UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

SISTEMÁTICA PARA O DESENVOLVIMENTO E MANUTENÇÃO DE INFERÊNCIAS

DISSERTAÇÃO DE MESTRADO

Pedro Victor José de Lima Santos

Porto Alegre

2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

SISTEMÁTICA PARA O DESENVOLVIMENTO E MANUTENÇÃO DE INFERÊNCIAS

Pedro Victor José de Lima Santos

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

Área de concentração: Pesquisa e Desenvolvimento de Processos

Linha de Pesquisa: Engenharia de Sistemas – Projeto, Simulação, Modelagem, Controle e Otimização de Processos

Orientadores:

Prof. Dr. Jorge Otávio Trierweiler

Prof. Dr. Marcelo Farenzena

Porto Alegre

2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, assinada em ata (ATA Nº 401), aprova a Dissertação *Sistemática Para o Desenvolvimento e Manutenção de Inferências*, elaborada por Pedro Victor José de Lima Santos, como requisito parcial para obtenção do Grau de Mestre em Engenharia.

Comissão Examinadora:

Dr. Antônio Carlos Zanin - ACELEN

Dr. Rafael R. Sencio - ACELEN

Profa. Dra. Viviane Rodrigues Botelho - UFSCPA

ATA N° 401

Ata da Reunião da Comissão Julgadora da Dissertação de Mestrado de

PEDRO VICTOR JOSÉ DE LIMA SANTOS
Graduado em Engenharia Química

Data: 08/12/2023

TÍTULO: "SISTEMÁTICA PARA O DESENVOLVIMENTO E MANUTENÇÃO DE INFERÊNCIAS"

Orientadores: Prof. Dr. Jorge Otávio Trierweiler – DEQUI/UFRGS

Prof. Dr. Marcelo Farenzena – DEQUI/UFRGS

Comissão Julgadora:

Dr. Antônio Carlos Zanin – ACELEN

Dr. Rafael R. Sencio - ACELEN

Profa. Dra. Viviane Rodrigues Botelho – UFCSPA

Julgamento:	Assinatura	Parecer (*)
		Aprovado
		Aprovado
		Aprovado

Parecer Final: Aprovado

Alterações sugeridas: Conforme comentários durante a arguição.

Data da entrega da versão final: 08/01/2024.

OBSERVAÇÃO: ***(*) Aprovado ou Reprovado - ART 39 do Regimento do PPGEQ.***

Resumo

Os processos industriais modernos buscam constantemente uma produção mais segura, mais limpa e mais eficiente em termos energéticos. Para isso, sistemas avançados de monitoramento e controle vêm ganhando destaque nas fábricas e refinarias de petróleo. No entanto, processos industriais enfrentam problemas na medição de algumas variáveis, como por exemplo, a qualidade dos produtos, concentração dos componentes. O uso de analisadores em linha ou de medições laboratoriais não viabiliza o controle direto, por conta do tempo de amostragem e incerteza das medições dos analisadores. Para contornar esse problema e finalmente gerar informações frequentes e confiáveis, inferências são utilizadas. Seu desenvolvimento e manutenção ainda podem ser complexos em muitos casos, sendo uma nova metodologia proposta neste trabalho. Algumas técnicas de *machine learning* são apresentadas e utilizadas, bem como uma nova metodologia para segregação de dados, que agrupa as amostras com o método *k-means* e as seleciona aplicando *y-rank* nos *clusters* gerados. Essa técnica, *k-rank*, se mostrou mais eficiente do que o *y-rank* em seleções de dados com multiplicidade de soluções. Em seguida, a metodologia proposta para o desenvolvimento das inferências é detalhada, a qual é separada em etapas, cujo objetivo é melhorar a qualidade do modelo final. A primeira etapa, pré-processamento dos dados, é responsável pela lapidação dos dados. Posteriormente os dados são separados em conjuntos de calibração e teste, utilizando a metodologia *k-rank.* Em seguida os modelos são construídos através de métodos (*Ridge, Lasso, Lars,* e algoritmos de busca) que realizam a seleção de variáveis, descartando as variáveis desnecessárias aos modelos, os quais são validados utilizando diversas métricas de avaliação. Essa metodologia é testada com dados de uma simulação rigorosa de uma unidade de separação de propeno/propano. Essa unidade tem o objetivo de produzir propeno com pureza de 99,6%, a partir de uma carga de GLP sendo composta por três colunas de destilação. Os resultados obtidos mostram que é possível estimar as concentrações dos componentes-chave com elevado grau de confiança para ajudar no controle do processo. Para isso foram criadas expansões polinomiais com as variáveis de processo disponíveis, e foram criados modelos com parâmetros lineares, porém com características não lineares. Baseado na metodologia proposta, conclui-se que foi possível estimar as concentrações chave para o controle da unidade. Entretanto, em algumas situações de concentrações muito baixas o erro foi demasiado, como na coluna T-03, para a estimativa de propano no topo. Mas percebe-se que para concentrações acima de 0,0002 kg/kg de propano no topo dessa coluna, pode-se estimar com um erro máximo deaproximadamente 16%. E como a coluna propõe uma pureza de 99,6% de propeno, o erro para concentrações próximas de 0,004 kg/kg de propano ainda é menor. Pois quanto mais baixa a concentração de propano, maior o erro da sua estimativa. Já para as colunas T-01 e T-02, as variáveis foram estimadas com boa acurácia, com erros percentuais médios abaixo de 2% para todas as concentrações estimadas.

Abstract

Modern industrial processes constantly seek safer, cleaner and more energy-efficient production. To this end, advanced monitoring and control systems have been gaining prominence in oil factories and refineries. However, industrial processes face problems in measuring some variables, such as product quality and component concentration. The use of in-line analyzers or laboratory measurements does not enable direct control, due to the sampling time and uncertainty of analyzer measurements. To overcome this problem and finally generate frequent and reliable information, inferences are used. Its development and maintenance can still be complex in many cases, and a new methodology is proposed in this work. Some machine learning techniques are presented and used, as well as a new methodology for data segregation, which groups samples with the k-means method and selects them by applying y-rank to the generated clusters. This technique, k-rank, proved to be more efficient than y-rank in data selections with a multiplicity of solutions. Next, the proposed methodology for developing inferences is detailed, which is separated into stages, the objective of which is to improve the quality of the final model. The first step, data pre-processing, is responsible for polishing the data. Subsequently, the data is separated into calibration and test sets, using the k-rank methodology. The models are then built using methods (Ridge, Lasso, Lars, and search algorithms) that perform the selection of variables, discarding unnecessary variables from the models, which are validated using various evaluation metrics. This methodology is tested with data from a rigorous simulation of a propylene/propane separation unit. This unit aims to produce propylene with a purity of 99.6%, from an LPG charge consisting of three distillation columns. The results obtained show that it is possible to estimate the concentrations of key components with a high degree of confidence to help control the process. For this, polynomial expansions were created with the available process variables, and models were created with linear parameters, but with non-linear characteristics. Based on the proposed methodology, it is concluded that it was possible to estimate the key concentrations for controlling the unit. However, in some situations of very low concentrations the error was too much, as in column T-03, for the estimate of propane at the top. But it can be seen that for concentrations above 0.0002 kg/kg of propane at the top of this column, it can be estimated with a maximum error of approximately 16%. And as the column proposes a purity of 99.6% propylene, the error for concentrations close to 0.004 kg/kg of propane is even smaller. Because the lower the propane concentration, the greater the error in your estimate. For columns T-01 and T-02, the variables were estimated with good accuracy, with average percentage errors below 2% for all estimated concentrations.

"E o que há algum tempo era jovem, novo, hoje é antigo".
(Antônio Carlos Belchior)

Agradecimentos

A Deus primeiramente pelo dom da vida, por estar sempre comigo e à frente de cada passo que eu dou. Também por ser meu guia e força para superar cada dificuldade.

Aos meus pais e familiares, principais incentivadores do meu estudo, pelo amor, incentivo e apoio incondicional.

À minha tia, mãe e anjo da guarda, Maria do Carmo, que fez tudo por mim em vida e agora assiste mais essa nossa vitória lá de cima. Te amo, meu amor.

À minha esposa, Layanne, que foi quem mais me incentivou de vir ao outro lado do país fazer o mestrado e, nos momentos mais difíceis, não me deixou fraquejar.

Ao Dr. Lucas Ranzan. Formamos uma boa dupla, “rapaize”.

Aos amigos da academia de jiu-jitsu MR-Centro, vocês são minha família gaúcha.

Ao meu orientador Jorge Otávio Trierweiler e ao meu coorientador Marcelo Farenzena, pelo suporte que me deram, pela paciência e pelas instruções essenciais.

Aos componentes da banca examinadora pelos elogios, críticas e sugestões que contribuíram para a melhora do meu trabalho.

Aos docentes que se fizeram presente na minha formação, bem como toda a estrutura da universidade e cada funcionário desta.

E a todos que direta ou indiretamente fizeram parte da minha formação,

o meu muito obrigado.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

NOTAÇÃO E SIMBOLOGIA

MAPE

Max e%

RMSE

AIC

BIC

PCA

KPCA

GLP

R^2

LASSO

LARS

LASSOLARS

PCR

ACO

PLS

GLRT

CSTR

PC

SNV

NIR

SSE

SOC

Capítulo 1 – **Introdução**

Primeiramente, será delineado o amplo contexto desta dissertação, destacando a relevância dos modelos de inferência na condução de processos industriais. Essa abordagem está diretamente alinhada com os objetivos do trabalho, cuja estrutura completa é apresentada no desfecho deste capítulo.

1.1 Motivação

A busca por padrões de qualidade sempre fez parte dos objetivos das indústrias. A competição entre essas faz com que seja necessário produzir mais e melhor com menos recursos. Tão relevante quanto a competitividade são as leis e regulamentações ambienteis que estão cada vez mais rígidas para emissões de efluentes, sejam eles gasosos ou líquidos, provocando a necessidade de um monitoramento periódico e confiável dos seus níveis de emissão. Aliado a isso, os consumidores estão, cada vez mais, exigindo produtos de melhor qualidade.

É notável, diante deste cenário, a necessidade de se monitorar e controlar o processo produtivo. Isso remete às grandes quantidades de variáveis de processo. Algumas delas são de fácil medição, porém outras exigem dispositivos de alto custo, o que muitas vezes inviabiliza sua implantação. Uma alternativa para essas variáveis de difícil medição é a implantação de sistemas que consistem em modelos matemáticos que produzem estimativas em tempo real e confiáveis das variáveis não mensuráveis usando sua correlação com os dados disponíveis. Tais sistemas são comumente chamados de *soft sensors*, VOA (*Virtual Online Analyzer*) ou simplesmente de inferências.

Os processos industriais modernos buscam constantemente uma produção mais segura, mais limpa e mais eficiente em termos energéticos. Para isso, sistemas avançados de monitoramento e controle vêm ganhando destaque em plantas químicas. No entanto, processos industriais enfrentam problemas na medição de algumas variáveis, como por exemplo, a qualidade dos produtos, concentração dos componentes. Em algumas situações, são utilizados analisadores em linha, porém são equipamentos caros e de difícil instalação e manutenção. Além disso, suas informações podem não ser confiáveis e os elevados tempos de análise e amostragem impactam negativamente no desempenho das malhas de controle.

Na maioria dos casos, essas variáveis de difícil medição são obtidas através de medições laboratoriais, o que leva a estimativas pouco frequentes dificultando sua utilização para o controle do processo. Tais dificuldades na obtenção de variáveis relacionadas com a qualidade resultam inevitavelmente no controle inadequado do processo, o que pode conduzir a um aumento dos custos de produção, redução da qualidade do produto final, ou até mesmo situações que afetam a segurança do processo.

Para contornar esse problema e finalmente gerar informações frequentes e confiáveis, inferências vêm sendo utilizadas na indústria de processos químicos.

1.2 Inferências

Inferências são modelos matemáticos construídos a partir de variáveis facilmente medidas em um sistema com o intuito de estimar variáveis de difícil medição direta. Esses modelos possuem diversas aplicações na indústria que vão desde o simples monitoramento, detecção de falhas, back-up para outros sensores, até o uso para o controle avançado. O desempenho do controle estará fortemente ligado à qualidade da inferência, sendo então imprescindível a manutenção da sua qualidade ao longo do tempo (QIN et al., 1997).

Na literatura internacional, o termo inferências é compreendido como *soft-sensors,* palavra gerada a partir da união entre o termo *software* e *sensor*. Isso remete diretamente à sua definição, considerando que o termo *software* se refere ao modelo matemático, nesse caso, um código de computador; já o termo *sensor* está ligado ao fato desse modelo servir como instrumento de medição. Essas ferramentas são baseadas em modelos matemáticos sofisticados, técnicas computacionais de última geração e estratégias algorítmicas rigorosas. Áreas tão diversas como física, engenharia elétrica, biologia, matemática e até economia influenciaram o seu desenvolvimento, tornando-as um verdadeiro exemplo de excelência interdisciplinar (TOMAŽIČ, 2023).

Em português, Facchin (2005) sugeriu uma diferenciação entre os termos inferências e analisadores virtuais. O primeiro seria o modelo, o qual já foi retratado aqui. O segundo engloba o modelo, isto é, a inferência, e o esquema de correção baseado no erro de predição do modelo e a medição da variável inferida, normalmente, com análises de laboratório.

Kadlec *et al.* (2009) propôs uma metodologia dividida em etapas de acordo com a Figura 1.1. Na primeira etapa, é realizada a primeira inspeção nos dados. O objetivo principal deste passo é obter uma visão geral da estrutura de dados e identificar quaisquer problemas óbvios que podem ser tratados nesta fase inicial (por exemplo, variáveis bloqueadas com valor constante). A segunda etapa consiste na seleção dos dados a serem utilizados para treinamento e avaliação do modelo. Em seguida, são identificados e selecionados os períodos estacionários dos dados. Na terceira etapa, o objetivo é lapidar os dados com o intuito de viabilizar a modelagem, removendo *outliers* e realizando sua normalização para média zero e desvio padrão unitário. A quarta etapa é crítica para o desenvolvimento da inferência final. Até agora, não há uma teoria unificada para esta tarefa e, portanto, o tipo de modelo e seus parâmetros. No entanto, apesar da falta de uma abordagem comum para a seleção do modelo, existem algumas técnicas que podem ser

adotadas para esta tarefa. Uma abordagem possível é começar com um modelo simples ou estrutura do modelo (por exemplo, modelo de regressão linear) e aumentar gradualmente a complexidade do modelo, desde que seja significativo. E, após uma comparação entre os modelos criados, adotar o melhor modelo como sendo o ótimo.

Depois de desenvolver e implantar um *soft sensor*, ele deve ser mantido e sintonizado regularmente. A manutenção é necessária devido aos desvios e outras mudanças nos dados que causam a deterioração do desempenho da inferência. Atualmente, a maioria dos *soft sensors* não fornece mecanismos automatizados para sua manutenção. Este fato, juntamente com a evidência discutida anteriormente de desvios e mudanças nos dados, resulta no requisito de controle de qualidade contínua e manutenção desses. Isso acaba sendo um fator de custo significativo para a aplicação de *soft sensors*. Entretanto, muitas vezes não há uma medida objetiva para avaliar o nível de qualidade do modelo e o julgamento depende da percepção subjetiva do operador com base na interpretação visual do desvio entre o valor alvo correto e sua predição (Kadlec *et al.*, 2009).

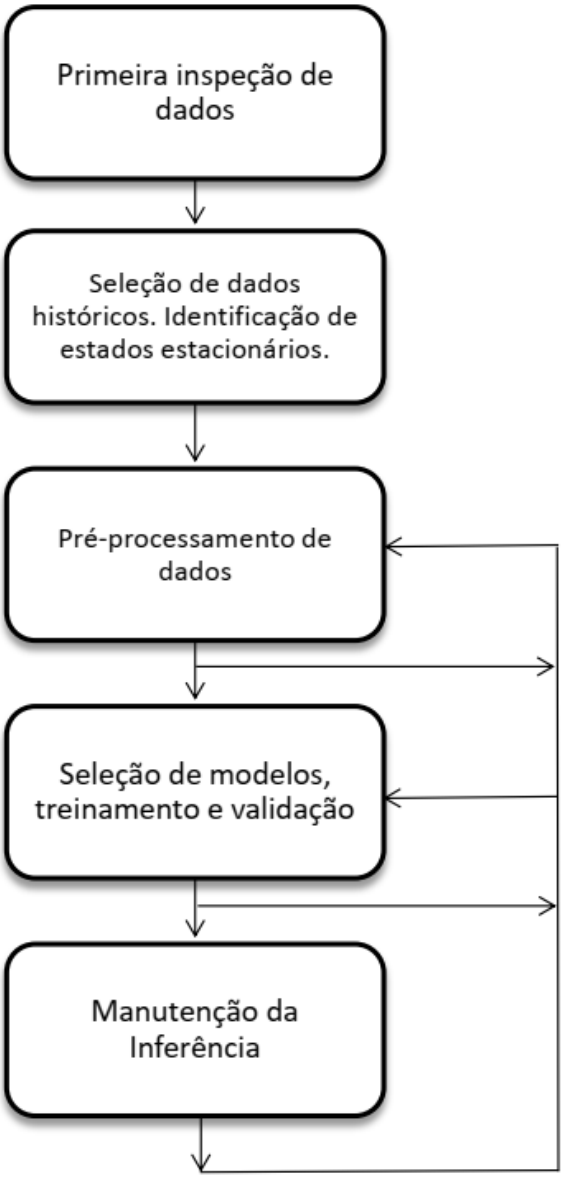

Figura 1.1: Metodologia para o desenvolvimento de inferências (Kadlec *et al.*, 2009).

1.3 Objetivo do trabalho

O objetivo geral deste trabalho é desenvolver uma sistemática para o desenvolvimento e manutenção de inferências, a partir de dados, atribuindo metodologias para casos lineares e casos não lineares. Para tanto os seguintes objetivos específicos nortearam a realização desta dissertação:

- Desenvolvimento de uma sistemática para geração de conjuntos de dados para calibração/treino e outro para validação/testes;
- Sistemática para escolha das variáveis empregando o método de otimização de colônia de formigas, bem como métodos de regressão com regularização;
- Sistemática para geração de modelos não lineares baseados na expansão das bases das variáveis de entrada para o modelo;
- Método estatístico de detecção de falhas para posterior manutenção das inferências.

1.4 Estrutura da dissertação

Neste capítulo foi realizada a apresentação geral desse trabalho, juntamente com os objetivos e motivações para o desenvolvimento do mesmo.

No capítulo 2 é feita uma revisão bibliográfica referente aos conceitos e informaçõesque serviram como alicerce para o andamento da pesquisa.

No capítulo 2 é mostrado, em formato de artigo, uma nova metodologia para segregação de dados em subconjuntos de calibração, validação e teste: o *k-rank*. Este capítulo foi necessário separá-lo do restante da dissertação por se tratar de uma nova metodologia desenvolvida. Entretanto, esta metodologia íntegra a metodologia do trabalho como um todo. Então o capítulo 3 faz uma revisão bibliográfica do que foi utilizado para o desenvolvimento do *k-rank*, explica sua metodologia e mostra os resultados com um exemplo hipotético.

O capítulo 4 detalha a metodologia para o desenvolvimento e manutenção de inferências.

O capítulo 5 apresenta o estudo de caso que será utilizado para aplicar a metodologia para o desenvolvimento e manutenção das inferências. Trata-se de uma unidade de separação de Propeno/Propano simulada em *Aspen Plus*.

O capítulo 6 apresenta os resultados obtidos utilizando as metodologias propostas aplicadas ao estudo de caso apresentado no capítulo 5.

O capítulo 7 apresenta as conclusões finais e sugestões de trabalhos futuros.

Capítulo 2 – **Revisão Bibliográfica**

Este capítulo apresenta um levantamento bibliográfico dos principais assuntos para o desenvolvimento e manutenção de inferências baseada em dados.

As etapas propostas por Kadlec*et al.* (2009), mencionadas no capítulo anterior, são seções deste capítulo, porém com algumas modificações para melhor organização do texto. Na primeira seção, são abordados os temas sobre tratamento de dados. Portanto, esta seção engloba as três primeiras seções apontadas por Kadlec *et al.* (2009). A segunda seção aborda a seleção dos dados. Na terceira seção, são mostradas técnicas de redução de dimensionalidade. Na quarta seção, são apresentadas técnicas de seleção de variáveis, construção e validação de modelos. Por fim, discute-se a manutenção da inferência.

2.1 Tratamento e Pré-Processamento de Dados

Nesta seção, são discutidas as técnicas essenciais de pré-processamento de dados que são empregadas na etapa de construção de modelos.

A qualidade dos dados e a quantidade de informações úteis que contém são fatores primordiais que determinam a qualidade do modelo final. Portanto, é imprescindível examinar e pré-processar o conjunto de dados antes de alimentá-lo para um algoritmo de aprendizagem (Raschka, 2015). Nesta seção, são discutidas as técnicas essenciais de pré-processamento de dados que foram empregadas na etapa de construção de modelos.

As plantas de processamento industrial podem apresentar muitos sensores. O objetivo principal dos sensores é fornecer dados para monitoramento e controle de processos. A grande quantidade desses dados também abre caminho para o desenvolvimento de *soft sensors.* Porém esses dados, quando coletados brutos, são necessários vários procedimentos para que se possa realizar a modelagem de forma correta. Em dados de processos, por exemplo, é importante identificar os estados estacionários para se realizar a modelagem. Também são comuns a presença de *outliers* eausência de dados (Kadlec *et al.*, 2009).

Detecção de Estados Estacionários

A identificação de estado estacionário em dados de processo é importante, pois os modelos de estado estacionário são amplamente utilizados no controle de processos, análise de processos *on-line* e otimização de processos. Se esses dados não apresentassem o efeito de distúrbios não medidos e ruídos, a identificação seria trivial, já que no estado estacionário não há alteração do valor do dado. Entretanto, em dados de processos, esses são corrompidos por ruídos, que podem ser atribuídos a vibrações mecânicas, interferências eletromagnéticas dispersas na transmissão de sinal, ruído eletrônico térmico, turbulência de fluxo, etc.(Rhinehart, 2013).

Uma implementação direta de um método para identificação de estado estacionário seria um teste estatístico da inclinação de uma tendência linear na série temporal de uma janela de dados em movimento. Aqui, em cada amostragem, usa-se regressão linear para determinar a melhor linha de tendência linear para os últimos N pontos de dados. Se o processo estiver em estado estacionário, então a inclinação da linha de tendência será idealmente zero. No entanto, devido à existência do ruído, a inclinação irá flutuar com valores próximos de zero. Portanto, um valor diferente de zero não é motivo para rejeitar a hipótese. Pode-se realizar um teste-t para a inclinação: se a inclinação de regressão dividida pelo erro padrão do coeficiente de inclinação excede o valor crítico, então há evidências suficientes para rejeitar com confiança a hipótese de estado estacionário. Alega-se então que se esteja provavelmente em estado transiente (Rhinehart, 2013).

Alternativamente, outra abordagem direta é avaliar o valor médio em janelas de dados sucessivas. Calcula-se a média e o desvio padrão dos dados em conjuntos sucessivos de dados de N amostras. Em seguida, comparam-se as médias com um teste-t. Se o processo estiver em estado estacionário, idealmente as médias são iguais. Porém o ruído fará com que as médias sequenciais flutuem. Se a flutuação for excessiva em relação à variabilidade dos dados, a estatística t (diferença na média dividida por erro padrão da média) excederá o valor crítico e se pode afirmar que é provavelmente um estado transiente (Rhinehart, 2013).

Cao e Rhinehart (1995) propuseram uma técnica baseada na comparação da variância dos dados estimada de duas maneiras distintas (Cao e Rhinehart, 1995). O autor Rhinehart esteve envolvido em muitas aplicações em escala piloto e em escala comercial. Por exemplo, Brown e Rhinehart (2000) demonstraram uma versão multivariável da técnica em um processo de destilação em escala piloto (Brown e Rhinehart, 2000). Huang e Rhinehart (2013) informam sobre uma aplicação no monitoramento do fluxo em uma unidade de absorção de gás em escala piloto.

Mejia et al. (2010) introduziram uma inovação na detecção de estados estacionários ao utilizar uma abordagem baseada no cálculo da correlação local amostral, denominada LOC. Este método demonstrou um desempenho superior em comparação com a técnica proposta por Cao e Rhinehart (1995). Ambos os métodos enfrentam desafios ao lidar com sinais altamente correlacionados, mas essas dificuldades podem ser superadas por meio da reamostragem adequada e da reconstrução dos modos de estado estacionário identificados na mesma escala de tempo (Mejia et al., 2010).

2.1.1 Normalização

Ao analisar duas ou mais variáveis, corriqueiramente é necessário normalizar os valores dessas, especialmente nos casos em que os valores são muito diferentes em escala (Zaki e Meira, 2014). Para tanto, duas formas de normalização são geralmente utilizadas: a) normalização por intervalo e b) normalização por desvio padrão.

Normalização por Intervalo – Range Normalization

Essa normalização transforma a variável x_i numa nova variável com um intervalo [0,1], de acordo com a equação 2.1 (Zaki e Meira, 2014).

$$x'_i = \frac{x_i - min_i\{x_i\}}{\hat{r}} = \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min_i\{x_i\}} \tag{2.1}$$

Normalização por desvio padrão – Standard Score Normalization

Também chamada de *z-normalization,* essa normalização transforma a variável em uma nova variável com média zero e desvio padrão unitário, calculada por Zaki e Meira (2014).

$$x'_i = \frac{x_i - \mu}{\sigma} \tag{2.2}$$

onde μ é uma estimativa da média e σ é uma estimativa do desvio padrão do vetor.

Há inúmeras outras formas de se normalizar dados. Em redes neuronais, por exemplo, é comum usar uma normalização para transformar as variáveis em um intervalo de 0,1 a 0,9 (Lee *et al.*, 2005). Em tratamentos de dados espectrais, é muito comum se utilizar formas de escalonar o espectro por completo, isto é, leva-se em consideração a matriz de dados completa, e não variável por variável (Fearn *et al.*, 2009). A normalização *SNV* (*Standard Normal Variate*) vem sendo amplamente utilizada e com bons resultados nos estudos de dados espectrais (Fearn *et al.*, 2009; Faassen e Hitzmann, 2015; Pessoa *et al.*, 2015; Ranzan, Ranzan*, et al.*, 2015; Ranzan, Trierweiler*, et al.*, 2015).

2.1.2 Tratamento de dados ausentes

É comum em aplicações do mundo real que, na matriz de dados originais, algumas amostras estejam com um ou mais valores em falta. Isso pode decorrer de várias situações como, por exemplo, falha no processo de coleta de dados ou defeitos nos intrumentos de medição. Normalmente esses valores em falta aparecem como espaços em branco no conjunto de dados ou como cadeias de caracteres, como *NaN* (*Not A Number*). Infelizmente, a maioria das ferramentas computacionais são incapazes de lidar com esses valores ausentes. Portanto, é crucial que haja um tratamento desses valores perdidos antes de prosseguir com análises adicionais (Raschka, 2015).

Há duas maneiras básicas de tratar esses valores ausentes: eliminando as amostras ou variáveis que os contêm, ou estimando os valores ausentes. A primeira alternativa é mais simples, porém podem-se perder informações valiosas. A segunda alternativa pode ser feita de várias maneiras. Neste caso, podem-se usar diferentes técnicas de interpolação para estimar os valores faltantes das outras amostras do conjunto de dados. Uma das técnicas de interpolação mais comuns é a imputação da média, onde simplesmente se

substituí o valor faltante pelo valor médio da variável (Raschka, 2015). Para casos em que se perde informação de determinadas variáveis em apenas uma fração do conjunto de dados total, é possível calibrar um modelo para preencher os dados faltosos.

2.1.3 Detecção e remoção de outliers

Estatisticamente, os *outliers* são definidos como dados que têm baixa probabilidade de serem consistentes com outros dados e potencialmente mascaram as características reais dos dados (Jeong *et al.*, 2017). Essas observações anormais podem, de alguma forma, conduzir adversamente a um modelo incoerente, estimativa de parâmetros tendenciosa e resultados incorretos. Portanto, é importante identificá-los antes de prosseguir com a modelagem (Williams *et al.*, 2002; Liu *et al.*, 2004; Maimon e Rokach, 2010).

Uma porção desses *outliers* pode ser facilmente identificada ao se analisar os dados em conjunto com o processo. Pode haver valores que violam limites técnicos ou físicos e esses podem ser descartados através de um processamento lógico simples (Fleck, 2012). A outra parte, de difícil identificação, exige procedimentos mais complexos que podem ser baseados em métodos univariados, propostos em trabalhos anteriores neste campo, e métodos multivariados que formam a maior parte do corpo atual de pesquisa (Maimon e Rokach, 2010). Métodos gráficos têm sido relatados como uma alternativa ineficaz para casos com múltiplos outliers, além de ficar sujeito à subjetividade do usuário (Jeong et al., 2017).

No trabalho de Lin *et al. (2007)* foi apontado um método baseado no desvio absoluto da mediana (MAD). Esse método é conhecido como identificador de *Hampel* e foi considerado por Lin *et al.* (2007) e Pearson (2002) como superior ao método 3σ, que remove os pontos que estão além de um limite superior ao desvio-padrão, dos dados, multiplicado por três (Pearson, 2002; Lin *et al.*, 2007). Ambos os métodos são baseados em estatísticas monovariável.

Existem técnicas que utilizam estatística multivariável, algumas baseadas na análise dos componentes principais (*PCA*) (Warne *et al.*, 2004).Os principais índices utilizados com os métodos *PCA* são a estatística de Hotelling, T^2, e a soma dos resíduos quadrados, *SPE* ou Q. A estatística T^2 é uma medida não negativa da variação capturada no modelo *PCA* e a estatística Q é uma medida da quantidade de variação não capturada pelo modelo *PCA*, também não negativa (Mansouri *et al.*, 2016).

A estatística de Hotelling, T^2, é uma análise multivariada dos dados, calculada para cada amostra através da seguinte equação (Hotelling, 1933):

$$T^2 = X^T \widehat{W} \hat{\Lambda}^{-1} \widehat{W}^T X \tag{2.3}$$

onde Λ = diag (λ_1, λ_2,...,λ_l) é a matriz diagonal contendo os autovalores dos *l* componentes principais retidos. A matriz *X* corresponde à matriz de dados centrados em zero e com desvio padrão unitário, e a matriz W corresponde aos autovetores, referentes aos maiores autovalores, associados à matriz de covariâncias de *X,* enquanto que $\widehat{W}$ corresponde à matriz contendo apenas os autovetores referentes aos *l* componentes principais retidos

(Russell *et al.*, 2000; Garcia-Alvarez e Fuente, 2011; Kruger e Xie, 2012; Mansouri *et al.*, 2016; Boullosa *et al.*, 2017).

Calculado o valor de T^2, resta então definir um valor limite para a caracterização dos dados como anômalos ou não. Uma falha é detectada quando a estatística de Hotelling excede esse valor limite, definido como T_α^2, i.e.

$$T_\alpha^2 = \frac{l(N-1)}{(N-l)} F_{l,N-l,\alpha} \tag{2.4}$$

onde α é o nível de significância (está geralmente entre 1% e 5%), que especificará o compromisso entre as falhas e falsos alarmes. N é o número de amostras no conjunto de treinamento que foram usadas no cálculo do *PCA*, l é o número de *PCs* retidos e $F_{l,N-l,\alpha}$ é o valor crítico da distribuição de Fisher com l e N - l graus de liberdade e nível de sinificância α (Russell *et al.*, 2000; Garcia-Alvarez e Fuente, 2011; Kruger e Xie, 2012; Mansouri *et al.*, 2016; Boullosa *et al.*, 2017).

A outra medida estatístia Q ou *SPE*, desenvolvida por Jackson e Mudholkar (1979), pode ser calculada (Jackson e Mudholkar, 1979; Russell *et al.*, 2000; Garcia-Alvarez e Fuente, 2011; Kruger e Xie, 2012; Mansouri *et al.*, 2016; Boullosa *et al.*, 2017).

$$Q = \left\|\tilde{X}\right\|^2 = \left\|\left(I - \hat{W}\hat{W}^T\right)X\right\|^2 \tag{2.5}$$

onde,

$$\tilde{X} = X - \hat{X} = \left(I - \hat{W}\hat{W}^T\right)X \tag{2.6}$$

Semelhante ao método de Hotelling, quando estatística Q excede um valor limite, definido como Q_a (equação 2.7), uma falta é detectada.

$$Q_\alpha = \varphi_1 \left[\frac{h_0 C_\alpha \sqrt{2\varphi_2}}{\varphi_1} + 1 + \frac{\varphi_2 h_0 (h_0 - 1)}{\varphi_1}\right] \tag{2.7}$$

onde,

$$\varphi_i\{i = 1,2,3\} = \sum_{j=i+1}^{m} \lambda_j^i \, , h_0 = 1 - \frac{2\varphi_1\varphi_3}{3\varphi_2^2}$$

e α é o nível de significância. Para os novos dados, a estatística Q é calculada e comparada com o valor limite Q_a(Jackson e Mudholkar, 1979; Mansouri *et al.*, 2016). Quando o limite de confiança é violado, uma falha é declarada. Vale salientar que o valor de Q_a é calculado com base no pressuposto de que as medições são independentes e seguem distribuição normal multivariada. Portanto, a estatística Q é altamente sensível aos erros de modelagem (Benaicha *et al.*, 2010; Mansouri *et al.*, 2016).

Boullosa *et al.* (2017) monitoraram o processo de lubrificação do cilindro de um motor marinho à diesel, para condições específicas de operação do navio. Um estudo com um conjunto de dados históricos para um trabalho ideal foi feito e com uma nova entrada de 23 amostras, as observações fora do controle estatístico foram detectadas usando a estatística de Hotelling T^2(Boullosa *et al.*, 2017).

Além das estatísticas T^2 e Q para identificação de *outliers* ou falhas no processo, outros métodos multivariados vêm sendo propostos na literatura. Métodos baseados em PCA

podem ser vistos no trabalho de Garcia-Alvarez e Fuente (2011)(Garcia-Alvarez e Fuente, 2011). Escobar *et al.* (2017) propuseram uma combinação de GTM (*generative topographic mapping*) e teoria gráfica, como uma abordagem não supervisionada. Eles compararam o método com métodos baseado em PCA (Escobar *et al.*, 2017).

Seleção de Dados

Uma importante etapa no desenvolvimento de inferências ou em modelagem em geral é a seleção do conjunto de pontos que serão empregados. Depois de realizado o tratamento dos dados, é necessário dividir o conjunto dos dados em subconjuntos de treinamento, validação (se necessário) e testes. O conjunto de treinamento é destinado à calibração dos parâmetros do modelo; é com esse conjunto que se calibra o modelo. O conjunto de validação, quando necessário (Massaron e Boschetti, 2016), é utilizado para seleção de modelos calibrados. Já com o conjunto de testes são realizados testes finais com o modelo selecionado, sem que os mesmos sejam utilizados em nenhuma etapa de geração dos modelos (Raschka, 2015; Kramer, 2016; Massaron e Boschetti, 2016). Um bom modelo deve ser capaz de generalizar dados que não foram usados em seu treinamento (Kramer, 2016).

Nesta etapa de seleção de dados, alguns pontos devem ser salientados. Ao deixar de lado parte das amostras, reduz-se o número de exemplos a serem aprendidos, enquanto que os modelos precisam de tantos quanto possível para reduzir a variância das estimativas, desambiguar variáveis colineares e modelar corretamente a não-linearidade. Outro ponto é que uma vez que esta etapa envolve sub-amostragem, consequentemente há um risco de desenhar conjuntos que são muito favoráveis ou desfavoráveis para treinamento e testes (Massaron e Boschetti, 2016).

Para resolver essas situações, diversas técnicas foram desenvolvidas com o intuito de realizar a melhor separação do conjunto inicial de dados. Neste item, serão apresentadas algumas das técnicas disponíveis na literatura para seleção de dados, bem como uma nova técnica desenvolvida e apresentada nesta dissertação de mestrado.

2.1.4 Seleção Aleatória

A seleção aleatória das amostras é o método mais simples de segregação de dados. A seleção é feita de forma randômica, sendo necessário apenas informar a proporção em cada subconjunto (por exemplo, 80% para calibração, 10% para validação e 10% para testes). Este método é bastante utilizado em validação cruzada, e, por sua natureza estocástica, deve ser repetido várias vezes para que possa ser considerado significativo. Do contrário, a seleção aleatória fica sujeita a uma má seleção dos dados, prejudicando o modelo final (Facchin, 2005).

2.1.5 K-fold

Na validação cruzada *k-fold*, divide-se aleatoriamente o conjunto de dados de treinamento em *k* subconjuntos sem repetição, onde (*k*-1) subconjuntos são usados para o treinamento do modelo e um subconjunto é usado para validação. Este procedimento é repetido *k* vezes para que se obtenham *k* modelos e estimativas de desempenho (Raschka, 2015).

O valor padrão para *k* é 10, que é tipicamente uma escolha razoável para a maioria das aplicações. No entanto, se os conjuntos de treinamento forem relativamente pequenos, pode ser útil aumentar o valor de *k*. Com isso, mais dados de treinamento serão usados em cada iteração. No entanto, valores grandes de *k* também aumentarão o tempo de execução do algoritmo de validação cruzada e produzirão estimativas com maior variação, uma vez que os subconjuntos de treinamento serão mais semelhantes entre si. Por outro lado, se o conjunto de dados for suficientemente grande, pode-se escolher um valor menor para *k*, por exemplo, *k* = 5, e ainda obter uma estimativa precisa do desempenho médio do modelo, reduzindo o custo computacional (Raschka, 2015). O problema com a validação cruzada é quando se tem uma modelagem custosa computacionalmente, pois um modelo precisa ser gerado para cada repetição do algoritmo.

2.1.6 y-rank

Outro método para seleção de dados é conhecido como *y-rank*. Este seleciona as amostras a partir dos valores da variável a ser modelada. Os dados são dispostos em ordem crescente do vetor de saída *y*. Em seguida, são determinadas as proporções de cada subconjunto de dados. Um exemplo: 50% para treinamento e 50% para testes. Então o *y-rank* adota uma espécie de DNA do tipo "calibração-teste", por exemplo, em que esse DNA se repete por todo o intervalo do conjunto de dados (Facchin, 2005; Fleck, 2012). Vale ressaltar que os valores extremos são mantidos no conjunto de calibração, para tentar evitar extrapolações. A Figura 2.1 ilustra esse exemplo de separação pelo algoritmo *y-rank* aplicado em um pequeno conjunto de dados hipotético.

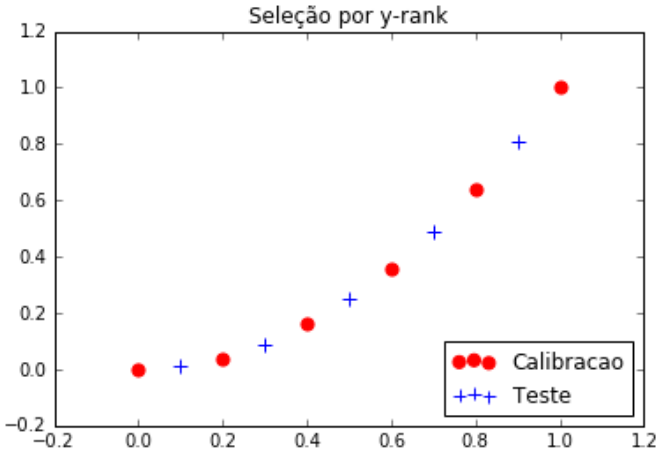

Figura 2.1: Seleção por *y-rank.*

2.2 Análise da Dimensionalidade do Problema

Nesta seção serão apresentadas técnicas de *PCA* e *Kernel-PCA* que podem ser utilizadas para uma análise dos dados já tratados. Pode-se retirar informações importantes através dessas técnicas e ajudar em decisões na etapa de seleção de variáveis.

