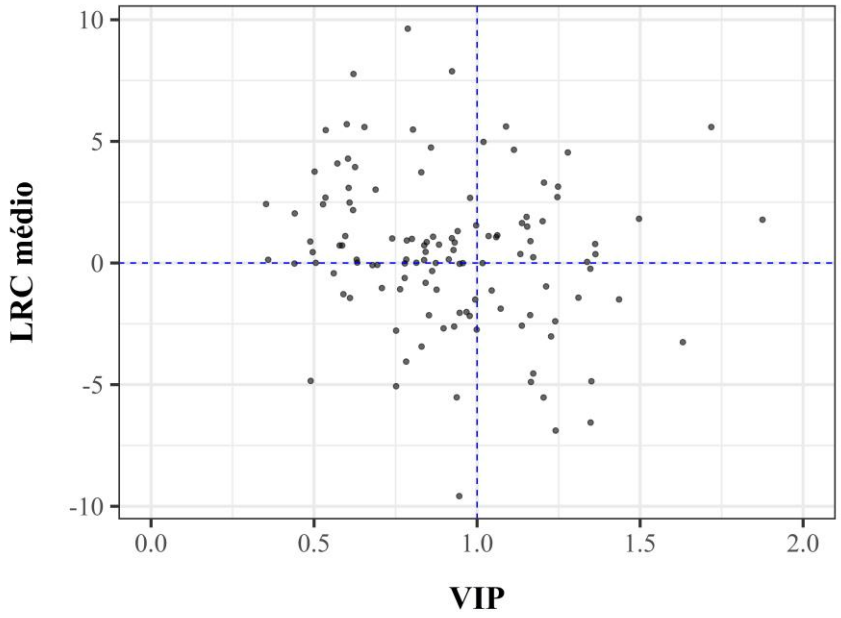
A variável de entrada que causou maior pico de performance em razão do aumento de custo numa das variáveis de saída foi a V075 (posição 11 na hierarquia do VIP), com LRC de 5,715 para a saída P-010, e valores positivos paras as demais, apesar de uma magnitude próxima a zero para P-100. Já a VP que apresentou LRC mais negativo foi a V033 (posição 98 na hierarquia VIP), com valor de -5,105 na P-000 e valores negativos nas demais saídas. Através da Tabela 7 no Apêndice B nota-se que a V075 (maior pico de LRC) possui VIP maior que a unidade, enquanto que a V033 (maior vale de LRC) possui VIP menor que um. De fato, espera-se que variáveis com maior VIP apresentarem LRCs positivos, enquanto que variáveis com menor VIP apresentarem LRCs negativos, já que, quanto maior o VIP, maior a importância da variável no modelo, o que indica que sua presença deve melhorar a performance deste.

Figura 16 - Fator LRC



Dessa forma, espera-se que a maioria dos pontos da Figura 17, onde o LRC médio de cada variável de entrada é detalhado em função do respectivo escore VIP, estejam posicionados nos quadrantes 1 e 3. Entretanto, esta não é a tendência observada.

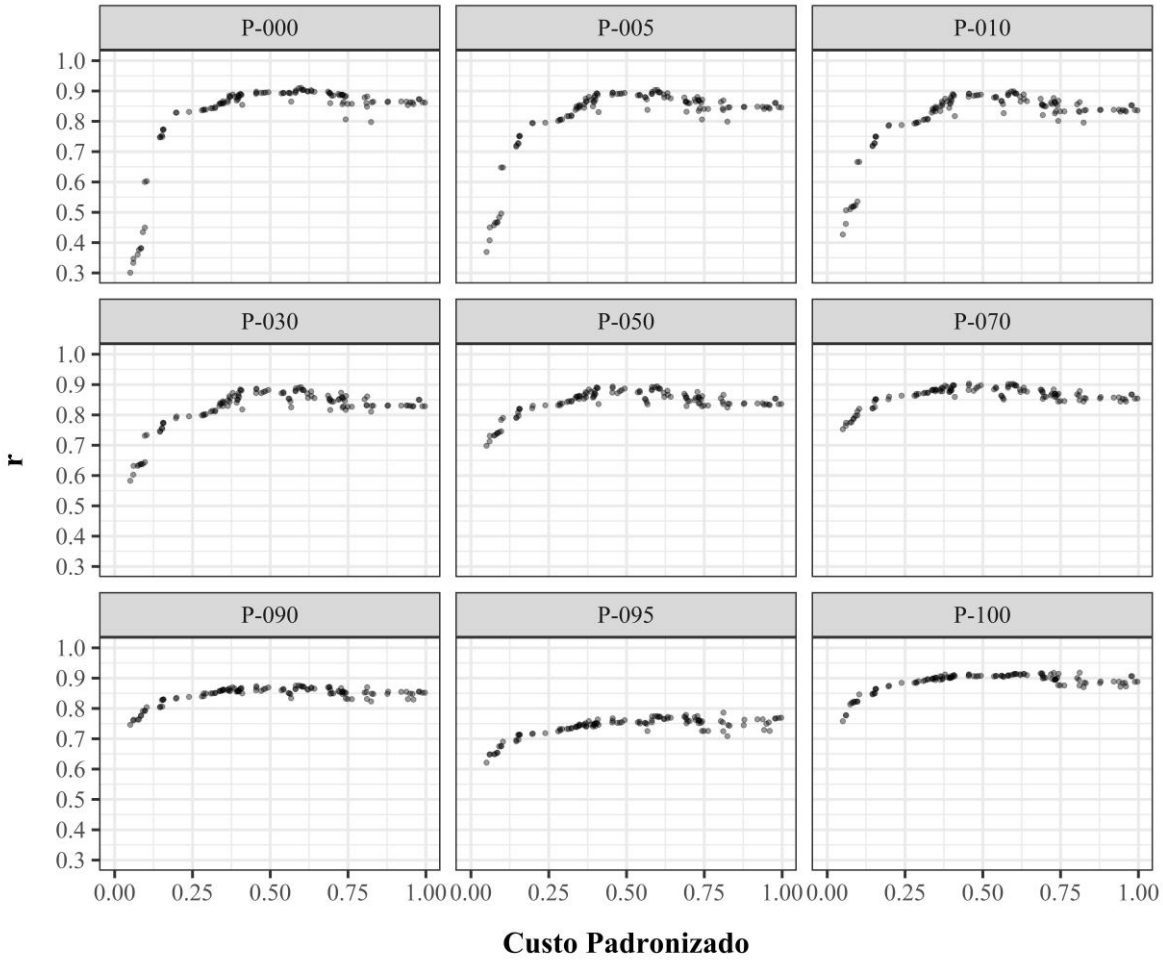
Figura 17 - LRC médio vs. VIP



Há, de fato, alta densidade de pontos no segundo quadrante da Figura 17, onde variáveis com baixo VIP possuem valores de LRC positivos. Apesar de parecer contra intuitivo, este fenômeno ocorre devido à natureza do fator RC, que é recursivo, ou seja, o valor do fator RC de uma variável de entrada n depende da performance do modelo construído com $n-1$ variáveis. Desta forma, caso uma variável cause uma piora visível na performance do modelo quando adiciona à entrada deste (como é o caso da V033), a variável seguinte na hierarquia tende a recuperar a capacidade preditiva do modelo, o que gera valores positivos de RC e, consequentemente, de LRC. Como exemplo, a V077 possui VIP de 0,610 e RCs altamente positivos (de acordo com a Tabela 8, o LRC médio é de 2,49). Isto ocorre devido ao fato de V077 ser adicionada logo após a V033, que apresenta os menores valores de LRC entre as 121 variáveis. Hierarquicamente consecutiva a V033 e com VIP de 0,609, a V091 possui valores de LRC negativos. Esta característica oscilatória nos LRCs pode ser graficamente observada na Figura 16, particularmente nos pontos mais à direita dos gráficos, quando os valores de escore VIP das variáveis adicionadas tornam-se menores.

