

UNESP
Faculdade de Engenharia do Campus de Guaratinguetá

Guaratinguetá
2012

LEANDRO VALIM DE FREITAS

ANÁLISE MULTIVARIADA NO MONITORAMENTO DO PROCESSO DE
PRODUÇÃO DE ÓLEO DIESEL NUMA REFINARIA DE PETRÓLEO

Dissertação apresentada à Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, para a obtenção do título de Mestre em Engenharia Mecânica na área de Gestão e Otimização.

Orientador: Prof. Dr. Messias Borges Silva
Co-orientador: Prof. Dr. Fernando Augusto Silva Marins

Guaratinguetá
Outubro 2012

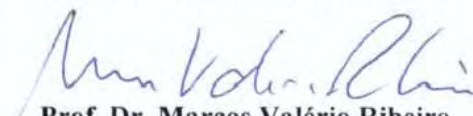
F866a	Freitas, Leandro Valim de Análise multivariada no monitoramento do processo de produção de óleo diesel numa refinaria de petróleo / Leandro Valim de Freitas – Guaratinguetá : [s.n], 2012. 100 f : il. Bibliografia: f. 93 Dissertação (Mestrado) – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2012. Orientador: Prof. Dr. Messias Borges Silva Coorientador: Prof. Dr. Fernando Augusto Silva Marins 1. Petróleo - Refinarias I. Título CDU 665.62(043)
-------	---

LEANDRO VALIM DE FREITAS

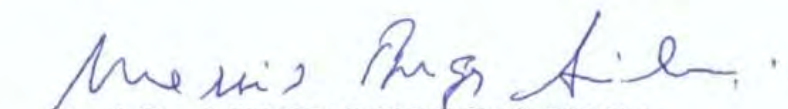
ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
“MESTRE EM ENGENHARIA MECÂNICA”

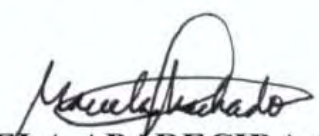
PROGRAMA: ENGENHARIA MECÂNICA
ÁREA: GESTÃO E OTIMIZAÇÃO

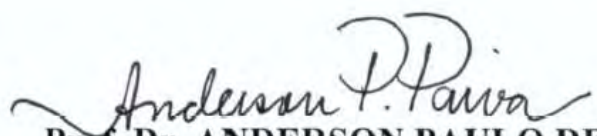
APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO


Prof. Dr. Marcos Valério Ribeiro
Coordenador

BANCA EXAMINADORA:


Prof. Dr. MESSIAS BORGES SILVA
Orientador / Unesp-Feg


Prof.ª. Dr.ª. MARCELA APARECIDA GUERREIRO
MACHADO DE FREITAS
Unesp - Feg


Prof. Dr. ANDERSON PAULO DE PAIVA
UNIFEI

DADOS CURRICULARES

LEANDRO VALIM DE FREITAS

NASCIMENTO	26.01.1986 – MOGI GUAÇU / SP
FILIAÇÃO	Carlos Alberto de Freitas Sandra Teresinha Valim de Freitas
2004/2009	Curso de Graduação em Engenharia Química. Escola de Engenharia Química de Lorena – Universidade de São Paulo, USP, Brasil.
2009/2009	Curso de Aperfeiçoamento em Métodos de Análise e Regressão Multivariada. Universidade Corporativa Petrobras – Petróleo Brasileiro SA, PETROBRAS, Brasil
2009/2010	Curso de Pós-Graduação em MBA Gestão da Produção. Faculdade de Engenharia de Guaratinguetá – Universidade Estadual Paulista, UNESP, Brasil.
2010/2010	Curso de Especialização em <i>Design of Experiments</i> . <i>Massachusetts Institute of Technology</i> , MIT, Estados Unidos.
2011/2012	Curso de Pós-Graduação nível Mestrado em Engenharia Mecânica. Faculdade de Engenharia de Guaratinguetá – Universidade Estadual Paulista, UNESP, Brasil.

AGRADECIMENTOS

À Deus, que me concedeu saúde e força de vontade por estar aqui.

Expresso também minha gratidão às pessoas que me incentivaram na elaboração da dissertação.