2.2.1 PCA: Principal Component Analysis

Análise de componentes principais (*principal component analysis, PCA*) é uma transformação linear não supervisionada com o objetivo de reduzir a dimensionalidade do problema. *PCA* serve como uma ferramenta para identificar padrões em dados com base na correlação entre as variáveis disponíveis e permite determinar quantas destas variáveis podem descrever o problema satisfatoriamente. Em suma, essa técnica visa encontrar as

direções da máxima variância nos dados e projeta-os para um novo espaço com a mesma ou menor dimensão. Os eixos ortogonais (componentes principais) podem ser interpretados como as direções de máxima variância, respeitando a restrição de que os novos eixos são ortogonais entre si. A Figura 2.2 mostra x_1 e x_2, que são características de um conjunto original de dados e PC1 e PC2 são os componentes principais (Raschka, 2015).

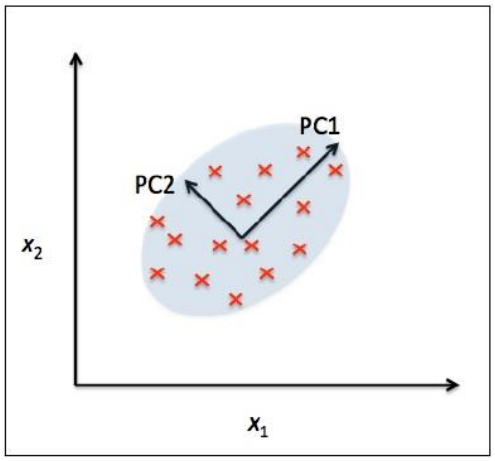

Figura 2.2: Explicação dos componentes principais de dados hipotéticos (Raschka, 2015).

Ao usar *PCA* para a redução de dimensionalidade, constrói-se uma matriz de transformação *W* de dimensão *d* x *k* que permite mapear um vetor de amostra *x* para um novo subespaço de dimensão *k* que tem dimensão menor que o espaço original dos dados. Para chegar a essa transformação, basta seguir os passos a seguir, descritos em Raschka (2015):

1. Normalizar o conjunto de dados de dimensão *d;*

2. Construir a matriz de covariância:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

A matriz ∑ seria para um caso em que *d = 3,* ou seja, a matriz de covariância gera uma matriz de dimensão *d* x *d.* E a covariância entre duas variáveis pode ser calculada da seguinte forma:

$$\sigma_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(x_j^{(i)} - \mu_j\right)\left(x_k^{(i)} - \mu_k\right) \quad (2.8)$$

Onde μ_j e μ_k são as médias das colunas *j* e *k,* respectivamente, que serão zero se estiverem normalizadas.

3. Decompor a matriz de covariâncias em seus autovetores e autovalores:

$$\sum v = \lambda_v \quad (2.9)$$

4. Selecionar *k* autovetores que correspondem aos *k* maiores autovalores, onde *k* é a dimensionalidade do novo subespaço.

5. Construir a matriz de projeção *W* a partir dos k autovetores selecionados.

6. Transformar o conjunto de dados *x* de dimensão *d* usando a matriz de projeção *W* para obter o novo subespaço *Z* de dimensão *k.*

$$x = [x_1, x_2, \cdots, x_d], x \in \mathbb{R}^d$$

$$\downarrow xW, W \in \mathbb{R}^d$$

$$Z = [z_1, z_2, \cdots, z_k], z \in \mathbb{R}^k \tag{2.10}$$

2.2.2 Kernel-PCA

Muitas das técnicas de *machine learning* assumem que os dados são linearmente separáveis. No entanto, nem todos os dados possuem características lineares. Logo, para esses casos não lineares, a redução da dimensionalidade utilizando *PCA* pode ser um caminho equivocado. A Figura 2.3 mostra a diferença entre dados que podem ser separados linearmente e dados que não podem (Raschka, 2015).

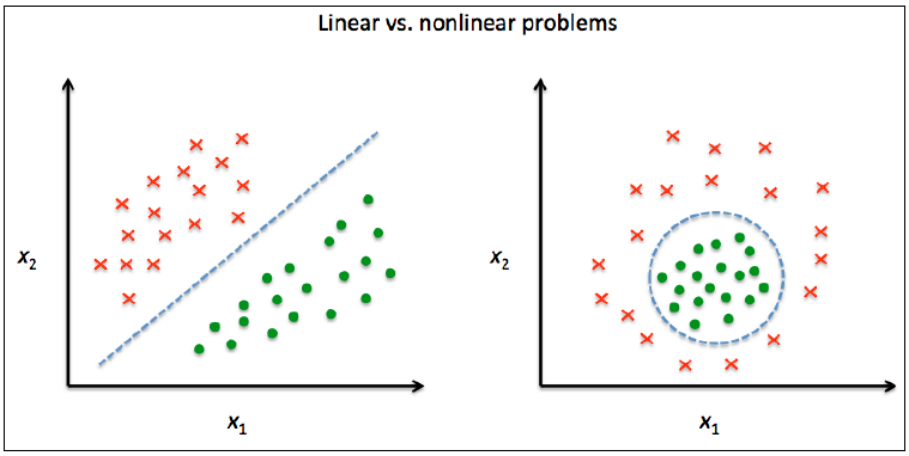

Figura 2.3: Problema linearmente separável versus problema não-linearmente separável (Raschka, 2015).

A ideia por trás do *kernel-PCA* é realizar um mapeamento não-linear que transforma os dados em um espaço de maior dimensão. Em seguida, usa-se *PCA* neste espaço de maior dimensão para projetar os dados de volta para um espaço de menor dimensão onde as amostras podem ser separadas linearmente (Raschka, 2015).

Scholkopf (1997) generalizou a abordagem *KPCA* alterando a forma como se calcula a matriz de covariâncias, trazendo agora uma não linearidade para o problema (Schölkopf *et al.*, 1997). Considerando que os dados foram normalizados e as variáveis possuem média zero, a nova matriz de covariâncias pode ser calculada da seguinte forma:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \emptyset\left(x^{(i)}\right) \emptyset\left(x^{(i)}\right)^T \tag{2.11}$$

Ø pode ser pensado como uma função que cria combinações não-lineares das variáveis originais para mapear o conjunto de dados original de dimensão *d* em um espaço de maior dimensão *k*.

Para obtenção dos autovetores, é necessário resolver a seguinte equação:

$$\sum v = \lambda_v$$

$$\rightarrow \frac{1}{n}\sum_{i=1}^{n} \emptyset\left(x^{(i)}\right)\emptyset\left(x^{(i)}\right)^T v = \lambda_v$$

$$\rightarrow v = \frac{1}{n\lambda}\sum_{i=1}^{n} \emptyset\left(x^{(i)}\right)\emptyset\left(x^{(i)}\right)^T v = \frac{1}{n}\sum_{i=1}^{n} a^{(i)}\emptyset\left(x^{(i)}\right) \quad (2.12)$$

A matriz de covariâncias pode ser escrita na notação de matrizes, onde Ø(X) corresponde a uma matriz de dimensão *n* x *k.*

$$\sum = \frac{1}{n}\sum_{i=1}^{n} \emptyset\left(x^{(i)}\right)\emptyset\left(x^{(i)}\right)^T = \frac{1}{n}\emptyset(X)^T\emptyset(X) \quad (2.13)$$

Agora o problema de autovetores pode ser escrito da seguinte forma:

$$v = \frac{1}{n}\sum_{i=1}^{n} a^{(i)}\emptyset\left(x^{(i)}\right) = \lambda\emptyset(X)^T a \quad (2.14)$$

E substituindo em ∑v = λv:

$$\frac{1}{n}\emptyset(X)^T\emptyset(X)\emptyset(X)^T a = \lambda\emptyset(X)^T a \quad (2.15)$$

Multiplicando ambos os lados por Ø(X) tem-se:

$$\frac{1}{n}\emptyset(X)\emptyset(X)^T\emptyset(X)\emptyset(X)^T a = \lambda\emptyset(X)\emptyset(X)^T a \quad (2.16)$$

$$\rightarrow \frac{1}{n}\emptyset(X)\emptyset(X)^T a = \lambda a$$

$$\rightarrow \frac{1}{n}Ka = \lambda a$$

$$K = \emptyset(X)\emptyset(X)^T$$

$$k\left(x^{(i)}, x^{(j)}\right) = \emptyset\left(x^{(i)}\right)^T\emptyset\left(x^{(j)}\right) \quad (2.17)$$

Onde *K* ou $k(x^{(i)}, x^{(i)})$ é chamada matriz de similaridade (*Kernel*). Basicamente, a função *kernel* (ou simplesmente *kernel*) pode ser entendida como uma função que calcula um produto ponto entre dois vetores - uma medida de similaridade. As funções *kernel* mais comuns são:

- *Kernel* polinomial:

$$k\left(x^{(i)}, x^{(j)}\right) = \left(x^{(i)T} x^{(j)} + \theta\right)^{p} \tag{2.18}$$

onde θ é o limite e *p* é a potência do polinômio; esses valores são definidos pelo usuário.

- *Kernel* tangente hiperbólica ou sigmoide:

$$k\left(x^{(i)}, x^{(j)}\right) = \tanh\left(\eta x^{(i)T} x^{(j)} + \theta\right) \tag{2.19}$$

- *Kernel* RBF (*Radial basis function):*

$$k\left(x^{(i)}, x^{(j)}\right) = e^{-\frac{\left\|x^{(i)} - x^{(i)}\right\|^2}{2\sigma^2}} \tag{2.20}$$

2.3 Seleção de Variáveis: Construindo Modelos

Os algoritmos de seleção de variáveis são familiares na literatura: *Forward Selection, Backward Elimination, All Subsets Regression*. Esses e alguns outros algoritmos são usados para gerar um modelo que irá prever uma resposta y com base em medidas covariadas x_1,x_2,...,x_i. A qualidade desses modelos é muitas vezes definida em termos de exatidão de predição. Todavia a parcimônia é outro critério importante: modelos mais simples são preferidos por causa do conhecimento sobre a relação x-y (Efron *et al.*, 2004). Além disso, a presença de variáveis redundantes ou desnecessárias pode ocasionar o sobreajuste do modelo (Massaron e Boschetti, 2016).

Regularização é uma maneira de modificar o papel de variáveis em um modelo de regressão para evitar o sobreajuste e para atingir formas mais simples de funções. Essa técnica utiliza uma penalização nos parâmetros para evitar modelos de dimensões elevadas diminuindo ou reduzindo a zero os coeficientes relativos às variáveis que são irrelevantes ou redundantes para o modelo. Uma vantagem da regularização é que não há a necessidade de manipular o conjunto original dos dados, podendo esta ser usada em inferências online ou modelos preditivos online, sem a intervenção humana (Massaron e Boschetti, 2016). Para se ter uma ideia mais clara, se discutirá a seguir as três principais formas de regularização.

2.3.1 Ridge (L2 Regularization)

A regressão *Ridge* é uma técnica de regularização utilizada em modelos de regressão linear para lidar com o problema de multicolinearidade (alta correlação entre variáveis independentes) e, ao mesmo tempo, evitar coeficientes de regressão muito grandes. Na regressão linear padrão, o objetivo é minimizar a soma dos quadrados dos resíduos entre os valores previstos pelo modelo e os valores reais. Já na regressão *Ridge*, a ideia é reduzir os coeficientes das variáveis que afetam o modelo, incluindo, na função objetivo, uma penalidade adicional baseada na magnitude dos coeficientes de regressão. Com isso, a contribuição dessas variáveis exerce pouca influência no resultado final do modelo (Massaron e Boschetti, 2016).

A função objetivo da regressão Ridge (equação 2.21) é o critério que o algoritmo de otimização tenta minimizar durante o treinamento do modelo. Na regressão Ridge, a função objetivo incorpora tanto o termo tradicional de erro quadrático médio (MSE) quanto a penalidade de regularização para os coeficientes. A função MSE (equação 2.22) é

uma medida comum de quão bem um modelo de regressão se ajusta aos dados. Ela quantifica a média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais observados. A segunda parte da função objetivo é a penalidade de regularização para os coeficientes. A inclusão deste termo ajuda a evitar coeficientes muito grandes, contribuindo para a estabilidade do modelo, especialmente em situações de multicolinearidade (Massaron e Boschetti, 2016).

$$J(\beta) = MSE + \alpha \sum_{i=1}^{p} \beta_i^2 \tag{2.21}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i')^2 \tag{2.22}$$

O parâmetro α controla a força da penalidade de regularização. Quando α é zero, a penalidade é inexistente, e a regressão ridge é equivalente à regressão linear padrão. À medida que α aumenta, a penalidade se torna mais pronunciada, levando a coeficientes mais suavizados. y_i é o valor real observado para a i-ésima observação. y_i' é o valor predito pelo modelo para a i-ésima observação. O objetivo do treinamento é encontrar os coeficientes β que minimizam essa função objetivo (equação 2.21). Isso geralmente é feito usando métodos de otimização, como gradiente descendente, ajustando iterativamente os coeficientes para encontrar o mínimo da função objetivo (Massaron e Boschetti, 2016). Unindo as equações 2.21 e 2.22 e fazendo algumas operações matemáticas, podemos reescrever a função objetivo como na equação 2.23.

$$\hat{\beta}^{Ridge} = arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \alpha \sum_{j=1}^{p} {\beta_j}^2 \right\} \tag{2.23}$$

2.3.2 LASSO (L1 Regularization)

A regularização *LASSO* (*Least Absolute Shrinkage and Selection Operator*), introduzida por Rob Tibshirani (1994), adiciona à função objetivo uma penalização absoluta dos coeficientes do modelo. Isso irá selecionar apenas as variáveis informativas, levando a zero as não informativas, trazendo mais clareza e utilizando apenas as variáveis realmente necessárias ao modelo (Tibshirani, 1994; Kramer, 2016; Massaron e Boschetti, 2016). A estimativa *LASSO* pode ser definida a seguir, onde t ≥ 0 é a restrição na soma do valor absoluto de todos os coeficientes β (Tibshirani, 1994; Cui e Wang, 2016).

$$\hat{\beta}^{Lasso} = arg\min_{\beta} \{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \} \tag{2.24}$$

Cuie Wang (2016) obtiveram êxito em usar um grupo de modelos *LASSO* com fator de regularização α diferentes como seletores de variáveis, com o intuito de gerar várias possibilidades de combinações de variáveis do conjunto de dados original. Este conjunto continha dados de experimentos com comparações sobre a estimativa do teor de proteína do leite a partir de seu espectro de resonância nuclear magnética (*NMR spectrum*). O conjunto de dados possui 31.570 variáveis (tamanho do espectro) e 120 amostras. A construção dos modelos, entretanto foi feita utilizando redes neurais, a partir das variáveis selecionadas pelo *LASSO* (Cui e Wang, 2016).

2.3.3 LARS

Least Angle Regression (*LARS*) é um algoritmo de regressão que, de forma rápida e inteligente, seleciona as melhores variáveis para usar no modelo. *LARS*, proposto por Efron *et al.*(2004), é uma evolução do algoritmo *Forward Selection*, também chamado de *Forward Stepwise Regression* (Weisberg, 1980), e do algoritmo de Regressão *Forward Stagewise* (Efron *et al.*, 2004). Também pode ser visto como uma versão vetorial do *LASSO* para acelerar os cálculos (Iturbide *et al.*, 2013).

LARS é um compromisso entre a rapidez do clássico *Forward Selection* e a cautela do *Forward Stagewise*, criando uma solução que é estável, não tão propensa a *overfitting* e rápida. O algoritmo é chamado de regressão de menor ângulo pelo seguinte motivo: a cada iteração, o algoritmo inclui no modelo a variável mais correlacionada com o vetor residual, isto é, a variável que gera o menor ângulo com o residual.

O algoritmo *LARS* pode ser resumido segundo o algoritmo a seguir (Massaron e Boschetti, 2016):

1. No modelo, cada variável tem um peso zero associado, isto é, $w_i = 0$ para cada variável i.

2. Dos possíveis preditores para o problema, aquele com a maior correlação absoluta com a variável alvo y é adicionado parcialmente ao modelo.

3. Manter o aumento do peso $w_j$de qualquer outro preditor (por exemplo, preditor j) que tenha tanta correlação com o vetor residual quanto o preditor atual tem.

4. Aumentar w_i e w_j simultaneamente até que outro preditor tenha tanta correlação com o vetor residual quanto os preditores atuais têm.

5. Continuar adicionando preditores e pesos até que todos os preditores estejam no modelo ou atenda a outro critério de terminação, como o número de iterações.

Como Iturbide *et al*. (2013) mencionou em seu trabalho, o algoritmo *LARS* prossegue na direção equiangular da variável mais correlacionada com o residual atual. Em cada estágio, uma variável é adicionada ao conjunto ativo, então este processo pode continuar até que todas as variáveis tenham sido adicionadas. Tanto o *LARS* como o *LASSO* produzem modelos mais simples e, portanto, mais facilmente interpretáveis. Em seu trabalho, Iturbide *et al.* (2013) comparou modelos gerados pelo *LARS* e *LASSO* para 4004 séries temporais diferentes. Ele concluiu que o *LARS* foi superior ao *LASSO* por uma pequena diferença, mas que os dois demonstraram uma boa alternativa para seleção de variáveis.

2.3.4 ACO-Ant Colony Optimization: Uma Alternativa para Seleção de Variáveis

O algoritmo *Ant Colony Optimization* é baseado no comportamento coletivo hipotético das formigas ao saírem em busca de fontes de alimentos. Durante essa busca, as formigas secretam feromônios para marcação do caminho que, todavia, evaporam ao longo do tempo. Na natureza, formigas que viajam pelo caminho mais curto em busca do alimento, retornam ao ninho mais rapidamente, de modo que o caminho percorrido por estes indivíduos tem uma maior concentração de feromônio. Esta trilha age como um chamariz

para outras formigas e, com o tempo, todos os indivíduos da colônia tendem a atravessar este ótimo (mais curto) caminho (Allegrini e Olivieri, 2011), como mostrado na Figura 2.4.

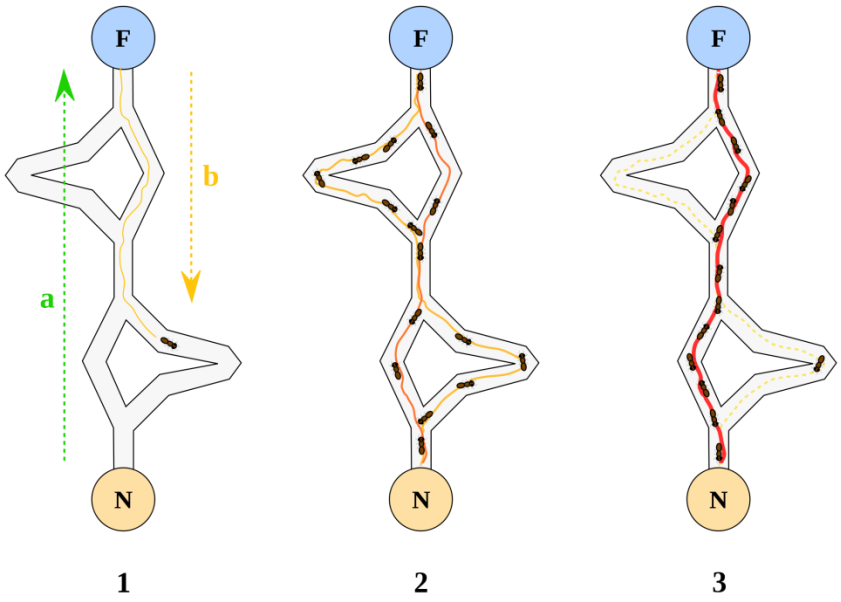

Figura 2.4: Formação da trilha de feromônio das formigas em busca de alimento
Fonte: (Toksari, 2016).

Dorigo e Gambardela (1997) desenvolveram a primeira versão do ACO buscando solucionar o problema do Caixeiro Viajante, um problema de busca de otimização combinatória no espaço de permutações (Dorigo e Gambardella, 1997; Dorigo e Blum, 2005; Dorigo *et al.*, 2006; Ranzan *et al.*, 2014). Atualmente, vários estudos têm sido publicados sobre a aplicação do método ACO para triagem de variáveis (Ranzan *et al.*, 2014), (Allegrini e Olivieri, 2011), (Hemmateenejad *et al.*, 2011), (Mullen *et al.*, 2009) e (Socha e Dorigo, 2008).

Embora os algoritmos seletores de variáveis mais simples e rápidos sejam bons, nem sempre conseguem resolver os problemas da melhor maneira, especialmente quando se trata de problemas com um número vasto de variáveis (Iturbide *et al.*, 2013).

Ranzan *et al.* (2014) aplicou a soma de erros quadrados (*SSE*) como critério para atualizar a trilha de feromônio e comparar modelos. O objetivo era prever o conteúdo de proteína em diferentes marcas de farinha com base em dados espectrais *NIR*. Os resultados mostraram que a utilização de *ACO* como ferramenta de filtragem possibilitou a seleção de importantes regiões espectrais, aumentando o coeficiente de determinação de modelos gerados em 60% em comparação com outros métodos que utilizaram todo o espectro, como *PCA* e *PCR*.

Sua metodologia associa o compromisso entre a densidade de feromônio que as variáveis possuem e a randomicidade do gatilho para a escolha das variáveis. Quanto mais feromônio uma variável possui, mais chances de ser escolhida. O algoritmo pode ser resumido de acordo com o esquema da Figura 2.5. Outros detalhes podem ser encontrados em Ranzan *et al.* (2014) e Mullen *et al.,* (2009).

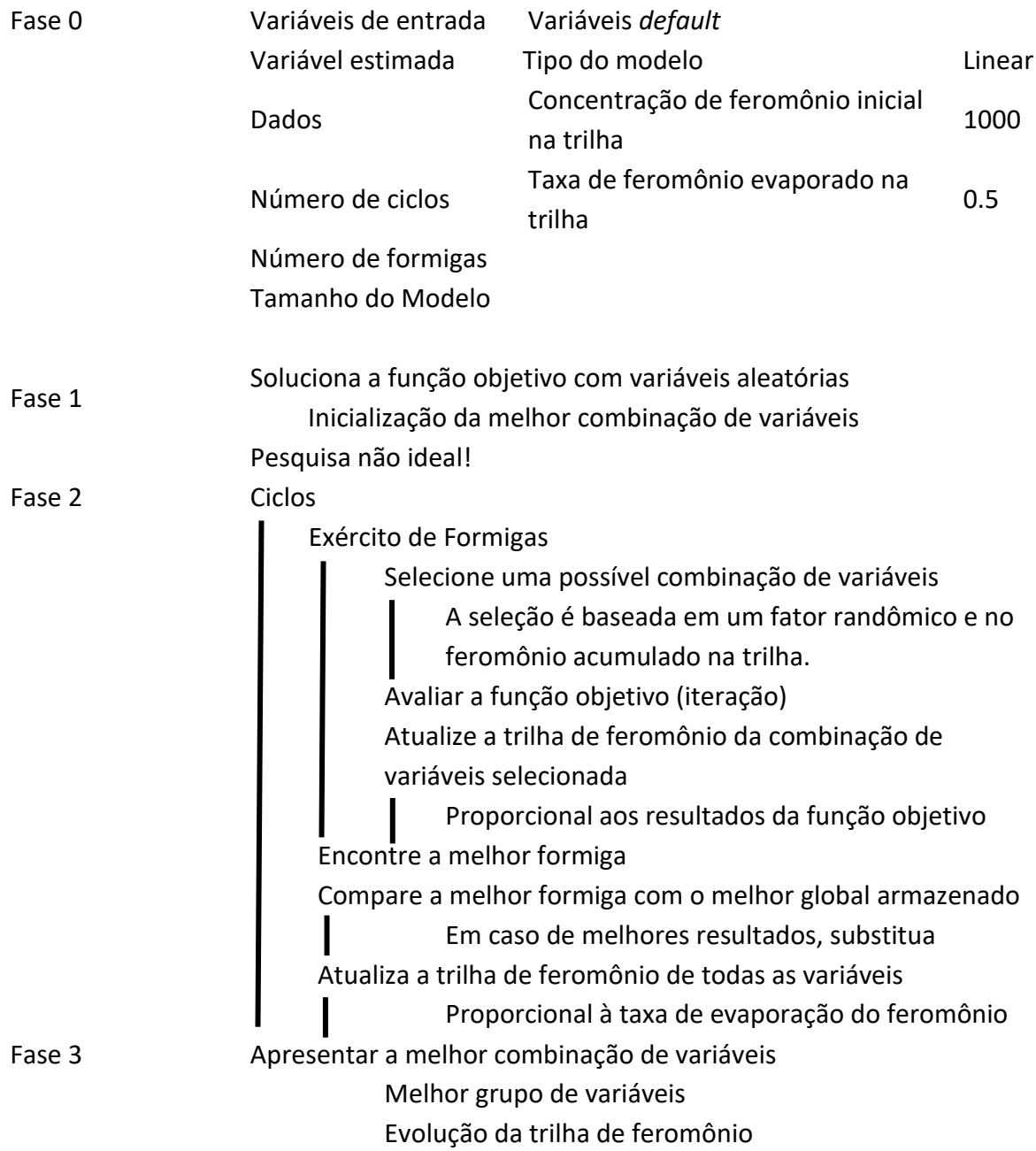

Figura 2.5: Implementação discreta do *ACO.* Fonte: (Ranzan *et al.*, 2014).

O *ACO* desenvolvido por Ranzan *et al.* (2014) aplicou a soma de erros quadrados (*SSE*) como critério para atualizar a trilha de feromônio e comparar modelos. Em seu método, foi utilizado a regressão linear para gerar modelos na região de busca. Dessa forma, as variáveis selecionadas recebem a mesma quantidade de feromônio, baseada no erro de predição (Ranzan *et al.*, 2014).

Uma melhoria do algoritmo foi proposta no trabalho de Pessoa *et al.* (2015), que realizou testes estatíscos com as variáveis selecionadas para verificar quais eram mais importantes para o modelo. As métricas individuais escolhidas foram o teste-t, associando

diretamente o valor t absoluto ao componente espectral, e o teste-F, associando o valor F de um submodelo ao componente espectral ausente nele (Pessoa *et al.*, 2015).

Em outras palavras, o método F-test funciona da seguinte forma: primeiro, o algoritmo escolhe variáveis dentre as disponíveis e gera um modelo. Então, um dos componentes é retirado do grupo original e outro modelo, com menor tamanho, é construído. O submodelo é então comparado ao modelo completo através do teste-F. O valor F está associado à variável não incluída no subgrupo. Portanto, as variáveis que são importantes para o modelo final terão uma estatística F mais alta, uma vez que não usá-las resulta em um modelo pior do que o completo. Essas variáveis com o valor de F maior receberão um incremento maior na trilha de feromônios (Pessoa *et al.*, 2015).

2.4 Critérios de Avaliação de Modelos

Segundo Greene (2002), não há um critério de avaliação absoluto para modelos de regressão linear. Entretanto, na comparação entre modelos, uma boa prática é utilizar as mesmas métricas, pois assim faz-se uma análise justa (Greene, 2002). A seguir, serão discutidos alguns critérios de avaliação de modelos que serão posteriormente utilizados na metodologia do trabalho.

2.4.1 R^2 e RMSE

O coeficiente de determinação (R^2) e a raiz do erro médio quadrático (*RMSE, root mean squared error*) são critérios bastante utilizados baseados na avaliação dos resíduos. Se todas as predições fossem perfeitas, os resíduos teriam valor zero, o coeficiente de determinação seria 1, e a raiz do erro médio quadrático seria 0. O coeficiente de determinação faz uma medida da proporção do quanto a variação em y pode ser explicada pela variação nos regressores, enquanto que o *RMSE* é apenas uma medida do erro de predição. A forma como calcular ambos os critérios é dada por (Greene, 2002).

$$R^2 = \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{[\sum_i (y_i - \bar{y})^2][\sum_i (\hat{y}_i - \bar{\hat{y}})^2]} \tag{2.25}$$

$$RMSE = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2} \tag{2.26}$$

2.4.2 R^2 Ajustado, AIC e BIC

O coeficiente de determinação R^2 pode ser útil para avaliar o ajuste sobre os dados utilizados na calibração, como também na previsão. Porém, quando o modelo é direcionado à predição, as informações no âmbito da calibração podem não ser necessariamente ideias. O R^2 pode não cair quando variáveis são adicionadas ao modelo. Entretanto sabe-se que, ao adicionar variáveis ao modelo, a variância do erro de previsão pode aumentar, podendo haver uma tendência ao sobreajuste (*overfitting*), mesmo que o modelo apresente um melhor ajuste aos dados de calibração. Com isso, outros critérios de avaliação de modelo têm sido sugeridos com o intuito de penalizar modelos com um maior número de variáveis (Greene, 2002).

Três critérios foram revisados para avaliação dos modelos: R² ajustado, critério de informação de *Akaike* (*AIC*) e critério de informação de *Bayesian* (*BIC*). Os três critérios são funções do R², do tamanho da amostra e do número de parâmetros ajustados. Entretanto o critério *BIC* penaliza mais fortemente modelos com mais parâmetros. Os cálculos de cada critério são realizados como:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K}\left(\frac{e'e}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right) \tag{2.27}$$

$$AIC(K) = s_y^2(1-R^2)e^{\frac{2K}{n}} = \log\left(\frac{e'e}{n}\right) + \frac{2K}{n} \tag{2.28}$$

$$BIC(K) = s_y^2(1-R^2)n^{\frac{K}{n}} = \log\left(\frac{e'e}{n}\right) + \frac{K\log n}{n} \tag{2.29}$$

onde *e'e* é a soma dos quadrados dos erros, *K* é o número de parâmetros e *n* é o número de amostras. Vale lembrar que o valor do R² ajustado será menor ou igual ao valor do R² e que quanto mais próximo de 1, melhor. Já os outros dois critérios podem assumir qualquer valor real e quanto menor esse valor, melhor o modelo (Greene, 2002).

2.4.3 Erros Percentuais

Os erros percentuais podem ajudar na análise dos modelos, de modo que se tenha conhecimento da capacidade preditiva desses. O erro absoluto médio percentual (*mean absolut percentual error, MAPE*) indica a média da soma de todos os erros relativos percentuais absolutos, como mostra a equação 2.30. Já o máximo erro relativo percentual indica o valor máximo dos erros relativos percentuais individuais, considerando o valor absoluto deles. A equação 2.31 mostra o cálculo do *Max e%.*

$$MAPE = \frac{100\%}{i} \cdot \sum_i \left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{2.30}$$

$$Max\ e\% = Max\left(100\% \cdot \left|\frac{y_i-\hat{y}_i}{y_i}\right|\right) \tag{2.31}$$

2.5 Manutenção de Inferências

A capacidade preditiva dos *soft sensors* diminui devido às mudanças nos processos das plantas químicas. Portanto faz-se necessário uma manutenção do modelo. A maneira com a qual está sendo proposta o desenvolvimento de inferências baseadas em dados facilita sua manutenção, pois a manutenção implica num desenvolvimento de uma nova inferência. A grande questão é como detectar que o modelo já não está mais adequado ao processo.

Os trabalhos voltados para identificação de *falhas* ou *outliers* para geração de alarmes focam em anomalias das variáveis facilmente medidas no processo (Russell *et al.*, 2000; Garcia-Alvarez e Fuente, 2011; Mansouri *et al.*, 2016; Escobar *et al.*, 2017). Para as inferências, geralmente essas variáveis seriam as variáveis de entrada no modelo. Dessa forma, ao identificar dados que estejam presentes no domínio do modelo, pode-se afirmar que não há confiança na predição, pois o modelo fará uma extrapolação. Além disso, tratando-se de um modelo caixa preta, sabe-se que é uma prática inadequada.

Contudo, há ainda outra forma do modelo se tornar inadequado ao processo. Podem as variáveis de entrada no modelo estar de acordo com o esperado, porém a predição estar

incorreta, isto é, a variável predita não estar de acordo com a real. Isso indica uma alteração no processo, que pode ser justificada, por exemplo, em mudanças no desempenho do catalisador. O problema é que para identificar tal situação, a variável inferida deve ser analisada de outra maneira, seja por analisador em linha, ou com análise laboratorial (Kaneko e Funatsu, 2016).

Mansouri *et al.* (2016) utilizou uma metodologia baseada em *kernelPCA* para detecção de falhas não lineares em processos químicos. Ele propôs um teste *GLRT* (*generalized likelihood ratio test*) baseado em *KPCA*. O problema de detecção de falhas foi abordado para que os dados sejam primeiro modelados usando o método *KPCA* e, em seguida, as falhas são detectadas usando *GLRT*. Os resultados demonstraram a eficácia do método na identificação de falhas não lineares em dois exemplos: um usando dados sintéticos e outro usando dados simulados de um reator *CSTR* (Mansouri *et al.*, 2016).

Godoy *et al.* (2017) usou uma técnica de detecção e diagnóstico de falhas, com base em uma decomposição parcial de mínimos quadrados (*PLS*) das medições de processo *on-line*, para desenvolver uma estratégia de auto-validação aprimorada capaz de confirmar, corrigir ou rejeitar as previsões do *soft sensor*. A eficácia da técnica proposta foi validada por meio de dois exemplos numéricos. Primeiro, um exemplo sintético foi usado para interpretar os fundamentos do método. Em seguida, a técnica foi aplicada a uma simulação do processo industrial de borracha de estireno-butadieno.

Capítulo 3 – **K-rank: Uma Nova Metodologia Para Segregação de Dados**

Nesse capítulo propõe-se um novo algoritmo, batizado de *k-rank,* para segregação de dados com o intuito de melhorar o método *y-rank.* Essa metodologia se baseia na utilização do algoritmo não supervisionado *k-means* para separar amostras similares e, dentro dos grupos formados pelo *k-means*, aplica-se o algoritmo *y-rank* para segregar os dados em subconjuntos de calibração, validação (se necessário), e teste.

Para demonstração da metodologia e comparação com o método *y-rank* convencional, será utilizado um estudo de caso em que há inversão do sinal do ganho estacionário, ou seja, da derivada da curva de soluções estacionárias. Neste capítulo, além da metodologia, se apresenta também os resultados seguindo o mesmo padrão adotado no artigo que foi publicado na *Brazilian Journal of Chemical Engineering* (SANTOS, P. V. J. L. *et al.*, 2019).

K-Means

K-means é um popular algoritmo de divisão em *clusters*. O algoritmo tem o objetivo de particionar um conjunto de dados em *k* grupos baseados na similaridade dos dados; vale lembrar que o valor de *k* é uma entrada para o algoritmo (Shamir *et al.*, 2005; Thalamuthu *et al.*, 2006; Raschka, 2015). Diferente do algoritmo *k-medoid*, no qual os *clusters* são representados por objetos pertecentes ao conjunto de dados, o algoritmo *k-means* representa os *clusters* pelos seus centróides. Esse algoritmo pode ser sumarizado da seguinte forma (Raschka, 2015):

1. Inicialização aleatória de *k* centróides, como centros dos *clusters* iniciais;
2. Assimilar cada ponto com o centróide mais próximo;
3. Mover os centróides para os centros dos pontos que lhe foram atribuídos;
4. Repetir os passos 2 e 3 até que os *clusters* não mudem mais, ou que seja respeitada uma tolerância, ou que seja realizado um número máximo de iterações.

A similaridade entre os pontos é definida como o oposto da distância. A métrica utilizada pelo algoritmo é o quadrado da distância euclidiana entre dois pontos *x* e *y* m-dimensionais. O índice *j* refere-se à coluna do conjunto dos dados, isto é, a Equação 3.1 mede a distância multivariável de 2 pontos (Raschka, 2015).

$$d(x,y)^2 = \sum_{j=1}^{m}(x_j - y_j)^2 = \|x - y\|_2^2 \quad (3.1)$$

Fundamentado na distância euclidiana, o algoritmo *k-means* está compreendido num simples algoritmo de otimização no qual a função de minimização é a soma dos erros quadrados dos pontos para os centroides (do inglês, *sum of squared errors*). $\mu^{(j)}$ representa o centróide do *cluster j*, e se $w^{(i,j)}$= 1, o ponto $x^{(i)}$ pertence ao *cluster*. Caso contrário $w^{(i,j)}$= 0.

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m} w^{(i,j)} \left\|x^{(i)} - \mu^{(j)}\right\|_2^2 \quad (3.2)$$

3.1.1 Silhouette analysis

Silhouette analysis ou análise silhueta é um método que analisa a qualidade dos *clusters*. A análise é feita através do cálculo do coeficiente silhueta (*silhouette coefficient*). O cálculo do coeficiente pode ser compreendido no seguinte algoritmo (Raschka, 2015):

1. Calcular a coesão dos *clusters* $a^{(i)}$ como distância média entre o ponto $x^{(i)}$ e todos os outros pontos do mesmo *cluster*.

2. Calcular a separação dos clusters $b^{(i)}$ como distância média entre o ponto $x^{(i)}$ e todos os outros pontos do *cluster* mais próximo.

3. Calcular o coeficiente silhueta como a diferença entre a coesão e a separação dos *clusters*, dividido pelo valor máximo de um dos dois.

$$S^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)}, a^{(i)}\}} \quad (3.3)$$

Interpretação do coeficiente de silhueta:

- S(i) ≈ 1: O ponto de dados está bem ajustado ao seu próprio cluster e mal ajustado a clusters vizinhos;

- S(i) ≈ 0: O ponto de dados está na fronteira entre dois clusters;

- S(i) ≈ −1: O ponto de dados pode ter sido atribuído ao cluster errado.

Na prática, ao aplicar a análise de silhueta, busca-se uma pontuação média de silhueta mais próxima de 1, indicando uma boa separação entre os clusters. O coeficiente médio de silhueta está compreendido entre -1 e 1. Se o coeficiente for zero, quer dizer que a coesão e a separação são iguais, ou seja, não haverá *clusters* bem separados, mas sim *clusters* desordenados e sobrepostos. Porém, quando o coeficiente se aproxima de 1, a separação assume um elevado valor, enquanto que a coesão assume valor próximo de zero. Nesse caso, os *clusters* estão bem definidos (Wang *et al.*, 2009; Raschka, 2015).

Wang *et al.* (2009), sugeriu, para a sua ferramenta de validação de clusters *CVAP* (*Cluster Validity Analysis Platform*), que um maior valor do índice de silhueta indica um melhor agrupamento dos dados. Em seu trabalho, ele comparou *K-means*, *Hierarchical Clustering* (*HC*), *partitioning around medoids* (*PAM*) e *self-organizing maps* (*SOM*) para agrupar um conjunto de dados de leveduras (Wang *et al.*, 2009).

3.1.2 Y-rank

A metodoloia *y-rank*, apresentada no capítulo anterior, apresenta problemas quando há regiões distintas que podem resultar num mesmo valor de y, ou numa mesma região de valores de *y*. Consequentemente, isso pode tornar falho o algoritmo *y-rank*, visto que, para um mesmo valor de *y*, podem existir soluções em regiões distintas. O exemplo ilustrado na Figura 3.1 pode demonstrar essa falha.

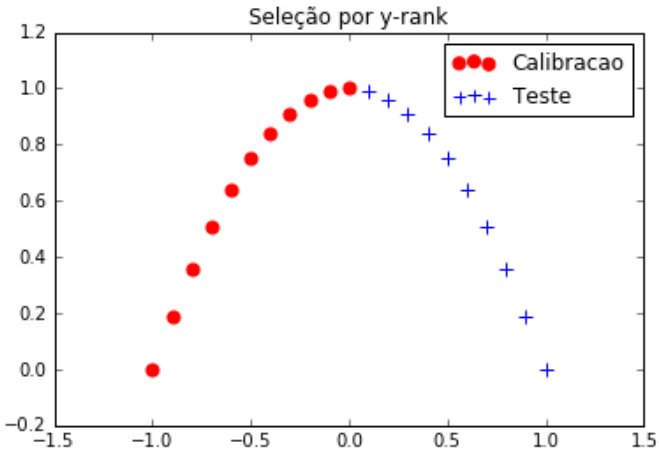

Figura 3.1: Seleção por *y-rank* de dados obtidos de uma função quadrática.