Apesar de indicar o *tradeoff* entre performance e custo, o fator LRC não indica se a capacidade preditiva do modelo e seu custo total são adequados. De fato, é necessário analisar os valores numéricos de r e de c_T. A Figura 18 detalha a relação entre esses dois parâmetros para cada uma das nove variáveis de saída. Ressalta-se que, segundo a Equação (15), o fator RC representa a inclinação da linha formada entre dois pontos consecutivos na Figura 18.

Figura 18 - r vs. c_T



Nota-se na Figura 18 que a correlação linear r entre os valores observados e os preditos para as nove saídas (figura de performance do modelo), tende a crescer com um aumento do número de VPs utilizadas nos modelos PLS até atingir um certo platô, a partir de quando os valores de r oscilam ao redor de uma certa constante. Apesar dessa tendência de crescimento, há visíveis quedas de performance em certos pontos, que são análogos ao pontos de vale (LRC negativos) na Figura 16. Como exemplo, os r da P-000, P-005 e P010 caem de 0,88, 0,86 e 0,85 para cerca de 0,80 nos três casos ao se adicionar a variável V033, que possui um c_T de

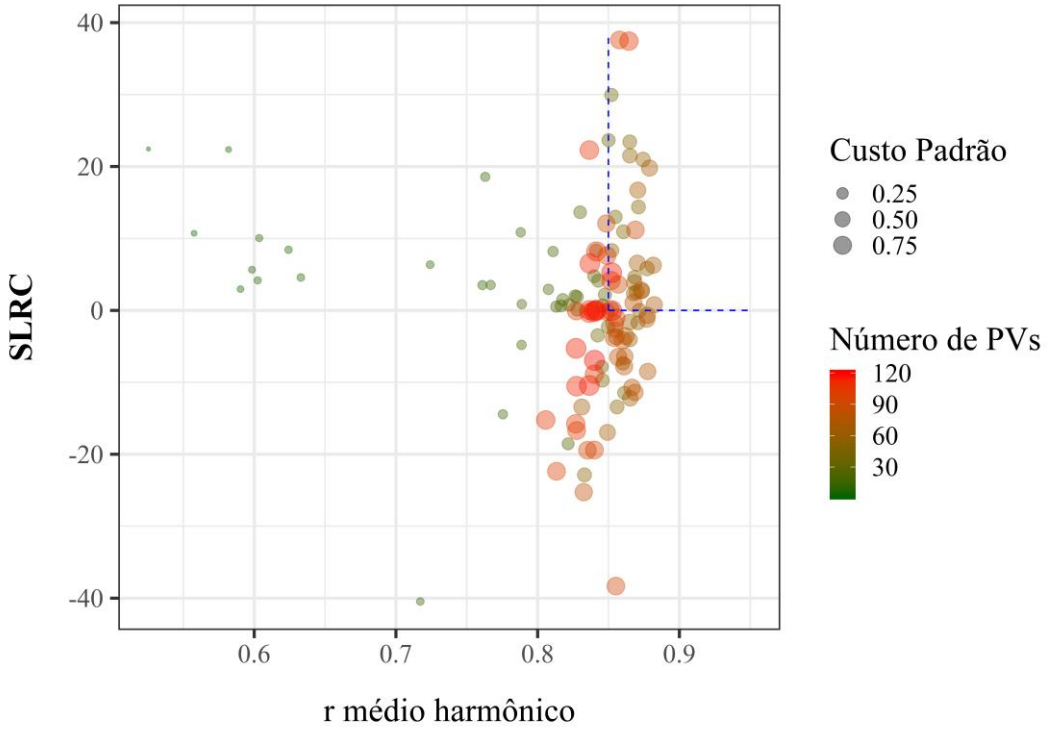
aproximadamente 0,74. A adição da variável em sequência, a V077, resulta num r de 0,88, 0,87 e 0,86 para a P-000, P-005 e P-010, respectivamente. Esse comportamento oscilatório de r reflete nos valores de LRC, sendo as razões que o explicam discutidas em parágrafos anteriores.

É importante salientar que o platô atingido por r nas variáveis de saída está ao redor de 0,90, com exceção da P-095, que possui comportamento anômalo e só consegue atingir um valor próximo de 0,80. Observa-se também que os modelos que estimam as temperaturas com baixo teor de vaporizado (P-000 a P-030) têm suas performances mais afetadas com a retirada de variáveis com alto VIP do que os modelos que estimam temperaturas com percentual de vaporizado maior (P-030 a P-100). Graficamente, esse fenômeno é representado por uma queda mais acentuada no valor de r à esquerda dos gráficos de P-000 a P-030 na Figura 18. Desta forma, há evidências que indicam que é necessário um número maior de variáveis com alto VIP para atingir um valor de r próximo ao platô nas temperaturas com baixo teor de vaporizado do que o necessário nos modelos com alto percentual de vaporizado. Este comportamento alinha-se com o fato das misturas com baixo teor de vaporizado apresentarem maior heterogeneidade em sua composição, e, desta forma, exigirem mais variáveis para modelar seu comportamento, observação semelhante à descrita em Gillon, Hossard e Joffre (1999).

Como observado na Figura 16 e na Figura 18, e discutido nos parágrafos anteriores desta seção, a adição de certas VPs pode causar uma melhora na estimação de certas variáveis de saída ao mesmo tempo em que prejudica o desempenho do modelo na estimação de outras. De forma a analisar o efeito global da adição de uma certa variável, define-se o fator SLRC como a soma dos fatores LRCs das nove variáveis de saída e $\overline{r_h}$ como a média harmônica entre os valores individuais de r, conceitos descritos nas Equações (17) e (18) da seção 4.5.

A Figura 19 detalha a relação entre o SLRC e o $\overline{r_h}$ para cada um dos modelos PLS construídos com as primeiras n VPs, e seus respectivos custos padrão. Nota-se que a maioria dos pontos de coloração verde (baixo de número de variáveis preditoras) possui SLRC positivo, porém $\overline{r_h}$ baixo. Isto é decorrente da adição de variáveis com alto VIP a modelos com poucos preditores causar uma melhora na performance em função do custo (LRC positivo), porém a capacidade preditiva alcançada ainda permanece baixa. Tal fenômeno também é perceptível, de maneira menos explícita, na Figura 18, particularmente na estimação das temperaturas com baixo teor de vaporizado (P-000 a P-030).