- A minha querida esposa Ana Paula que tanto apóia o desenvolvimento da minha carreira acadêmica.
- Aos meus pais que sempre me lembram da importância da paciência e da humildade.
- Aos orientadores Prof. Dr. Messias Borges Silva e Prof. Dr. Fernando Augusto Silva Marins do Departamento de Produção da FEG-UNESP pela paciência, boa vontade em ajudar e por estarem sempre me motivando durante nossas conversas.
- Aos Prof. Dr. Valério Salomon e Prof. Dr. Fernando Branco pelos incentivos durante nossas conversas no Departamento de Produção da UNESP.
- Ao gerente Pedro Luiz Corrêa Garcia da Petrobras pela oportunidade de utilizar dados reais de produção e pela compreensão das vezes que estive ausente do trabalho para cursar disciplinas.
- Ao Consultor Aerenon F. Bueno da Petrobras e ao Prof. Dr. Célio Pasquini da UNICAMP pelos ensinamentos de Análise Multivariada e Quimiometria.
- Aos professores que participaram da banca examinadora, pelas críticas e sugestões que certamente contribuiram para melhorar o conteúdo deste trabalho.

FREITAS, L. V. **Análise multivariada no monitoramento do processo de produção de óleo diesel numa refinaria de petróleo.** 2012. 100 f. Dissertação (Mestrado em Engenharia Mecânica) – Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, Guaratinguetá, 2012.

RESUMO

Este trabalho tem por objetivo desenvolver, validar e testar modelos matemáticos multivariados com a finalidade de monitorar em tempo real a qualidade no processamento do óleo diesel numa refinaria de petróleo. Os produtos derivados têm especificações baseadas em propriedades físico-químicas que podem variar de maneira significativa com a modificação do elenco de petróleos durante o processamento, mantendo-se as mesmas condições de controle de produção, o que compromete os padrões de qualidade dos seus produtos. Isso leva à necessidade de determiná-las ou prevê-las com a maior frequência possível em detrimento às análises pontuais tradicionais em laboratório. Foi possível modelar matematicamente as propriedades temperaturas de destilação, índice de cetano, ponto de fulgor, viscosidade cinemática e densidade do óleo diesel pela técnica multivariada - Mínimos Quadrados Parciais (PLS – *Partial Least Squares*) a partir de dados experimentais. Os modelos desenvolvidos foram utilizados em um processo real de produção permitindo agilidade na tomada de decisão no sistema de *blending*, redução da variabilidade do processo e aumento da rentabilidade da produção, eliminação de reprocessamentos, sem geração de resíduos químicos.

PALAVRAS-CHAVE: Mínimos quadrados parciais. Monitoramento. Refinaria de petróleo.

FREITAS, L. V. **Multivariate Analysis in monitoring the production of diesel oil in a petroleum refinery.** 2012. 100 f. Dissertation (Master in Mechanical Engineering) - Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, Guaratinguetá, 2012.

ABSTRACT

This study aims to develop, validate and test multivariate mathematical models in order to monitor real-time processing on the quality of diesel oil in a petroleum refinery. Their products have specifications based on physicochemical properties that can vary significantly with the change of cast during the processing of oil, keeping the same conditions of production control, which compromises the quality standards of their products. This leads to the need to determine them or provides them as often as possible in lieu of the traditional point analysis in the laboratory. It was possible to model the properties distillation, cetane index, flash point, kinematic viscosity and density of diesel by multivariate analysis - Partial Least Squares from experimental data. The models developed were used in an actual production process allowing flexibility in decision-making system in blending, reduction of process variability, increase the profitability of production, elimination of reprocessing without chemical residues.