Na Figura 3.1, uma simples função quadrática pode gerar sérios problemas com o algoritmo *y-rank*, pois, nessa situação, o mesmo iria calibrar apenas um lado da curva. Isso ocorreu porque quando o algoritmo dispôs os dados em ordem crescente do vetor de saída, o DNA, isto é, o padrão de escolha, "calibração-teste" fez com que o algoritmo pusesse cada lado da curva para cada subconjunto.

Numa situação real, com a presença de muitas variáveis, essa visualização não seria possível, e muito provavelmente o erro do algoritmo não seria tão drástico como no caso do exemplo acima que foi utilizado para demonstrar didaticamente a falha do algoritmo. Para contornar essa situação, pode-se fazer uma validação cruzada variando o DNA, mas isso implica em um maior tempo computacional.

3.2 Estudo de Caso: Tanque de Aquecimento

Para o estudo em questão, foram utilizados dados de simulação de um sistema hipotético de aquecimento (Figura 3.2), proposto como exemplo motivacional descrito pelo balanço de energia.

$$\frac{dT}{dt} = \frac{Fin}{V}(Tin - T) + \frac{10x}{1+20x^2}(Ta - T) \tag{3.4}$$

onde T é a temperatura do tanque de aquecimento, Fin a vazão de entrada, V o volume do tanque, Tin a temperatura da corrente de entrada, Ta a temperatura do fluido de

aquecimento e *x* a abertura da válvula. As soluções estacionárias do sistema podem ser obtidas por:

$$T_{ss} = \frac{Df \cdot Tin + 20Df \cdot Tin \cdot x^2 + 10 \cdot x \cdot Ta}{Df + 20Df \cdot x^2 + 10x} \tag{3.5}$$

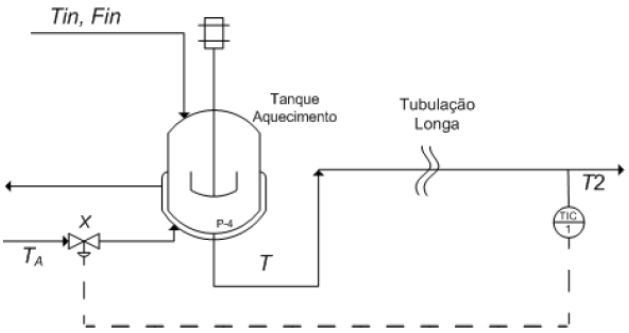

Figura 3.2: Sistema do tanque de aquecimento.

Como exemplo motivador para a demonstração do método, utilizou-se Df = 2 (Df=Fin/V), Tin = 10, Ta = 80. Com esses valores, as soluções da temperatura estacionária do tanque em função da abertura da válvula podem ser visualizadas na Figura 3.3.

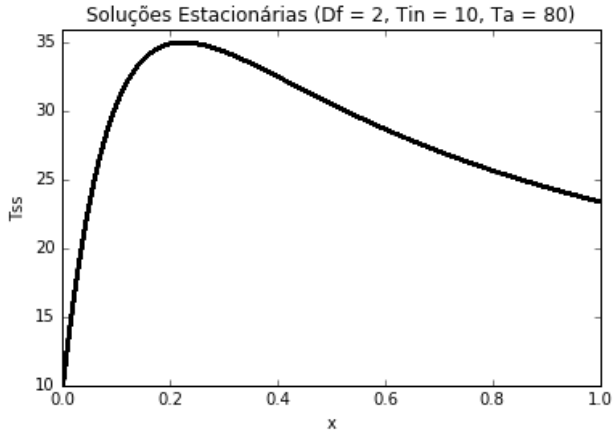

Figura 3.3: Soluções estacionárias do sistema tanque de aquecimento para Df = 2, Tin = 10, Ta = 80.

A partir do gráfico da Figura 3.3, fica visível que, para determinados valores de T_{ss}, i.e., temperatura estacionária, podem existir dois valores distintos de *x* e esses valores pertencem a regiões distintas do gráfico, ou regiões de operação distintas. A ideia por trás do algoritmo *k-rank* consiste na aplicação do *y-rank* após o algoritmo *k-means* separar essas regiões, excluindo a possibilidade do *y-rank* ser injusto com alguma região de operação.

3.3 Desenvolvimento do Método

A metodologia foi desenvolvida em *Python*, fazendo uso principalmente do pacote *scikit-learn*. O pacote é uma biblioteca de aprendizado de máquina de código aberto escrito em *Python* e possui um leque de métodos para classificação, regressão, estimativa da

matriz de covariância, redução de dimensionalidade, pré-processamento dos dados, entre outros (Kramer, 2016).

O resumo da metodologia proposta pode ser visualizado no fluxograma da Figura 3.4. O caminho da esquerda do fluxograma aplica o *y-rank* nos *clusters* gerados pelo *k-means*, ou seja, aplica o método *k-rank*. Em seguida, ele une o que foi alocado para calibração num conjunto de treino, e o que foi alocado para testes noutro. No fim, têm-se conjuntos de treinamentos e testes do mesmo tamanho para ambos os caminhos, porém com uma seleção de dados diferentes. Vale salientar que o algoritmo *y-rank* não está disponível em *Python*, portanto foi necessário implementá-lo.

Foram gerados 21 valores randômicos entre 0 e 1 para a variável x e, utilizando a equação 3.5 para Df = 2, Tin = 10 e Ta = 80, calcularam-se 21 valores para a temperatura do tanque T. Em seguida, os dados foram padronizados de acordo com a equação 3.6. Com isso, tem-se um conjunto de dados que é uma matriz de dimensão (21,2), onde a primeira coluna representa a abertura da válvula x e a segunda coluna representa a temperatura do tanque T.

$$\left(z_i^k\right)_N = \frac{z_i^k}{z_{max}^k} \tag{3.6}$$

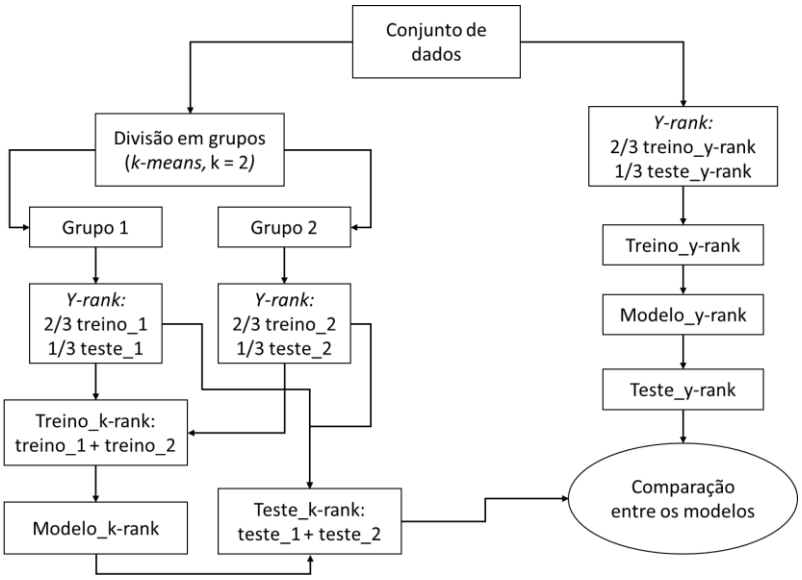

Figura 3.4: Fluxograma simplificado da metodologia proposta (lado esquerdo) comparada ao y-rank tradicional (lado direito).

O próximo passo foi calcular o número de *clusters* utilizando a análise silhueta. Calculou-se o coeficiente silhueta (equação 3.3) variando o número de *clusters* de 2 a 7 e atribuiu-se, ao valor de k, aquele com o coeficiente mais próximo de 1. O número máximo de clusters igual à 7 está baseado na quantidade de pontos. Embora não haja a garantia de que todos os *clusters* terão pelo menos 3 pontos, se fosse utilizado um número máximo de 8 *clusters*, teriam pelo menos 2 *clusters* com 1 ou 2 pontos. Concluiu-se então que não era necessário calcular o coeficiente de silhueta para 8 ou mais *clusters*.

Tendo determinado o número de *clusters*, resta agora gerar os modelos usando a seleção dos dados do *y-rank* aplicado ao conjunto de dados inicial e do *y-rank* aplicado nos *clusters* gerados pela função *KMeans*, disponível no *scikit-learn*. Vale lembrar que se trata de um modelo não linear. Logo, pela simplicidade e eficácia para o estudo em questão,

optou-se por expandir a variável *x* em um polinômio de quarta ordem (equação 3.7). Dessa forma, foi possível ajustar 5 parâmetros com regressão linear, e adequá-los à curva T(x).

$$\text{novo\_x} = [1, x, x^2, x^3, x^4] \tag{3.7}$$

$$T(x) = \alpha + \beta x + \gamma x^2 + \rho x^3 + \varphi x^4. \tag{3.8}$$

Com os modelos gerados, é necessário compará-los. Tal comparação será analisada levando em consideração principalmente a capacidade de estimar novos dados, já que com os dados de treinamento a regressão linear força a minimizar o erro da curva ajustada em relação aos dados de calibração utilizados. Portanto o interesse maior está em obter um modelo capaz de generalizar novos dados.

Para a comparação entre os métodos, foi feito um loop de 10 mil repetições, gerando 10 mil conjuntos de dados distintos e 10 mil modelos distintos para cada método. Para comparar os modelos, foram utilizadas três métricas difundidas na literatura: *MAPE* (equação 2.30), R^2 (equação 2.25) e *RMSE* (equação 2.26) (Greene, 2002).

Uma segunda análise também será feita: a incidência de pontos preditos discrepantes. Ao analisar tal incidência, podem-se descartar modelos que não ajustaram bem alguma região da curva, comparando os valores preditos com os valores reais no conjunto de testes. Tal análise foi feita ponto a ponto analisando o erro percentual relativo de acordo com a Equação 3.9.

$$e\% = \left|\frac{y_{predito} - y_{real}}{y_{real}}\right| \cdot 100\% \tag{3.9}$$

3.4 Resultados e Discussões

Foram gerados pontos da curva, de acordo com a metodologia supracitada, dentro de um loop de repetição para avaliar a análise silhueta e verificar o número de *clusters* que melhor agrupa os dados em questão. Percebeu-se que houve uma variação desse número dentro das possibilidades de 2 a 7, que dependia da disposição dos dados gerados. A Figura 3.5 mostra dois casos extremos em que, numa simulação do estudo de caso, a divisão em 2 *clusters* seria a ótima, com S(2) = 0,78. Já numa outra simulação, 7 *clusters* seria considerado o ótimo, com S(7) = 0,73, como mostra a Figura 3.6.

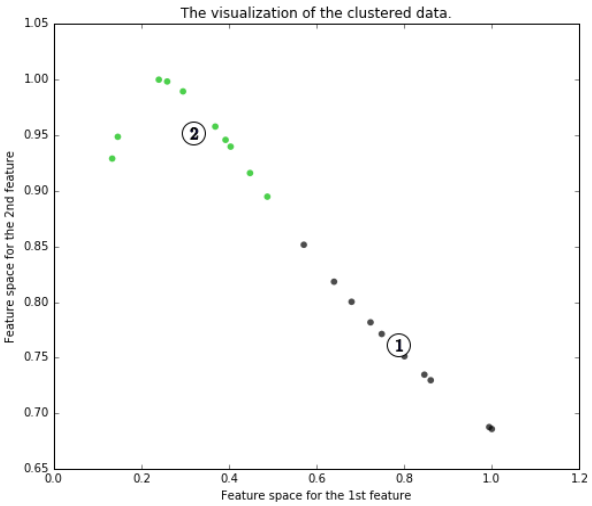

Figura 3.5: *Silhouette analysis*: $k_{ótimo}$ = 2, S(2) = 0,78.

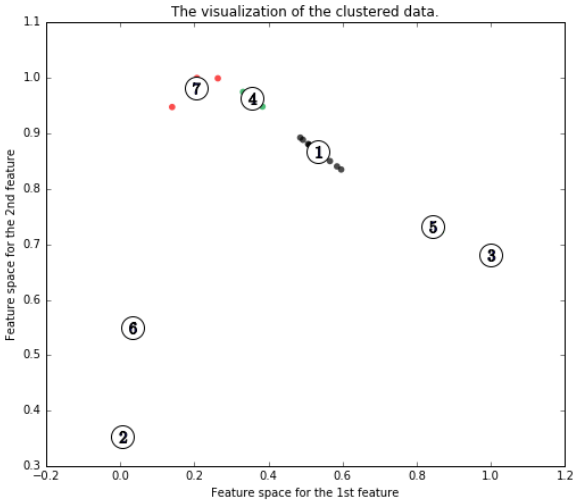

Figura 3.6: *Silhouette analysis*: $k_{ótimo}$ = 7, S(7) = 0,73.

Os clusters mostrados na Figura 3.6 e enfatizados pelos seus centros dificultam a utilização do *y-rank*, pois há clusters com um ponto apenas. Isso não seria problema, para o algoritmo em si, mas sim para a comparação que este trabalho propôs. Pois, neste caso, haveria menos pontos para testes que não foram usados na calibração do que se fosse utilizado o *y-rank* comum. Então para evitar essas desigualdades, foram realizados testes apenas com k = 2, mesmo que este não fosse o ótimo da análise silhueta. Vale ressaltar que o coeficiente de silhueta sugere qual é o melhor valor para k, mas não impede que seja utilizado outro valor.

3.4.1 Demonstração dos resultados em um exemplo

Foram gerados 21 pontos aleatoriamente e utilizou-se k = 2 na divisão dos *clusters*, isto é, os dados foram divididos em 2 *clusters*. Em seguida, aplicou-se o *y-rank* nos *clusters* e nos dados gerados. A seleção dos dados pelos dois métodos pode ser visualizada nas Figuras 3.7 e 3.8. Os pontos em azul são os dados selecionados para teste e os pontos em vermelho para calibração.

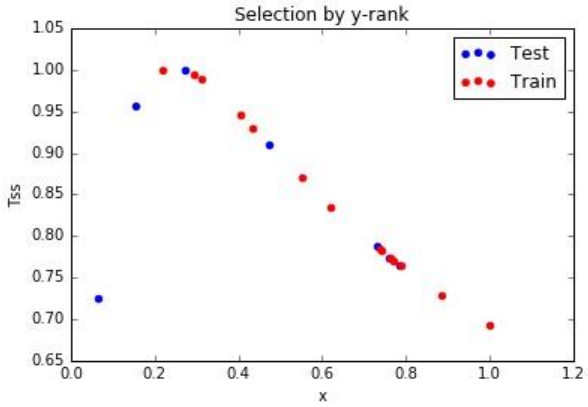

Figura 3.7: Seleção dos dados pelo *y-rank*.

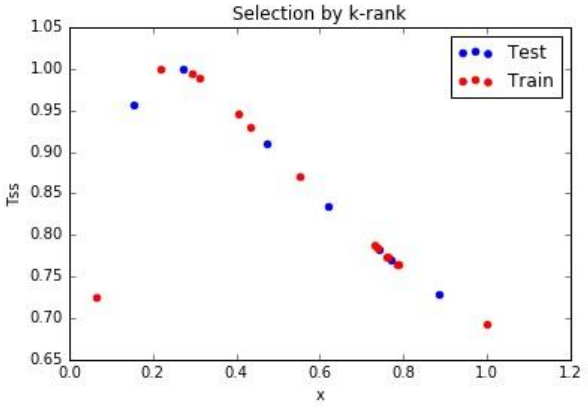

Figura 3.8: Seleção dos dados pelo *k-rank*.

Algumas observações e previsões já podem ser feitas diante das seleções dos dados acima. O *y-rank* escolhe seus dados em ordem crescente do valor de *y*. Pode-se perceber pelo gráfico da figura 3.7 que ele selecionou os primeiros pontos da esquerda para teste, de modo a cometer um erro grave. Ao excluir da calibração o primeiro ponto da esquerda, ele exige que o modelo faça extrapolações ao testar com aqueles dois dados, o que não é viável para modelos empíricos.

Todavia a seleção feita pelo método defendido aqui neste trabalho se mostrou mais interessante. Ocorre que é informado, ao *y-rank*, que o primeiro e o último dado devem ser para calibração, de modo a evitar extrapolações. E, já que o *y-rank* foi aplicado aos *clusters* separadamente, ele garantiu que esses dados seriam usados para o ajuste do modelo.

Em seguida, testaram-se os dois modelos com seus respectivos dados de teste. O modelo feito com o *y-rank* obteve os resultados das Figuras 3.9 e 3.10. Os resultados para o modelo feito com o *k-rank* está ilustrado nas Figuras 3.11 e 3.12. A Figura 3.13 mostra o ajuste de cada um dos modelos e todo o conjunto de dados.

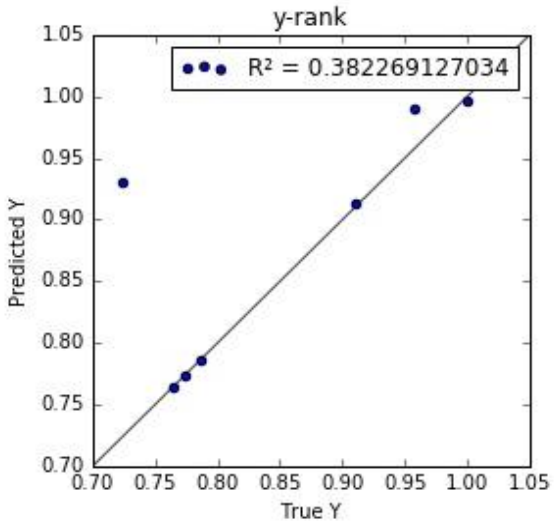

Figura 3.9: Predição pelo modelo gerado com *y-rank*, $R^2 = 0.38$.

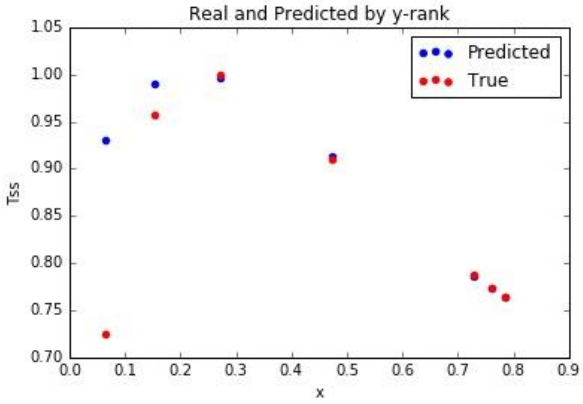

Figura 3.10: *Y-rank*: visualização dos pontos preditos e reais.

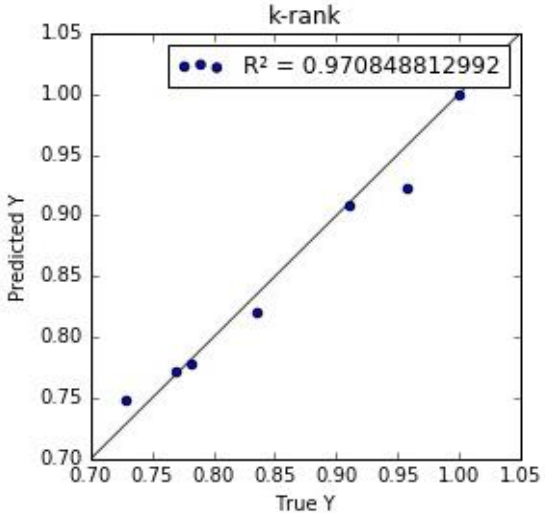

Figura 3.11: Predição pelo modelo gerado com *k-rank*, $R^2 = 0.97$.

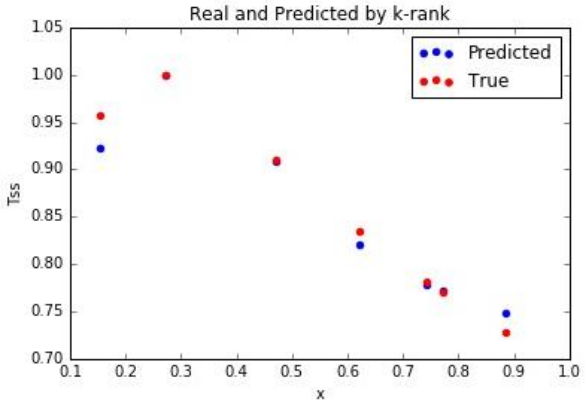

Figura 3.12: *K-rank*: visualização dos pontos preditos e reais.

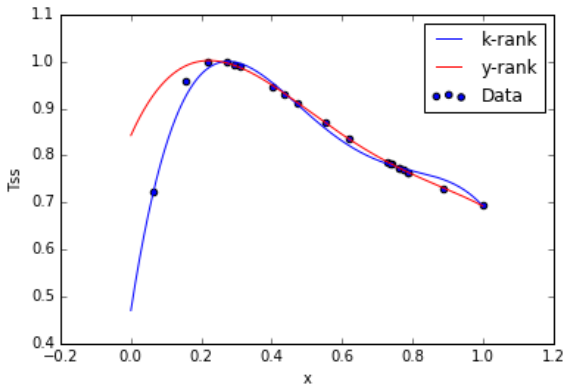

Figura 3.13: Aderência dos modelos aos dados simulados.

Diante das curvas ajustadas, é possível perceber a falha do modelo produzido pelo y-rank aplicado ao conjunto inicial de dados. A falha está localizada justamente onde ele não selecionou dados para a calibração, como já era esperado.

3.4.2 Comparação entre os métodos para um loop de 10 mil repetições

Para fazer uma comparação e analisar o ganho em utilizar o método defendido pelo trabalho, foi feito um loop de 10 mil repetições em que cada iteração era informado um novo conjunto de dados e ao final, comparado os dois modelos, de acordo com a metodologia apresentada. Somou-se uma unidade a uma variável contadora para o modelo que obtivesse os melhores resultados nas três métricas (*MAPE, R^2, RMSE*) para o conjunto de teste. Se houvesse pelo menos uma das métricas escolhendo um modelo enquanto as outras optassem pelo outro, era considerado um empate entre os modelos. O resultado pode ser visto na tabela 3.1:

Tabela 3.1:Resultado da disputa entre os métodos.

Total	*y-rank*	*k-rank*	*Empate*
10000	1764	5912	2324

Em seguida, foram analisadas as falhas das seleções diante de situações de multiplicidade de soluções. Então para avaliar a ocorrência de modelos descartáveis, foi considerado um erro máximo de 15% para cada valor predito pelos modelos. Caso o modelo estimasse pelo menos um valor com uma diferença igual ou maior do que 15% do valor real, de acordo com a equação 3.7, o modelo era contabilizado como descarte.

O valor de 15% foi atribuído de maneira que não se descartassem modelos pela incapacidade de a curva aderir perfeitamente aos dados. Entretanto com um erro superior a 15%, notou-se que este era causado pela má seleção de dados, perdendo informações de regiões importantes da curva. Com isso, foi feito um loop de 10 mil repetições e, utilizando o critério para o descarte, foram contabilizados os modelos descartáveis e anotou-se na tabela 3.2.

Tabela 3.2: Descartes de modelos para o loop de 10 mil repetições.

Total	*Modelos descartados pelo y-rank*	*Modelos descartados pelo k-rank*	*Modelos descartados em comum*
10000	1187	17	12

Pode-se concluir a partir desses resultados que houve uma incidência de aproximadamente 12% de modelos descartáveis pelo *y-rank*. Isto aconteceu porque o *y-rank* perdeu informações por selecionar mal os dados para a calibração. Enquanto que apenas 0,17% dos modelos foram descartados pelo *k-rank*, mostrando uma superioridade na segregação de dados com multiplicidade de soluções. Ainda dos 17 modelos descartados pelo *k-rank*, 12 foram descartados em comum com o *y-rank*.

3.5 Conclusões

Os resultados mostraram que o *y-rank* aplicado a um conjunto de dados em que há multiplicidade de soluções pode haver perdas de informações ao selecionar conjuntos com uma distribuição injusta de dados. Essa má distribuição pode tornar o modelo ruim em determinadas regiões da função, geralmente num dos extremos da curva, para o caso em questão. No exemplo mostrado, o *y-rank* falhou no extremo esquerdo da curva, mas em outras ocasiões, o mesmo ocorreu para o extremo direito. Isso porque o *y-rank* garante que o primeiro ponto (menor valor de *y*) e o último ponto (maior valor de *y*) serão destinados à calibração. Porém, neste caso, os pontos num dos extremos da curva podem assumir valores intermediários de *y*, deixando o algoritmo *y-rank* vulnerável.

Diante de tal situação, o *k-means* se mostrou eficiente em dividir o conjunto de dados em clusters, melhorando a qualidade da seleção dos dados pelo *y-rank*. Vale ressaltar que

o *k-means* pode ser usado em funções multivariáveis, mas é um algoritmo sensível ao escalonamento destas variáveis. Logo é essencial a padronização dos dados antes da aplicação da técnica.

Os resultados mostraram uma superioridade do método *k-rank* em ambas as comparações. Ao comparar o que obteve o melhor ajuste para novos dados (conjunto de testes) usando três métricas diferentes (*MAPE, RMSE*, R^2), os modelos produzidos com a seleção de *k-rank* foram superiores em 59,12% dos casos. 23,24% foram considerados empates e 17,64% ganham o *y-rank*. No entanto, o resultado a destacar é o descarte de modelos discrepantes. Apenas 17 modelos de 10.000 foram descartados pela seleção do *k-rank*, enquanto que para a seleção do *y-rank*, 1187 foram descartados.

Outro fator interessante é que, caso houvesse um mau agrupamento dos *clusters* pelo *k-means*, o *y-rank* ainda estaria sendo aplicado de maneira semelhante ao método convencional. Todavia a análise silhueta se mostrou eficaz em qualificar os clusters.

Capítulo 4 – **Metodologia para o Desenvolvimento e Manutenção de Inferências**

Este capítulo apresenta a metodologia proposta para desenvolvimento e manutenção de inferências. Toda a metodologia foi implementada em *Python versão 3.5.1.2,* utilizando um computador *intel core™ i5*com 6GB de memória RAM. O esquema da Figura 4.1 mostra as etapas que farão parte da metodologia e em seguida cada uma será detalhada.

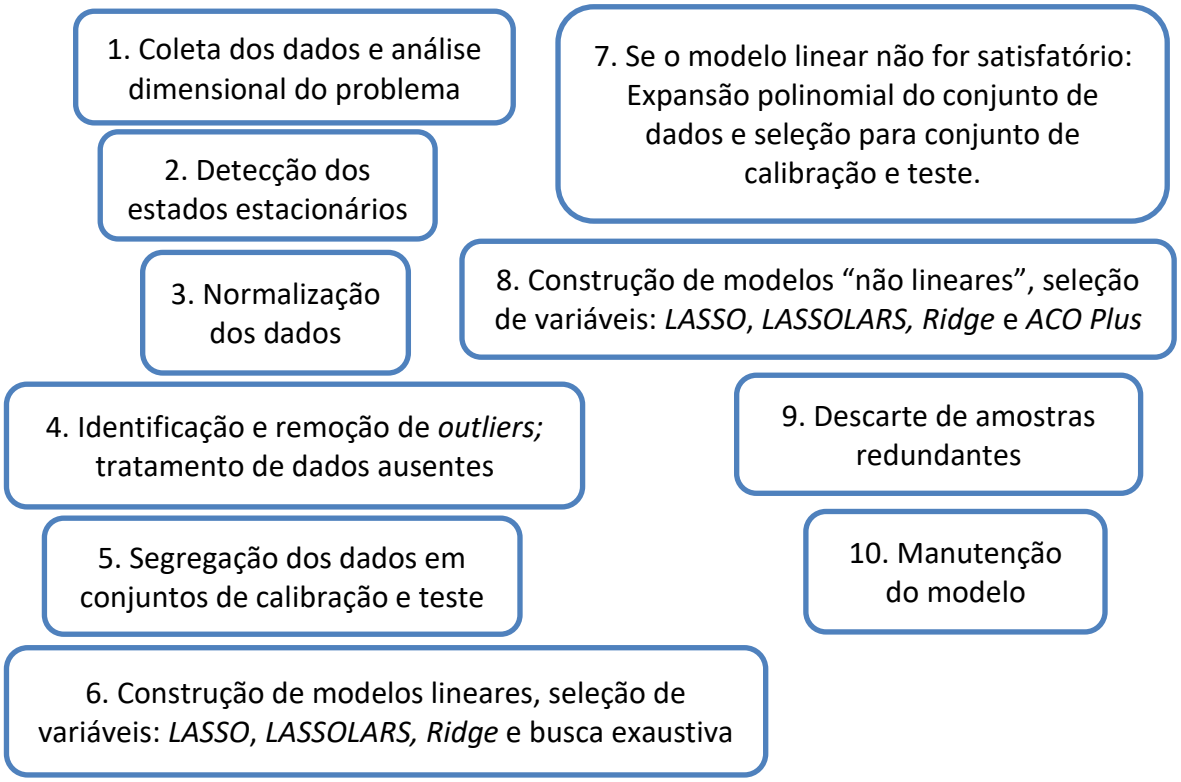

Figura 4.1: Esquema simplificado da metodologia proposta para o desenvolvimento e manutenção de inferências.

4.1 Análise da Dimensionalidade do Problema

De posse dos dados, pode-se fazer uma análise da dimensionalidade do problema, estimar o número de variáveis que explicam os dados, a partir de uma análise dos componentes principais. Os métodos *PCA* e *Kernel-PCA*, apresentados no capítulo 3, são úteis neste sentido e podem ser utilizados em situações linearmente separáveis (*PCA* ou *KPCA*) ou situações não linearmente separáveis (*KPCA*).

Esta etapa serve apenas como um indicativo da quantidade de variáveis necessárias ao modelo. Não necessariamente será encontrado o número ideal de variáveis, mas já se pode ter uma noção de busca quando se fizer uma opção pelo *ACO Plus* como seletor de variáveis. O *ACO Plus* consiste na união entre o *ACO*, utilizado por Ranzan et al. (2014), e o *LASSOLARS* e será explicada na seção 4.4 deste capítulo. Além disso, é possível uma visualização prévia dos dados, através dos componentes principais.

4.2 Pré-processamento dos Dados

A metodologia se inicia realmente no pré-processamento dos dados. Nesta primeira etapa, é ideal que se faça uma pré-análise dos dados obtidos para que uma limpeza prévia seja realizada. Em dados de processo, podem ocorrer situações de congelamento de variáveis, ou valores fisicamente impossíveis, como, por exemplo, vazões negativas. Então é interessante remover esses valores, através de processos lógicos simples, pois eles não ajudam no desenvolvimento da inferência.

Em seguida, inicia-se o tratamento dos dados com a identificação dos estados estacionários, normalização dos dados, tratamento de dados ausentes, identificação e remoção de *outliers*, e, também, pode-se fazer um descarte de dados redundantes com um processamento lógico simples. Cada etapa do tratamento é atribuída quando realmente necessária. Por exemplo, numa simulação estacionária não precisa identificar os estados estacionários.

A etapa de identificação de estados estacionários não será mostrada, pois os dados utilizados para demonstração dos resultados foram obtidos de uma simulação estática. Todavia, para a metodologia proposta, atribui-se o método de identificação por linha de tendência, que é um método simples e direto. Em cada amostragem, usa-se regressão linear para determinar a melhor linha de tendência linear para os últimos N pontos de dados. Se o processo estiver em estado estacionário, então a inclinação da linha de tendência será bem próxima de zero (Rhinehart, 2013).

Em seguida, normalizam-se os dados através do método *z-normalization,* que atribui, aos dados, média zero e desvio padrão unitário. Esse método é bastante utilizado em dados de processo, mas isso não impede que outras normalizações sejam testadas, desde que não mude as informações contidas nos dados.

Para a identificação de *outliers* será utilizado a estatística de Hotelling, por se tratar de uma estatística multivariável. Além disso, trata-se de um método bastante difundido na literatura. Já a remoção de *outliers* e de dados ausentes pode ser feita através de um processamento lógico simples, se essa for a melhor opção. Se caso for necessário estimar

os valores ausentes, pode-se usar diferentes técnicas de interpolação para estimar os valores faltantes das outras amostras do conjunto de dados.

4.3 Segregação dos Dados

Para a seleção dos dados e separação nos conjuntos de calibração, validação e testes, ou apenas calibração e testes, o método *k-rank* se mostrou superior ao método y-rank, como anteriormente apresentado, então este será adotado.

Todavia, em casos em que o tempo computacional não é limitante, e se houver necessidade, pode-se realizar uma validação cruzada usando o método k-fold, já mencionado na revisão bibliográfica.

4.4 ACO Plus: A União entre o ACO e o LASSOLARS

A versão do *ACO* implementada neste estudo é uma modificação da utilizada por Ranzan et al. (2014), que se baseia na evolução do rastro de feromônios durante a varredura das variáveis. Inicialmente, todas as variáveis são marcadas com a mesma concentração de feromônio. A rotina *ACO* seleciona variáveis aleatoriamente para compor um grupo que é avaliado usando a função objetivo para a predição da variável do processo. Com base no erro de predição, a concentração de feromônio, associada a cada variável, é atualizada. Para a seleção subsequente, a seleção aleatória escolhe variáveis que associam o mesmo gatilho aleatório e uma densidade cumulativa de feromônio para toda a gama de variáveis. Esta associação traz evidências de elementos significativos dentro da região de busca e, após algumas corridas iterativas, um perfil de feromônio é estabelecido e as regiões com alta densidade de feromônio destacam as variáveis mais significativas para a predição da variável do processo.

Dentro do *ACO Plus*, o algoritmo utilizado para a calibração dos modelos utiliza a metodologia *Lars*, porém na função objetivo utiliza-se a penalização da metodologia *LASSO*. Então, a cada iteração, o algoritmo inclui no modelo a variável mais correlacionada com o vetor residual, isto é, a variável que gera o menor ângulo com o residual. Porém se esta variável não servir ao modelo, ela já pode ser descartada pela penalização incluída na função objetivo. Em seguida, dentre os modelos gerados, é escolhido aquele que tiver o menor valor do critério *BIC*. Esse algoritmo pode ser encontrado na função *LassoLarsIC*, que é uma ferramenta disponível no *Scikit Learn* para criação de modelos. Nessa função pode ser utilizado também o critério *AIC*, mas, por levar a maiores reduções no número de parâmetros ajustados, optou-se pelo critério *BIC*, já que a intenção é zerar as variáveis dispensáveis para melhorar a trilha.

Dessa forma, essa nova versão do *ACO* assume o compromisso entre uma boa trilha de feromônio e um baixo tempo computacional, visto que não será adicionado feromônio nas variáveis que a função *LassoLarsIC* atribuir o valor zero aos coeficientes. Assim a trilha terá mais influência na busca das variáveis, pois as variáveis descartáveis dificilmente irão receber feromônio a cada iteração.

4.5 Seleção de Variáveis: Construindo Modelos

Nesta etapa serão construídos vários modelos, utilizando metodologias diferentes. Como não se sabe previamente se os dados se ajustarão melhor em modelos lineares ou não lineares, o ideal é que se comece com modelos mais simples (lineares), e se estes não

forem satisfatórios, inicia-se a busca por modelos não lineares. Esses, por sua vez, serão fundamentados nas variáveis disponíveis e em expansões polinomiais delas. Este método é simples e de baixo custo computacional.

Independentemente de se estar buscando um modelo linear ou não linear, a construção deles já se dará por métodos que evitam *overfitting,* apresentados na revisão bibliográfica. São eles: *Ridge Regression, LASSO* e *LARS.* Para a metodologia *LARS*, será utilizado o mesmo algoritmo de regressão utilizado no *ACO Plus*: *LassoLarsIC.* Por conta disso, nos resultados o método será chamado de *LASSOLARS.* O *ACO* proposto neste trabalho também será utilizado para a construção dos modelos. Para a regressão *LASSO* e *Ridge*, serão utilizadas as funções *LassoCV* e *RidgeCV* do *Scikit Learn*. Essas funções fazem um *cross validation* com os parâmetros associados às penalizações L1 e L2.

Por fim, todos os modelos serão postos a testes para que se avalie o melhor modelo criado. Serão analisados alguns critérios de avaliação de modelo, já relatados anteriormente. A comparação não será baseada apenas nos erros de predição, mas também no tamanho do modelo, pois, do ponto de vista prático, quanto mais simples for o modelo, menos problemas podem surgir por conta de medições errôneas de instrumentos mal calibrados, ou até mesmo falha desses instrumentos.

4.6 Descarte de Amostras Redundantes

Quando se analisa o conjunto de dados por completo, muitas vezes não é possível identificar que existem amostras redundantes naquele conjunto, visto que podem estar presentes variáveis que não interferem nas variáveis do modelo, nem nas variáveis de entrada, nem na variável de saída. Entretanto, após a seleção das variáveis, já com o modelo criado, o conjunto de dados diminui em termos de variáveis, mas se mantém na quantidade das amostras. Nesse momento, agora tratando-se apenas das variáveis envolvidas no modelo, pode-se ter amostras redundantes, isto é, valores iguais ou praticamente iguais para todas as variáveis (as de entrada e a de saída). Entretanto, isso não é bom para o desenvolvimento do modelo.

Os modelos baseados em regressão forçam a redução do erro entre o valor predito e o valor real da variável a ser modelada. Com isso, na presença de amostras redundantes, desloca-se o modelo para próximo delas, pois assim o erro total será menor. Então uma boa prática será a redução no número dessas amostras redundantes na calibração do modelo para que se tenha uma melhor aproximação da realidade, e não um menor erro em relação a essas amostras redundantes.

Então, após a seleção do modelo, serão descartadas as amostras redundantes para a calibração de um novo modelo apenas com as amostras significativas. De modo que, o modelo será mais representativo em relação a todos os dados, e não em relação às amostras redundantes. O descarte será feito de acordo com o algoritmo a seguir:

1. Com os dados normalizados com média zero e desvio padrão unitário, arredonda-se os valores de todos os dados da matriz para uma casa decimal;

2. Descartam-se as linhas em que todos os valores (de todas as variáveis) se repetem;

3. Coleta, da matriz original sem arredondamento, apenas uma amostra para cada grupo de amostras redundantes, isto é, apenas uma amostra irá representar o grupo de amostras que foram descartadas;

4. Calibra-se o modelo para os dados armazenados na nova matriz, testa-o com todos os outros dados que foram descartados e analisa sua qualidade. Se o modelo for representativo para todos os dados, adota-o. Se não for, volta ao passo 1 e aumenta o número de casas decimais com o intuito de descartar menos amostras e consequentemente conseguir com que o modelo represente melhor todos os dados.

4.7 Manutenção do Modelo

Para a manutenção do modelo gerado, deve-se ater às duas possíveis situações. A primeira situação, onde ocorre uma variação na região de operação da planta, a metodologia utilizada será com a finalidade de identificar os dados anômalos utilizando a estatística T^2 de Hotelling. Dessa forma, sabe-se que o modelo não irá funcionar bem para os novos dados que não se enquadrarem aos dados utilizados na calibração da inferência.

Nesse caso, será necessário averiguar se foi apenas uma situação transitória da planta, ou se é uma região de operação não identificada quando se desenvolveu a inferência. De modo que, pode-se então coletar dados dessa região e desenvolver um novo modelo. Se a planta operar em regiões distintas constantemente, é possível ainda gerar mais de um modelo para se trabalhar com modelos locais. A metodologia da geração dos modelos, nesse caso, teria que separar os dados previamente por métodos de agrupamento ou *clustering analysis.* O método *k-means*, relatado no capítulo 3, pode ser usado para esse fim, ou ainda pode-se separar por métodos supervisionados, caso o *k-means* não consiga separar as regiões satisfatoriamente.