Figura 19 - SLRC vs. r médio harmônico

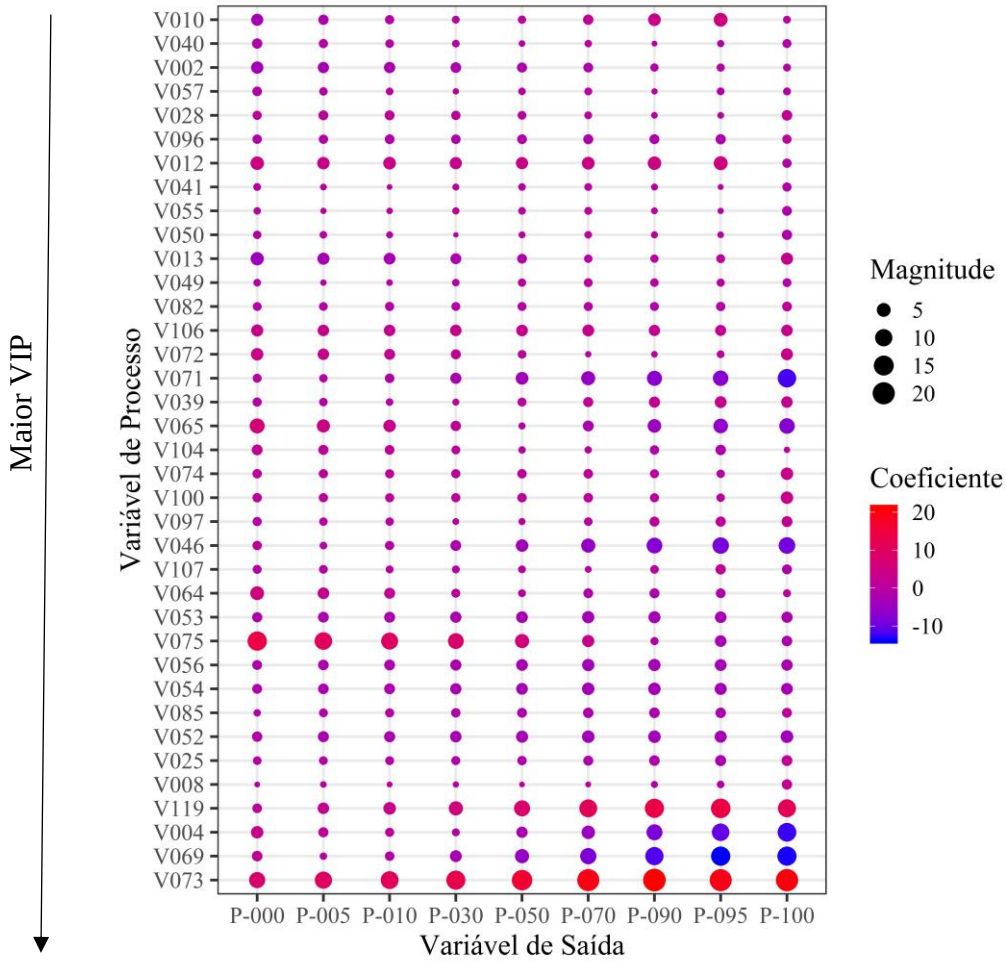


Além disso, observa-se que há uma quantidade expressiva de modelos com um número intermediário/alto de VPs que apresentam alto $\overline{r_h}$ e SLRC negativo. Nestes casos, os valores de r estão oscilando ao redor de um platô e a adição de uma nova VP pode ter efeito negativo na capacidade preditiva, porém essa ainda se mantem relativamente alta (com valores de r entre 0,80 e 0,90), um cenário já discutido em parágrafos anteriores desta seção. Linhas pontilhadas de coloração azul delimitam a zona de aceitação na Figura 19, com $\overline{r_h} > 0{,}85$ e SRLC positivo. Um dos três critérios de decisão descritos na seção 4.7 (menor custo associado, maior $\overline{r_h}$ ou maior SLRC) deve ser aplicado para que as variáveis de processo de um dos modelos PLS dentro da zona de aceitação sejam selecionadas. Como o método de seleção de variáveis proposto por este trabalho possui uma abordagem econômica, deduz-se que o critério mais adequado ao contexto seja o (i), que considera o modelo com menor custo padrão associado. Na Figura 19, esse modelo é representado pelo ponto com menor tamanho e coloração mais para o tom de verde dentro da zona de aceitação. Neste modelo, utiliza-se 37 das 121 variáveis de processo disponíveis como entrada e o custo padrão é de 0,34. A próxima seção detalha as variáveis selecionadas e analisa a performance do modelo escolhido.

6.6 MODELO PLS COM SELEÇÃO DE VARIÁVEIS

A partir das variáveis latentes do modelo PLS escolhido na seção anterior é possível determinar um modelo linear do tipo $\boldsymbol{Y}_{(9\times1)} = \boldsymbol{B}_{(9\times37)}\boldsymbol{X}_{(37\times1)}$, onde $\boldsymbol{Y}$ é a matriz coluna com as nove variáveis de saída e $\boldsymbol{X}$ a matriz coluna com as 37 variáveis de entrada selecionadas. A matriz $\boldsymbol{B}$ contém os respectivos coeficientes lineares de cada uma das 37 VPs (após escalonamento) em cada uma das nove saídas. É possível verificar o valor numérico desses coeficientes na Tabela 9 do Apêndice D. A Figura 20 os representa graficamente.

Figura 20 - Coeficientes Lineares das VPs (escalonadas)

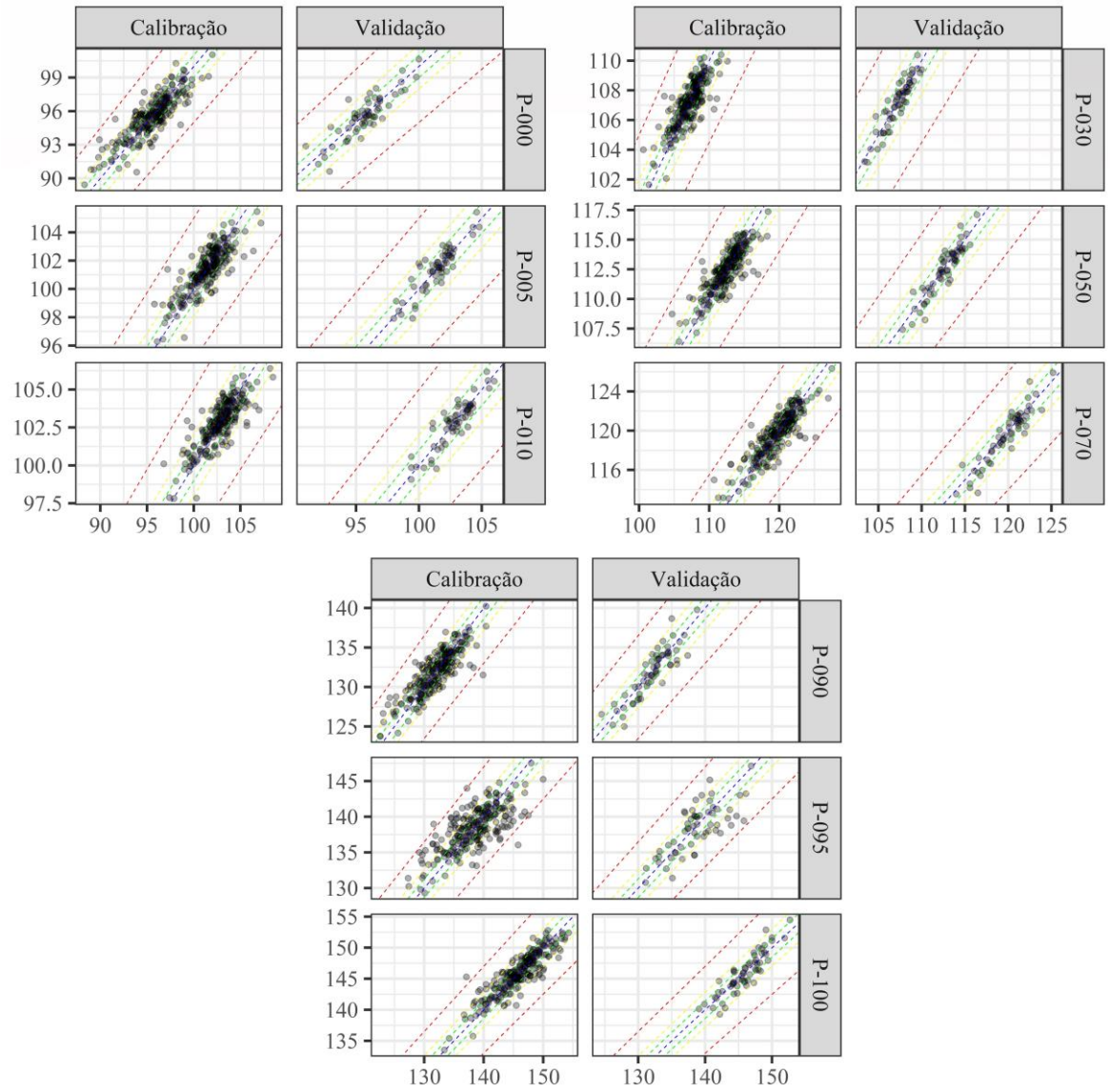


Na Figura 20, é possível perceber que há uma tendência geral de crescimento dos coeficientes das quatro variáveis de processo com maior VIP em função do aumento no teor de vaporizado na variável de saída, já que o tamanho das circunferências nas quatro primeiras linhas tende a crescer da esquerda para a direita. Esse mesmo aumento é verificado em algumas outras variáveis, como a V046, V039 e V071, mas a mesma tendência de crescimento não é

sempre observada nas demais VPs. Este crescimento nas quatro primeiras linhas indica que as VPs de maior VIP carregam um peso maior na explicação do comportamento das saídas de alto teor de vaporizado quando comparadas aos modelos que estimam as saídas com menor teor de vaporizado, um fenômeno discutido na seção anterior e que está relacionado ao fato de saídas com baixo teor de vaporizado apresentarem maior heterogeneidade em sua composição.