KEYWORDS: Partial least squares. Monitoring. Oil refinery

LISTA DE FIGURAS

FIGURA 1 – Etapas da pesquisa	16
FIGURA 2 – Esquema da tecnologia para aquisição do banco de dados	22
FIGURA 3 – Conjunto de espectros de absorção do óleo diesel	23
FIGURA 4 – Seleção de variáveis para regressão por MLR	27
FIGURA 5 – Conjunto de amostras no espaço tridimensional	32
FIGURA 6 – Primeiro componente principal plotado no espaço tridimensional	33
FIGURA 7 – Projeção das amostras no primeiro componente principal	33
FIGURA 8 – Segundo componente principal plotado no espaço tridimensional	34
FIGURA 9 – Projeção das amostras no segundo componente principal	34
FIGURA 10 – Interpretação geométrica dos <i>loadings</i>	35
FIGURA 11 – Variáveis transformadas em combinações lineares, relacionadas com Y por uma equação de regressão	36
FIGURA 12 – Fenômenos <i>underfitting</i> e <i>overfitting</i>	39
FIGURA 13 – Esquema de produção nível 1	41
FIGURA 14 – Esquema de produção nível 2	41
FIGURA 15 – Esquema de produção nível 3	42
FIGURA 16 – Esquema de produção nível 4	42
FIGURA 17 – Exemplo de ganho de margem de refino	45
FIGURA 18 – Estrutura de consumo de derivados no Brasil 2005-2030	46
FIGURA 19 – Fluxograma simplificado da produção de diesel numa refinaria de petróleo	47
FIGURA 20 – Arranjo simplificado para monitoramento <i>on-line</i>	48
FIGURA 21 – Histograma da propriedade temperatura de destilação 10% rec.	51
FIGURA 22 – Histograma da propriedade temperatura de destilação 50% rec.	51
FIGURA 23 – Histograma da propriedade temperatura de destilação 85% rec.	52
FIGURA 24 – Histograma da propriedade temperatura de destilação 90% rec.	52
FIGURA 25 – Histograma da propriedade viscosidade cinemática	53
FIGURA 26 – Histograma da propriedade densidade	53
FIGURA 27 – Histograma da propriedade ponto de fulgor	54
FIGURA 28 – Histograma da propriedade índice de cetano	54

FIGURA 29 – Scores dos dois primeiros componentes principais da PCA	55
FIGURA 30 – X-loading da PCA	56
FIGURA 31 – Espectros das amostras de óleo diesel	56
FIGURA 32 – Variância explicada acumulada em função do número de componentes principais	57
FIGURA 33 – Variância explicada versus número de variáveis latentes para destilação 10% rec.....	59
FIGURA 34 – Variância explicada versus número de variáveis latentes sem utilizar validação cruzada.....	60
FIGURA 35 – Resíduos de Student versus leverage.	60
FIGURA 36 – Valores preditos versus valores medidos para temperatura de destilação 10% rec.	61
FIGURA 37 – Teste de Normalidade para propriedade temperatura de destilação 10% rec.	62
FIGURA 38 – Histograma do resíduo padronizado para temperatura de destilação 10% rec.	63
FIGURA 39 – Variância explicada versus número de variáveis latentes para destilação 50% rec.	64
FIGURA 40 – Valores preditos versus valores medidos para temperatura de destilação 50% rec.	65
FIGURA 41 – Teste de Normalidade para propriedade temperatura de destilação 50% rec.	66
FIGURA 42 – Histograma do resíduo padronizado para temperatura de destilação 50% rec.	67
FIGURA 43 – Variância explicada versus número de variáveis latentes para temperatura de destilação 50% rec.	68
FIGURA 44 – Valores preditos versus valores medidos para temperatura destilação 85% rec.	69
FIGURA 45 – Teste de Normalidade para propriedade temperatura de destilação 85% rec.	70
FIGURA 46 – Histograma do resíduo padronizado para temperatura de destilação 85% rec.	71

FIGURA 47 – Variância explicada versus número de variáveis latentes para temperatura de destilação 90% rec.	72
FIGURA 48 – Valores preditos versus valores medidos para temperatura de destilação 90% rec.	73
FIGURA 49 – Teste de Normalidade para propriedade temperatura de destilação 90% rec.	74
FIGURA 50 – Histograma do resíduo padronizado para temperatura de destilação 90% rec.	75
FIGURA 51 – Variância explicada versus número de variáveis latentes para índice de cetano	76
FIGURA 52 – Valores preditos versus valores medidos para índice de cetano	77
FIGURA 53 – Teste de Normalidade para propriedade índice de cetano	78
FIGURA 54 – Histograma do resíduo padronizado para índice de cetano	79
FIGURA 55 – Variância explicada versus número de variáveis latentes para viscosidade cinemática	80
FIGURA 56 – Valores preditos versus valores medidos para viscosidade cinemática	80
FIGURA 57 – Teste de Normalidade para propriedade viscosidade cinemática	82
FIGURA 58 – Histograma do resíduo padronizado para viscosidade cinemática	82
FIGURA 59 – Variância explicada versus número de variáveis latentes para ponto de fulgor	83
FIGURA 60 – Valores preditos versus valores medidos para ponto de fulgor	84
FIGURA 61 – Teste de Normalidade para propriedade ponto de fulgor	85
FIGURA 62 – Histograma do resíduo padronizado para ponto de fulgor	86
FIGURA 63 – Variância explicada versus número de variáveis latentes para densidade ..	87
FIGURA 64 – Valores preditos versus valores medidos para densidade	87
FIGURA 65 – Teste de Normalidade para propriedade densidade	89
FIGURA 66 – Histograma do resíduo padronizado para densidade	89