A segunda possível situação traz limitações para os métodos estatísticos. Pois trata-se de casos em que a variável estimada não condiz com a variável real. Isto é, os valores das variáveis de entrada estão dentro da região de calibração do modelo, mas o valor real da variável de saída sofreu alteração, indicando uma alteração no processo, que pode ser justificada, por exemplo, em mudanças no desempenho do catalisador. Nessa situação, a identificação da falha se dará por métodos laboratoriais ou analisadores em linha. E novamente, poderá se desenvolver um novo modelo, baseado em novos dados com as mudanças da planta.

Capítulo 5 – **Estudo de Caso: Unidade de Separação de Propeno/Propano**

Com o intuito de sistematizar a metodologia e demonstrar sua eficácia, foi escolhido um estudo de caso de uma unidade de separação de propeno. Esta unidade foi desenvolvida baseada numa unidade real em operação (Schultz, 2015). Schultz (2015) simulou esta unidade em Aspen Plus e a utilizou para estudar tétnicas de *SOC* (*self-optimizing control*). A seguir será descrita a unidade e como foi feita a simulação dos dados.

5.1 Unidade de Separação de Propeno/Propano

A unidade possui o objetivo de produzir uma corrente de propeno (C3-) com elevado grau de pureza (99,6%) a partir de uma corrente de gás liquefeito de petróleo (GLP). O processamento é realizado por três colunas de destilação em série, sendo o GLP alimentado na primeira coluna (T-01). Nesta coluna, os compostos pesados (C4+) são removidos pelo fundo, enquanto que a corrente de topo, rica em propeno, alimenta a segunda coluna (T-02). Esta coluna extrai, pelo topo, uma corrente rica em etano (C2) e, pelo fundo, a corrente rica em propano (C3+) e propeno (C3-) que irá alimentar a terceira coluna (T-03). Nesta última, por sua vez, a corrente rica em propeno é extraída pelo topo. A T-03 utiliza uma bomba de calor, onde a corrente de topo é utilizada como fluido de aquecimento do refervedor após passar por uma etapa de compressão. O fluxograma simplificado do processo é apresentado na Figura 5.1, e a terminologia utilizada é apresentada na Tabela 5.1. A Tabela 5.2 discrimina os equipamentos da unidade.

Tabela 5.1: Lista de equipamentos da unidade (Schultz, 2015).

Representação	*Descrição*
	Limite de bateria – Alimentação da unidade
	Limite de Bateria – Correntes de produtos
	Número para identificação da corrente
AR	Água de resfriamento
CB	Condensado de baixa pressão
DR	Sistema de drenagem
VB	Vapor de baixa pressão

Tabela 5.2: Lista de equipamentos da unidade (Schultz, 2015).

Nome	**Descrição**
B-01	Bomba de refluxo da coluna T-01
B-02	Bomba de alimentação da coluna T-02
B-03	Bomba de refluxo da coluna T-02
B-04	Bomba de produto de fundo da coluna T-03
C-01	Compressor da corrente de topo da coluna T-03
P-01	Condensador da coluna T-01
P-02	Refervedor da coluna T-01
P-03	Condensador da coluna T-02
P-04	Refervedor da coluna T-02
P-05	Refervedor da coluna T-03
P-06	Condensador da coluna T-03
P-07	Resfriador da corrente de produto especificado
T-01	Coluna de destilação para remoção de C4+
T-02	Coluna de destilação para remoção de etano
T-03	Coluna de destilação para separação propeno/propano
V-01	Vaso de acúmulo de condensado da T-01
V-02	Vaso de acúmulo de produto de topo da T-01
V-03	Vaso de acúmulo de condensado da T-02
V-04	Vaso de acúmulo de condensado da T-03

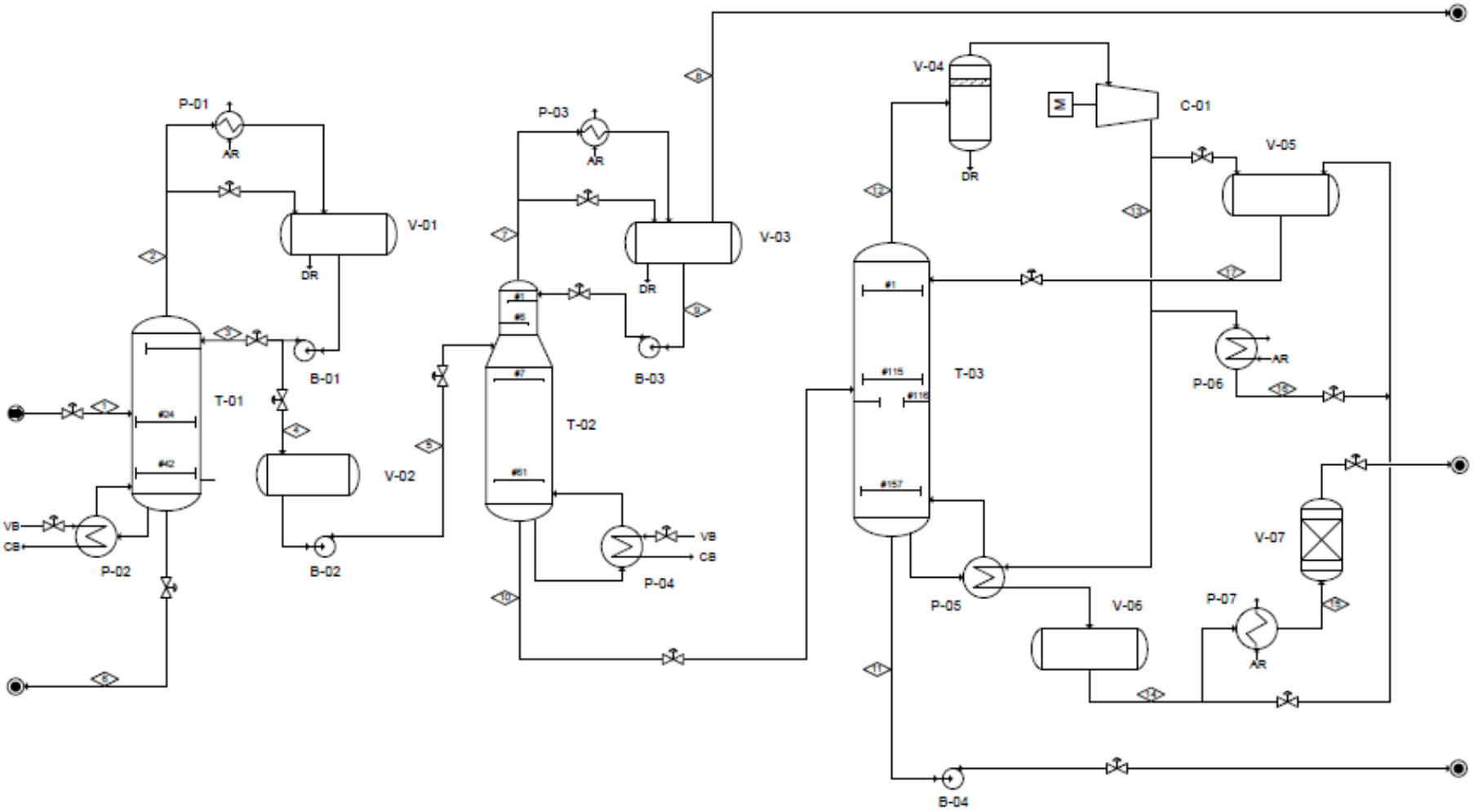

Figura 5.1: Fluxograma simplificado da uniade de separação de propeno (Schultz, 2015).

Os dados construtivos das colunas foram omitidos por Schultz (2016) e Haykin (2001) devido ao sigilo industrial dessas informações. A alimentação da unidade é composta por uma corrente de GLP, cujas especificações são apresentadas na Tabela 5.3.

Tabela 5.3: Especificação da corrente de alimentação, composta por GLP(Schultz, 2015).

Especificação	*Valor*
Vazão (kmol/h)	1296,84
Vazão (kg/h)	63000
Temperatura (°C)	66
Pressão (kgf/cm²g)	17,9
Composição	Valor molar (kmol/h)
Água	1,90
Etano	44,33
Propano	127,75
Propeno	531,20
Isobutano	119,65
Isobuteno	155,35
1-Buteno	83,72
1-3 Butadieno	3,83
Butano	40,82
trans-2-Buteno	103,9
cis-2-Buteno	77,11
Isopentano	3,44
n-Pentano	2,79
Hexano	1,05

5.2 Modelagem da Unidade

O modelo estacionário da unidade simulado em Aspen Plus versão 7.2 pode ser visto na Figura 5.2. Na modelagem da unidade, Schultz (2015) fez as seguintes considerações:

- Cada coluna de destilação possui dois graus de liberdade estacionários;
- O distúrbio da unidade consiste na corrente de alimentação com variação na vazão de entrada e na concentração de propeno, sendo que a pressão e a temperatura são controladas;
- A pressão de topo de cada coluna é mantida constante;
- A temperatura de entrada de cada coluna é mantida controlada;
- Os vasos do processo não foram simulados, visto que não influenciam no resultado da simulação estacionária;
- Foi utilizado o modelo termodinâmico de Peng-Robinson para calcular as propriedades físico-químicas das correntes, devido às correntes serem compostas por hidrocarbonetos;
- Foram desconsideradas as perdas de carga das tubulações do processo;
- A alimentação de cada coluna possui pressão constante, controlada por uma válvula, sendo que esta foi modelada de forma a fornecer uma pressão de saída especificada;
- Os trocadores de calor, com exceção do P-05, foram modelados apenas para o cálculo da troca térmica necessária, desconsiderando os limites mecânicos dos equipamentos;
- O trocador P-05 foi modelado como um casco e tubo com coeficiente global de transferência de calor constante no valor de 932 kcal/(h·m²·°C) e área total de 2168 m²;
- O compressor foi considerado isentrópico, calculando a energia requerida para manter uma pressão de descarga especificada.

Os graus de liberdade utilizados na modelagem de cada coluna são descritos a seguir:

- Coluna T-01: Razão de refluxo (RR1) e razão mássica entre a vazão de destilado (corrente 4) e a vazão de entrada (corrente 1), sendo que essa nova variável será chamada de D/F1.
- Coluna T-02: Razão de refluxo (RR2) e a razão mássica entre a vazão de fundo (corrente 10) e a vazão de entrada da coluna (corrente 5), sendo que essa nova variável será chamada de B/F2.

- Coluna T-03: fração da corrente que sai do compressor que será utilizada como fluido de aquecimento do refervedor, que será chamada de FA3. Além disso, será utilizada a fração da corrente que sai do refervedor (corrente 14) que retorna para a coluna como refluxo para a coluna, que será chamada de FR3.

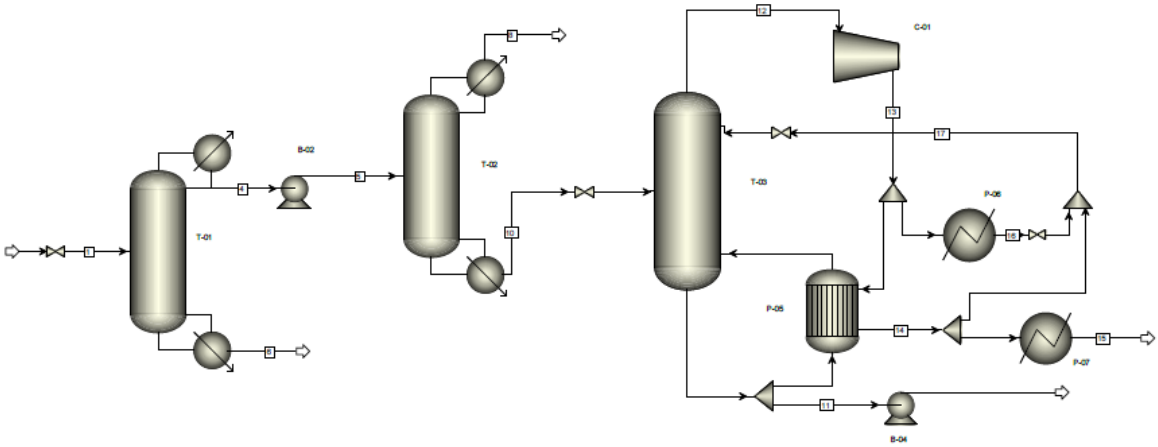

Figura 5.2: Modelo da unidade criado no Aspen Plus(Schultz, 2015).

As variáveis calculadas e o diagrama de entradas e saídas do modelo de cada coluna são apresentados na Figura 5.3.

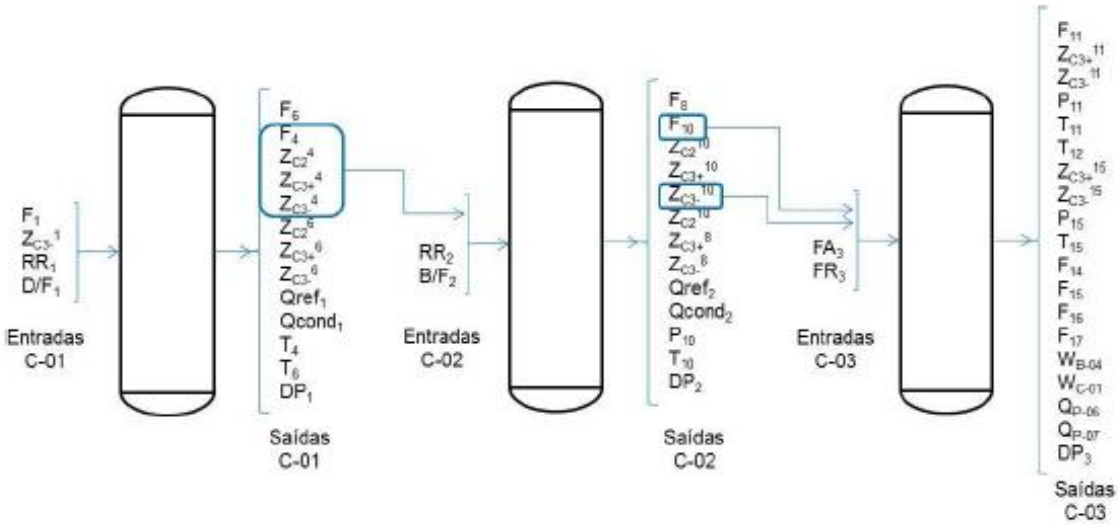

Figura 5.3: Modelo de entradas e saídas do processo (Schultz, 2015).

Para facilitar a representação das variáveis do processo, por haver um grande número dessas, foi criada uma notação para representá-las, conforme descrito na Tabela 5.4. Já na Tabela 5.5 se descrevem as principais correntes do processo, de modo a facilitar a compreensão.

Tabela 5.4: Notação das variáveis do processo.

Notação	*Descrição*
DP_x	Diferença de pressão entre o topo e o fundo da coluna T-0X
F_x	Vazão mássica da corrente x
Fm_x	Vazão molar da corrente x
P_x	Pressão da corrente x
$Qcond_x$	Calor trocado no condensador da coluna T-0X
$Qref_x$	Calor trocado no refervedor da coluna T-0X
Q_x	Calor trocado no trocador P-0X
RR_x	Razão de refluxo da coluna T-0X
T_x	Temperatura da corrente x
W_x	Trabalho realizado no equipamento x
Z_x^y	Fração mássica do componente x na corrente y

Tabela 5.5: Principais correntes da unidade.

Variável	*Corrente*
F_1	Vazão de alimentação da unidade
F_4	Vazão de topo da coluna T-01
F_6	Vazão de fundo da coluna T-01
F_8	Vazão de topo da coluna T-02
F_{10}	Vazão de fundo da coluna T-02
F_{11}	Vazão de fundo da coluna T-03
F_{14}	Vazão de saída do refervedor da coluna T-03
F_{15}	Vazão de propeno produzido na unidade
F_{16}	Vazão de saída do condensador da coluna T-03
F_{17}	Vazão de refluo da coluna T-03

A simulação da unidade possui uma convergência lenta devido à coluna com bomba de calor. Schultz (2016) verificou que em algumas simulações foram necessários mais de cinco minutos para a convergência, utilizando um computador comum com processador Intel® CoreTM i5 e 6GB de memória RAM. Por esse motivo ele optou por utilizar um modelo caixa preta, neste caso redes neuronais, para representar a unidade, visto o grande número de avaliações necessárias e a não-linearidade do sistema (Haykin, 2001).

5.3 Modelo Caixa Preta

Schultz (2016) gerou as redes neuronais a partir de uma análise de sensibilidade com base nos resultados das simulações do simulador. Para isso o modelo foi dividido em três partes, cada uma representando uma coluna. Cada coluna foi simulada separadamente para um conjunto de dados de entrada, uniformemente espaçados, de forma a se obter uma rede neural com validade em toda a região de operação.

As redes neuronais utilizadas foram compostas de duas camadas, sendo que na primeira camada os neurônios utilizados foram do tipo tangente sigmoidal. Na segunda camada foi utilizada a função tangente sigmoidal apenas para as redes relacionadas ao cálculo das concentrações; para as demais foi utilizada uma função linear. Os números de neurônios utilizados em cada rede foram de: 12 para a T-01, 20 para ambas as redes da T-02 e 20 para ambas as redes da T-03.

Foram utilizados mais neurônios nas colunas T-02 e T-03 para melhorar a precisão do modelo, sendo que foram sendo adicionados neurônios até ser obtido um valor do coeficiente de correlação para os dados de validação superior a 0,98. Os dados disponíveis foram divididos de forma randômica, sendo que 75% dos pontos foram utilizados para treinamento e os demais para a validação do modelo.

Para a coluna T-01 foi realizada a variação nos distúrbios da unidade e nos graus de liberdade da coluna (RR_1 e D/F_1). Como distúrbios foram consideradas variações na vazão mássica de alimentação (F_1) e na fração mássica de propeno (Z_{C3-}^{1}), sendo que para a alteração da fração molar do propeno foi mantida a mesma proporção da especificação nominal da corrente para os demais componentes, conforme a Tabela 5.3. Os valores nominais das variáveis, seus limites e o número de pontos utilizados para treinar a rede neural da coluna T-01 são apresentados na Tabela 5.6.

Tabela 5.6:Região utilizada para treinar a rede neural da coluna T-01 (Schultz, 2015).

Varável	*Valor nominal*	*Limites*	*Número de Pontos*
F_1	63000	59850-66150	5
Z_{C3-}	0,355	0,319-0,390	6
RR_1	1,98	1,00-4,00	6
D/F_1	0,459	0,250-0,650	5
Total de Pontos			9000

A mesma metodologia foi utilizada para a geração da rede neuronal para a coluna T-02. Porém, foram utilizadas duas redes para simular o comportamento da coluna ao invés de apenas uma. Foi gerada uma rede para calcular as concentrações das correntes de saída e uma segunda rede para as demais variáveis. Foi necessário realizar essa divisão para melhorar os resultados do modelo, que apresentavam erros significativos nos cálculos das concentrações quando apenas uma rede era utilizada.

Foi utilizado o mesmo conjunto de dados para o treinamento das duas redes da coluna T-02, sendo que esses foram gerados a partir de seis variáveis de entrada do modelo: vazão mássica de entrada da coluna (F_5), fração mássica de eteno na entrada (Z_{C2}^{5}), fração mássica de propano na entrada (Z_{C3+}^{5}), fração mássica de propeno na entrada (Z_{C3-}^{5}), razão de refluxo da coluna (RR_2) e razão entre a vazão de fundo e a vazão de entrada (B/F_2). Os limites de variação da composição de entrada e vazão de alimentação da T-02 foram obtidos a partir dos resultados de saída calculados para a T-01. Os valores nominais dessas variáveis, os limites utilizados e o número de pontos são apresentados na Tabela 5.7.

Tabela 5.7: Região utilizada para treinar a rede neural da coluna T-02 (Schultz, 2015).

Variável	*Valor nominal*	*Limites*	*Número de Pontos*
F_5	30064	11656 – 51630	10
Z_{C2}^5	0,0460	0,0171 – 0,109	8
Z_{C3-}^5	0,765	0,449 – 0,941	9
Z_{C3+}^5	0,188	0,0575 – 0,320	9
RR_2	12,23	8,00 – 15,00	8
B/F_2	0,896	0,600 – 0,910	8
Total de Pontos			414720

A coluna T-03 foi modelada de forma semelhante à coluna T-02, utilizando-se duas redes neuronais, sendo uma para as concentrações de saída da coluna e outra para as demais variáveis. As variáveis utilizadas como entradas do modelo foram as seguintes: vazão mássica de entrada da coluna (F_{10}), fração mássica de propeno na entrada (Z_{C3-}^{10}), e os graus de liberdade descritos anteriormente FA_3 e FR_3, sendo que os limites de variação da composição de entrada e vazão de alimentação da coluna T-03 foram obtidos a partir dos resultados de saída calculados para a T-02. Os valores nominais, os limites e o número de pontos utilizados são apresentados na Tabela 5.8.

Tabela 5.8:Região utilizada para treinar a rede neural da coluna T-03(Schultz, 2015).

Variável	*Valor nominal*	*Limites*	*Número de Pontos*
F_{10}	26458	15874 – 37041	7
Z_{C3-}^{10}	0,798	0,594 – 0,953	7
FA_3	0,873	0,600 – 0,890	6
FR_3	0,792	0,642 – 0,942	6
Total de Pontos			1764

Schultz (2016) ainda realizou a otimização da unidade para definição do ponto de operação e posterior aplicação do método exato local, para encontrar o melhor conjunto de variáveis individuais que minimizam a perda (isto é, que aumentam o lucro); e o método do espaço nulo para avaliar o melhor conjunto de combinações lineares. Mas isso foge do objetivo desta dissertação, e o que interessa realmente são os pontos simulados por ele.

Capítulo 6 – **Desenvolvimento das Inferências para a Unidade de Separação de Propeno/Propano**

Neste capítulo, a unidade de separação de Propeno/Propano será utilizada como estudo de caso para o desenvolvimento de inferências. Como a unidade é composta por três colunas, os resultados obtidos serão mostrados separadamente para as três colunas. As variáveis de entrada e saída do modelo serão reorganizadas neste capítulo. No capítulo anterior, as variáveis de entrada eram as variáveis apontadas antes da simulação, enquanto que as de saída eram resultados da simulação. Mas para o desenvolvimento das inferências, as variáveis de entrada agora serão as variáveis facilmente medidas, como temperatura, vazão e pressão. Entretanto, outras variáveis que também podem ser facilmente estimadas serão utilizadas como variáveis de entrada, como por exemplo: DP_x (diferença de pressão entre o topo e o fundo da coluna T-0X), RR_x (razão de refluxo da coluna T-0X), $Qcond_x$ (calor trocado no condensador da coluna T-0X), $Qref_x$ (calor trocado no refervedor da coluna T-0X), W_x (trabalho realizado no equipamento x). Vazões mássicas e razões entre vazões mássicas também serão utilizadas se necessário, pois essas variáveis possuem alta correlação com as concentrações das correntes, e podem ser medidas e estimadas com o uso de medidores de densidade baseados no princípio de Coriolis.

6.1 Coluna T-01

Para a coluna T-01, a variável chave para ser estimada é a concentração dos pesados na corrente de topo, tida como impureza dessa corrente que seguirá para a coluna T-02. A estimativa dessa corrente é importante para o controle do processo, pois essa impureza deve se manter no mínimo, considerando as condições termodinâmicas e econômicas da planta. Os valores da média, desvio padrão, mínimo e máximo, 1º, 2º e 3º quartis dessa concentração, para os dados disponíveis estão dispostos na Tabela 6.1.

Tabela 6.1: Descricação da variável de saída Z_{C4+}^{4}
(concentração dos pesados no topo da coluna T-01).

Variável	*Nº de Amostras*	*Média*	*Desvio Padrão*	*Mínimo*	*25%*	*50%*	*75%*	*Máximo*
Z_{C4+}^{4}	900	0,1067	0,1555	4,9e-4	7,1e-4	8,7e-4	0,170	0,5255

6.1.1 Pré-Processamento dos Dados da Coluna T-01

Em posse dos dados da unidade, o desenvolvimento da inferência para a coluna T-01 se inicia então com o pré-processamento dos dados. Como os dados são oriundos de uma simulação estática, a etapa de detecção de estados estacionários será omitida, como já falado anteriormente. Porém a detecção de *outliers* não será, uma vez que os dados foram gerados a partir de um modelo caixa preta que foi ajustado a dados da simulação em Aspen Plus. Portanto, existe a possibilidade de haver inconsistências.

As variáveis de processo disponíveis no conjunto de dados estão descritas na Tabela 6.1. Já aqui, pode-se concluir que a matriz de entrada terá dimensão 900x9, que são 900 amostras representadas pelas linhas da matriz versus 9 colunas que são as variáveis de entrada.

Tabela 6.2: Variáveis disponíveis, da coluna T-01,
que podem ser utilizadas como entrada para o modelo.

Variável	*Descrição*
D/F_1	Razão mássica entre a vazão de destilado (corrente 4) e a vazão de entrada (corrente 1)
RR_1	Razão de refluxo da coluna T-01
$Qref_1$	Calor trocado no refervedor da coluna T-01
$Qcond_1$	Calor trocado no condensador da coluna T-01
F_4	Vazão mássica de topo da coluna T-01
F_6	Vazão mássica de fundo da coluna T-01
T_4	Temperatura no topo da coluna T-01
T_6	Temperatura no fundo da coluna T-01
DP_1	Diferença de pressão na coluna T-01

Foi feita uma análise prévia na matriz de dados e foi constatado que não havia dados anômalos no conjunto. Com isso, 3,33% desses dados foram modificados afim de se criar dados anômalos. Então, das 900 amostras, 30 foram escolhidas igualmente espaçadas dentro do conjunto (isto é, as amostras 30ª, 60ª, ..., 870ª, 900ª) e foram acrescidas num valor de 10%. Com esse novo conjunto, na presença de *outliers*, dá-se início à metodologia proposta.

A metodologia se inicia com a normalização desses dados, atribuindo média zero e desvio padrão unitário. Em seguida, com os dados normalizados, aplica-se PCA para uma

visualização das amostras e já se pode usar a matriz de covariâncias para estimar o vetor T² de Hotelling. A variância explicada em cada componente principal, bem como a variância acumulada nos componentes pode ser visualizada no gráfico da Figura 6.1.

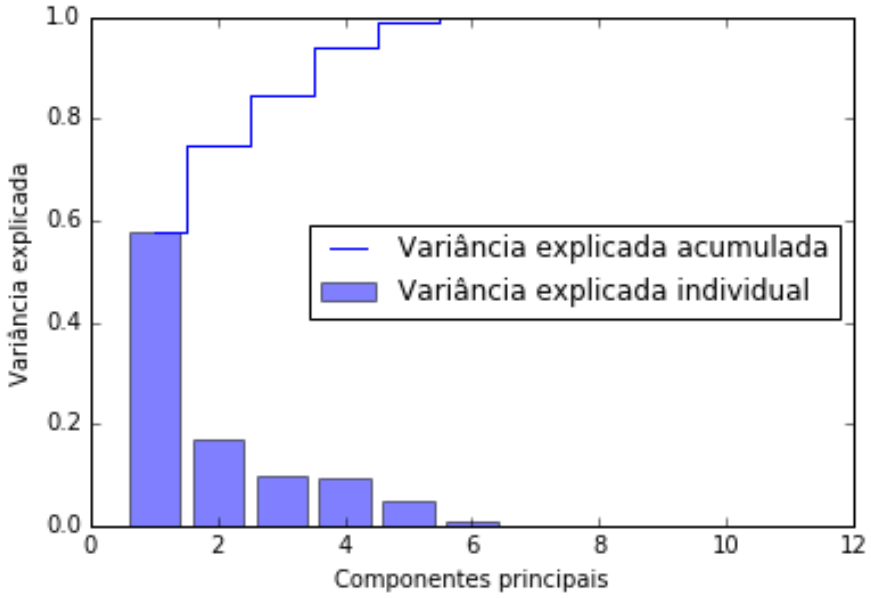

Figura 6.1: Variância explicada acumulada e individual para as variáveis de entrada da coluna T-01.

O gráfico da Figura 6.2 mostra os dados distribuídos nos primeiros dois componentes principais; esses componentes explicam mais de 70% da variância dos dados (cf. Figura 6.1).

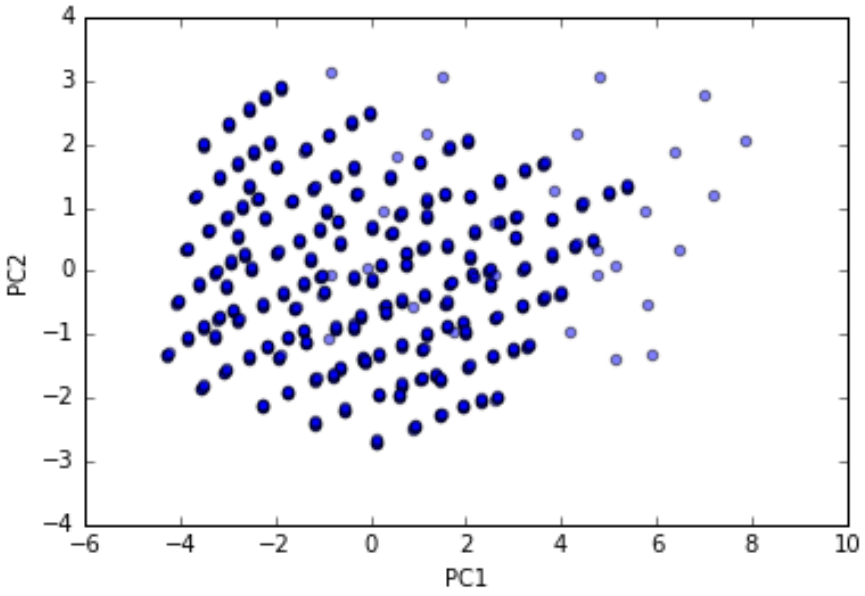

Figura 6.2: Visualização das amostras, da coluna T-01, com adição de *outliers* espalhadas nos primeiros dois componentes principais (os pontos mais claros são *outliers*).

Já é possível visualizar alguns dados anômalos ao restante do conjunto. Todavia, uma forma mais direta e eficiente de identificar os *outliers* é por meio do método estatístico T² de Hotelling. O cálculo do T² foi obtido através da equação 2.3 e o limite T_α^2 foi obtido

através da equação 2.4. Para T_{α}^{2}, foram usados os valores α = 1%, N = 900 (amostras), l = 9 (variáveis). O resultado pode ser visualizado no gráfico da Figura 6.3.

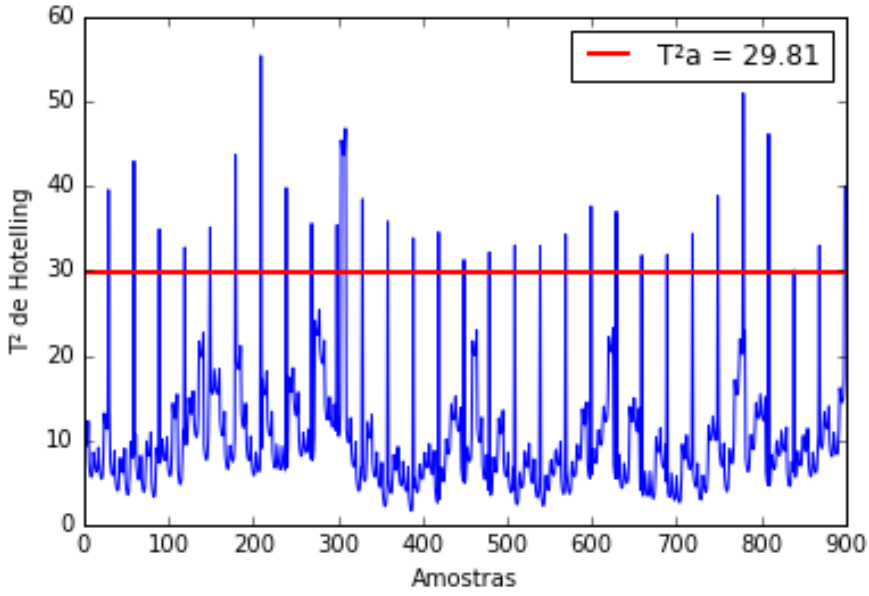

Figura 6.3: Gráfico do T² versus amostras da coluna T-01.

O método selecionou 36 pontos anômalos. A seleção de 6 pontos, além dos 30 pontos alterados propositalmente, pode ter sido falha do método, porém trata-se de uma falha aceitável, visto que o método selecionou 6 pontos dentre 870 não anômalos, isto é, um erro de 0,69%. Existe também a possibilidade desses pontos serem anômalos de fato, devido à não convergência da simulação ou ao ajuste inadequado da rede neural. Para verificar que o método foi eficiente na detecção dos *outliers*, aplica-se PCA novamente e plota-se PC1 versus PC2 com os *outliers* removidos. O resultado está ilustrado no gráfico da Figura 6.4.

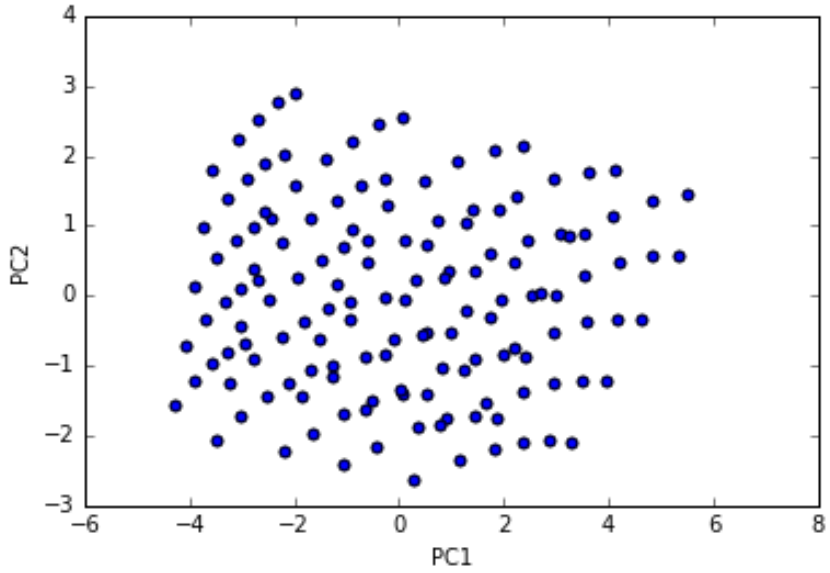

Figura 6.4: Visualização das amostras da coluna T-01 livres de *outliers* espalhadas nos primeiros dois componentes principais.

Os dados agora estão livres de *outliers*, consequentemente o modelo não será prejudicado por conta da presença de dados incoerentes. Desta forma, os dados estão prontos para seguir adiante na metodologia e fazer a seleção dos conjuntos que serão utilizados para calibração e teste.

6.1.2 Segregação de Dados

A segregação dos dados nos conjuntos de calibração e teste se dará pelo método *k-rank*, o qual foi apresentado no capítulo 3. O método propõe uma divisão das amostras em *clusters* utilizando a metodologia *k-means.* Entretanto, faz-se necessário identificar o número de *clusters* ideal, já que o algoritmo *k-means* não o faz. Para isso, utiliza-se a análise de silhueta (*sillouette analysis*), que analisa a qualidade dos *clusters* gerados. Dessa forma, analisou-se a qualidade dos *clusters* gerados pelo *k-means*, para k = 2, 3, ..., 19, 20. Os valores dos coeficientes de silhueta podem ser vistos no gráfico da Figura 6.5.

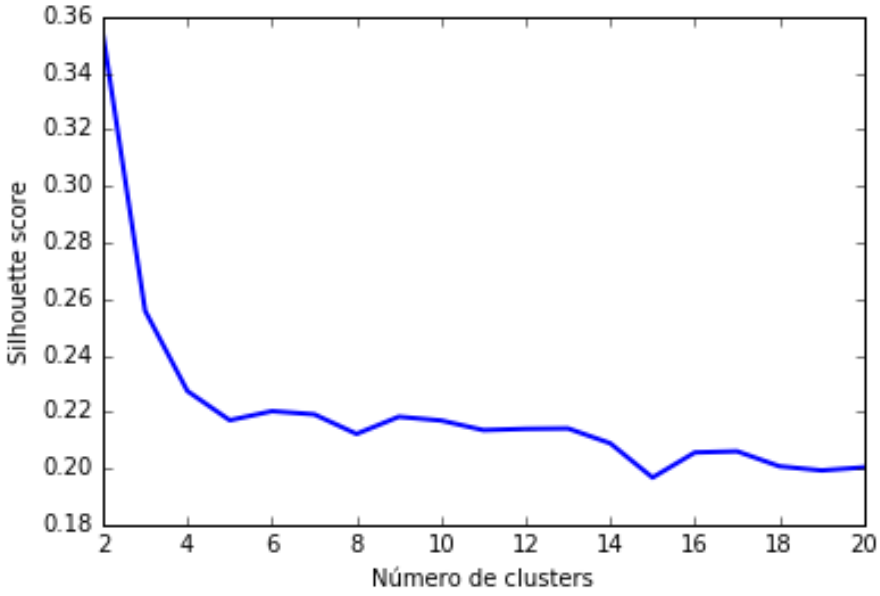

Figura 6.5: Valores dos coeficientes de silhueta para *k* = 2, 3, ..., 19, 20, para a coluna T-01.

O gráfico mostra que aumentando o número de *clusters* não há uma tendência de melhorar o agrupamento dos dados, visto que quanto mais próximo de 1 estiver o coeficiente, mais bem definidos são os *clusters,* sendo que o maior valor é atingido para k=2. Logo, o melhor agrupamento utilizando *k-means* se dá para *k* = 2, com o valor do coeficiente de silhueta de S_2 = 0,35.

Com isso, utilizou-se o método *k-means* com *k* = 2 *clusters* para agrupar as amostras. Como o conjunto de dados possui 9 variáveis, fica difícil a visualização nos eixos das variáveis. Porém, uma alternativa é utilizar os primeiros dois componentes principais do PCA para ilustrar o agrupamento feito pelo *k-means.* A Figura 6.6, ilustra a separação dos dois *clusters* sendo mostrados nos eixos PC1 e PC2.

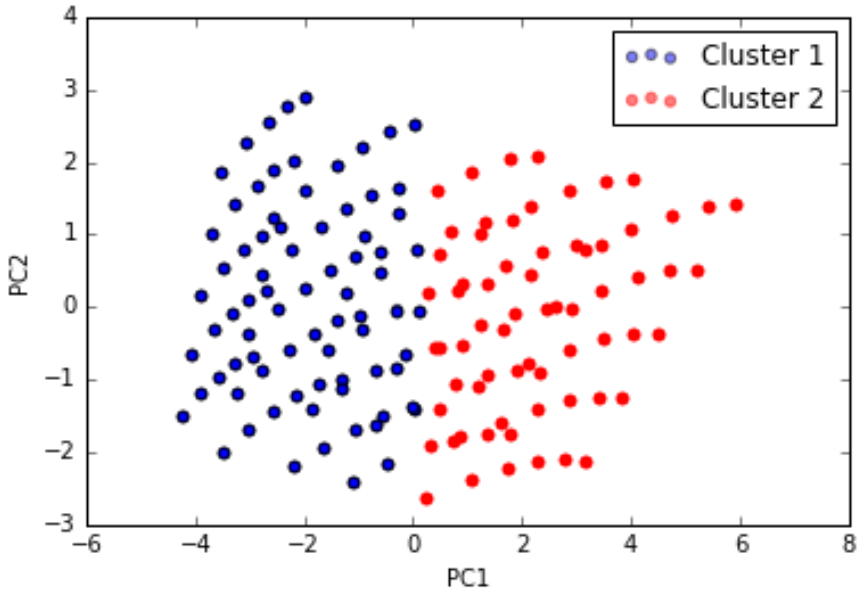

Figura 6.6: Visualização das amostras da coluna T-01 agrupadas pelo *k-means* com *k* = 2.