A Figura 21 detalha a relação entre os valores observados das noves temperaturas de saída e o predito pelo modelo com seleção de variáveis. As retas de suporte azul, verde, amarela e vermelha representam, respectivamente, 0%, 1%, 2% e 5% de erro.

Figura 21 - Predito vs. Observado - Modelo com seleção de variáveis



Comparada à Figura 13, que representa a relação entre os valores observados de temperatura e os preditos através do modelo PLS sem seleção de variáveis, a Figura 21 indica uma tendência geral de maior espalhamento dos pontos ao redor da reta de suporte azul. Apesar disto, a maior porção destes ainda se encontra entre as retas de suporte amarela (2% de erro), com exceção da P-095, onde há uma proporção elevada de pontos entre as retas de suporte amarela e vermelha (entre 2% e 5% de erro). Este espalhamento maior sugere uma possível perda de capacidade preditiva do modelo, também indicada por uma leve redução nos valores do r dos conjuntos de validação, cujos valores estão descritos na Tabela 6. No entanto, há uma redução nos valores numéricos do RMSEP, um tipo de erro descrito na Tabela 1, cujos valores numéricos também estão descritos na Tabela 6. Como o RMSEP indica a qualidade de predição do conjunto de calibração, uma redução no RMSEP sem o respectivo aumento do r_{val} sugere que a seleção de variáveis causou um leve sobre-ajuste no modelo, um fenômeno indesejado conhecido como *overfitting*. Entretanto, a diferença ínfima entre os valores de r_{val} dos modelos com e sem seleção de variáveis na Tabela 6 e os gráficos dos conjuntos de validação na Figura 21 descartam a possibilidade de sobre-ajuste significativo. Além disso, os valores de r_{val} do modelo com seleção de variáveis indicam que não houve comprometimento da capacidade deste em estimar pontos alheios a sua calibração. Desta forma, a abordagem sugerida por este trabalho selecionou variáveis que atuam como preditores adequados à estimação dos pontos de temperatura necessários à análise da qualidade da nafta em estudo.

Tabela 6 - RMSEP e r: modelos com e sem seleção de variáveis

		P-000	**P-005**	**P-010**	**P-030**	**P-050**	**P-070**	**P-090**	**P-095**	**P-100**
121 VPs	**RMSEP** / °C	1,416	1,226	1,244	1,291	1,482	1,788	2,245	3,235	2,148
	$\boldsymbol{r_{val}}$	0,877	0,898	0,903	0,935	0,929	0,935	0,902	0,764	0,909
37 VPs	**RMSEP** / °C	1,327	1,184	1,196	1,235	1,365	1,584	1,986	3,108	1,876
	$\boldsymbol{r_{val}}$	0,875	0,897	0,901	0,924	0,923	0,932	0,893	0,753	0,901

7. CONCLUSÕES E SUGESTÕES

Este trabalho não somente propôs uma nova abordagem para a seleção de variáveis em modelos inferenciais, como também avaliou a qualidade do modelo desenvolvido a partir da abordagem proposta para estimar nove pontos de temperatura (ponto de bolha, 5%, 10%, 30%, 50, 70%, 90%, 95% de vaporizada, e ponto orvalho) de uma nafta média, os quais são utilizados para avaliar a qualidade desta. O modelo foi construído a partir de uma regressão por mínimos quadrados parciais (PLS), e apresentou RMSEPs que variam entre 1,184 °C e 3,108 °C para saídas de temperatura cujos valores pertencem ao intervalo entre 90 °C e 155 °C. Conjuntos de validação utilizados para avaliar a capacidade preditiva do modelo apresentaram coeficientes de correlação linear entre os valores observados de temperatura e os preditos pelo modelo (r_{val}) que variam entre 0,875 e 0,932, com exceção da temperatura com 95% de vaporizado, cujo r_{val} foi de 0,753. Apesar dessa performance menor em uma das saídas, o modelo apresentou capacidade preditiva adequada à utilização num analisador virtual que estime de maneira confiável os nove pontos de temperatura necessários para avaliação da qualidade da nafta.

Além disso, o modelo PLS desenvolvido com as variáveis selecionadas através da nova abordagem proposta inclui 37 das 121 variáveis de processo (VPs) originalmente sugeridas para inclusão no modelo. Desta forma, o custo total associado, que mensura os esforços financeiros para que os sensores que medem as VPs estejam disponíveis e funcionando num intervalo de tempo de cinco anos, representa 34% do custo total associado ao modelo onde não há seleção de variáveis, e 88% do custo total associado ao modelo onde há seleção de variáveis por VIP, uma técnica de seleção comumente aplicada a modelos PLS. Ressalta-se, porém, que a não seleção de uma VP não implica que a mesma não será mais monitorada pelo grupo de operação da planta, já que esta pode ser importante para auxiliar alguma tomada de decisão em relação ao processo. Entretanto, a possível falha do sensor que mede esta variável implicaria na impossibilidade de estimar a saída num modelo que a tem como entrada, diferente de um modelo em que a mesma não é necessária à estimação, o que torna o último mais robusto.

Por fim, ressalta-se a possibilidade de utilização dos fatores RC e LRC na seleção de variáveis durante o desenvolvimento de quaisquer modelos onde haja possibilidade de hierarquização destas. A introdução de um desses fatores acrescenta um traço econômico num processo de seleção que costuma ser quase que puramente estatístico, o que contribui para

procedimentos mais orientados por dados durante o processo de tomada de decisões em ambientes industriais complexos, particularmente os da petroquímica. Ressalta-se, porém, que a aplicação da abordagem sugerida neste trabalho é condicionada à disponibilidade de dados de manutenção dos sensores utilizados no monitoramento das varáveis de processo, o que nem sempre é possível, já que a existência destes dados em si exige esforços que não costumam ser priorizados no ambiente industrial. Além disso, uma das limitações do fator sugerido é que este não considera o possível reparo destes sensores, prática que pode ser comum a depender do custo de substituição do sensor. Desta forma, sugere-se dois complementos à abordagem sugerida neste trabalho para apreciação em pesquisas futuras:

- seleção de variáveis através da aplicação do fator LRC em modelos empíricos desenvolvidos por outros métodos de regressão além do PLS, tais como PCR (Regressão por Componentes Principais), MLR (Regressão Linear Múltipla), ICA (Análise Canônica Independente), SSI (Identificação Subespacial), ANN (Redes Neurais Artificiais), além de outros métodos onde a hierarquização de variáveis seja possível;
- análise do efeito do reparo de sensores na função de custo total associado, e suas respectivas consequências no fator LRC e na seleção de variáveis no modelo final.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDI, H. Partial least squares regression and projection on latent structure regression (PLS Regression). **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, n. 1, p. 97–106, 2010.

ANDERSEN, C. M.; BRO, R. Variable selection in regression-a tutorial. **Journal of Chemometrics**, v. 24, n. 11–12, p. 728–737, 2010.

BAKHTADZE, N. N. Virtual analyzers: Identification approach. **Automation and Remote Control**, v. 65, n. 11, p. 1691–1709, 2004.

BRERETON, R. G. **Applied chemometrics for scientist**s, John Wiley & Sons, 193–195, 2007.

BREUNING, M. M., KRIEGEL, H.P., RAYMOND, T.N., SANDER, Density-Based Local Outliers. **International Conference of Management of Data.** Dallas: October 2017, p. 1–22, 2012.

CLARK, M. **An Introduction to Machine Learning with applications in R**. Center for Social Research – University of Notre Dame. 2013.

DEEN, M.M. **Process Modelling**. Longman Inc.: Nova York, 1986

FACCHIN, S. **Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais**. 2005. Dissertação (Mestrado em Engenharia Química) – Escola de Engenharia – Universidade Federal do Rio Grande do Sul

FERREIRA, M. M., ANTUNES, A. M., MELGO, M. S., & VOLPE, P. L. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, 22(5), 724-731, 1999.

GALVÃO, R. K. H.; ARAUJO, M. C. U.; JOSÉ, G. E.; PONTOES, M. J C.; SILVA, E. C.; SALDANHA, T. C. B. A method for calibration and validation subset partitioning. **Talanta**, v. 67, n. 4, p. 736–740, 2005.

GILLON, D.; HOUSSARD, C.; JOFFRE, R. Using near-infrared reflectance spectroscopy to predict carbon, nitrogen and phosphorus content in heterogeneous plant material. **Oecologia**, v. 118, n. 2, p. 173–182, 1999.

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, n.1, p. 159–171, 1996.

HARROU, F.; NOUNOU, M. N.; NOUNOU, H. N.; MADAKYARU, M. PLS-based EWMA fault detection strategy for process monitoring. **Journal of Loss Prevention in the Process Industries**, v. 36, p. 108-119, 2015

HUG, C.; SIEVERS, M.; OTTERMANNS, R.; HOLLERT, H.; BRACK, W.; KRAUSS, M. Linking mutagenic activity to micropollutant concentrations in wastewater samples by partial least square regression and subsequent identification of variables. **Chemosphere**, v. 138, p. 176-182, 2015

KANO, M.; FUJIWARA, K. Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. **Journal of Chemical Engineering of Japan**, v. 46, n. 1, p. 1–17, 2013.

KANO, M.; OGAWA, M. The state of the art in chemical process control in Japan: Good practice and questionnaire survey. **Journal of Process Control**, v. 20, n. 9, p. 969–982, 2010.

KENNARD, R. W.; STONE, L. A. Technometrics Computer Aided Design of Experiments. **Technometric**, v. 11, n. 1, p. 137–148, 1969.

KISTER, H. Z. **Distillation Operation**. Cap. 18, p. 545–576. McGraw-Hill: Nova York, 1990.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics and Intelligent Laboratory Systems**, v. 28, n. 1, p. 3–21, 1995.

LI, Z. F.; XU, G.J.; WANG, J.J, DU, G.R. CAI, W.S. Outlier Detection for Multivariate Calibration in Near Infrared Spectroscopic Analysis by Model Diagnostics. **Chinese Journal of Analytical Chemistry**, v. 44, n. 2, p. 305–309, 2016.

MAITRA, S.; YAN, J. Principle component analysis and partial least squares: Two-dimension reduction techniques for regression. **Applying multivariate statistical models**, v. 79, p. 79-90, 2008.

MASSA, A. R. C. Analisador Virtual para a determinação do Teor de Contaminantes MAPD em um reator Trickle-bed. Dissertação (Mestrado em Engenharia Industrial) – Escola Politécnica, Universidade Federal da Bahia, 2017.

MORELLATO, S. A. **Modelos de Regressão PLS com Erros Heteroscedásticos**. Dissertação (Mestrado) - Universidade Federal de São Carlos, 2010.

NOGUEIRA, I.; FONTES, C.; SARTORI, I.; PONTES, K.; EMBIRUÇU, M. A model-based approach to quality monitoring of a polymerization process without online measurement of product specifications. **Computers and Industrial Engineering**. v. 106, p. 123–136, 2017.

PARRISH, J.R.; BROSILOW, C.B. Inferential control applications. **Automatica**, v. 21, n.5, p. 527–538, 1985

RAMIREZ-LOPEZ, L. SCHMIDT, K.; BEHRENS, T.; van WESEMAEL, B.; DEMATTÊ, J. A. M.; SCHOLTEN, T. Sampling optimal calibration sets in soil infrared spectroscopy. **Geoderma**, v. 226–227, n. 1, p. 140–150, 2014.

RITTER, A.; MUÑOZ-CARPENA, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. **Journal of Hydrology**, v. 480, p. 33–45, 2013.

RODRIGUES, B. S. Métodos de Construção de Analisadores Virtuais para Estimação de Teor de Enxofre de Hidrocarbonetos. Monografia (Especialização em Automação Industrial) – Escola de Engenharia, Universidade Federal de Minas Gerias, 2014.

ROMIA, M. B.; BERNARDEZ, M. A. Multivariate Calibration for Quantitative Analysis. **Infrared Spectroscopy for Food Quality Analysis and Control**, v. 1, p. 51-82, Elsevier Inc.: 2009

SILVA, N. C.D. Uso da espectroscopia NIR no desenvolvimento de um simulador para gasolina e na transferência de calibração entre instrumentos de bancada e portátil. Tese de doutorado: Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, Recife, Pernambuco, 2017

SODERSTROM, T. On some adaptive controllers for stochastic systems with slow output sampling. In. UNBEHAVEN, H. (ed.) **Methods and Applications in Adaptive Control**. Springer-Verlag: Berlim, 1980

THAM, M. T. MONTAGUE, G. A.; MORRIS, A. J.; LANT, P. A. Soft-sensors for process estimation and inferential control. **Journal of Process Control**, [s. l.], v. 1, n. 1, p. 3–14, 1991.

URHAN, A. INCE, N. G.; BONDY, R.; ALAKENT, B. **Soft-Sensor Design for a Crude Distillation Unit Using Statistical Learning Methods**. Elsevier Masson SAS, v.44, 2018.

VLASCICI, D. et al. Thiocyanate and fluoride electrochemical sensors based on nanostructurated metalloporphyrin systems. **Journal of Optoelectronics and Advanced Materials**, v. 10, n. 9, p. 2303–2306, 2008.

WALACH, J.; FILZMOSER, P.; HRON, K. **Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences**. 1. ed., Elsevier B.V., 2018. v. 82

WANG, Z. X.; HE, Q. P.; WANG, J. Comparison of variable selection methods for PLS-based soft sensor modeling. **Journal of Process Control**, v. 26, n. 2015, p. 56–72, 2015.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109–130, 2001.