Com os dados separados nos dois *clusters*, aplica-se o método *y-rank* em cada *cluster* gerado. Vale ressaltar que os dados serão dispostos em ordem crescente do vetor y.

6.1.3 Inferindo a Concentração dos componentes pesados (C4+) na Corrente de Topo da Coluna T-01

Utilizando a metodologia *k-rank* e uma proporção de 2:1 dos conjuntos de calibração e teste, isto é, 2/3 dos dados separados para calibração e 1/3 separados para teste, selecionaram-se os dados contendo as 9 variáveis de entrada e a variável de saída $y = Z_{c4+}^{4}$.

Do gráfico da variância explicada versus os componentes principais da Figura 6.1, sabe-se que 94% da variância é explicada através de 4 PCs, e 99% através de 5 PCs. Ou seja, a análise dos componentes principais informa que, embora haja 9 variáveis, algumas informações podem ser redundantes. Têm-se vazões de topo e fundo que são bastante correlacionadas; têm-se temperaturas de topo e fundo que também são bem correlacionadas. Entretanto, não se pode afirmar que um modelo com 5 variáveis será o melhor modelo possível, pois o método *PCA* não possui informações da variável de saída e isso depende fortemente da relação dela com as variáveis de entrada.

A busca pelo modelo para estimar a concentração de pesados na corrente de topo se inicia com uma busca por modelos lineares. Como não há, nesse caso, uma preocupação com o tempo computacional, já que são apenas 9 variáveis, será utilizado, além dos métodos *Ridge, LASSO e LASSOLARS*, o método busca exaustiva, também conhecido por força bruta, que nada mais é do que um método que busca o melhor modelo, dentre todos os possíveis. Entretanto, como, nesse caso, busca-se também uma seleção de variáveis, o critério de avaliação utilizado será o BIC, já que este penaliza mais fortemente modelos maiores.

Construíram-se os modelos utilizando os métodos supracitados e os resultados dos critérios de avaliação para o conjunto de calibração estão na Tabela 6.3. Para os dados alocados no conjunto de teste, os resultados estão na Tabela 6.4.

Tabela 6.3: Resultados dos critérios de avaliação para o conjunto de calibração – inferência dos pesados no topo da coluna T-01.

Método	*Quantidade de Variáveis Selecionadas*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	7	0,98	0,0215	1410,9%	9382,5%	-2771
LASSOLARS	7	0,98	0,0213	1416,9%	10573,9%	-2782
Ridge	9	0,99	0,0170	11.65,5%	9890,9%	-3032
Busca exaustiva	8	0,99	0,0147	882,3%	8926,7%	-3210

Tabela 6.4: Resultados dos critérios de avaliação para o conjunto de teste – inferência dos pesados no topo da coluna T-01.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,98	0,022	1421,9%	9353,0%
LASSOLARS	0,98	0,0218	1427,9%	10459,7%
Ridge	0,99	0,0175	1170,1%	9778,4%
Busca exaustiva	0,99	0,0152	897,8%	8926,1%

Para visualizar a capacidade de predição dos modelos, foram gerados gráficos dos valores preditos versus os valores reais, tanto para o conjunto de calibração, como para o conjunto de teste. Esses gráficos estão dispostos nas Figuras 6.7, 6.8, 6.9, 6.10, e a diagonal que corta o gráfico representa a reta em que o valor predito seria idêntico ao valor real, ou seja, quanto mais afastado dessa reta, pior a predição do modelo.

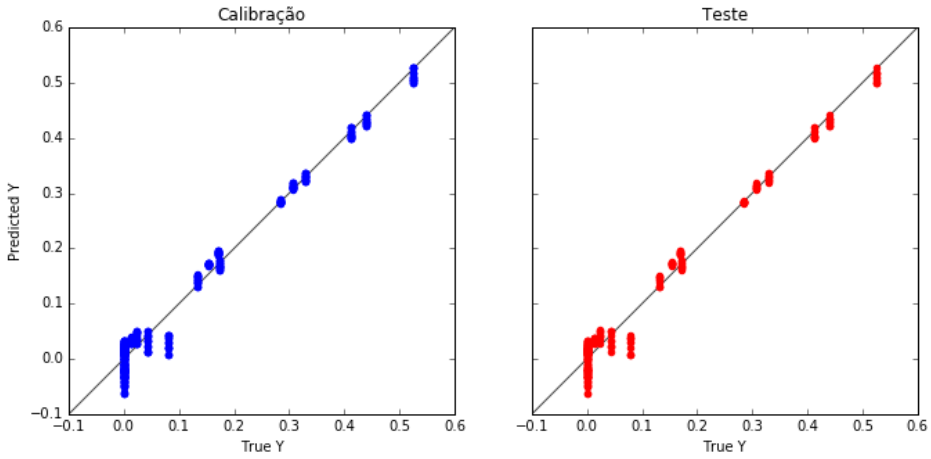

Figura 6.7: Predição do modelo criado pelo método *LASSO* – inferência dos pesados no topo da coluna T-01.

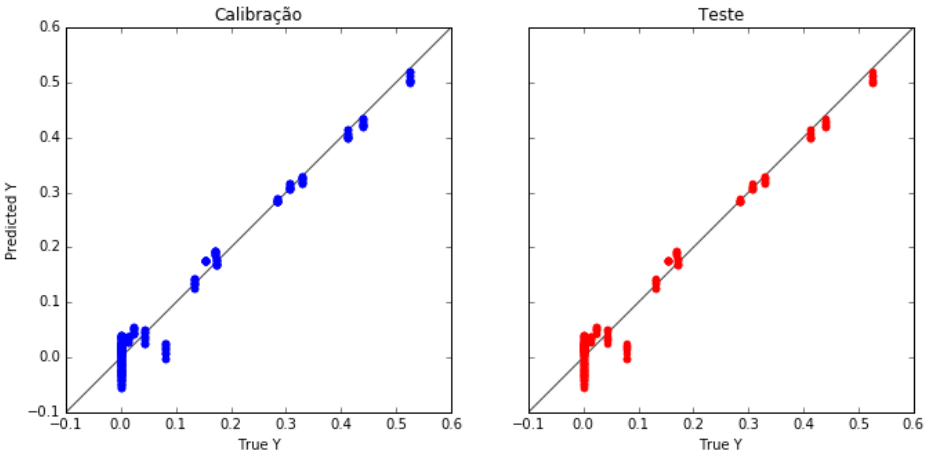

Figura 6.8: Predição do modelo criado pelo método *LASSOLARS* – inferência dos pesados no topo da coluna T-01.

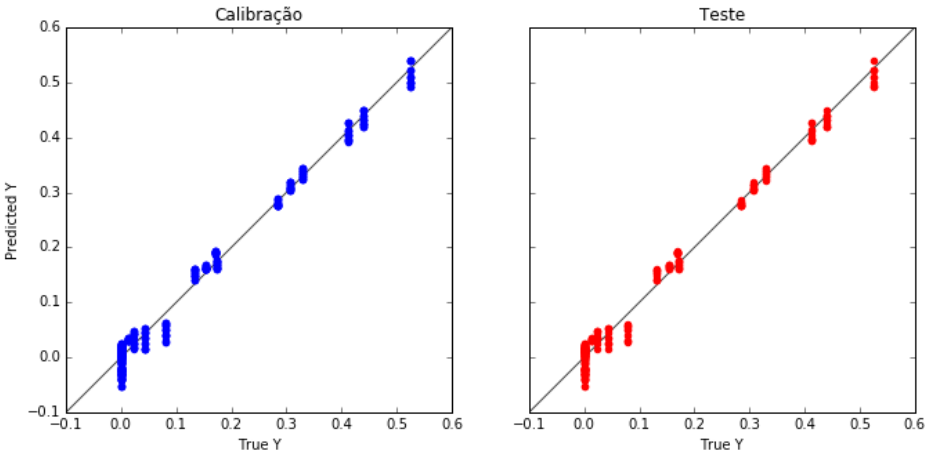

Figura 6.9: Predição do modelo criado pelo método *Ridge* – inferência dos pesados no topo da coluna T-01.

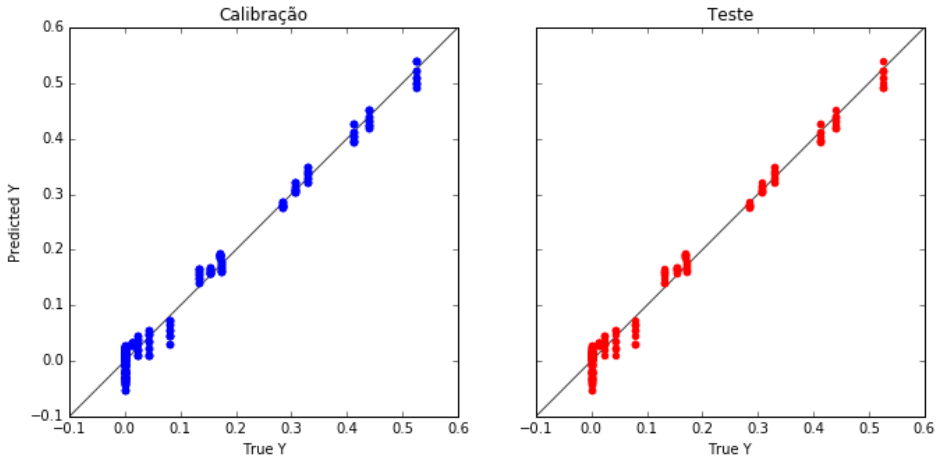

Figura 6.10: Predição do modelo criado pelo método busca exaustiva.

Os gráficos das Figuras 6.7, 6.8, 6.9 e 6.10, bem como os valores dos critérios de avaliação dos modelos, tanto para calibração, como para teste, mostram os valores preditos em função dos valores reais da variável modelada (concentração dos pesados no topo da coluna T-01). Pode-se concluir a partir dos valores do critério *MAPE* e Max e% que esses modelos não estão bons. É notório também que na região onde os dados são muito próximos de zero, o modelo obteve os piores resultados, e, nessa região, os erros relativos são amplificados. Portanto, de acordo com a metodologia proposta, inicia-se a busca por modelos não lineares, utilizando como estratégia a expansão polinomial das variáveis disponíveis. Todavia, fica inviável o uso da busca exaustiva, e com isso, adota-se o método proposto *ACO Plus* como algoritmo de busca de variáveis.

6.1.4 Expansão das Variáveis disponíveis

Como os dados já foram pré-processados, será feito apenas a expansão dos dados em terceira ordem. Essa expansão consiste em combinar as variáveis entre elas mesmas, de modo que sejam criadas variáveis de ordem 2 ($x_i \cdot x_j$) e ordem 3 ($x_i \cdot x_j \cdot x_k$), para $i = j = k = 1, \dots, 9$. Com isso, foram criadas 210 novas variáveis, totalizando 219 variáveis de entrada.

6.1.5 Segregação de Dados

Novamente se avaliou a qualidade dos *clusters* utilizando a análise de silhueta. Os resultados estão no gráfico da Figura 6.11. O melhor agrupamento ocorre para *k* = 2, com o valor do coeficiente de silhueta de S_2 = 0,63. A separação pode ser vista no gráfico da figura 6.12, no qual os dados foram processados utilizando Kernel PCA, com o Kernel sendo RBF (*Radial basis function*), visto que com o PCA a visualização não ficou nítida – as amostras ficaram sobrepostas.

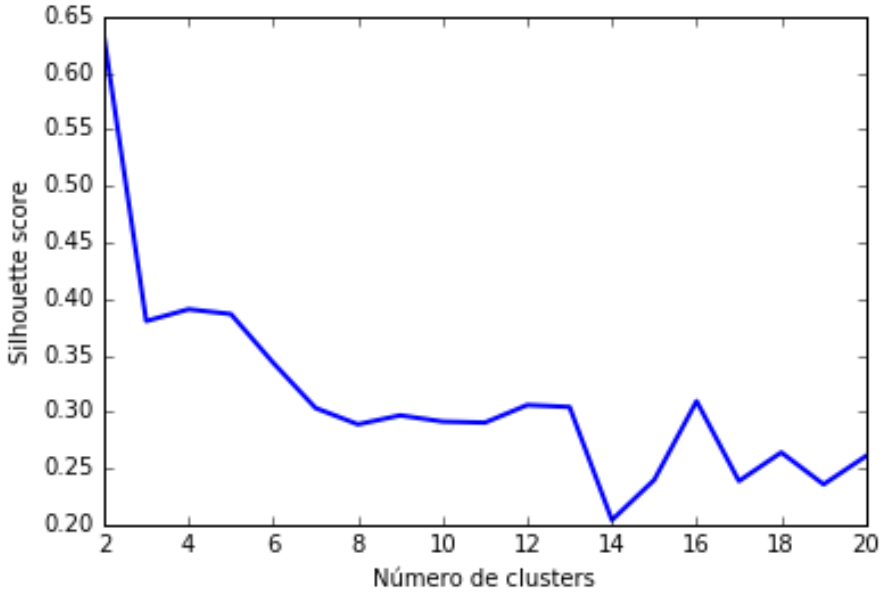

Figura 6.11: Valores dos coeficientes de silhueta para *k* = 2, 3, ..., 19, 20, para a coluna T-01.

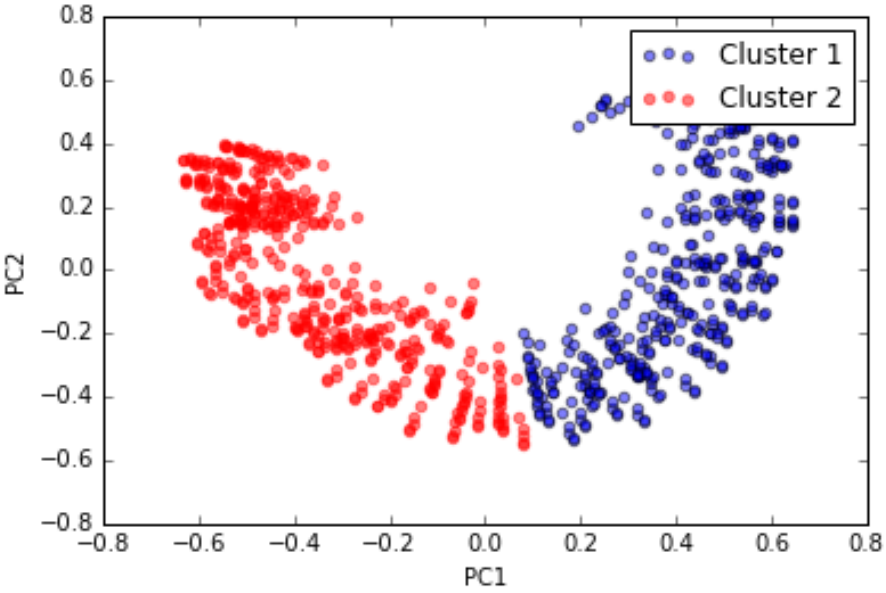

Figura 6.12: Visualização das amostras, da coluna T-01, agrupadas pelo *k-means* com *k* = 2.

Separados os *clusters*, aplica-se *y-rank* em cada grupo e se separa as amostras para calibração e teste.

6.1.6 Inferindo a Concentração de Pesados na Corrente de Topo da Coluna T-01 (Modelo com Expansão Polinomial de Ordem 3)

Utilizando a mesma proporção de 2:1 com a metodologia *k-rank* separaram-se os conjuntos em calibração e teste. Em seguida, foram construídos modelos com o conjunto de calibração, utilizando metodologias de seleção de variáveis de forma semelhante ao procedimento feito para o caso linear, todavia será utilizado o *ACO Plus* como algoritmo de busca. A Tabela 6.5 mostrará os resultados das métricas de avaliação para o conjunto de calibração e a Tabela 6.6 para o conjunto de teste.

Tabela 6.5: Resultados dos critérios de avaliação para o conjunto de calibração – inferência dos pesados no topo da coluna T-01.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a Expansão Polinomial*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	8	22	0,998	0,0073	420%	1784%	-3934
LASSOLARS	9	59	0,999	0,0014	77%	311,4%	-5582
Ridge	9	219	0,999	0,0011	65,9%	330%	-4889
ACO Plus	9	24	0,999	0,0018	80%	481%	-5521

Tabela 6.6: Resultados dos critérios de avaliação para o conjunto de teste – inferência dos pesados no topo da coluna T-01.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,998	0,0074	420,6%	1792,3%
LASSOLARS	0,999	0,0015	79%	304,1%
Ridge	0,999	0,0012	72,3%	329,5%
ACO Plus	0,999	0,0019	78%	482%

Como é necessário informar o tamanho do modelo para o *ACO Plus*, este algoritmo foi executado variando o número de variáveis e a Figura 6.13 mostra o decrescimento do erro em função do tamanho do modelo. Escolheu-se fazer modelos com até 30 variáveis, visto que o erro passou a diminuir pouco com o aumento do número de variáveis. Vale ressaltar que o método que o *ACO Plus* utiliza dentro do seu algoritmo de seleção de variáveis é o *LASSOLARS*, ou seja, ao permitir que o *ACO Plus* busque por tamanhos maiores, a tendência é que ele se aproxime ainda mais do método *LASSOLARS*. No gráfico, o menor erro está para o modelo com 26 variáveis. Este foi o modelo escolhido, porém, como o *ACO Plus* utiliza o método *LASSOLARS* em sua busca, 2 das 26 variáveis ficaram com o coeficiente iguais à zero.

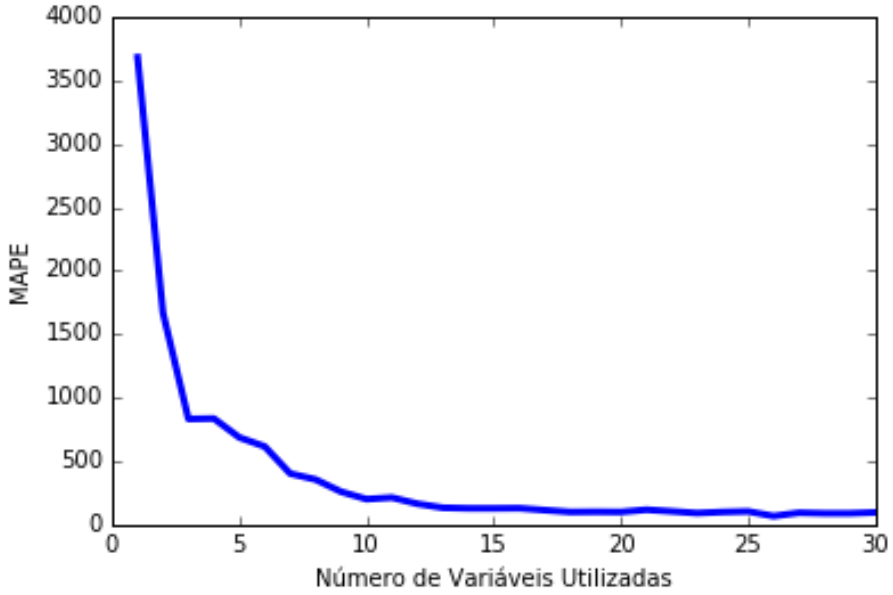

Figura 6.13: Decrescimento do erro médio percentual em função do número de variáveis utilizadas na regressão – inferência dos pesados no topo da coluna T-01.

Com os modelos, puderam-se construir os gráficos dos valores preditos versus os valores reais da variável de saída y (concentração dos pesados na corrente de topo da coluna T-01).

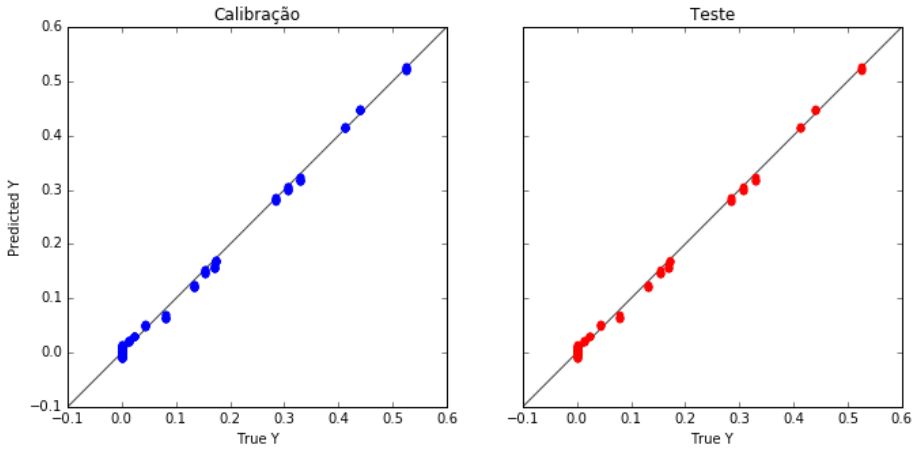

Figura 6.14: Predição do modelo criado pelo método *LASSO*– inferência dos pesados no topo da coluna T-01.

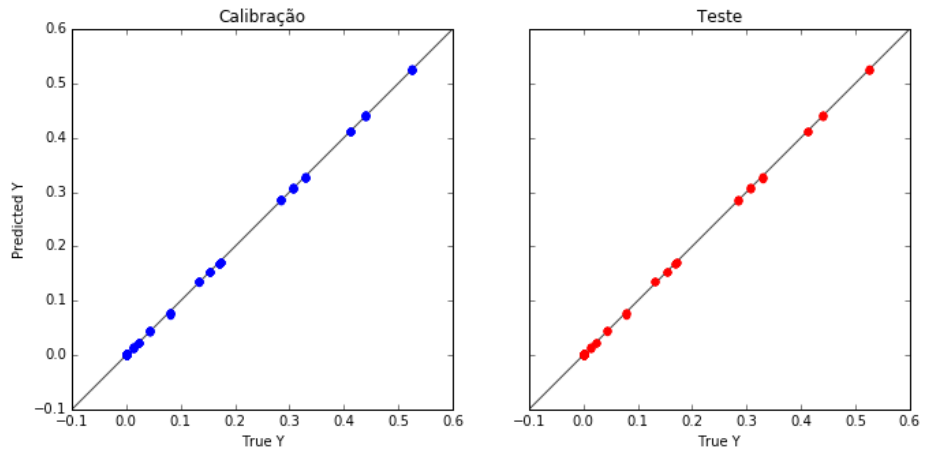

Figura 6.15: Predição do modelo criado pelo método *LASSOLARS*– inferência dos pesados no topo da coluna T-01.

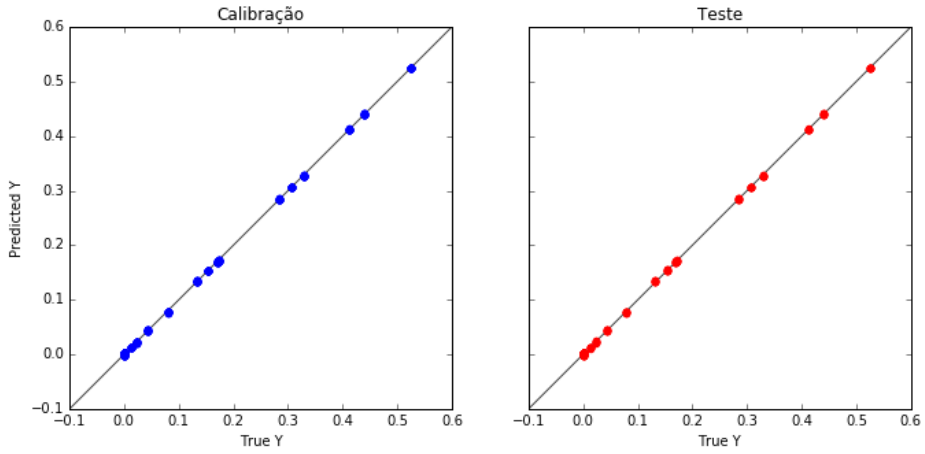

Figura 6.16: Predição do modelo criado pelo método *Ridge*– inferência dos pesados no topo da coluna T-01.

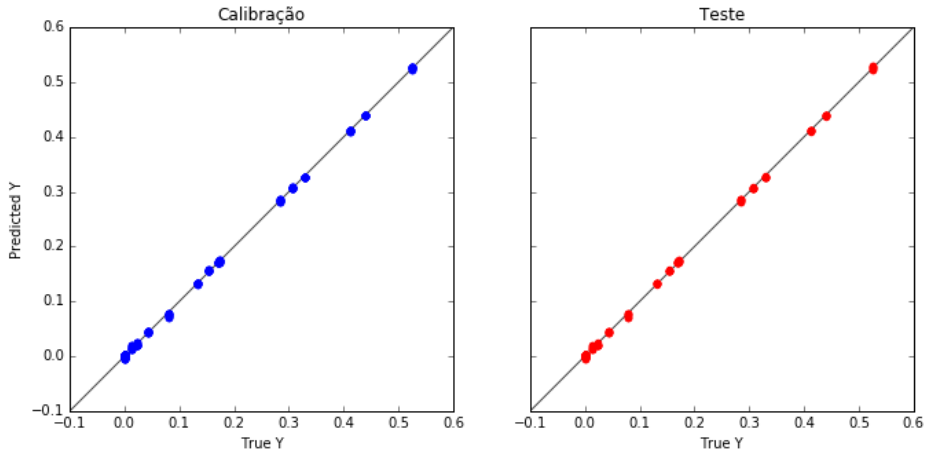

Figura 6.17: Predição do modelo criado pelo método *ACO Plus*– inferência dos pesados no topo da coluna T-01.

Analisando os gráficos de predição e os resultados das métricas de avaliação, percebe-se que houve uma melhora significativa em utilizar a expansão polinomial. Entretanto, ainda há um erro elevado que é provavelmente oriundo das concentrações muito próximas de zero.

Para analisar esse fato, escolheu-se o modelo do método *Ridge*, o qual obteve os melhores resultados e fez-se um *plot* de todos os valores reais e todos os valores preditos em função das amostras. Isso irá ajudar a enxergar a aderência do modelo aos dados.

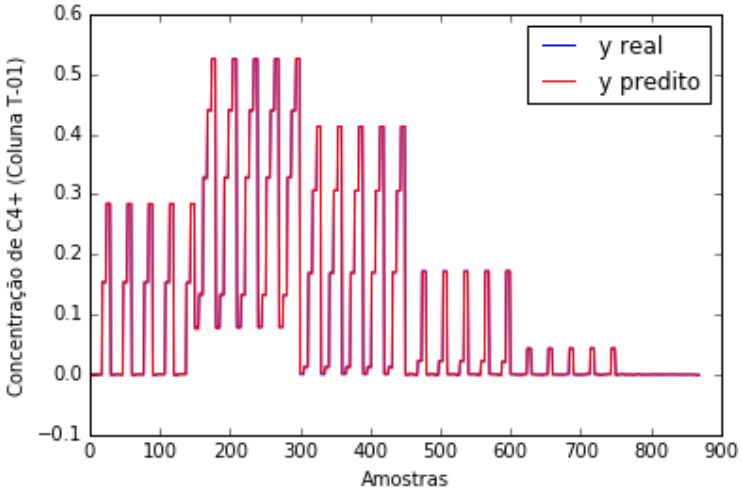

Figura 6.18: Aderência do modelo criado pelo método *Ridge* às amostras – inferência dos pesados no topo da coluna T-01.

Ainda não é possível enxergar no gráfico acima os erros demasiados. Portanto, separaram-se os dados em duas porções: concentrações maiores que 0,01 kg/kg e concentrações menores ou igual a 0,01 kg/kg. Isso facilitará a visualização do erro. O gráfico da Figura 6.19 mostra a aderência para valores em que a concentração de pesados se mantém acima de 0,01 kg/kg. Em seguida, o gráfico da Figura 6.20 mostra a região das baixas concentrações.

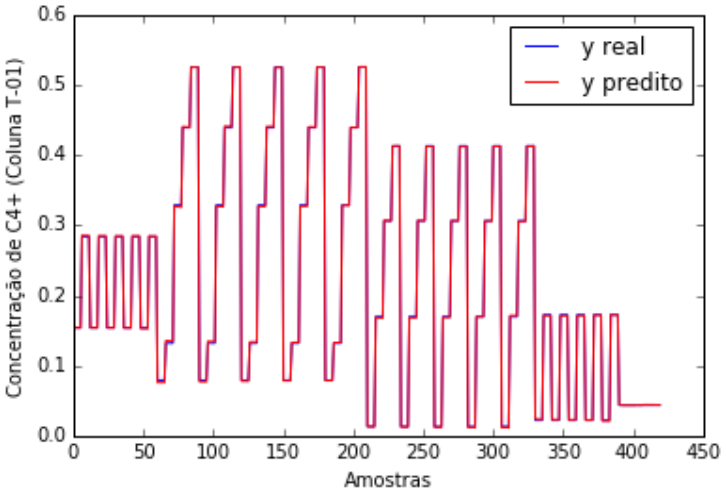

Figura 6.19: Aderência do modelo criado pelo método *Ridge* às amostras com concentrações acima de 0,01 kg/kg de pesados.

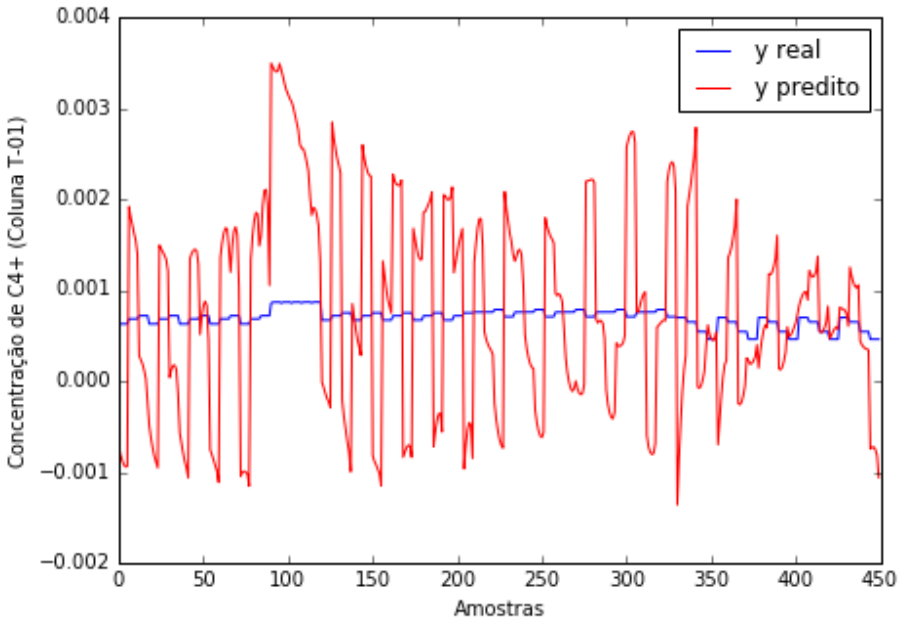

Figura 6.20: Aderência do modelo criado pelo método *Ridge* às amostras com concentrações abaixo de 0,01 kg/kg de pesados.

Com o gráfico acima, que mostra a discrepância de modelo nas baixas concentrações, é possível identificar onde o modelo está inferindo mal. Entretanto, para as concentrações acima de 0,01 kg/kg, o modelo se ajustou bem aos dados. Nesse caso, será necessário trabalhar com modelos locais para conseguir inferir bem as duas regiões dos dados.

Os modelos criados acima servirão como classificadores de regiões, isto é, com as variáveis de entrada, utiliza-se dos modelos acima para identificar se aquelas condições pertencem à região de baixas concentrações (menor que 0,01 kg/kg) ou à região com concentração acima de 0,01 kg/kg. Todos os modelos conseguem identificar as regiões com 100% de acuracidade, consequentemente escolheu-se o modelo desenvolvido pelo *ACO Plus* para realizar essa tarefa, por ser um modelo mais simples, podendo ainda utilizar-se de mais de um modelo para que um confirme a resposta do outro.

O método escolhido para se trabalhar com os modelos locais será o método de partição rígida, onde o modelo global é a união dos submodelos locais (Fernandes, 2001). No caso, um modelo irá estimar a região de concentrações acima de 0,01 kg/kg e o outro a região de menores concentrações. A união desses dois modelos locais resulta no modelo global. A resposta será condicionada, entretanto, à classificação realizada pelo modelo calibrado para as duas regiões. De modo que esse modelo escolhe para qual dos modelos locais devem ir os dados.

6.1.7 Inferindo a Concentração de Pesados em Baixas Concentrações (Modelo com Expansão Polinomial de Ordem 3)

Para inferir a concentração dos pesados na região de baixas concentrações, separaram-se as regiões utilizando-se do modelo classificador. O conjunto de dados para essa região possui 448 amostras e utilizou-se da mesma metodologia proposta com

expansão polinomial de ordem 3, ou seja, o conjunto de dados ficou com dimensão 448*x*220, sendo 219 variáveis de entrada e a variável de saída.

A segregação dos dados em calibração e teste foi feita utilizando *k-rank* com *k* = 2 e a proporção de 2:1 para os conjuntos de calibração e teste. Em seguida, construiram-se os modelos utilizando as metodologias *LASSO*, LASSO*LARS* e *Ridge*, além da busca com o *ACO Plus*. Os resultados para a calibração desses métodos estão dispostos na Tabela 6.7, e teste na Tabela 6.8.

Tabela 6.7: Resultados dos critérios de avaliação para o conjunto de calibração – inferência dos pesados, em baixas concentrações, no topo da coluna T-01.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a Expansão Polinomial*	*R²*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	0	0	0	9,5e-5	11,2%	50,5%	-4728
LASSOLARS	6	14	0,920	2,8e-5	3,3%	9,32%	-5382
Ridge	9	219	0,999	1,8e-6	0,2%	0,9%	-5862
ACO Plus	7	11	0,950	2,2e-5	2,26%	7,40%	-5566

Tabela 6.8: Resultados dos critérios de avaliação para o conjunto de teste – inferência dos pesados, em baixas concentrações, no topo da coluna T-01.

Método	*R²*	*RMSE*	*MAPE*	*Max e%*
LASSO	0	9,3e-5	10,8%	50,5%
LASSOLARS	0,92	2,9e-5	3,3%	9,32%
Ridge	0,999	2e-6	0,23%	0,96%
ACO Plus	0,951	2,1e-5	2,14%	7,40%

A Figura 6.21 mostra o decrescimento do erro em função do tamanho do modelo. No gráfico, o menor erro está para o modelo com 12 variáveis. Este foi o modelo escolhido, porém, um dos coeficientes foi zerado pelo método *LASSOLARS*.

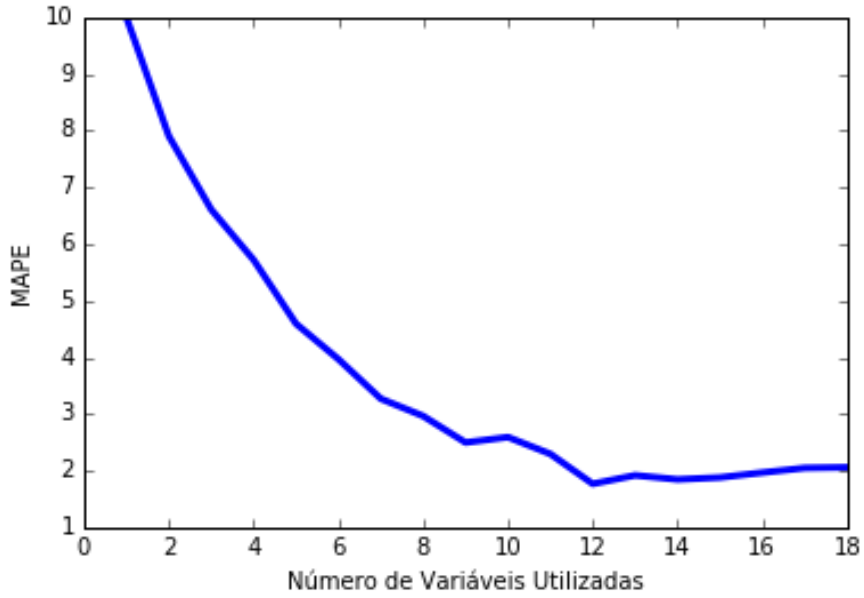

Figura 6.21: Decrescimento do erro médio percentual em função do número de variáveis utilizadas na regressão – inferência dos pesados, em baixas concentrações, no topo da coluna T-01.

Diante dos resultados, pode-se perceber que o método *LASSO* não selecionou nenhuma variável e o modelo criado por este método foi a média dos dados. Os outros métodos obtiveram resultados bons, sendo o método Ridge o que obteve os melhores resultados, porém utilizou todas as variáveis possíveis e suas expansões. Para ilustrar a capacidade preditiva dos modelos, os gráficos da aderência dos dados preditos aos dados reais serão mostrados nas Figuras 6.22, 6.23, 6.24 e 6.25.

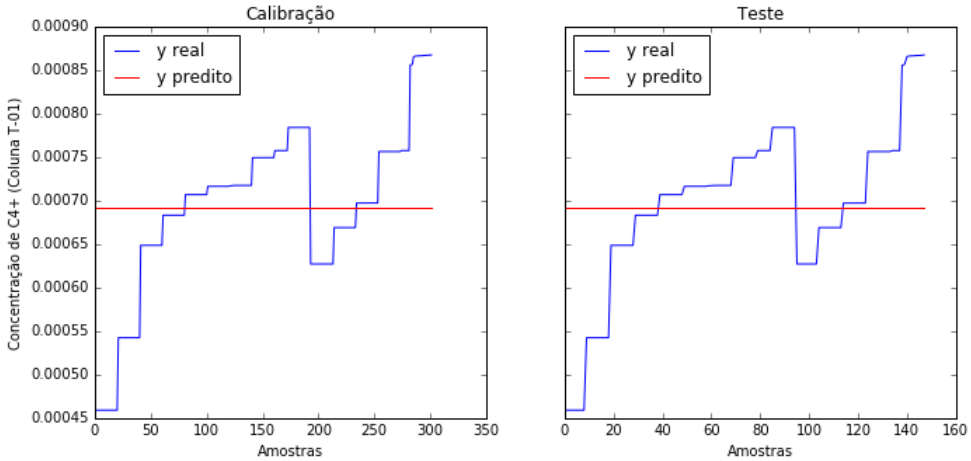

Figura 6.22: Aderência do modelo criado pelo método *LASSO* às amostras com concentrações abaixo de 0,01 kg/kg de pesados.

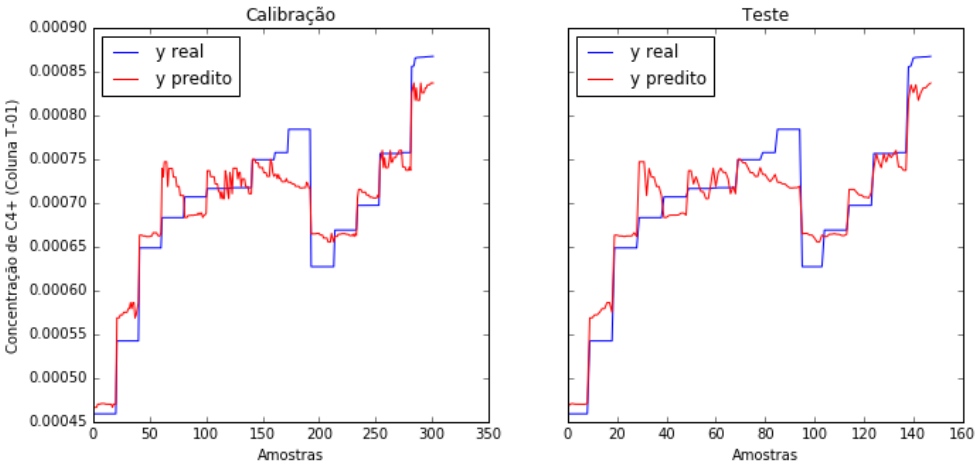

Figura 6.23 Aderência do modelo criado pelo método *LASSOLARS* às amostras com concentrações abaixo de 0,01 kg/kg de pesados.