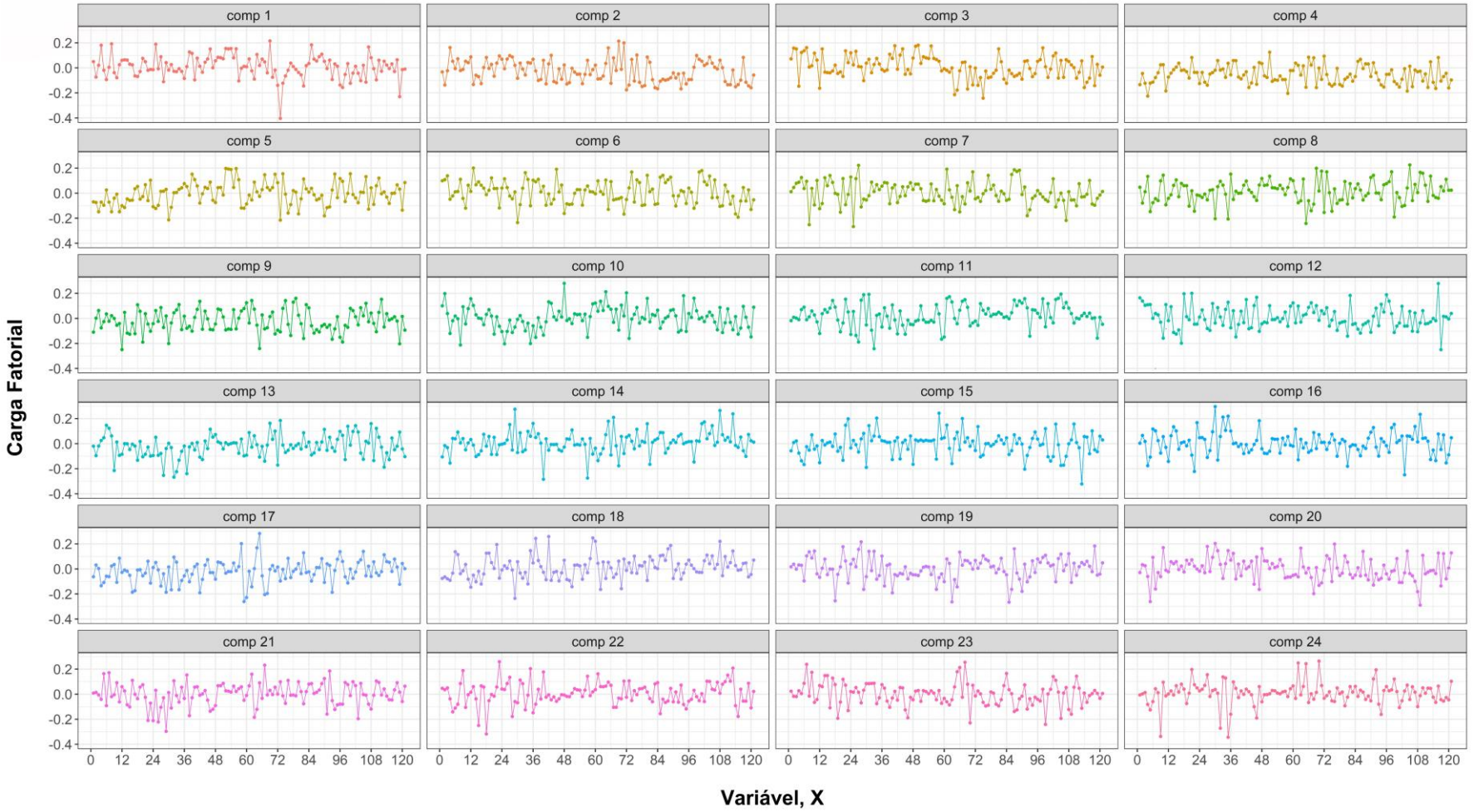
YADYKIN, I. B.; GROBOVOY, A. A.; ISKAKOV, A. B.; KATAEV, D. E.; KHMELIK, M. S. Stability analysis of electric power systems using finite Gramians. **IFAC-PapersOnLine**, v. 48, n. 30, p. 548–553, 2015.

YUN, Y. H.; LI, H. D. DENG, B.C.; CAO, D.S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. **TrAC - Trends in Analytical Chemistry**, v. 113, p. 102–115, 2019.

YUN, Y. H.; BIN, J.; LIU, D.L.; XU, L.; YAN, T.L.; CAO, D.S.; XU, Q.S. A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration. **Analytica Chimica Acta**, v. 1058, p. 58–69, 2019.

APÊNDICE A – CARGAS FATORIAIS EM X

Figura 22 - Cargas Fatoriais em X



APÊNDICE B – ESCORES VIP

Tabela 7 – Escores VIP

V001	0,873	**V032**	0,353	**V063**	0,883	**V094**	0,842
V002	1,089	**V033**	0,620	**V064**	1,311	**V095**	0,778
V003	0,979	**V034**	0,631	**V065**	1,212	**V096**	1,137
V004	1,719	**V035**	0,590	**V066**	0,931	**V097**	1,246
V005	0,846	**V036**	0,752	**V067**	0,783	**V098**	0,945
V006	1,059	**V037**	0,490	**V068**	0,604	**V099**	0,654
V007	0,865	**V038**	0,842	**V069**	1,876	**V100**	1,241
V008	1,497	**V039**	1,205	**V070**	0,535	**V101**	0,778
V009	0,441	**V040**	1,073	**V071**	1,204	**V102**	0,740
V010	1,062	**V041**	1,152	**V072**	1,201	**V103**	0,830
V011	0,923	**V042**	0,502	**V073**	2,899	**V104**	1,227
V012	1,138	**V043**	0,997	**V074**	1,240	**V105**	0,600
V013	1,164	**V044**	0,994	**V075**	1,347	**V106**	1,172
V014	0,606	**V045**	0,999	**V076**	0,862	**V107**	1,279
V015	0,488	**V046**	1,248	**V077**	0,610	**V108**	0,694
V016	0,527	**V047**	0,440	**V078**	0,829	**V109**	0,505
V017	0,804	**V048**	0,689	**V079**	0,800	**V110**	0,947
V018	0,708	**V049**	1,165	**V080**	0,852	**V111**	0,956
V019	0,359	**V050**	1,163	**V081**	1,045	**V112**	0,814
V020	0,838	**V051**	0,787	**V082**	1,172	**V113**	0,626
V021	0,897	**V052**	1,363	**V083**	0,859	**V114**	1,036
V022	0,837	**V053**	1,338	**V084**	0,968	**V115**	0,977
V023	0,783	**V054**	1,351	**V085**	1,362	**V116**	0,937
V024	0,923	**V055**	1,154	**V086**	0,621	**V117**	0,940
V025	1,436	**V056**	1,348	**V087**	0,572	**V118**	1,017
V026	0,913	**V057**	1,113	**V088**	0,679	**V119**	1,631
V027	0,928	**V058**	0,876	**V089**	0,785	**V120**	0,946
V028	1,133	**V059**	0,578	**V090**	0,764	**V121**	0,536
V029	0,632	**V060**	0,496	**V091**	0,609		
V030	0,930	**V061**	0,586	**V092**	0,752		
V031	0,596	**V062**	0,560	**V093**	1,020		

APÊNDICE C – FATORES LRC

Tabela 8 – Fatores LRC

		Fator LRC								
Hierarquia VIP	Variável	P-000	P-005	P-010	P-030	P-050	P-070	P-090	P-095	P-100
1	V073	–	–	–	–	–	–	–	–	–
2	V069	1,943	2,120	2,249	2,532	2,700	2,770	2,762	2,592	2,776
3	V004	1,462	1,599	1,520	1,123	0,892	0,781	0,915	1,325	1,101
4	V119	3,007	4,103	4,153	3,728	3,301	2,659	1,306	0,111	0,000
5	V008	0,833	0,531	0,293	0,004	0,058	0,106	0,058	0,067	1,485
6	V025	1,214	0,831	0,802	0,400	0,251	0,146	0,041	0,000	0,576
7	V052	0,637	0,000	0,004	0,258	0,659	1,026	1,127	0,562	0,051
8	V085	0,199	1,069	1,148	0,885	0,601	0,583	0,521	0,464	1,807
9	V054	2,286	1,296	0,399	0,000	0,437	0,959	1,278	1,493	0,014
10	V056	1,178	1,070	1,073	0,650	0,372	0,196	0,012	0,000	0,020
11	V075	5,706	5,715	5,562	5,160	4,325	3,275	0,143	1,514	0,115
12	V053	0,388	0,001	0,096	0,379	0,750	0,957	1,128	1,237	1,631
13	V064	1,508	0,986	0,814	0,236	0,000	0,028	0,000	0,023	0,002
14	V107	1,247	1,953	1,373	1,415	1,210	0,300	1,051	2,186	1,620
15	V046	0,041	0,605	0,755	0,771	0,599	0,411	0,168	0,101	0,080
16	V097	0,940	0,504	0,087	0,500	1,515	1,975	2,053	1,825	0,251
17	V100	3,241	3,403	3,357	3,085	2,725	2,224	0,945	-0,009	2,967
18	V074	-0,402	-0,970	-0,571	-0,852	-0,625	-0,352	-0,088	-0,208	-0,044
19	V104	0,005	0,053	0,000	0,024	0,005	0,047	0,241	0,055	0,110
20	V065	0,869	0,722	0,630	0,320	0,051	0,001	0,058	0,062	0,188
21	V039	-0,220	0,459	-1,307	-1,994	-2,260	-2,310	-1,385	-0,618	-1,117
22	V071	0,050	0,040	0,002	0,004	0,033	0,065	0,067	0,032	0,205
23	V072	0,110	0,126	0,112	0,067	0,004	0,000	0,027	0,126	0,001
24	V106	0,734	0,994	0,419	0,119	0,012	0,184	0,577	1,029	0,006
25	V082	-0,037	-0,300	-0,548	-0,707	-1,002	-0,948	-1,249	-0,841	-0,360
26	V049	0,056	0,068	0,075	0,010	-0,002	0,001	0,001	0,011	0,382
27	V013	0,280	0,546	0,421	0,443	0,286	0,129	0,003	0,030	0,007
28	V050	-0,046	-0,026	0,253	0,354	0,315	0,228	0,058	0,013	0,334
29	V055	0,064	-0,021	-0,089	-0,192	-0,122	-0,054	0,105	0,164	0,218
30	V041	-1,656	-1,577	-1,478	-1,567	-1,253	-0,812	-0,280	-0,772	-1,044
31	V012	0,520	0,814	0,894	0,724	0,505	0,435	0,341	0,221	-0,023
32	V096	0,724	1,703	1,876	1,806	1,449	0,736	0,502	0,156	0,746
33	V028	0,812	1,836	1,821	1,404	0,886	0,185	0,422	0,635	0,700
34	V057	0,049	-0,018	0,046	0,030	0,004	-0,021	0,144	0,201	0,128
35	V002	-0,163	-0,776	-0,827	-0,627	-0,204	-0,018	-0,854	-1,100	-0,565
36	V040	-0,533	-1,349	-1,830	-1,854	-1,574	-1,149	-0,538	-0,766	-1,517
37	V010	-2,135	-2,323	-2,746	-2,805	-2,443	-1,870	-1,152	-1,728	-2,174
38	V006	-0,169	-0,433	-0,586	-0,636	-0,469	-0,266	-0,144	-0,197	-0,360

Hierarquia VIP	Variável	P-000	P-005	P-010	P-030	P-050	P-070	P-090	P-095	P-100
39	V081	1,235	1,670	1,978	2,081	1,733	1,210	0,703	0,886	1,722
40	V114	2,787	2,602	1,379	-1,992	-2,518	-2,373	-2,269	-2,431	-2,684
41	V093	0,594	0,558	1,210	1,616	1,689	1,407	0,595	0,011	-0,100
42	V118	1,213	2,002	1,674	1,684	1,341	1,255	1,808	2,156	1,897
43	V045	0,417	0,615	0,877	0,955	0,830	0,634	0,422	0,720	0,730
44	V043	-0,602	-0,806	-0,733	-0,818	-0,634	-0,430	-0,404	-0,857	-0,718
45	V044	-1,200	-1,552	-1,559	-1,590	-1,476	-1,372	-1,051	-0,779	-0,364
46	V003	1,099	2,161	2,188	1,972	1,575	1,278	0,855	1,463	-0,037
47	V115	1,403	1,270	0,902	-0,512	-1,179	-1,242	-0,469	-0,157	-0,093
48	V084	1,282	1,817	1,861	1,977	1,742	1,377	0,460	0,997	0,556
49	V111	-1,933	-2,427	-2,590	-2,713	-2,860	-2,449	-1,483	-0,874	-2,217
50	V110	-0,773	-1,249	-1,169	-1,208	-1,113	-0,878	-0,560	-1,007	-0,601
51	V120	4,057	4,757	4,820	4,937	4,806	4,543	4,003	4,042	2,565
52	V098	0,325	-0,029	0,678	0,555	0,604	0,038	0,377	-0,051	-1,030
53	V117	-0,530	-0,730	-0,558	-0,281	0,010	0,030	0,587	1,331	0,321
54	V116	-2,170	-2,695	-2,853	-2,785	-2,629	-2,322	-1,674	-1,677	-0,645
55	V066	0,640	0,908	0,998	0,948	0,848	0,699	0,436	0,399	0,126
56	V030	-1,564	-2,614	-2,910	-3,250	-3,544	-3,625	-3,187	-3,178	-2,347
57	V027	1,839	1,883	2,129	2,420	2,549	2,614	2,071	1,666	1,433
58	V011	-0,234	-0,241	-0,374	-0,653	-0,806	-0,834	-0,575	-0,399	-0,262
59	V024	-0,017	0,089	0,237	0,466	0,715	0,847	0,614	0,094	-0,162
60	V026	0,091	0,032	0,056	0,252	0,371	0,431	0,238	0,177	0,131
61	V021	0,118	0,242	0,210	0,370	0,465	0,452	0,371	0,569	0,093
62	V063	-0,090	-0,182	-0,218	-0,233	-0,265	-0,289	-0,206	-0,154	-0,048
63	V058	0,181	0,210	0,287	0,379	0,468	0,622	0,451	0,245	0,176
64	V001	-0,131	-0,513	-0,695	-0,876	-0,888	-0,761	-0,673	-1,212	0,040
65	V007	0,039	-0,320	-0,540	-0,760	-0,906	-0,941	-0,544	0,096	0,025
66	V076	0,246	0,204	0,518	0,568	0,911	0,894	0,474	-0,224	-0,225
67	V083	-2,094	-2,431	-2,383	-2,067	-1,596	-1,316	-1,611	-2,043	-0,655
68	V080	2,523	2,726	2,599	2,189	1,355	1,082	1,898	2,373	0,409
69	V005	0,098	0,825	1,107	1,233	1,344	1,292	0,752	0,306	0,151
70	V094	1,249	1,042	1,140	1,229	1,317	1,207	1,523	1,985	0,060
71	V038	-2,228	-2,642	-2,649	-2,974	-3,108	-3,197	-3,145	-3,259	1,084
72	V020	0,825	0,886	0,845	0,822	0,783	0,834	0,834	0,932	0,123
73	V022	0,152	0,073	0,061	0,196	0,221	0,079	-0,196	-0,027	0,219
74	V103	-1,416	-1,522	-1,422	-1,607	-1,457	-1,218	-0,642	-0,898	0,596
75	V078	-0,155	-0,230	-0,194	-0,245	-0,163	-0,110	-0,092	0,281	-0,036
76	V112	-0,105	-0,219	-0,338	-0,270	-0,230	-0,174	-0,129	0,205	-0,048
77	V017	-0,252	-0,704	-0,963	-0,953	-0,980	-0,834	-0,472	-0,400	-0,194
78	V079	0,367	2,324	2,751	2,658	2,573	2,009	0,700	0,410	1,133
79	V051	0,347	0,393	0,648	0,609	0,626	0,652	0,389	0,141	0,160
80	V089	-0,909	-1,825	-2,039	-2,050	-1,922	-1,660	0,713	1,206	-0,091