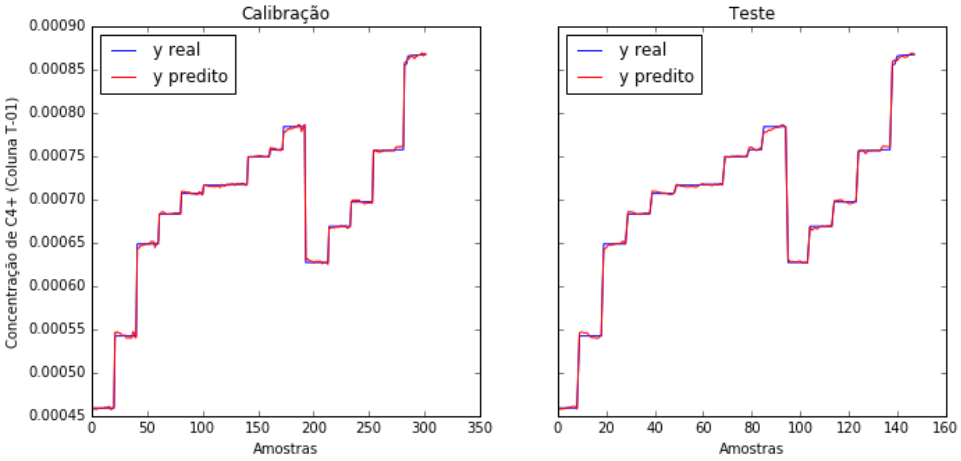

Figura 6.24: Aderência do modelo criado pelo método *Ridge* às amostras com concentrações abaixo de 0,01 kg/kg de pesados.

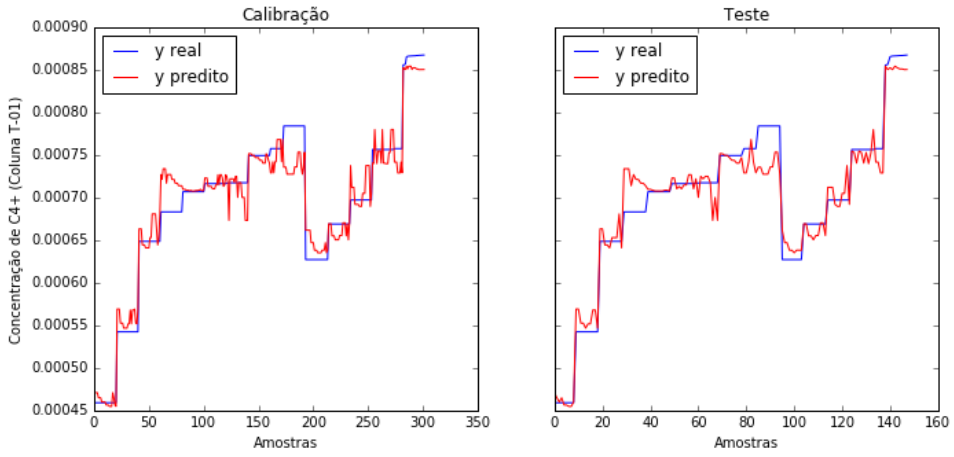

Figura 6.25: Aderência do modelo criado pelo método *ACO Plus* às amostras com concentrações abaixo de 0,01 kg/kg de pesados.

6.1.8 Inferindo a Concentração de Pesados em Concentrações Superiores a 0,01 kg/kg (Modelo com Expansão Polinomial de Ordem 3)

O segundo modelo local, que será utilizado em concentrações superiores a 0,01 kg de pesados por kg de mistura, foi construído utilizando o restante do conjunto de dados que possui 416 amostras. Seguindo a mesma metodologia proposta com expansão polinomial de ordem 3. A segregação dos dados em calibração e teste foi feita utilizando *k-rank* com *k* = 2 e a proporção de 2:1 para os conjuntos de calibração e teste. Em seguida, construíram-se os modelos utilizando as metodologias *LASSO, LASSOLARS* e *Ridge*, além da busca com o *ACO Plus*. Os resultados para a calibração desses métodos estão dispostos na tabela 6.9, e teste na tabela 6.10.

Tabela 6.9: Resultados dos critérios de avaliação para o conjunto de calibração – inferência dos pesados, concentrações superiores à 0,01 kg/kg, no topo da coluna T-01.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a Expansão Polinomial*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	7	13	0,997	0,0085	17,3%	136,9%	-1818
LASSOLARS	9	33	0,999	1,5e-4	0,13%	1,74%	-3971
Ridge	9	219	0,999	2,9e-4	0,33%	3,74%	-2562
ACO Plus	6	11	0,999	2,6e-4	0,25%	1,81%	-3779

Tabela 6.10: Resultados dos critérios de avaliação para o conjunto de teste – inferência dos pesados, concentrações superiores à 0,01 kg/kg, no topo da coluna T-01.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,1997	0,0084	16,5%	136,8%
LASSOLARS	0,999	1,7e-4	0,16%	1,74%
Ridge	0,999	3,2e-4	0,43%	4,29%
ACO Plus	0,999	2,7e-4	0,24%	1,81%

Diante dos resultados, o método que alcançou os melhores resultados com as métricas foi o método *LASSOLARS*, porém o método *ACO Plus* obteve resultados bem próximos com

um menor número de variáveis. O modelo criado pelo *ACO* ficou com apenas 11 variáveis, pois a diminuição do erro em relação ao aumento de variáveis já era irrelevante. Para tamanhos maiores de modelos, o *ACO* ultrapassaria ou igualaria as metas alcançadas pelo método *LASSOLARS*. Entretanto, preza-se por modelos mais simples. A Figura 6.26 mostra a diminuição do erro percentual médio para o conjunto de calibração do *ACO* em função do crescimento do modelo.

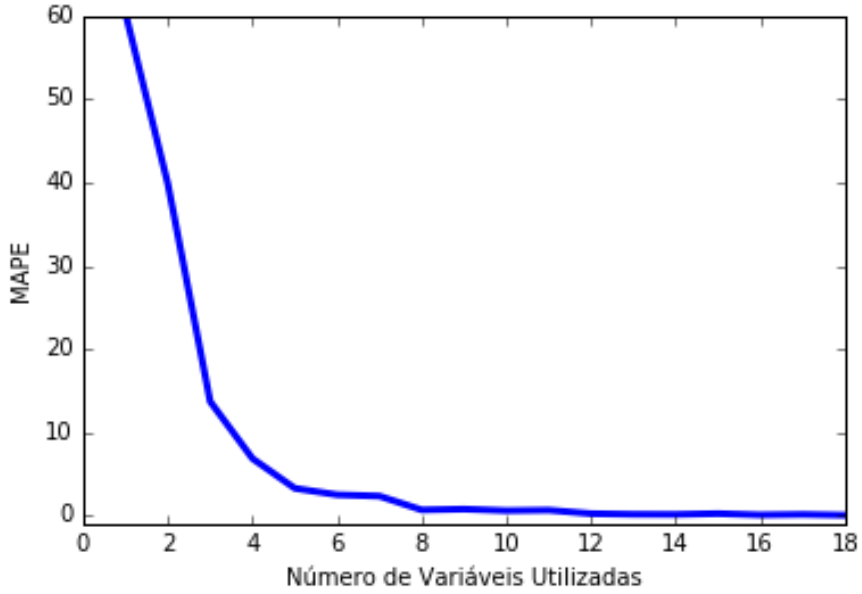

Figura 6.26: Decrescimento do erro médio percentual em função do número de variáveis utilizadas na regressão – inferência dos pesados, concentrações superiores à 0,01 kg/kg, no topo da coluna T-01.

Com os modelos construídos e com a escolha do modelo do *ACO* Plus com 11 variáveis no modelo, o que na verdade são 6 variáveis e algumas expansões polinomiais dessas, fez-se os gráficos da aderência dos dados preditos aos dados reais para o conjunto de calibração e teste. Os gráficos estão ilustrados nas Figuras 6.27, 6.28, 6.29, 6.30 e percebe-se que o único gráfico que os dados preditos não aderiram completamente aos dados reais foi o gráfico do método *LASSO*, justificado nos erros percentuais do método.

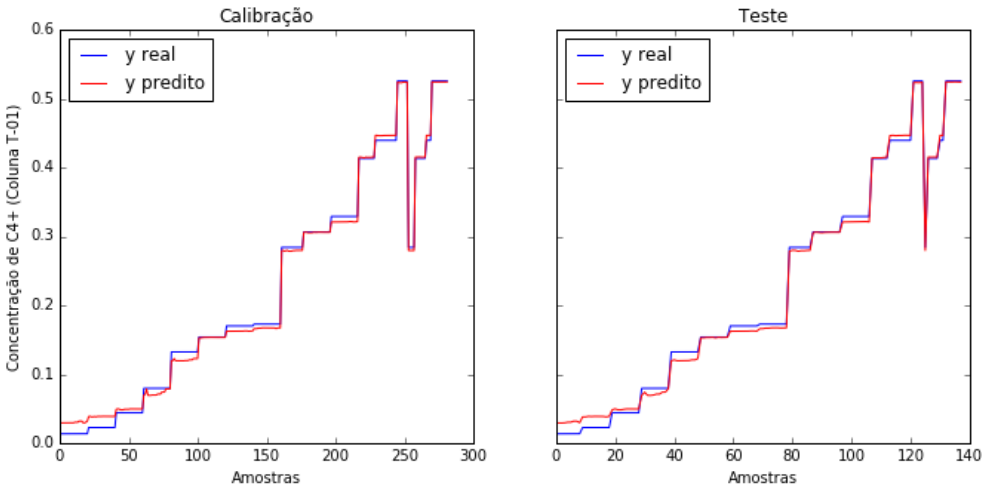

Figura 6.27: Aderência do modelo criado pelo método *LASSO* às amostras com concentrações acima de 0,01 kg/kg de pesados.

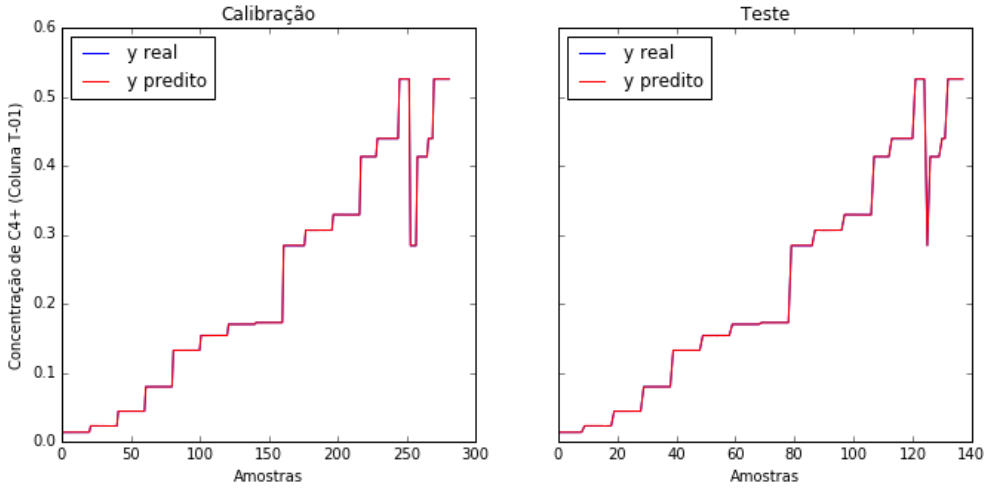

Figura 6.28: Aderência do modelo criado pelo método *LASSOLARS* às amostras com concentrações acima de 0,01 kg/kg de pesados.

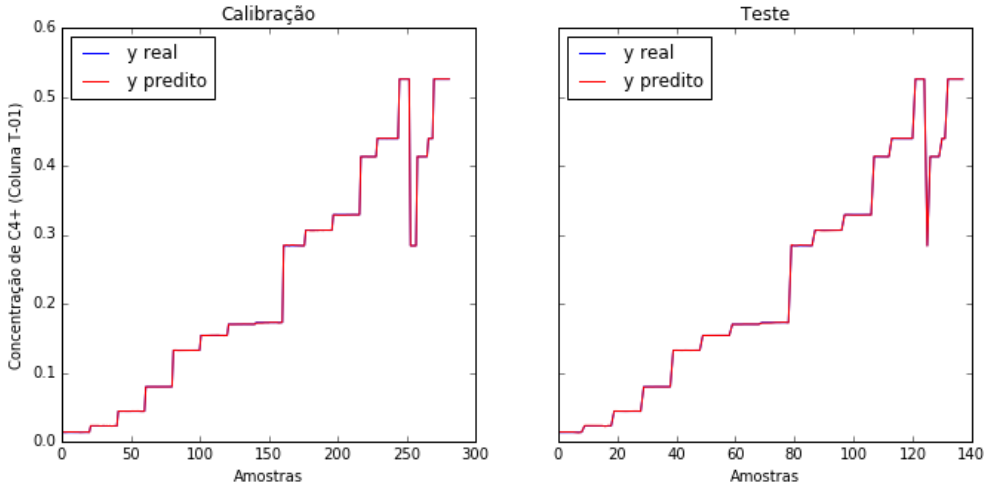

Figura 6.29: Aderência do modelo criado pelo método *Ridge* às amostras com concentrações acima de 0,01 kg/kg de pesados.

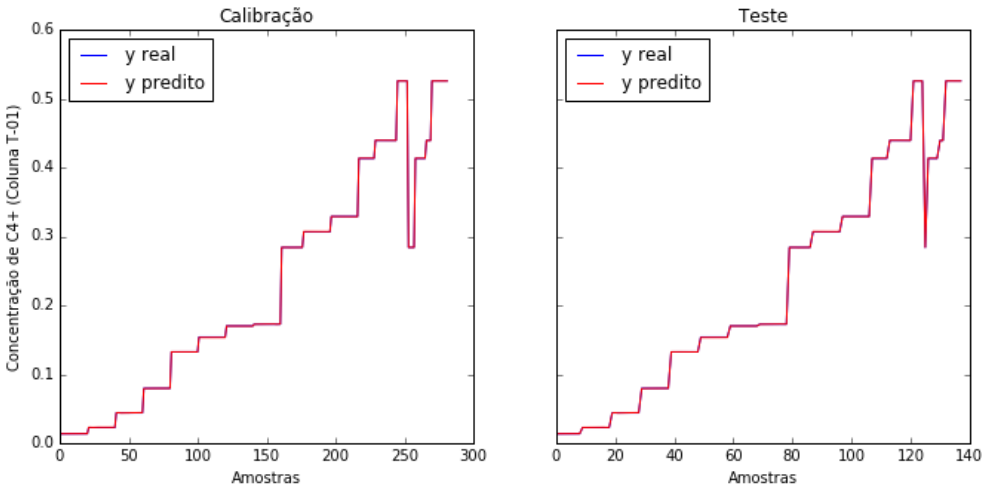

Figura 6.30: Aderência do modelo criado pelo método *ACO Plus* às amostras com concentrações acima de 0,01 kg/kg de pesados.

Os métodos *LASSOLARS*, *Ridge* e *ACO Plus* conseguiram ajustar bons modelos para as duas regiões. Já o método *LASSO* não obteve êxito. Embora o método *Ridge* tenha sido superior na região de baixas concentrações, o erro obtido pelo *ACO Plus* ainda é baixo, tendo um erro médio percentual da ordem de 2,1% e um erro máximo percentual da ordem de 7,4%. Do ponto de vista prático, esses erros são aceitáveis. Entretanto, os modelos podem ser usados em comunhão, podendo o melhor modelo, para cada caso, ser usado como inferência e os demais como ferramenta de validação da inferência.

6.1.9 Descarte de Amostras Redundantes

Como mencionado na metodologia desse trabalho, ao descartar amostras redundantes, pode-se aumentar a precisão do modelo para o total do conjunto de dados. Isso porque os pontos redundantes podem estar localizados de maneira desigual nos dados, diminuindo o erro próximos a eles e aumentando em regiões mais distantes. Por conta disso, será feito um descarte desses pontos redundantes para que se possa melhorar o modelo. Neste caso, será escolhido o modelo construído pelo *ACO Plus*, pois trata-se do modelo que possui um menor número de variáveis selecionadas.

Utilizando a metodologia proposta para o descarte de amostras redundantes, percebeu-se que, para as concentrações de pesados abaixo de 0,01 kg/kg, havia 248 amostras redundantes e, com o descarte dessas, o modelo ganhou mais precisão, passando de um erro máximo de 7,40% para 7,30%, e de um erro médio de 2,26% para 1,97%. Já para as concentrações de pesados acima de 0,01 kg/kg, havia 90 amostras redundantes e, com o descarte dessas, o modelo passou de um erro máximo de 1,81% para 0,79%, e de um erro médio de 0,25% para 0,14%.

6.1.10 Inferindo as Demais Concentrações na Corrente de Topo da Coluna T-01 (Modelo com Expansão Polinomial de Ordem 3)

Seguindo a mesma metodologia para a inferência da concentração dos pesados no topo da coluna T-01, desenvolveram-se as inferências da concentração de propeno, propano e etano no topo dessa mesma coluna. Essas concentrações serão informações úteis para a coluna T-02. Para resumir os resultados, optou-se por utilizar apenas o método *ACO Plus* como regressor, já que esse provou ser o método com a maior eficiência em função da simplicidade do modelo.

Como o tratamento dos dados já foi feito para inferi-la a concentração dos pesados, será mostrado apenas os resultados para o conjunto de calibração e teste (*k-rank*; *k* = 2; proporção 2:1) dos modelos desenvolvidos pelo método *ACO Plus* para inferir as concentrações de propeno, propano e etano na corrente de topo da coluna T-01. Os resultados serão mostrados nessa ordem. As tabelas 6.11, 6.12 e 6.13 apresentam os resultados das métricas de avaliação.

Tabela 6.11: Resultados das métricas de avaliação para o método *ACO Plus* – inferência de propeno no topo da coluna T-01.

ACO Plus (9 variáveis; 23 variáveis expandidas selecionadas)	*Nº de Amostras*	*R^2*	*RMSE*	*MAPE*	*Max e%*
Calibração	576	0,999	0,0040	0,47%	1,60%
Teste	288	0,999	0,0041	0,48%	1,57%

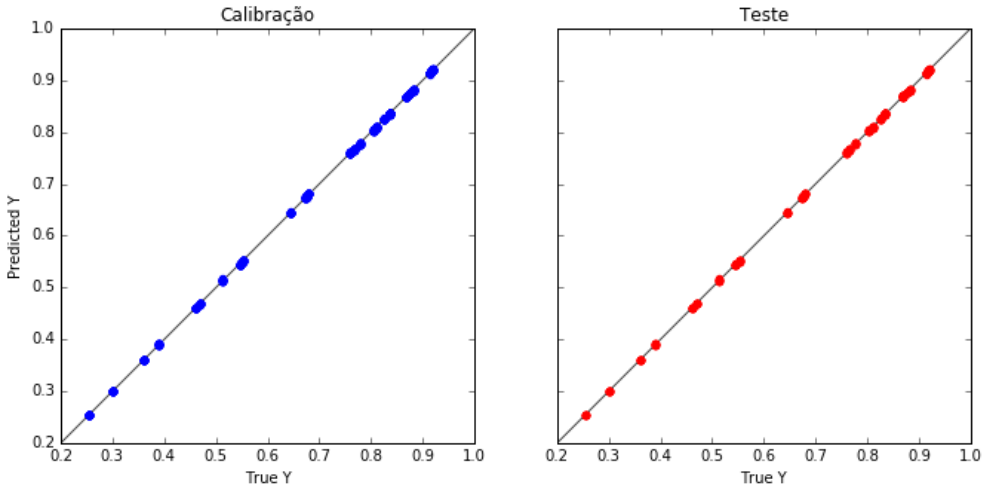

Figura 6.31: Predição do modelo criado pelo método *ACO Plus* – a inferência de propeno no topo da coluna T-01.

Tabela 6.12: Resultados das métricas de avaliação para o método *ACO Plus* – inferência de propano no topo da coluna T-01.

ACO Plus (7 variáveis; 16 variáveis expandidas selecionadas)	*Nº de Amostras*	*R^2*	*RMSE*	*MAPE*	*Max e%*
Calibração	576	0,999	0,0016	0,56%	1,91%
Teste	288	0,999	0,0016	0,57%	1,90%

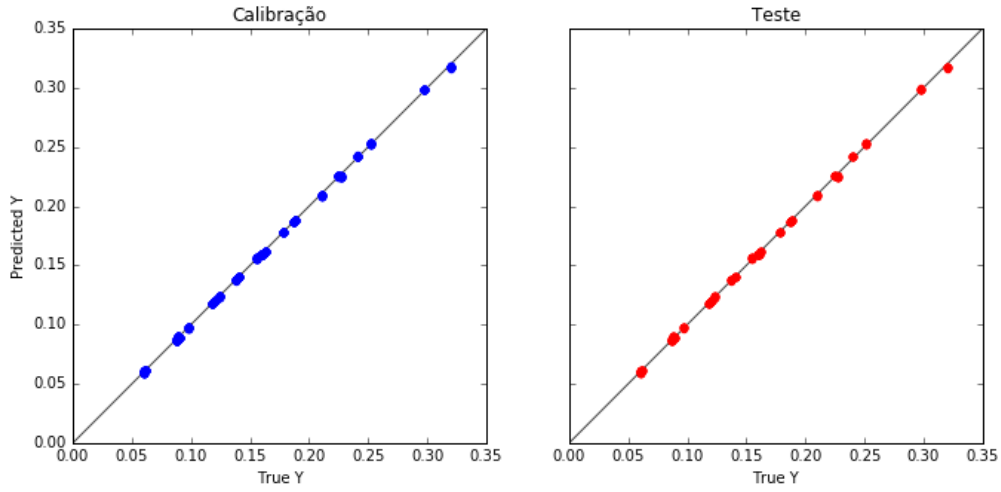

Figura 6.32: Predição do modelo criado pelo método *ACO Plus* – a inferência de propano no topo da coluna T-01.

Tabela 6.13: Resultados das métricas de avaliação para o método *ACO Plus* – inferência de etano no topo da coluna T-01.

ACO Plus (9 variáveis; 27 variáveis expandidas selecionadas)	*Nº de Amostras*	*R^2*	*RMSE*	*MAPE*	*Max e%*
Calibração	576	0,999	0,00053	0,93%	3,24%
Teste	288	0,999	0,00058	0,96%	3,18%

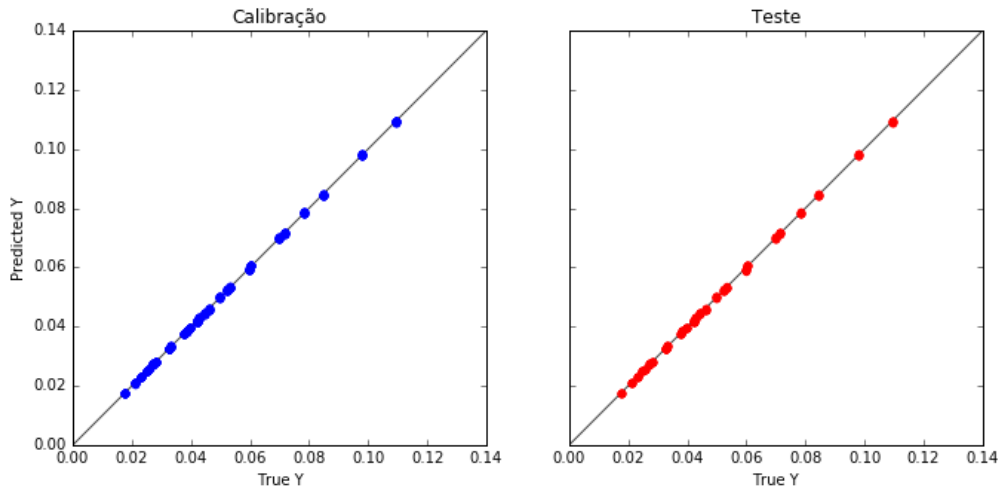

Figura 6.33: Predição do modelo criado pelo método *ACO Plus* – a inferência de etano no topo da coluna T-01.

Todas as concentrações puderam ser estimadas satisfatoriamente no topo da coluna T-01, podendo essas serem utilizadas como variáveis de entrada na coluna T-02.

6.2 Coluna T-02

Na coluna T-02, é economicamente importante reduzir a perda de propeno no topo da coluna. Portanto, a tarefa agora é extrair a informação dessa concentração a partir das variáveis de processo disponíveis, bem como das concentrações que estão vindo da coluna T-01. Essas concentrações foram inferidas com ótima precisão e podem ser utilizadas como variáveis de entrada para a coluna T-02. Os valores da média, desvio padrão, mínimo e máximo, 1º, 2º e 3º quartis da concentração de propeno no topo da coluna T-02, para os dados disponíveis estão dispostos na Tabela 6.14.

Tabela 6.14: Descricação da variável de saída Z_{C3-}^{8} (concentração de propeno no topo da coluna T-02).

Variável	*Nº de Amostras*	*Média*	*Desvio Padrão*	*Mínimo*	*25%*	*50%*	*75%*	*Máximo*
Z_{C3-}^{8}	414720	0,6176	0,1702	0,0128	0,5209	0,6516	0,7445	0,9396

6.2.1 Pré-Processamento dos Dados da Coluna T-02

Para a coluna T-02, as variáveis disponíveis que podem ser utilizadas como variáveis de entrada estão dispostas na tabela 6.15. Já aqui, pode-se concluir que a matriz de entrada terá dimensão 414720x12, que são 414720 amostras representadas pelas linhas da matriz versus 12 colunas que são as variáveis de entrada.

As concentrações de propeno, propano e etano serão utilizadas, pois foram inferidas satisfatoriamente no topo da coluna T-01. Infelizmente, por conta do custo computacional, as simulações foram feitas separadamente para cada coluna (Schultz, 2015). Por conta disso, não será usado os valores dessas concentrações estimadas, mas sim os valores simulados, o que não desmerece o método, já que esses valores foram estimados com uma precisão elevada, como mostrado anteriormente.

Tabela 6.15: Variáveis, da coluna T-02, que podem ser utilizadas como entrada para o modelo.

Variável	*Descrição*
Z_{C3-}^{5}	Concentração de propeno na corrente de entrada da coluna T-02
Z_{C3+}^{5}	Concentração de propano na corrente de entrada da coluna T-02
Z_{C2}^{5}	Concentração de etano na corrente de entrada da coluna T-02
RR_2	Razão de refluxo da coluna T-02
B/F_2	Razão mássica entre a vazão de fundo (corrente 10) e a vazão de entrada da coluna T-02 (corrente 5)
F_8	Vazão de topo da coluna T-02

F_{10}	Vazão de fundo da coluna T-02
P_{10}	Pressão da corrente de fundo da coluna T-02
T_{10}	Temperatura da corrente de fundo da coluna T-02
$Qcond_2$	Calor trocado no condensador da coluna T-02
$Qref_2$	Calor trocado no refervedor da coluna T-02
DP_2	Diferença de pressão na coluna T-02

Os dados simulados com as redes neuronais possuíam erros de simulação, por consequência, há uma incidência de *outliers* que devem ser identificados e removidos. A metodologia inicia-se com a normalização dos dados, atribuindo média zero e desvio padrão unitário. Em seguida, com os dados normalizados, aplica-se PCA para uma visualização das amostras e já se pode usar a matriz de covariâncias para estimar o vetor T^2 de Hotelling. A variância explicada em cada componente principal, bem como a variância acumulada nos componentes pode ser visualizada no gráfico da Figura 6.34.

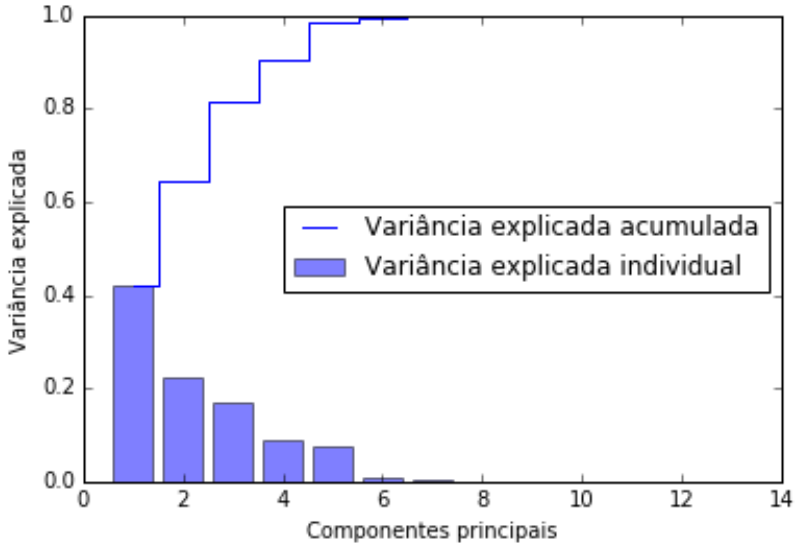

Figura 6.34: Variância explicada acumulada e individual para as variáveis de entrada da coluna T-02.

O gráfico da Figura 6.35 mostra os dados espalhados nos primeiros dois componentes principais; esses componentes explicam 65,4% da variância dos dados.

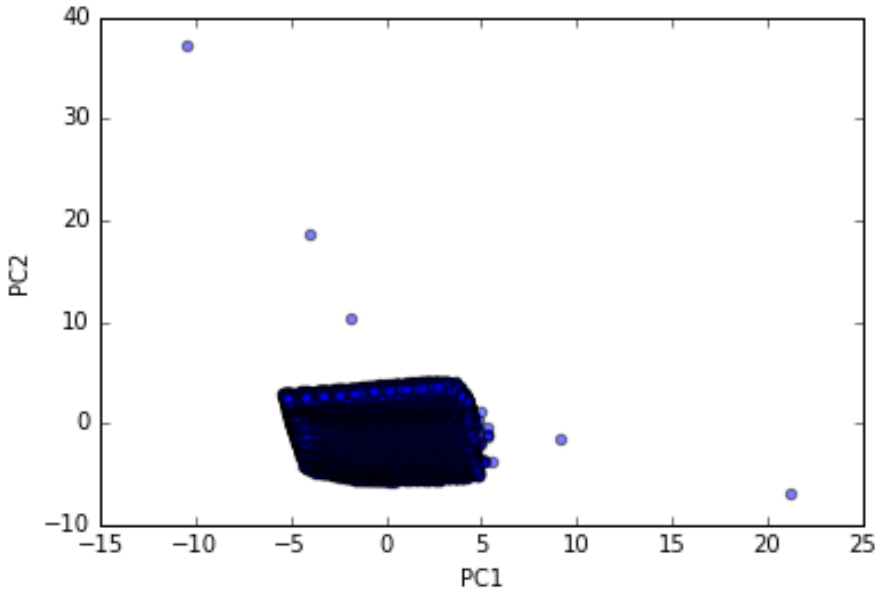

Figura 6.35: Visualização das amostras da coluna T-02 espalhadas nos primeiros dois componentes principais.

É notória a presença de dados anômalos no gráfico acima. Novamente utilizando o método T² de Hotelling, os dados anômalos foram identificados. Para T_{α}^{2}, foram usados os valores α = 1%, N = 414720 (amostras), l = 12 (variáveis). Para a visualização dos resultados do método T² de Hotelling, foi necessário colocar os resultados na escala logarítmica, visto que os picos dos *outliers* eram valores muito elevados e dificultariam a visualização gráfica. O gráfico da Figura 6.36 mostra a identificação dos *outliers* pelo método.

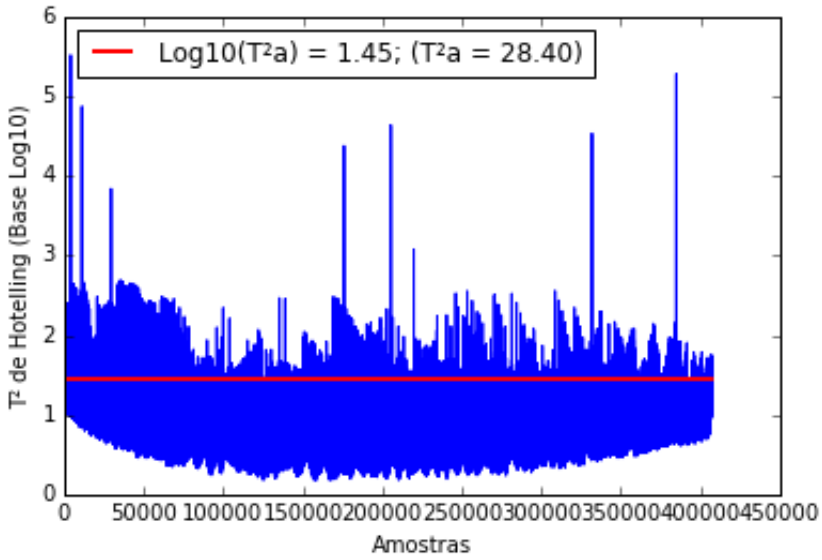

Figura 6.36: Gráfico do T² versus as amosrtras da coluna T-02.

O método selecionou 21903 pontos anômalos, que correspondem a pouco mais de 5% do total do conjunto de dados. Para verificar o novo conjunto de dados, agora com a remoção dos *outliers*, fez-se um novo *plot* das amostras espalhadas no PC1 e PC2.

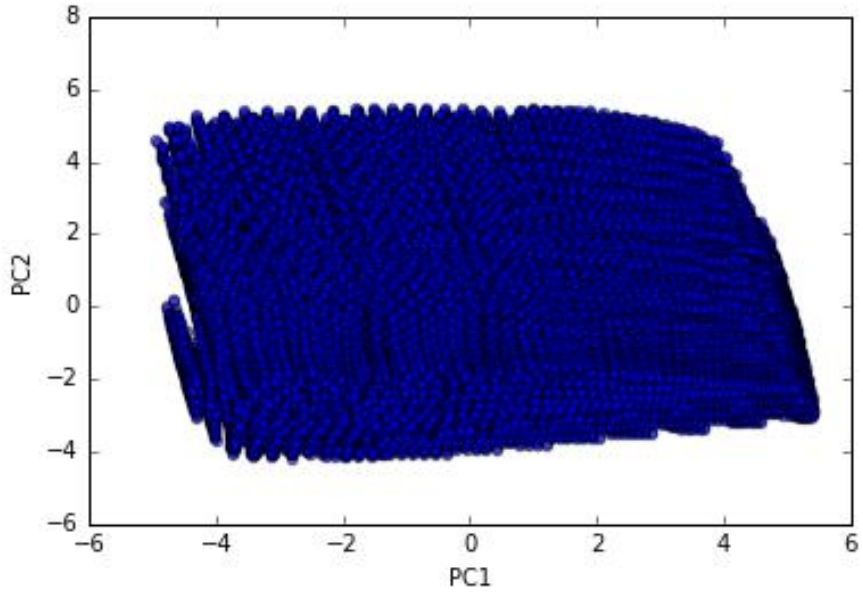

Figura 6.37: Visualização das amostras da coluna T-02 livres de *outliers* espalhadas nos primeiros dois componentes principais.

Os dados agora estão livres de *outliers*, consequentemente o modelo não será prejudicado por conta da presença de dados incoerentes. Portanto, os dados estão prontos para seguir adiante na metodologia e fazer a seleção dos conjuntos que serão utilizados para calibração e teste.

6.2.2 Segregação de Dados

A segregação dos dados nos conjuntos de calibração e teste se dará pelo método *k-rank*. Para isso, utiliza-se a análise de silhueta (*sillouette analysis*), que analisa a qualidade dos *clusters* gerados, para k = 2, 3, ..., 19, 20. Os valores dos coeficientes de silhueta podem ser vistos no gráfico da Figura 6.38.

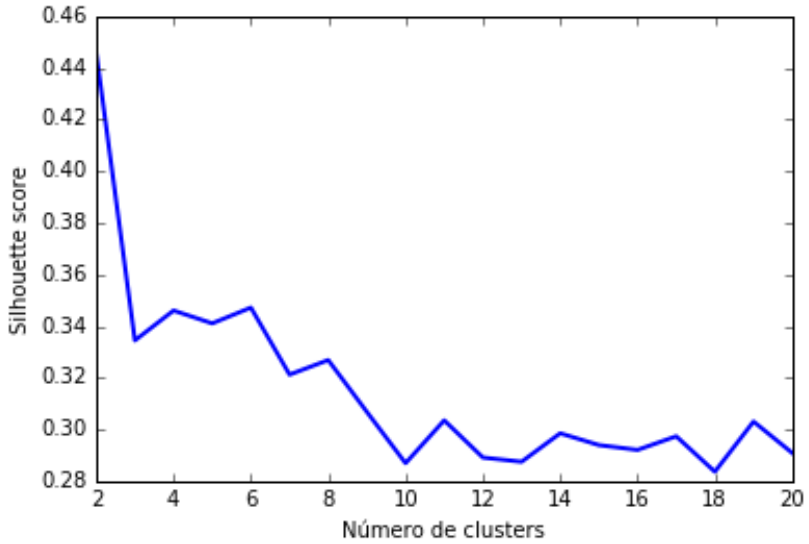

Figura 6.38: Valores dos coeficientes de silhueta para *k* = 2, 3, ..., 19, 20, para a coluna T-02.

O gráfico mostra que aumentando o número de *clusters* não há uma tendência de melhorar o agrupamento dos dados. Logo, o melhor agrupamento utilizando *k-means* se

dá para k = 2, com o valor do coeficiente de silhueta de S_2 = 0,44. A Figura 6.39, ilustra a separação dos dois *clusters* sendo mostrados nos eixos PC1 e PC2.

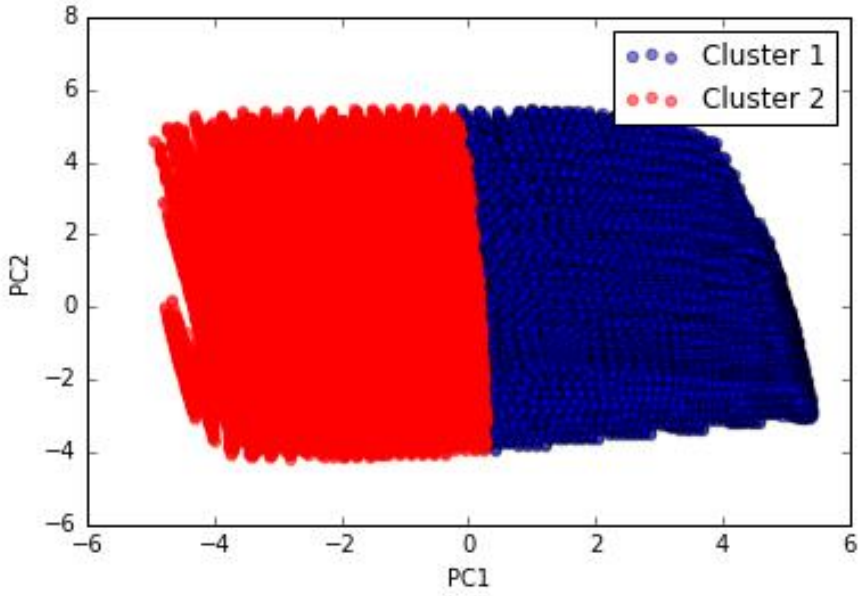

Figura 6.39: Visualização das amostras da coluna T-02 agrupadas pelo *k-means* com *k* = 2.

Com os dados separados nos dois *clusters*, aplica-se o método *y-rank* em cada *cluster* gerado.

6.2.3 Inferindo a Concentração de Propeno na Corrente de Topo da Coluna T-02 (Modelo com Expansão Polinomial de Ordem 3)

Após a remoção dos *outliers*, restaram 392817 amostras em 12 variáveis de entrada. Com a expansão polinomial de ordem 3, aumentou-se de 12 para 454 variáveis. Então, utilizando *k-rank* para a seleção dos conjuntos de calibração e teste, foi necessário apenas 0,5% dos dados para a calibração, pois essa quantidade (1965 amostras) já foi suficiente para descrever o restante dos dados. Em seguida, construiram-se modelos com o conjunto de calibração, utilizando as metodologias de seleção de variáveis. A Tabela 6.16 mostrará os resultados das métricas de avaliação para o conjunto de calibração e a Tabela 6.17 para o conjunto de teste.

Tabela 6.16: Resultados dos critérios de avaliação para o conjunto de calibração – inferência de propeno no topo da coluna T-02.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a Expansão Polinomial*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	12	42	0,998	0,0097	6,52%	101%	-5798

LASSOLARS	12	105	0,999	0,0030	1,61%	21,50%	-7551
Ridge	12	454	0,999	0,0023	1,37%	20,37%	-5654
ACO Plus	12	32	0,999	0,0074	3,61%	38,15%	-6372

Tabela 6.17: Resultados dos critérios de avaliação para o conjunto de teste – inferência de propeno no topo da coluna T-02.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,989	0,0171	2,43%	103,8%
LASSOLARS	0,996	0,0097	1,05%	40,6%
Ridge	0,997	0,0087	0,95%	33,53%
ACO Plus	0,991	0,0162	2,11%	45,6%

As Figuras 6.40, 6.41, 6.42 e 6.43 mostram os valores preditos versus os valores reais para os conjuntos de calibração e teste para os métodos utilizados.

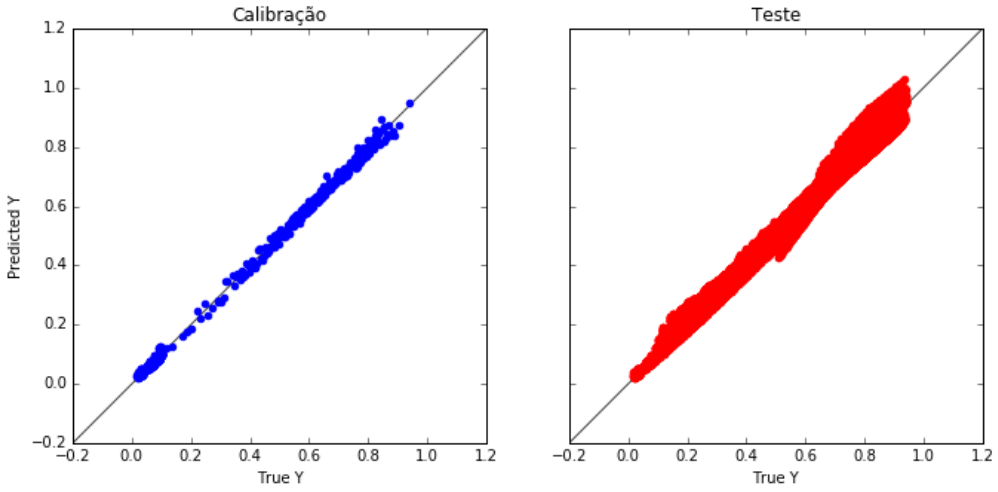

Figura 6.40: Predição do modelo criado pelo método *LASSO*– inferência de propeno no topo da coluna T-02.

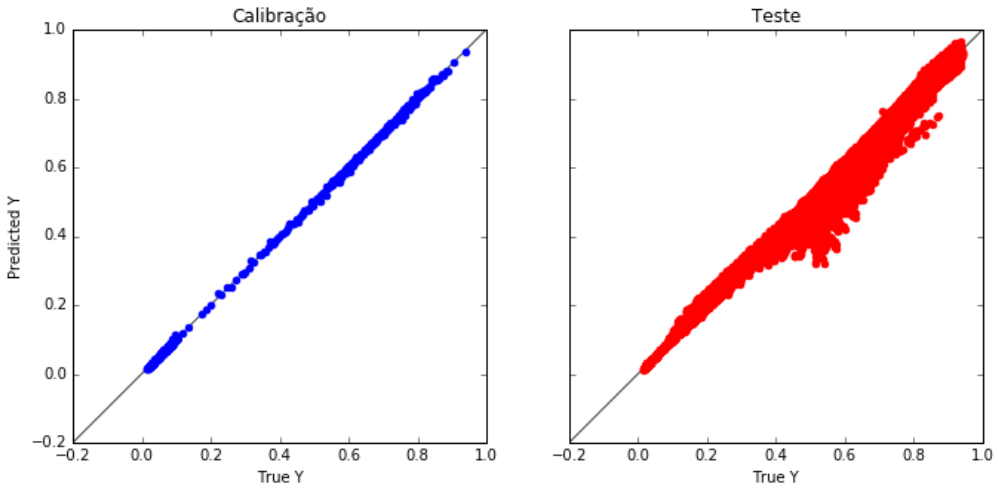

Figura 6.41: Predição do modelo criado pelo método *LASSOLARS*– inferência de propeno no topo da coluna T-02.

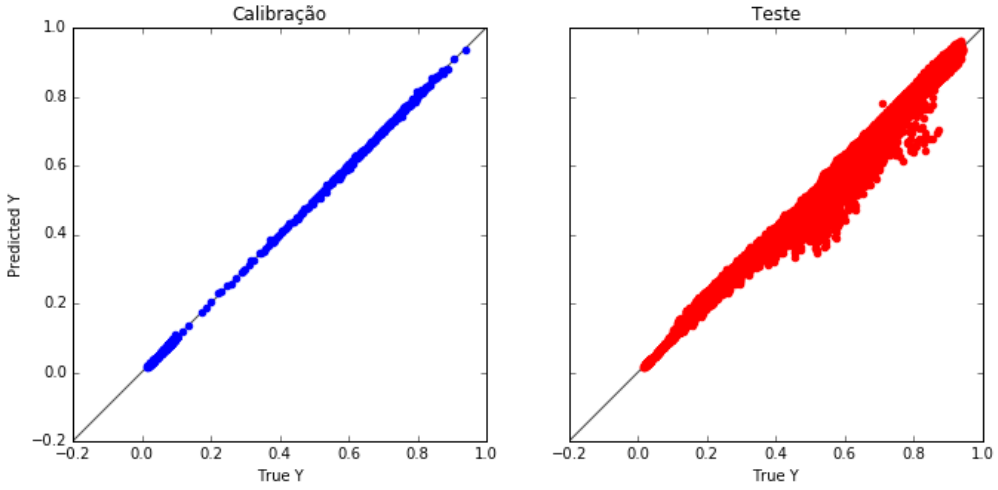

Figura 6.42: Predição do modelo criado pelo método *Ridge*– inferência de propeno no topo da coluna T-02.

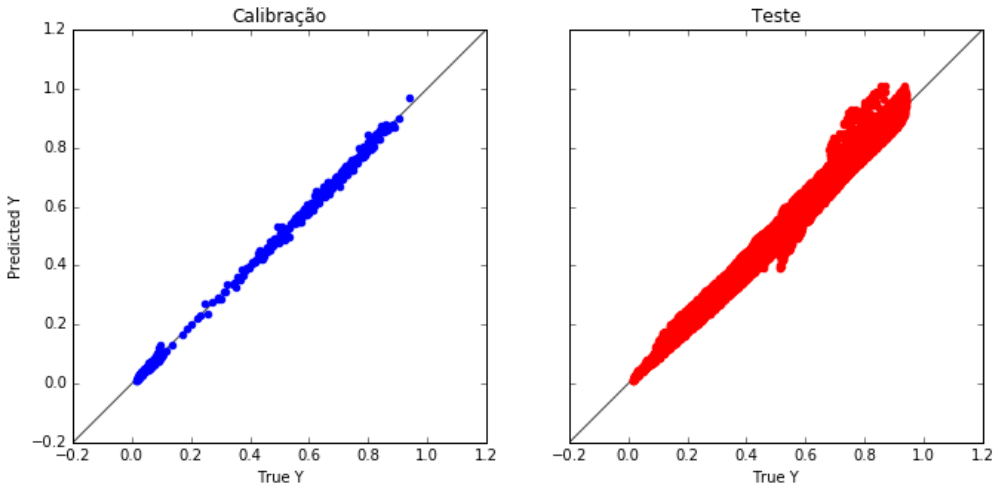

Figura 6.43: Predição do modelo criado pelo método *ACO Plus*– inferência de propeno no topo da coluna T-02.

Nota-se, pelos gráficos gerados, que os modelos criados pelos métodos *Ridge* e *LASSOLARS* transferiram o erro para a região de maiores concentrações e por isso os erros relativos desses métodos foram menores. Os modelos criados pelo *LASSO* e pelo *ACO Plus* distribuíram melhor seus erros, porém obtiveram um erro máximo percentual elevado. Como a intenção é manter essa concentração baixa, pode-se optar pelos modelos construídos pelos métodos *Ridge* e *LASSOLARS*.

6.2.4 Inferindo a Concentração de Propeno na Corrente de Fundo da Coluna T-02 (Modelo com Expansão Polinomial de Ordem 3)

Seguindo a mesma metodologia para a inferência da concentração do propeno no topo da coluna T-02, desenvolveu-se a inferência da concentração de propeno no fundo dessa mesma coluna. Essa concentração será uma informação útil para a coluna T-03. Para resumir os resultados, optou-se por utilizar apenas o método *LASSO* como regressor.

Como o tratamento dos dados já foi feito para a inferir a concentração de propeno no topo, será mostrado apenas os resultados para o conjunto de calibração e teste (*k-rank*; k = 2; 0,5% para calibração e o restante para teste) do modelo desenvolvido pelo método *LASSO* para inferir a concentração de propeno no fundo da coluna T-02. A Tabela 6.18 mostra os resultados das métricas de avaliação, e os gráficos da Figura 6.19 apresenta os valores preditos versus os valores reais, tanto para o conjunto de calibração como para o conjunto de teste.

Tabela 6.18: Resultados das métricas de avaliação para o método *LASSO* – inferência de propeno no fundo da coluna T-02.

LASSO (10 variáveis; 27 variáveis expandidas selecionadas)	*Nº de Amostras*	*R^2*	*RMSE*	*MAPE*	*Max e%*
Calibração	1965	0,999	0,0015	0,14%	1,31%
Teste	390852	0,999	0,0017	0,18%	2,63%

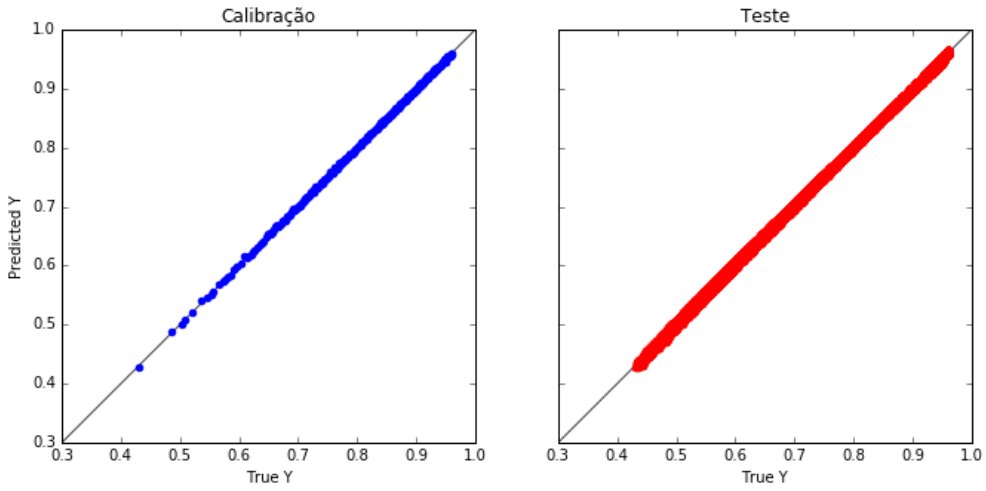

Figura 6.44: Predição do modelo criado pelo método *LASSO* – inferência de propeno no fundo da coluna T-02.

6.3 Coluna T-03

Na coluna T-03, é importante controlar a especificação da concentração de propeno no topo da coluna. Para isso, faz-se necessário inferir a concentração de propano, componente que torna impura a corrente de propeno. Portanto, a tarefa agora é extrair a informação dessa concentração a partir das variáveis de processo disponíveis, bem como das concentrações de propeno e propano que estão vindo da coluna T-02, pois essas concentrações foram inferidas com ótima precisão e podem ser utilizadas como variáveis de entrada para a coluna T-03. Os valores da média, desvio padrão, mínimo e máximo, 1º, 2º e 3º quartis da concentração de propano no topo da coluna T-03, para os dados disponíveis estão dispostos na Tabela 6.19.

Tabela 6.19: Descricação da variável de saída Z_{C3+}^{15} (concentração de propano no topo da coluna T-03).

Variável	*Nº de Amostras*	*Média*	*Desvio Padrão*	*Mínimo*	*25%*	*50%*	*75%*	*Máximo*
Z_{C3+}^{15}	1764	0,1094	0,1064	1e-6	0,0029	0,0692	0,1953	0,3292

6.3.1 Pré-Processamento dos Dados da Coluna T-03

Para a coluna T-03, as variáveis disponíveis que podem ser utilizadas como variáveis de entrada estão dispostas na tabela 6.20. Já aqui, pode-se concluir que a matriz de entrada terá dimensão 1764x17, que são 1764 amostras representadas pelas linhas da matriz versus 17 colunas que são as variáveis de entrada. A concentração de propeno será utilizada também como entrada, pois foi inferida satisfatoriamente no fundo da coluna T-02.

Tabela 6.20: Variáveis, da coluna T-03, que podem ser utilizadas como entrada para o modelo.

Variável	*Descrição*
Z_{C3-}^{10}	Concentração de propeno na corrente de entrada da coluna T-03
FR_3	Fração da corrente que sai do refervedor (corrente 14) que retorna para a coluna como refluxo para a coluna
FA_3	Fração da corrente que sai do compressor que será utilizada como fluido de aquecimento do refervedor
F_{11}	Vazão de fundo da coluna T-03
P_{11}	Pressão da corrente de fundo da coluna T-03
F_{15}	Vazão de topo da coluna T-03
T_{15}	Temperatura da corrente de topo da coluna T-03
F_{14}	Vazão que sai do refervedor (corrente 14) e retorna para a coluna como refluxo
F_{14}	Vazão que sai do condensador (corrente 16) e retorna para a coluna como refluxo
$W_{C\text{-}01}$	Energia requerida pelo compressor (C-01)
$W_{B\text{-}04}$	Energia requerida pela bomba (B-04)
Q_{P06}	Energia requerida pelo condensador P-06
Q_{P07}	Energia requerida pelo trocador P-07
T_{TOPO}	Temperatura no topo da coluna T-03
T_{FUNDO}	Temperatura no fundo da coluna T-03
P_{FUNDO}	Pressão no fundo da coluna T-03
V_{REF}	Fração vaporizada que sai do refervedor

Os dados simulados com as redes neuronais possuíam erros de simulação, por consequência, há uma incidência de *outliers* que devem ser identificados e removidos. A metodologia inicia-se com a normalização dos dados, atribuindo média zero e desvio padrão unitário. Em seguida, com os dados normalizados, aplica-se PCA para uma visualização das amostras e já se pode usar a matriz de covariâncias para estimar o vetor T^2 de Hotelling. A variância explicada em cada componente principal, bem como a variância acumulada nos componentes pode ser visualizada no gráfico da Figura 6.45.

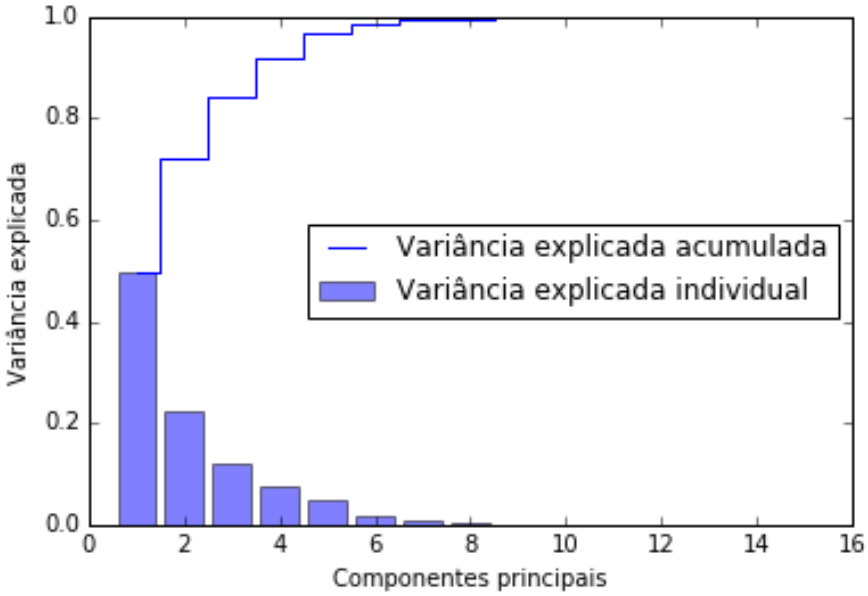

Figura 6.45: Variância explicada acumulada e individual para as variáveis de entrada da coluna T-03.

O gráfico da figura 6.46 mostra os dados espalhados nos primeiros dois componentes principais; esses componentes explicam 72,0% da variância dos dados.

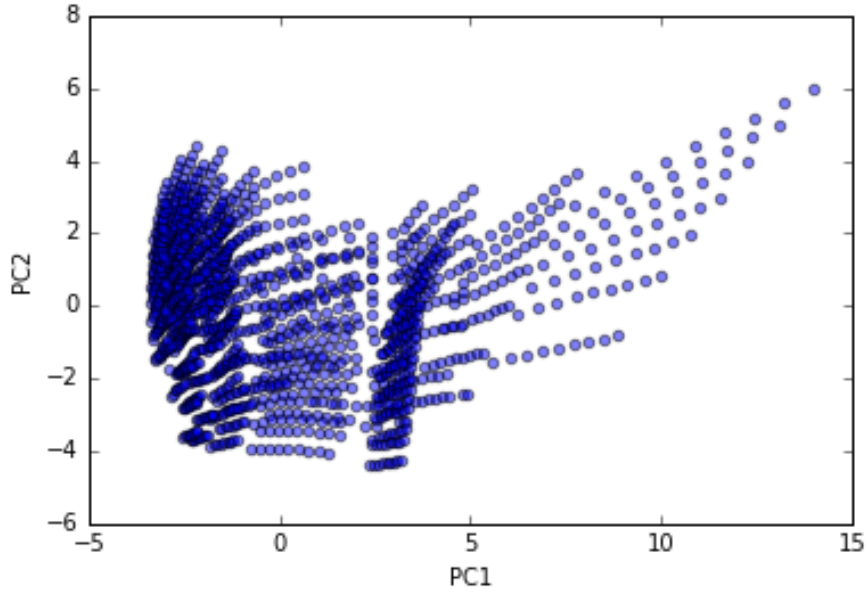

Figura 6.46: Visualização das amostras da coluna T-03 espalhadas nos primeiros dois componentes principais.

Dessa vez não ficou tão evidente a presença de *outliers.* Porém, utilizando o método T² de Hotelling, foram identificados alguns *outliers*. Para T_{α}^{2}, foram usados os valores α = 1%, N = 1764 amostras, l = 17 variáveis. O gráfico da Figura 6.47 mostra a identificação dos *outliers* pelo método.

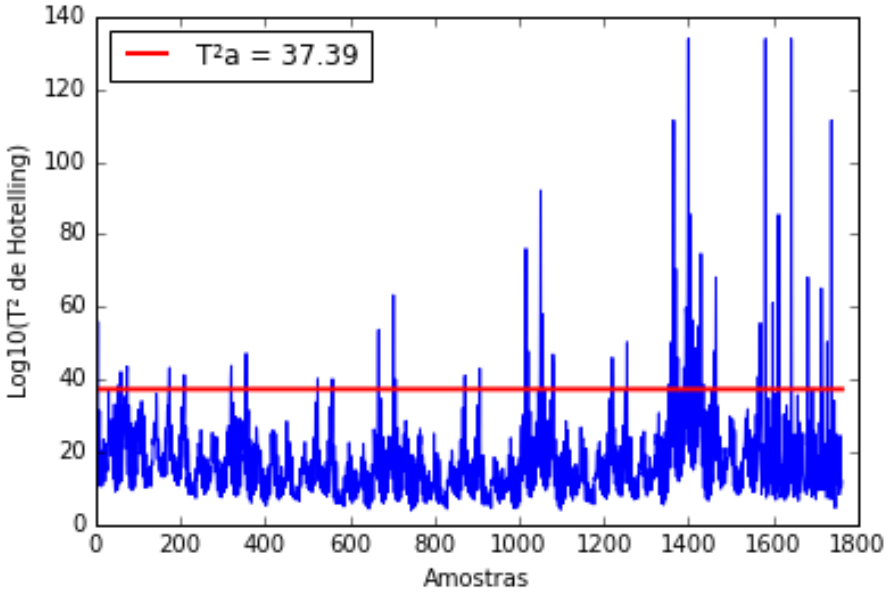

Figura 6.47: Gráfico do T^2 versus as amosrtras da coluna T-03.

O método selecionou 85 pontos anômalos, que correspondem a pouco menos de 5% do total do conjunto de dados. Para verificar o novo conjunto de dados, agora com a remoção dos *outliers*, fez-se um novo *plot* das amostras espalhadas no PC1 e PC2.

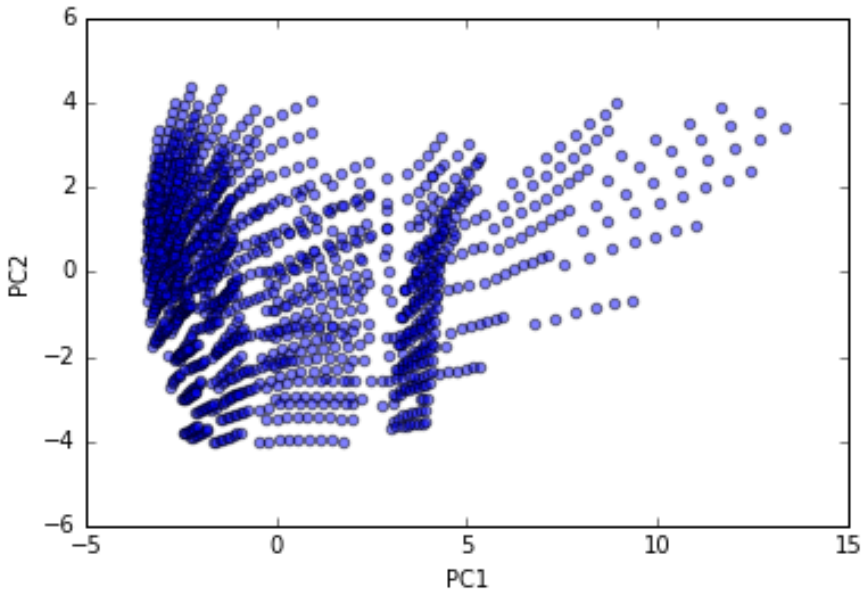

Figura 6.48: Visualização das amostras da coluna T-03 livres de *outliers* espalhadas nos primeiros dois componentes principais.

Os dados agora estão livres de *outliers* e estão prontos para seguir adiante na metodologia e fazer a seleção dos conjuntos que serão utilizados para calibração e teste.

6.3.2 Segregação de Dados

Novamente se avaliou a qualidade dos *clusters* utilizando a análise de silhueta. Os resultados estão no gráfico da figura 6.49. O melhor agrupamento ocorre para k = 2, com

o valor do coeficiente de silhueta de S_2 = 0,41. A figura 6.50 mostra o agrupamento feito pelo *k-means*.

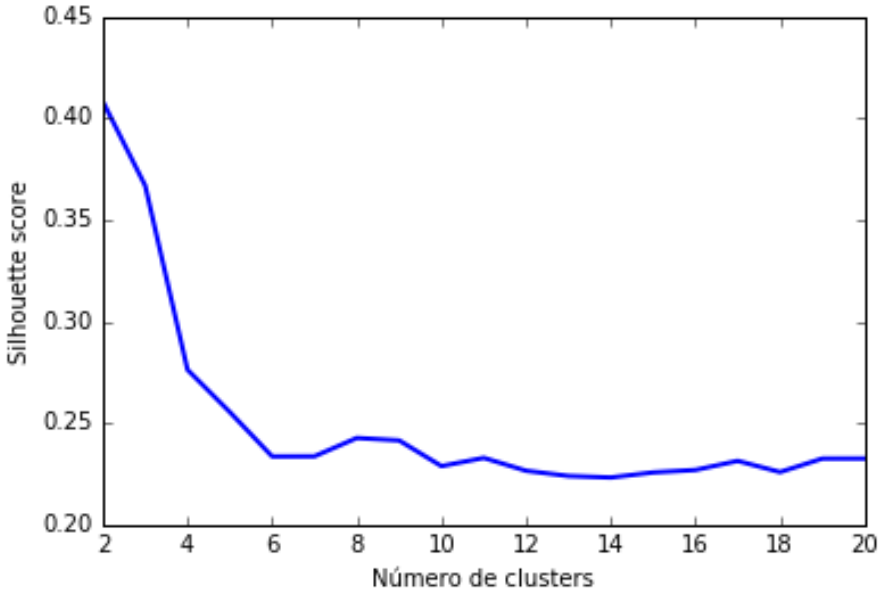

Figura 6.49: Valores dos coeficientes de silhueta para *k* = 2, 3, ..., 19, 20, para a coluna T-03.

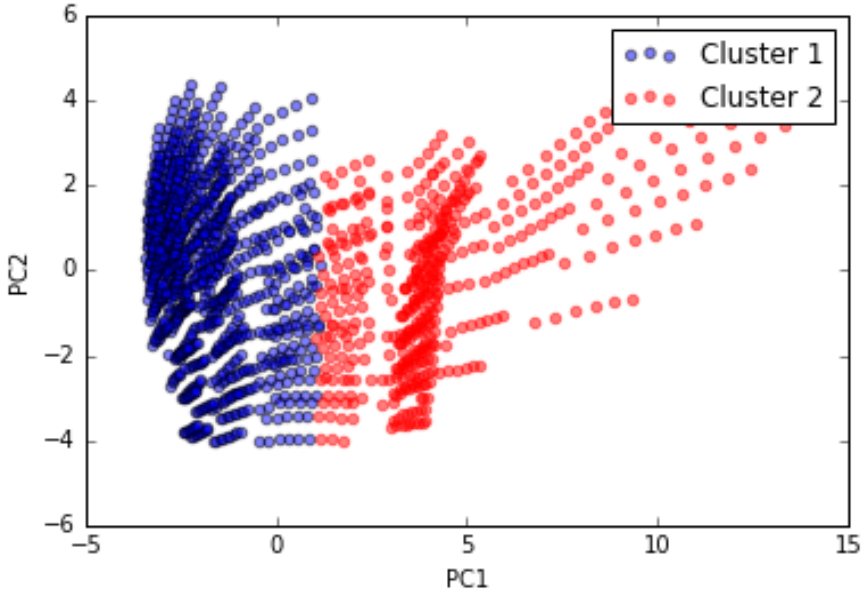

Figura 6.50: Visualização das amostras da coluna T-03 agrupadas pelo *k-means* com *k* = 2.

6.3.3 Inferindo a Concentração de Propeno na Corrente de Topo da Coluna T-03 (Modelo com Expansão Polinomial de Ordem 3)

Para essa coluna, será necessário adotar uma estratégia diferente das outras colunas. A concentração de propano no topo da coluna T-03 é extremamente baixa. A especificação da planta é manter um grau de pureza de no mínimo 99,6% de propeno no topo, isto é, 0,4% (0,004 kg/kg) apenas de propano. Para inferir o propano, será necessária novamente

a divisão dos dados para que se trabalhe com modelos locais. Além disso, a informação do propeno no topo é imprescindível. Todavia não será utilizado, na calibração do modelo, a informação do propeno real no topo da coluna, mas sim a informação do propeno estimado no topo da coluna. Isto é, um modelo servirá de informação para o outro modelo. Essa foi a alternativa para conseguir estimar essa impureza utilizando a metodologia proposta.

Comecemos então por estimar a concentração do propeno no topo. O conjunto de dados possui 1679 amostras (livres dos *outliers*) e 17 variáveis. Com a expansão polinomial de ordem 3, esse conjunto passou a ter 1139 variáveis, sendo as 17 iniciais e mais suas combinações. Para a segregação dos dados, será utilizado o método *k-rank,* com *k* = 2 e uma proporção de 2:1 para calibração e teste. Em seguida, será estimado o propeno utilizando a metodologia *LASSOLARS*, a qual obteve os melhores resultados para essa variável e dispensou a necessidade de utilizar o *ACO Plus*, visto que selecionou apenas 5 variáveis e 6 variáveis expandidas. Seguem os resultados dispostos na tabela 6.21 e a visualização dos valores preditos versus os valores reais na figura 6.51.

Tabela 6.21: Resultados das métricas de avaliação para o método *LASSOLARS* – inferência de propeno no topo da coluna T-03.

LASSOLARS (5 variáveis; 6 variáveis expandidas selecionadas)	*Nº de Amostras*	*R^2*	*RMSE*	*MAPE*	*Max e%*
Calibração	1119	0,999	0,0019	0,18%	0,78%
Teste	560	0,999	0,0019	0,17%	0,80%

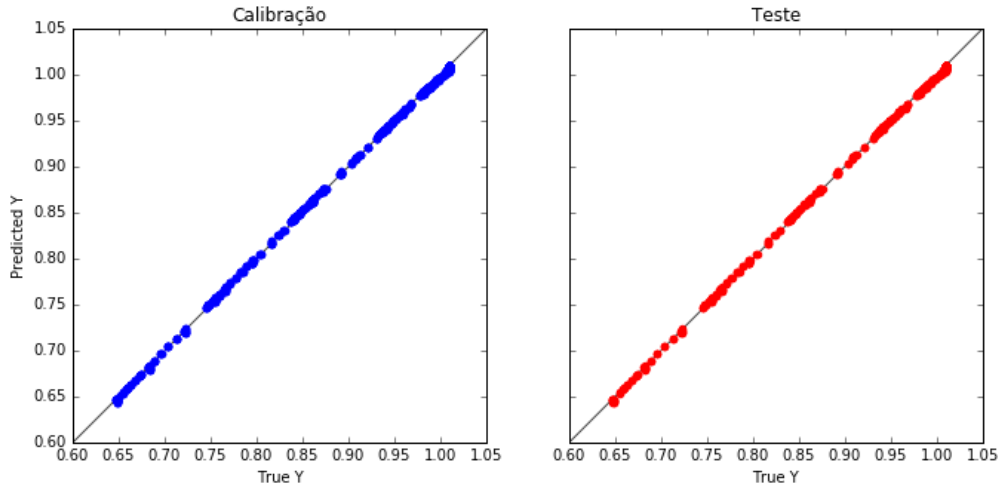

Figura 6.51: Predição do modelo criado pelo método *LASSOLARS* – inferência de propeno no topo da coluna T-03.

6.3.4 Inferindo a Concentração de Propano na Corrente de Topo da Coluna T-03 (Modelo com Expansão Polinomial de Ordem 3)

Com o modelo criado para inferir a concentração de propeno no topo da coluna, pode-se utilizar a informação do modelo como entrada para o modelo que irá inferir a concentração do propano no topo dessa mesma coluna. Mas antes é necessário um modelo

que classifique as regiões com concentrações maiores de propano e de menores concentrações. Para isso, será utilizado o modelo criado pelo método *Ridge*, pois este teve 100% de acuracidade na identificação das regiões. O limite entre as regiões será semelhante ao utilizado para os pesados no topo da coluna T-01: 0,01 kg/kg. Então será construído um modelo local para concentrações acima de 0,01 kg/kg e outro para concentrações abaixo disso. Segue abaixo o gráfico dos valores preditos e reais em função das amostras para o conjunto de calibração e teste. O modelo construído com o método *Ridge* utiliza todas as variáveis e suas combinações na regressão, porém são penalizadas as variáveis menos importantes ao modelo, de modo que não ocorre *overfitting.* Em seguida, a tabela 6.22 mostra os valores máximos e mínimos, preditos e reais, do limite entre as regiões.

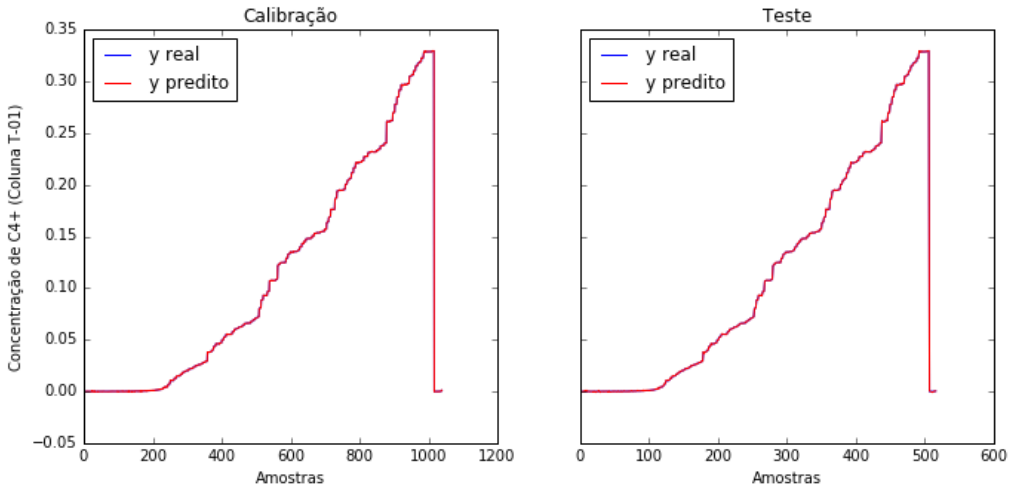

Figura 6.52: Aderência do modelo criado pelo método *Ridge* às amostras de propano no topo da coluna T-03.

Tabela 6.22: Valores máximos e mínimos, preditos e reais, do limite entre as regiões.

	Concentrações < 0,01 kg/kg		*Concentrações > 0,01 kg/kg*	
	Máximo Real	Máximo Predito	Mínimo Real	Mínimo Predito
Calibração	0,0074576	0,0076278	0,0102697	0,0104679
Teste	0,0077484	0,0079098	0,0103732	0,0106792

Pelos valores da tabela acima, pode-se concluir que o modelo consegue separar essas duas regiões perfeitamente. Embora o gráfico da figura 6.52 faça parecer que o modelo não está errando demasiadamente, o erro percentual médio absoluto é da ordem de 140% para calibração e 230% para o teste. Isso está ligado ao fato de que se o modelo fizer uma predição de 0,0002 e o valor real for 0,0001, quer dizer que o modelo errou 100%. Espera-

se, entretanto, que trabalhando com modelos locais, melhore a predição na região de concentrações muito baixas.

6.3.5 Inferindo a Concentração de Propano em Concentrações Superiores à 0,01 kg/kg (Modelo com Expansão Polinomial de Ordem 3)

O conjunto de dados para a região de concentrações superiores à 0,01 kg/kg possui 1140 amostras e utilizou-se da mesma metodologia proposta com expansão polinomial de ordem 3, ou seja, o conjunto de dados ficou com dimensão 1140*x*1329, sendo, dessas 1329, 18 variáveis e suas expansões. A segregação dos dados foi feita utilizando o método *k-rank* com $k = 2$ e proporção 2:1 para calibração e teste. Com esse conjunto de dados, fez-se a regressão utilizando os métodos de seleção de variáveis e os resultados das avaliações dos métodos estão dispostos na tabela 6.23, para calibração, e 6.24, para o conjunto de teste.

Tabela 6.23: Resultados dos critérios de avaliação para o conjunto de calibração.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a Expansão Polinomial*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*
LASSO	7	15	0,999	0,0009	1,54%	27,9%	-8536
LASSOLARS	0	0	0	0,0978	155,8%	1354%	-1390
Ridge	18	1329	0,999	1,5e-4	0,14%	2,6%	-1147
ACO Plus	13	11	0,999	2,2e-4	0,22%	2,8%	-9998

Tabela 6.24: Resultados dos critérios de avaliação para o conjunto de teste.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,999	0,0009	1,48%	26,7%
LASSOLARS	0	0,0973	151,6%	1333%
Ridge	0,999	2,4e-4	0,21%	5,93%
ACO Plus	0,999	2,4e-4	0,20%	2,24%

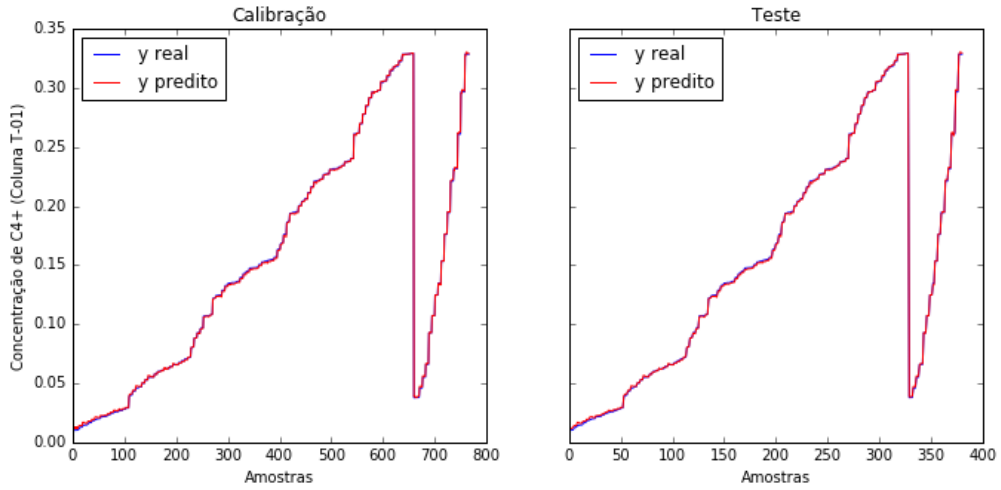

Figura 6.53: Aderência do modelo criado pelo método *LASSO* às amostras de propano no topo da coluna T-03 em concentrações acima de 0,01 kg/kg.

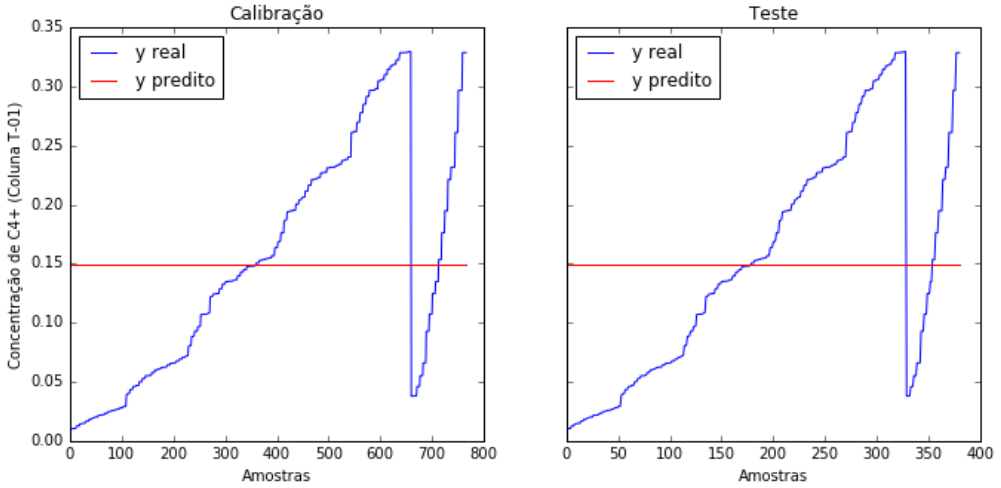

Figura 6.54: Aderência do modelo criado pelo método LASSOLARS às aostras de propano no topo da coluna T-03 emconcentrações acima de 0,01 kg/kg.