Hierarquia VIP	Variável	P-000	P-005	P-010	P-030	P-050	P-070	P-090	P-095	P-100
81	V023	-0,178	-0,561	-0,706	-0,667	-0,674	-0,439	-0,473	-0,356	-0,452
82	V067	0,007	0,053	0,144	0,158	0,183	0,131	0,092	0,074	0,113
83	V101	2,218	3,158	3,435	2,898	2,448	1,126	0,467	1,554	1,688
84	V095	-0,439	0,055	-0,330	-0,761	-1,246	-1,444	-1,465	-1,230	-1,198
85	V090	0,438	-0,283	-0,463	-0,672	-0,377	0,461	1,282	1,549	1,192
86	V036	3,682	3,626	3,633	3,600	3,443	3,321	3,288	3,477	3,011
87	V092	2,503	2,601	2,643	2,612	2,157	1,421	0,420	1,451	0,567
88	V102	0,033	0,073	-0,021	-0,106	-0,231	-0,182	-0,042	0,063	0,067
89	V018	0,286	0,570	0,551	0,309	0,188	0,253	0,401	0,587	0,552
90	V108	-1,030	-1,359	-1,046	0,885	1,382	1,175	-0,769	-1,581	-1,975
91	V048	0,261	1,150	1,340	1,345	1,337	1,104	1,025	1,045	1,338
92	V088	-2,231	-2,578	-2,668	-2,628	-2,407	-2,013	-1,826	-1,985	-1,929
93	V099	2,843	2,992	3,025	2,921	2,479	1,872	1,584	1,839	0,351
94	V029	2,189	2,146	2,455	2,041	0,399	-1,402	-1,846	-1,874	-2,270
95	V034	-0,359	-1,145	-1,374	-1,318	-0,781	0,102	0,843	0,497	1,073
96	V113	-1,723	-1,765	-1,804	-1,778	-1,439	-0,967	-0,166	-0,267	-0,398
97	V086	2,230	2,320	2,306	2,284	1,890	1,165	-1,262	-1,484	1,394
98	V033	-5,105	-4,800	-4,711	-4,188	-3,902	-3,860	-3,651	-4,090	-4,006
99	V077	3,076	2,920	2,932	2,582	2,291	2,082	1,821	2,256	2,396
100	V091	-2,928	-3,091	-3,211	-3,225	-3,099	-2,878	-2,714	-3,120	-3,289
101	V014	0,014	0,006	-0,013	0,011	0,006	-0,008	-0,037	0,000	-0,033
102	V068	0,404	0,433	0,439	0,453	0,443	0,418	0,412	0,568	0,455
103	V105	-2,202	-2,337	-2,366	-2,245	-1,784	-0,991	0,411	-0,975	-1,248
104	V031	-1,471	-0,902	-0,617	-0,214	-1,129	-1,669	-1,986	-2,236	-1,850
105	V035	2,471	2,437	2,453	2,361	2,380	2,384	2,605	3,040	2,697
106	V061	-2,383	-2,223	-2,189	-1,922	-1,766	-1,621	-1,866	-2,313	-1,888
107	V059	2,851	2,532	2,400	1,754	1,418	1,349	2,004	2,317	1,552
108	V087	0,227	0,244	0,218	0,061	-0,103	-0,187	-0,234	-0,301	-0,240
109	V062	-0,022	0,012	0,005	-0,005	-0,008	0,000	0,019	0,032	0,003
110	V121	-0,563	-0,165	-0,184	-0,364	-0,837	-1,109	-1,460	-2,258	-1,248
111	V070	0,009	0,009	0,002	-0,011	-0,020	-0,027	-0,013	0,031	-0,001
112	V016	-0,028	-0,017	0,016	0,018	0,026	0,029	0,013	0,002	-0,018
113	V109	1,883	1,497	1,417	0,043	1,231	2,019	2,550	2,896	2,233
114	V042	0,554	0,387	0,345	-0,066	0,166	0,529	0,881	0,948	0,684
115	V060	-0,844	-0,759	-0,730	-0,186	-0,498	-1,059	-1,626	-1,776	-1,217
116	V037	-2,782	-2,483	-2,282	1,386	-1,379	-2,956	-3,893	-4,289	-3,409
117	V015	0,425	0,566	0,639	0,754	0,700	0,584	0,306	0,498	0,763
118	V009	-0,020	-0,021	-0,101	-0,103	-0,025	0,099	0,160	0,120	-0,125
119	V047	-1,441	-1,722	-1,805	-2,052	-1,938	-1,715	-0,823	0,420	-1,930
120	V019	-0,129	-0,088	-0,102	0,003	0,047	0,057	0,038	0,044	0,011
121	V032	2,363	2,429	2,555	2,627	2,613	2,496	2,274	2,334	2,158

APÊNDICE D – COEFICIENTES DO MODELO SELECIONADO

Tabela 9 - Coeficientes das VPs (centralizadas) no modelo PLS

	Coeficiente linear, b_{nj}								
VP (X_j)	**P-000**	**P-005**	**P-010**	**P-030**	**P-050**	**P-070**	**P-090**	**P-095**	**P-100**
V073	8,23	9,82	11,00	13,18	16,19	19,75	21,05	19,67	20,21
V069	1,88	-0,29	-1,22	-2,93	-5,48	-8,33	-11,70	-13,75	-13,53
V004	3,46	1,59	0,96	-0,44	-2,46	-4,70	-8,05	-10,44	-12,59
V119	1,26	2,64	3,60	5,45	7,96	11,05	13,53	14,21	11,29
V008	-0,02	-0,12	-0,02	0,06	0,01	-0,03	-0,22	-0,34	1,76
V025	-0,76	-0,81	-0,86	-0,88	-1,16	-1,54	-1,90	-2,24	2,07
V052	-1,74	-2,21	-2,33	-2,68	-3,14	-3,69	-3,60	-3,16	-3,79
V085	-0,30	-0,85	-0,96	-1,13	-1,41	-1,79	-1,92	-1,85	1,44
V054	-1,42	-1,94	-2,10	-2,43	-2,89	-3,46	-3,50	-3,22	-2,84
V056	-1,41	-1,92	-2,07	-2,39	-2,82	-3,35	-3,33	-2,99	-2,53
V075	13,68	10,44	9,58	7,75	5,47	3,31	-0,53	-2,74	-1,92
V053	-1,63	-2,07	-2,19	-2,45	-2,82	-3,29	-3,15	-2,77	-2,31
V064	4,99	2,70	2,17	0,92	-0,42	-1,38	-1,52	-1,38	0,47
V107	-0,91	-0,76	-0,45	-0,46	-0,42	-0,15	0,74	1,73	-1,45
V046	1,17	-0,38	-1,01	-2,08	-3,65	-5,54	-7,66	-8,84	-8,96
V097	-1,05	-0,64	-0,61	-0,15	0,22	0,75	1,47	1,65	2,16
V100	1,13	1,02	0,95	1,02	1,12	1,21	0,96	0,69	3,50
V074	1,16	1,04	0,96	1,03	1,12	1,20	0,93	0,65	3,49
V104	2,07	1,54	1,31	0,85	0,28	-0,26	-1,07	-1,81	0,07
V065	6,54	4,24	3,52	1,75	-0,20	-2,19	-4,73	-6,01	-7,40
V039	-1,03	-0,66	-0,32	0,18	0,91	1,62	2,37	2,91	2,61
V071	0,87	-0,64	-1,24	-2,34	-3,87	-5,33	-6,76	-7,23	-12,20
V072	3,49	2,63	2,19	1,51	0,74	0,07	0,11	0,40	3,11
V106	3,16	2,63	2,63	2,66	2,76	2,98	2,49	2,28	2,62
V082	-0,83	-0,76	-0,99	-0,82	-0,92	-0,98	-0,87	-1,07	1,34
V049	-0,37	-0,07	0,12	0,34	0,59	0,83	0,66	0,52	-0,84
V013	-4,33	-3,20	-2,90	-2,15	-1,33	-0,63	0,59	0,86	3,20
V050	-0,64	-0,35	-0,17	0,00	0,18	0,34	0,20	0,10	-1,62
V055	-0,31	-0,05	0,09	0,23	0,36	0,44	0,15	-0,05	-1,37
V041	-0,42	-0,13	0,01	0,17	0,32	0,42	0,18	0,02	-1,02
V012	4,70	3,55	3,80	3,32	3,33	3,85	4,59	5,16	-1,07
V096	-1,16	-1,11	-1,37	-1,22	-1,38	-1,52	-1,45	-1,66	1,05
V028	1,02	1,41	1,40	1,10	0,78	0,39	0,11	-0,12	1,92
V057	-1,27	-0,61	-0,39	-0,08	0,34	0,44	-0,05	-0,39	-0,46
V002	-3,42	-2,48	-2,41	-1,98	-1,58	-1,32	-0,59	-0,54	0,44
V040	-1,72	-1,01	-0,75	-0,39	0,11	0,31	-0,02	-0,23	-0,91
V010	-3,22	-1,65	-1,01	-0,41	0,58	1,74	3,77	5,06	0,43

UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA

PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA INDUSTRIAL - PEI

Rua Aristides Novis, 02, 6º andar, Federação, Salvador BA
CEP: 40.210-630
Telefone: (71) 3283-9800
E-mail: pei@ufba.br
Home page: http://www.pei.ufba.br