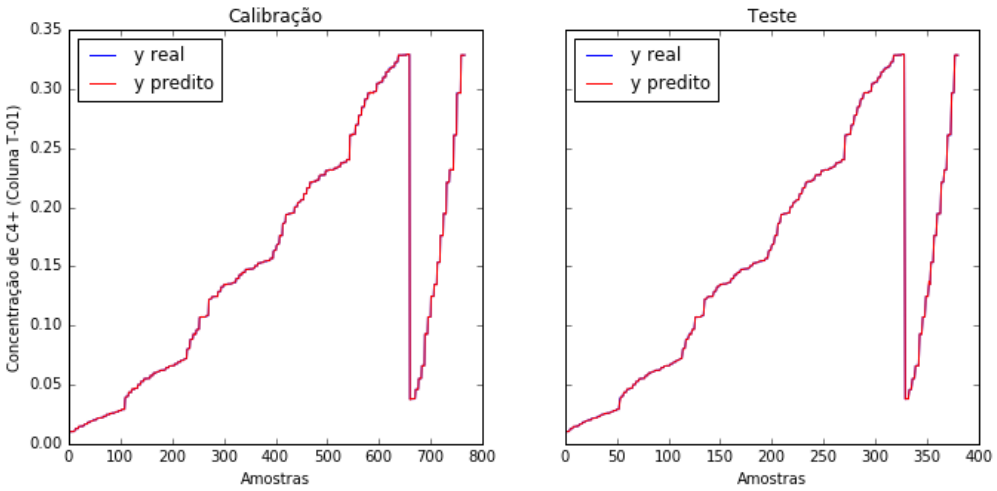

Figura 6.55: Aderência do modelo criado pelo método *Ridge* às amostras de propano no topo da coluna T-03 em concentrações acima de 0,01 kg/kg.

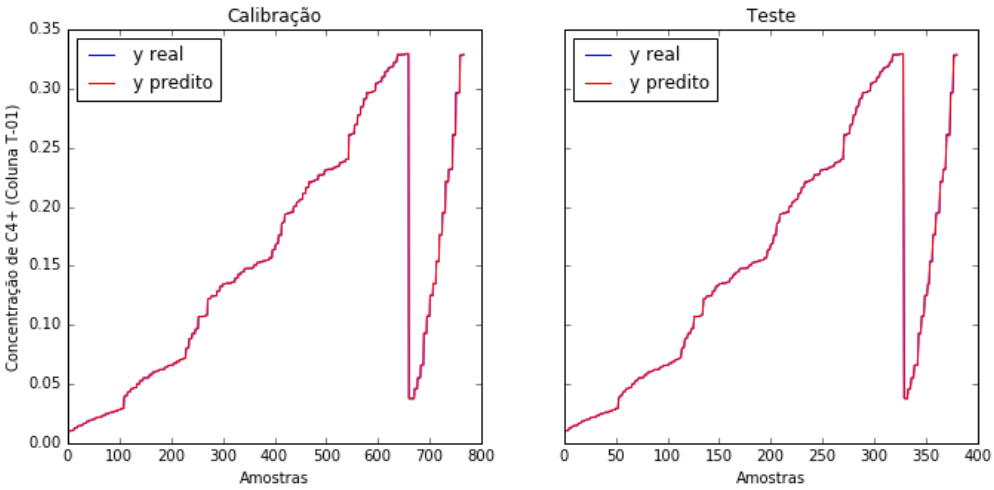

Figura 6.56: Aderência do modelo criado pelo método *ACO Plus* às amostras de propano no topo da coluna T-03 em concentrações acima de 0,01 kg/kg.

6.3.6 Inferindo a Concentração de Propano em Concentrações Inferiores à 0,01 kg/kg (Modelo com Expansão Polinomial de Ordem 3)

O conjunto de dados para a região de concentrações inferiores à 0,01 kg/kg é constituído pelo restante das amostras do conjunto total. A segregação dos dados foi feita utilizando o método *k-rank* com k = 2 e proporção 2:1 para calibração e teste. Com esse conjunto de dados, fez-se a regressão utilizando os métodos de seleção de variáveis e os resultados das avaliações dos métodos estão dispostos na tabela 6.25, para calibração, e 6.26, para o conjunto de teste.

Tabela 6.25: Resultados dos critérios de avaliação para o conjunto de calibração.

Método	*Quantidade de Variáveis Selecionadas Originais*	*Quantidade de Variáveis Selecionadas Após a*	*R^2*	*RMSE*	*MAPE*	*Max e%*	*BIC*

		Expansão Polinomial					
LASSO	18	117	0,94	3,6e-4	583%	14165%	-2920
LASSOLARS	0	0	0	0,0015	3607%	61188%	-2813
Ridge	18	1329	0,999	4,7e-5	120%	2948%	-3945
ACO Plus	10	17	0,999	1,5e-5	48,8%	1400%	-5220

Tabela 6.26:Resultados dos critérios de avaliação para o conjunto de teste.

Método	*R^2*	*RMSE*	*MAPE*	*Max e%*
LASSO	0,949	0,0003	530%	10298%
LASSOLARS	0	0,0013	2898%	33122%
Ridge	0,996	8,5e-5	180%	6336%
ACO Plus	0,999	2,1e-5	83%	2910%

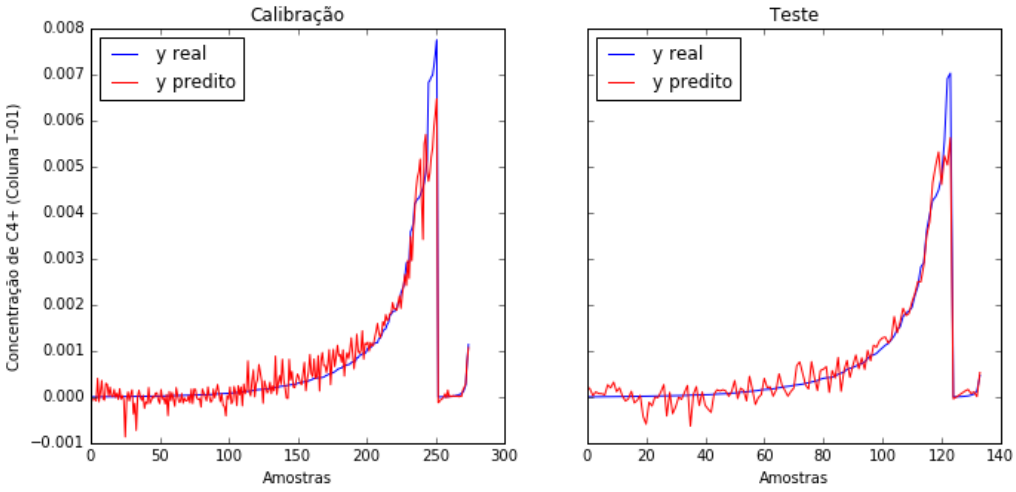

Figura 6.57: Aderência do modelo criado pelo método *LASSO* às amostras de propano no topo da coluna T-03 em concentrações abaixo de 0,01 kg/kg.

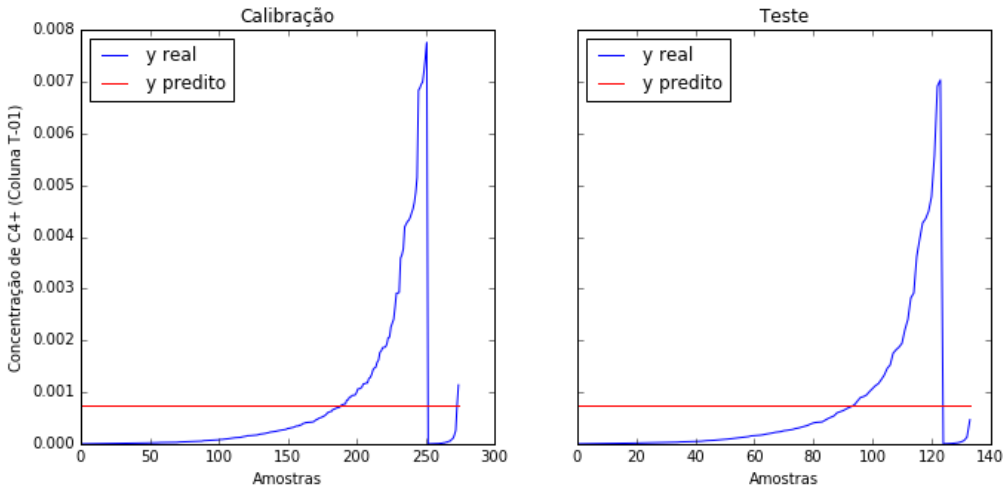

Figura 6.58: Aderência do modelo criado pelo método *LASSOLARS* às amostras de propano no topo da coluna T-03 em concentrações abaixo de 0,01 kg/kg.

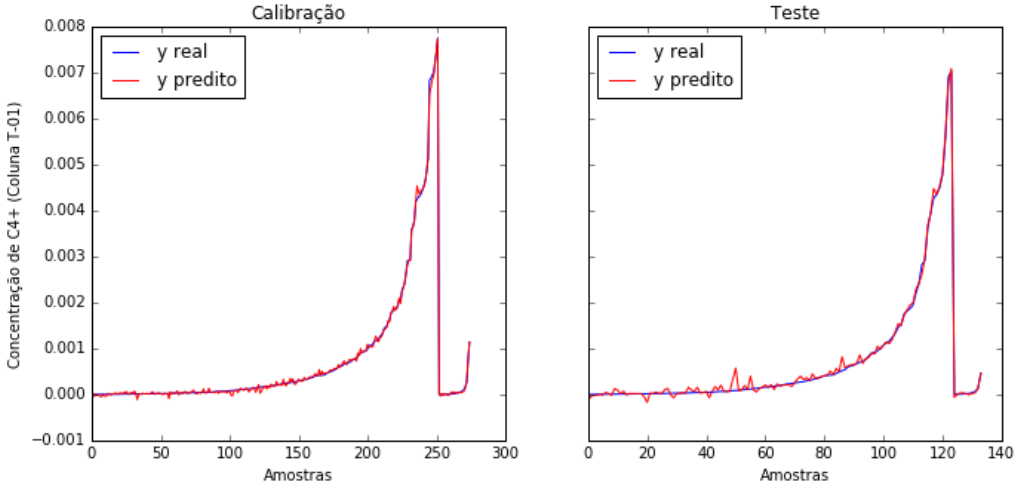

Figura 6.59: Aderência do modelo criado pelo método *Ridge* às amostras de propano no topo da coluna T-03 em concentrações abaixo de 0,01 kg/kg.

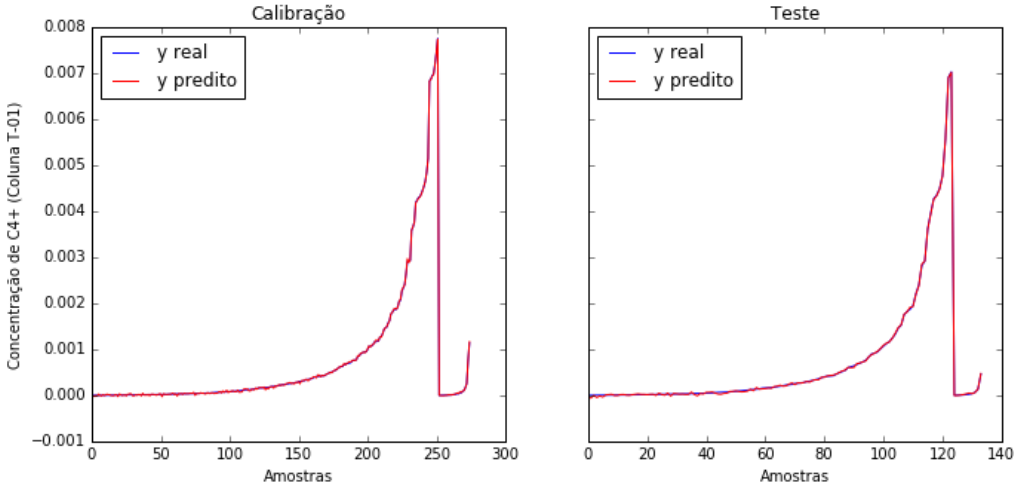

Figura 6.60: Aderência do modelo criado pelo método *ACO Plus* às amostras de propano no topo da coluna T-03 em concentrações abaixo de 0,01 kg/kg.

Mesmo com a utilização dos modelos locais, o erro ainda fica demasiado para a região de baixas concentrações. Entretanto, há uma região de confiança na estimativa do propano, mesmo que em baixas concentrações. O erro relativo cresce à medida que a concentração de propano se aproxima de zero. Para analisar esse fato, foi escolhido o modelo criado pelo método *ACO Plus*, o qual obteve os melhores resultados e foram feitos gráficos dos erros relativos em função das amostras para valores da concentração maiores e menores que 0,0002 kg/kg.

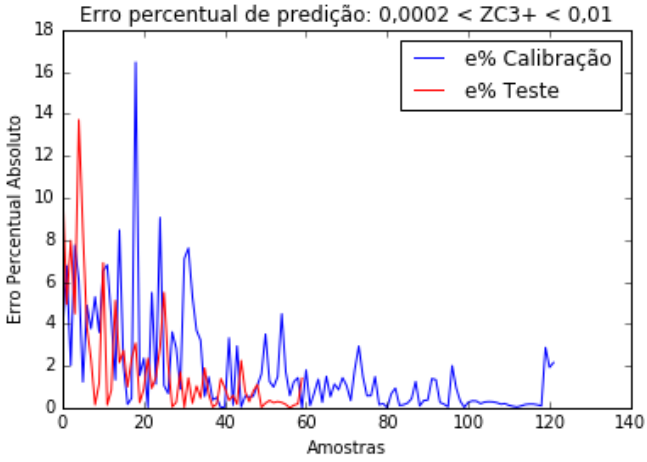

Figura 6.61: Erro percentual de predição do modelo criado pelo método *ACO Plus* para 0,0002 kg/kg <Z_{C3+}^{15}< 0,01 kg/kg.

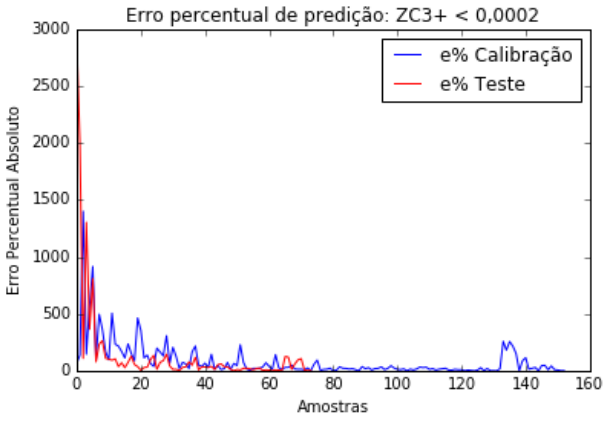

Figura 6.62: Erro percentual de predição do modelo criado pelo método *ACO Plus* para Z_{C3+}^{15}<0,0002 kg/kg.

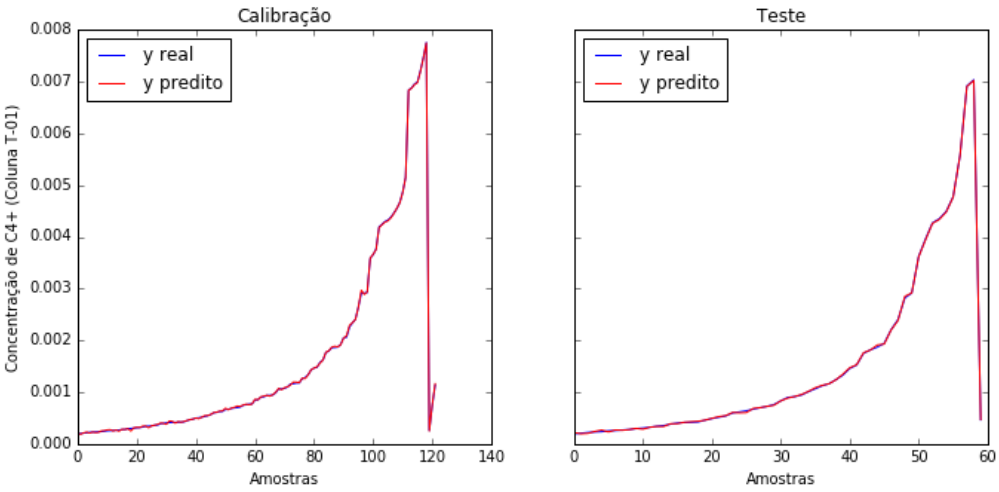

Figura 6.63: Aderência do modelo criado pelo método *ACO Plus* às amostras de propano no topo da coluna T-03 para 0,0002 kg/kg <Z_{C3+}^{15}< 0,01 kg/kg.

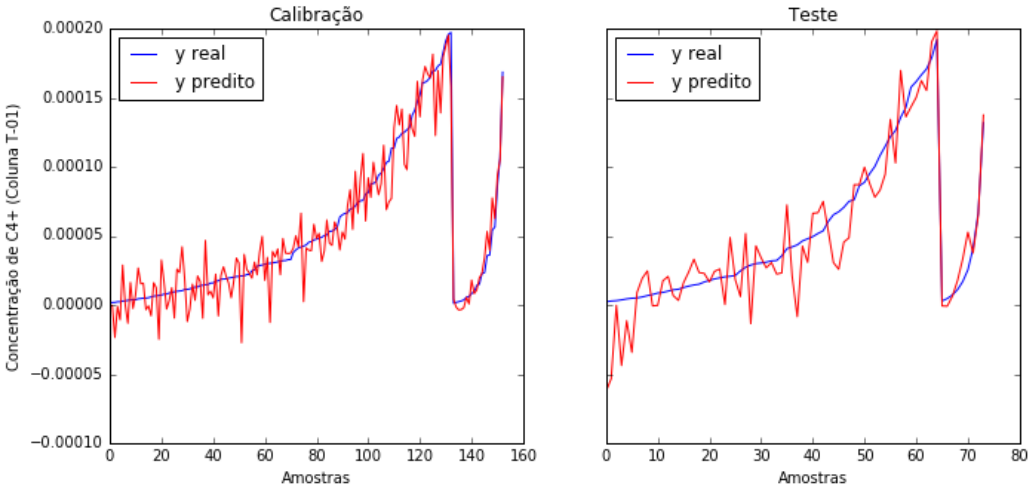

Figura 6.64: Aderência do modelo criado pelo método *ACO Plus* às amostras de propano no topo da coluna T-03 para Z_{C3+}^{15}< 0,0002 kg/kg.

Os gráficos mostram que para concentrações acima de 0,0002 kg/kg de propano o erro é baixo, sendo o erro máximo de aproximadamente 16%. Portanto o crescimento do erro máximo e do erro médio percentual ocorre devido às estimativas para concentrações menores que 0,0002 kg/kg de propano. Isto é, em concentrações extremamente baixas, onde o propeno já possui uma pureza de, no mínimo, 99,98%. Todavia, como a planta requer uma concentração de 99,6% de propeno, a inferência mostrou-se eficaz nessas condições operacionais e poderia ser utilizada para auxiliar no controle do processo.

6.4 Conclusões

Baseado na metodologia proposta, conclui-se que foi possível estimar as concentrações chave para o controle da unidade. Entretanto, em algumas situações de concentrações muito baixas o erro foi demasiado, como na coluna T-03, para a estimativa de propano no topo. Todavia percebe-se que para concentrações acima de 0,0002 kg/kg de propano no topo dessa coluna, pode-se estimar com um erro máximo de aproximadamente 16%. E como a coluna propõe uma pureza de 99,6% de propeno, o erro

para concentrações próximas de 0,004 kg/kg de propano ainda é menor. Pois quanto mais baixa a concentração de propano, maior o erro da sua estimativa. Já para as colunas T-01 e T-02, as variáveis foram estimadas com boa acuracidade.

Quanto aos métodos utilizados para se construir os modelos, não houve um método que se saiu melhor em todas as situações. Portanto, não se pode defender a utilização de apenas um método, mas sim a comparação deles e a validação do melhor método. Há, porém, uma vantagem em se utilizar o ACO como seletor de variáveis, pois, nesse método, a relação entre qualidade e tamanho de modelo é melhor em relação aos demais. Vale lembrar que o tamanho do modelo é escolhido pelo usuário, enquanto nos demais métodos não.

Além das conclusões acima, alguns aspectos relacionados à apresentação dos resultados podem ser salientados. Alguns resultados foram mostrados apenas para a primeira coluna para que não ficasse repetitivo no decorrer do texto. Por exemplo, para as colunas T-02 e T-03, não foram mostrados os resultados para os modelos lineares, visto que, por essa via, não foram construídos bons modelos.

Outro fato que merece destaque está relacionado ao número de pontos utilizado na calibração da inferência da coluna T-02. Essa coluna possuía um grande número de amostras, porém optou-se por utilizar uma proporção bem pequena para a calibração. Isso é justificado no fato de que os algoritmos ficam mais pesados quando se tem mais amostras para a calibração. Pois a função objetivo dos algoritmos de regressão minimiza o erro das amostras em relação ao modelo criado. Então, quando se tem muitos pontos, se tem mais erros para se calcular, consequentemente mais tempo computacional será gasto. E como a quantidade de pontos escolhida já era suficiente para estimar todo o conjunto de dados, não foi necessário gastar tempo computacional desnecessariamente.

6.5 Manutenção das Inferências

A etapa de manutenção das inferências não será mostrada em resultados, mas explicada como proceder em cada situação. Ficou claro no texto como o método estatístico T^2 de *Hotelling* identifica os *outliers*. Para fins de manutenção, não se pode realizar o mesmo procedimento que foi feito com os dados na etapa de pré-processamento.

O que deve ser feito para dados desconhecidos é uma espécie de predição, tendo em vista que o método T^2 de *Hotelling* foi calibrado para os dados utilizados na construção do modelo. Então, com a inserção de novos dados, a normalização é feita baseada na média e desvio padrão dos dados utilizados na calibração. Então, com o novo dado normalizado, utiliza-se a equação 2.3 para calcular a estimativa T^2, sendo os valores da diagonal dos autovalores e os autovetores iguais aos utilizados com os dados antigos, isto é, com os dados que foram utilizados na construção do modelo. Dessa forma, pode-se identificar dados que não estão na região de confiança da calibração do modelo.

Para o outro caso, em que a estimativa da variável de saída é feita incorretamente, mesmo com os dados dentro da região de confiança, a identificação da falha da inferência é feita por medições laboratoriais ou por analisadores em linha. Isso pode ocorrer por uma variação nas condições operacionais, por um desgaste de catalisador, por exemplo, ou

ainda por falhas nos sensores de medição das variáveis de processo. Pode acontecer de o sensor descalibrar e informar valores incoerentes aos valores reais. Nesse caso, o modelo fará uma predição incorreta. Então pode-se fazer um diagnóstico da falha utilizando as medições laboratoriais ou de analisadores em linha da variável de saída. Com essas medições, faz-se a análise das variáveis de entrada, utilizando o modelo da inferência. Mas agora a variável de saída servirá para identificar o erro numa das variáveis de entrada. Isto é, para aquele dado valor de y, as variáveis de entrada devem respeitar os seus respectivos valores. E, diante de uma falha, algumas das variáveis pode estar incoerente e assim ela pode ser identificada.

Em qualquer uma das situações que será necessário a manutenção do modelo, será necessária a coleta de novos dados. Porém, para a construção de uma nova inferência, a etapa de seleção das variáveis será desnecessária, visto que essa já foi realizada anteriormente. Logo, basta realizar a regressão utilizando as variáveis selecionadas pelo modelo antigo, utilizando os novos dados tratados.

Capítulo 7 – **Considerações Finais**

7.1 Conclusões

A indústria hoje compreende a necessidade por uma produção mais segura, mais limpa e mais eficiente. Consequentemente os sistemas avançados de monitoramento e controle vêm ganhando destaque nos processos industriais. No entanto, algumas variáveis ainda enfrentam problemas quando se fala em medições confiáveis e rápidas. Os analisadores em linha se apresentam como uma alternativa, mas ainda os tempos de análise e amostragem não são apropriados para o controle direto, além disso são equipamentos caros, que exigem manutenção especializada e suas informações não pode são confiáveis. As medições laboratoriais também passam pelo problema do tempo de amostragem. Com elas, não se consegue controlar automaticamente os processos.

Diante das dificuldades em se monitorar e mensurar variáveis relacionadas com a qualidade dos produtos, o presente trabalho propôs uma sistemática para o desenvolvimento de inferências, com a finalidade de estimar variáveis de difícil medição, seja para fins de simples monitoramento até o controle do processo.

Compreende-se que, ao se trabalhar com dados e com métodos de regressão, não há um método suficientemente bom para todos os possíveis casos, especialmente quando se trata de situações não lineares. Para isso, a metodologia foi construída de modo a permitir que diferentes alternativas de algoritmos de regressão possam ser testados e comparados.

Tal como os métodos de regressão, é válida a utilização de várias métricas de avaliação para que se possa validar com mais clareza a qualidade dos modelos. Essa necessidade pode ser vista, por exemplo, na análise do coeficiente de determinação R^2 dos modelos construídos nos estudos de caso. Esse critério não é suficiente para determinar qual dos modelos generaliza melhor os dados; os critérios de erros percentuais (médio e máximo)

se mostraram mais determinantes para a escolha dos modelos. Entretanto é interessante que se utilize várias métricas em conjunto.

Por fim, a principal conclusão que se pode ter é que, para o estudo de caso em questão (a unidade de separação de propeno/propano), utilizando a metodologia proposta, foi possível construir as inferências necessárias satisfatoriamente. Vale lembrar que as variáveis inferidas possuíamum alto grau de dificuldade nas situações de elevada pureza (ou baixa concentração de impureza), devido a elevada não linearidade, o que acarretouuma maior dificuldade para se obter en modelo adequado. Foi necessário, inclusive, se trabalhar com modelos locais em algumas situações. Mas, de modo geral, pôde-se aferir a capacidade da sistemática proposta e as inferências desenvolvidas obtiveram bons resultados para as necessidades da planta.

7.2 Sugestões para trabalhos futuros

A seguir são listadas algumas sugestões para trabalhos futuros:

- Utilizar composição logaritmica para se melhorar os modelos na faixa de elevada pureza;
- Estudar metodologias de modelos locais;
- Estudar a modelagem com dados dinâmicos e desenvolver a etapa que une a inferência estática aos dados dinâmicos de processos;
- Aplicar a metodologia à dados reais de processos;
- Aprimorar a metodologia de segregação de dados *k-rank,* de modo que o agrupamento dos dados (realizado pelo *k-means*) possa ser feito utilizando outros métodos, inclusive métodos supervisionados;

Referências

ALLEGRINI, F.; OLIVIERI, A. C. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. **Analytica Chimica Acta,** v. 699, n. 1, p. 18-25, 8/5/ 2011. ISSN 0003-2670. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0003267011006209>.

BENAICHA, A. et al. **New pca-based methodology for sensor fault detection and localization**. MOSIM'10. Hammamet - Tunisia. 2010.

BOULLOSA, D. et al. Monitoring through T2 Hotelling of cylinder lubrication process of marine diesel engine. **Applied Thermal Engineering,** v. 110, p. 32-38, 2017/01/05/ 2017. ISSN 1359-4311. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1359431116314089>.

BROWN, P. R.; RHINEHART, R. R. **Automated steady-state identification in multivariable systems**. 2000. 79-83.

CAO, S.; RHINEHART, R. R. An efficient method for on-line identification of steady state. **Journal of Process Control,** v. 5, n. 6, p. 363-374, 1995/12/01 1995. ISSN 0959-1524. Disponível em: <http://www.sciencedirect.com/science/article/pii/095915249500009F>.

CUI, C.; WANG, D. High dimensional data regression using Lasso model and neural networks with random weights. **Information Sciences,** v. 372, p. 505-517, 12/1/ 2016. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025516306314>.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. **Theoretical Computer Science,** v. 344, n. 2, p. 243-278, 2005/11/17 2005. ISSN 0304-3975. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304397505003798>.

DORIGO, M.; GAMBARDELLA, L. M. Ant Colonies for the Tavelling Salesman Problem. **Biosystems,** v. 43, p. 73 - 81, 1997.

DORIGO, M. et al. **Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006**. 2006.

EFRON, B. et al. Least angle regression. p. 407-499, 2004/04 2004. ISSN 0090-5364. Disponível em: <http://projecteuclid.org/euclid.aos/1083178935>.

ESCOBAR, M. S.; KANEKO, H.; FUNATSU, K. On Generative Topographic Mapping and Graph Theory combined approach for unsupervised non-linear data visualization and fault identification. **Computers & Chemical Engineering,** v. 98, p. 113-127, 2017/03/04/ 2017. ISSN 0098-1354. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098135416304033>.

FAASSEN, S. M.; HITZMANN, B. Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. **Sensors (Switzerland),** v. 15, n. 5, p. 10271-10291, 2015. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84929379514&partnerID=40&md5=c4f9eaa86e540219a7f7a76bc647ef67>.

FACCHIN, S. **Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais**. UFRGS: PPGEQ (Dissertação de Mestrado). Porto Alegre, RS, Brasil. 2005.

FEARN, T. et al. On the geometry of SNV and MSC. **Chemometrics and Intelligent Laboratory Systems,** v. 96, n. 1, p. 22-26, 2009/03/15/ 2009. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169743908002098>.

FERNANDES, P. R. **Rede de Modelos Termodinâmicos Locais**. UFRGS: PPGEQ (Dissertação de Mestrado). Porto Alegre, RS, Brasil. 2001.

FLECK, T. D. **Nova Metodologia para Desenvolvimento de Inferências Baseadas em Dados**. UFRGS: PPGEQ (Dissertação de Mestrado). Porto Alegre, RS, Brasil. 2012.

GARCIA-ALVAREZ, D.; FUENTE, M. J. Estudio comparativo de técnicas de detección de fallos basadas en el Análisis de Componentes Principales (PCA). **Revista Iberoamericana de Automática e Informática Industrial RIAI,** v. 8, n. 3, p. 182-195, 2011/07/01/ 2011. ISSN 1697-7912. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1697791211000070>.

GREENE, W. H. **Econometric analysis**. Upper Saddle River, N.J.: Pearson Education, 2002. ISBN 0130661899 9780130661890 0131108492 9780131108493.

HEMMATEENEJAD, B. et al. Building optimal regression tree by ant colony system–genetic algorithm: Application to modeling of melting points. **Analytica Chimica Acta,** v. 704, n. 1–2, p. 57-62, 10/17/ 2011. ISSN 0003-2670. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0003267011011081>.

HOTELLING, H. **Analysis of a complex of statistical variables into principal components**. Baltimore: Warwick & York, 1933. 48 p. Disponível em: < //catalog.hathitrust.org/Record/006826924 >.Disponível em: <https://hdl.handle.net/2027/wu.89097139406>.

HUANG, T.; RHINEHART, R. R. Steady state and Transient State Identification for flow rate on a pilot-scale absorption column. 2013 American Control Conference, 2013, 17-19 June 2013. p.4498-4503.

ITURBIDE, E.; CERDA, J.; GRAFF, M. A Comparison between LARS and LASSO for Initialising the Time-Series Forecasting Auto-Regressive Equations. **Procedia Technology,** v. 7, p. 282-288, 2013/01/01/ 2013. ISSN 2212-0173. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2212017313000364>.

JACKSON, J. E.; MUDHOLKAR, G. S. Control Procedures for Residuals Associated with Principal Component Analysis. **Technometrics,** v. 21, n. 3, p. 341-349, // 1979.

JEONG, J. et al. Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. **Journal of Hydrology,** v. 548, p. 135-144, 5// 2017. ISSN 0022-1694. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002216941730135X>.

KADLEC, P.; GABRYS, B.; STRANDT, S. Data-driven Soft Sensors in the process industry. **Computers & Chemical Engineering,** v. 33, n. 4, p. 795-814, 4/21/ 2009. ISSN 0098-1354. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098135409000076>.

KANEKO, H.; FUNATSU, K. Ensemble locally weighted partial least squares as a just-in-time modeling method. **AIChE Journal,** v. 62, n. 3, p. 717-725, 2016. ISSN 1547-5905. Disponível em: <http://dx.doi.org/10.1002/aic.15090>.

KRAMER, O. **Machine Learning for Evolution Strategies**. Switzerland: Springer, 2016.

KRUGER, U.; XIE, L. **Statistical monitoring of complex multivariate processes : with applications in industrial process control**. Chichester: John Wiley, 2012. ISBN 9780470028193 (hbk.) : ¹55.00
047002819X (hbk.) : ¹55.00.

LEE, K.-I. et al. Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant E coli fermentation processes. **Journal of Chemical Technology & Biotechnology,** v. 80, n. 9, p. 1036-1045, 2005. ISSN 1097-4660. Disponível em: <http://dx.doi.org/10.1002/jctb.1281>.

LIN, B. et al. A systematic approach for soft sensor development. **Computers & Chemical Engineering,** v. 31, n. 5–6, p. 419-425, 5// 2007. ISSN 0098-1354. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098135406001293>.

LIU, H.; SHAH, S.; JIANG, W. On-line outlier detection and data cleaning. **Computers & Chemical Engineering,** v. 28, n. 9, p. 1635-1647, 2004/08/15/ 2004. ISSN 0098-1354. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098135404000249>.

MAIMON, O. Z.; ROKACH, L. **Data mining and knowledge discovery handbook**. 2nd ed. New York ; London: Springer, 2010. ISBN 9780387098227 (hbk.) : ¹179.50

0387098224 (hbk.) : ¹179.50.

MANSOURI, M. et al. Kernel PCA-based GLRT for nonlinear fault detection of chemical processes. **Journal of Loss Prevention in the Process Industries,** v. 40, p. 334-347, 3// 2016. ISSN 0950-4230. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950423016300110>.

MASSARON, L. A.; BOSCHETTI, A. A. **Regression Analysis with Python**. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016. ISBN 9781783980741 Electronic book (EPUB format).

MEJIA, R. I. G. *et al.* Novo Método para a Identificação de Estado Estacionário Baseada na Estimativa da Autocorrelação Local. **IN Sba, S. B. D. A.-. (Ed.) XVIII Congresso Brasileiro de Automática - CBA2010. Bonito, MS**, [*s. l.*], p. 4083–4088, 2010.

MULLEN, R. J. et al. A review of ant algorithms. **Expert Systems with Applications,** v. 36, n. 6, p. 9608-9617, 8// 2009. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417409000384>.

PEARSON, R. K. Outliers in process modeling and identification. **IEEE Transactions on Control Systems Technology,** v. 10, n. 1, p. 55-63, 2002. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0036207952&doi=10.1109%2f87.974338&partnerID=40&md5=a8f20ac5508eeb7548a7777fbbfd43da>.

PESSOA, C. M. et al. Development of Ant Colony Optimization (ACO) Algorithms Based on Statistical Analysis and Hypothesis Testing for Variable Selection. **IFAC-PapersOnLine,** v. 48, n. 8, p. 900-905, 2015/01/01/ 2015. ISSN 2405-8963. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2405896315011659>.

QIN, S Joe; YUE, Hongyu; DUNIA, Ricardo. Self-Validating Inferential Sensors with Application to Air Emission Monitoring. **Industrial and Engineering Chemistry Research**, [*s. l.*], v. 36, n. 5, p. 1675–1685, 1997. Disponível em: https://doi.org/10.1021/ie960615y

RANZAN, C. et al. Sulfur Determination in Diesel using 2D Fluorescence Spectroscopy and Linear Models. **IFAC-PapersOnLine,** v. 48, n. 8, p. 415-420, // 2015. ISSN 2405-8963. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2405896315010848>.

_____. Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems,** v. 132, p. 133-140, 3/15/ 2014. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169743914000203>.

_____. NIR pre-selection data using modified changeable size moving window partial least squares and pure spectral chemometrical modeling with ant colony optimization for wheat flour characterization. **Chemometrics and Intelligent Laboratory Systems,** v. 142, p. 78-86, 2015/03/15/ 2015. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016974391500009X>.

RASCHKA, S. A. **Python machine learning**. Birmingham B3 2PB, UK.: Packt Publishing Ltd., 2015. ISBN 9781783555147 (PDF ebook) : ¹22.99.

RHINEHART, R. R. **Automated steady and transient state identification in noisy processes**. 2013. 4477-4493 ISBN 978-1-4799-0177-7.

RUSSELL, E. L.; CHIANG, L. H.; BRAATZ, R. D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. **Chemometrics and Intelligent Laboratory Systems,** v. 51, n. 1, p. 81-93, 2000/05/08/ 2000. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169743900000587>.

SANTOS, P. V. J. L.; RANZAN, L.; FARENZENA, M.; TRIERWEILER, J. O. K-RANK: AN EVOLUTION OF Y-RANK FOR MULTIPLE SOLUTIONS PROBLEM. **Brazilian Journal of Chemical Engineering**, v. 36, n. 1, p. 409–419, jan. 2019.

SCHULTZ, E. S. **A importância do ponto de operação nas técnicas de Self-optimizing Control**. Porto Alegre, RS, Brazil, 2015.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Kernel principal component analysis. In: GERSTNER, W.;GERMOND, A*., et al* (Ed.). **Artificial Neural Networks — ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p.583-588. ISBN 978-3-540-69620-9.

SHAMIR, R. et al. EXPANDER--an integrative program suite for microarray data analysis. **BMC Bioinformatics,** v. 6, p. 232, Sep 2005. ISSN 1471-2105. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/16176576>.

SOCHA, K.; DORIGO, M. Ant colony optimization for continuous domains. **European Journal of Operational Research,** v. 185, n. 3, p. 1155-1173, 3/16/ 2008. ISSN 0377-2217. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0377221706006333>.

THALAMUTHU, A. et al. Evaluation and comparison of gene clustering methods in microarray analysis. **Bioinformatics,** v. 22, n. 19, p. 2405-12, Oct 2006. ISSN 1367-4811. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/16882653>.

TIBSHIRANI, R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B, 1994. p.267-288.

TOMAŽIČ, Simon. 2023. Intelligent Soft Sensors. *Sensors. 20*23, no. 15: 6895. Disponível em: https://doi.org/10.3390/s23156895.

TOKSARI, M. D. A hybrid algorithm of Ant Colony Optimization (ACO) and Iterated Local Search (ILS) for estimating electricity domestic consumption: Case of Turkey. **International Journal of Electrical Power & Energy Systems,** v. 78, p. 776-782, 2016/06/01/ 2016. ISSN 0142-0615. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0142061515005840>.

WANG, K.; WANG, B.; PENG, L. CVAP: VALIDATION FOR CLUSTER ANALYSIS. **Data Science Journal,** v. 8, p. 6, 2009.

WARNE, K. et al. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. **Engineering Applications of Artificial Intelligence,** v. 17, n. 8, p. 871-885, 12// 2004. ISSN 0952-1976. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0952197604000971>.

WEISBERG, S. **Applied linear regression**. NY: Wiley, 1980. ISBN 0471044199.

WILLIAMS, G. et al. A comparative study of RNN for outlier detection in data mining. 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 2002, 2002. p.709-712.

ZAKI, M. J.; MEIRA, W. **Data mining and analysis : fundamental concepts and algorithms**. New York, NY 10013-2473, USA: Cambridge University Press, 2014. ISBN 9780521766333 (hbk.) : ¹35.00.