LISTA DE TABELAS

TABELA 1 – Cálculos intermediários para obtenção do coeficiente de autocorrelação (amostragens sucessivas)	49
TABELA 2 – Cálculos intermediários para obtenção do coeficiente de autocorrelação (amostragens após troca de carga de carga da refinaria)	50
TABELA 3 – Resumo dos resultados de modelagem e validação	58
TABELA 4 – Variância explicada acumulada versus número de variáveis latentes – temperatura de destilação (10% rec.)	59
TABELA 5 – Conjunto de validação externa – temperatura de destilação (10% rec.)	61
TABELA 6 – Variância explicada acumulada versus número de variáveis latentes – temperatura de destilação (50% rec.)	63
TABELA 7 – Conjunto de validação externa – temperatura de destilação (50% rec.)	65
TABELA 8 – Variância explicada acumulada versus número de variáveis latentes – temperatura de destilação (85% rec.)	67
TABELA 9 – Conjunto de validação externa – temperatura de destilação (85% rec.)	69
TABELA 10 – Variância explicada acumulada versus número de variáveis latentes – temperatura de destilação (90% rec.)	71
TABELA 11 – Conjunto de validação externa – temperatura de destilação (90% rec.)	73
TABELA 12 – Variância explicada acumulada versus número de variáveis latentes – índice de cetano	75
TABELA 13 – Conjunto de validação externa – índice de cetano	79
TABELA 14 – Variância explicada acumulada versus número de variáveis latentes – viscosidade cinemática	81
TABELA 15 – Conjunto de validação externa – viscosidade cinemática	83
TABELA 16 – Variância explicada acumulada versus número de variáveis latentes – ponto de fulgor	84
TABELA 17 – Conjunto de validação externa – ponto de fulgor	86
TABELA 18 – Variância explicada acumulada versus número de variáveis latentes – densidade	88

LISTA DE ABREVIATURAS E SIGLAS

ANN	- Artificial Neural Network
ANP	- Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
API	- American Petroleum Institute
ASTM	- American Society for Testing Materials
CV	- Cross Validation
DEA	- Data Envelopment Analysis
DOE	- Design of Experiments
GA	- Genetic Algorithm
GAIA	- Geometrical Analysis for Interactive Assistance
MCDM	- Multiple Criteria Decision Making
MLR	- Multiple Linear Regression
NIPALS	- Nonlinear Iterative
PAT	- Process Analytical Technology
PC	- Principal Component
PCA	- Principal Component Analysis
PCR	- Principal Component Regression
PLS	- Partial Least Squares
PROMETHEE	- Preference Ranking Organization Method for Enrichment Evaluations
RMSECV	- Root Mean Square Error of Cross Validation
RMSEP	- Root Mean Square Error of Prediction
RSM	- Response Surface Method
SIMCA	- Soft Independent Modeling of Class Analogy
SPA	- Sucessive Projections Algorithm

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Contextualização do Problema	14
1.2	Objetivos e Justificativas	14
1.2.1	Objetivo Geral	14
1.2.2	Objetivos Específicos.....	15
1.2.3	Justificativas	15
1.3	Métodos de Pesquisa	16
1.4	Delimitação do Trabalho	18
1.5	Estrutura do Trabalho.....	18
2	REFERENCIAL TEÓRICO	19
2.1	Aplicações da Análise Multivariada	19
2.2	REVISÃO SOBRE OS MÉTODOS UTILIZADOS	22
2.2.1	Ciência e Tecnologia envolvidas na mineração dos dados	22
2.2.2	Análise Multivariada.....	25
2.2.2.1	Considerações Iniciais.....	25
2.2.2.2	Regressão Linear Múltipla	26
2.2.2.3	Componentes Principais e Variáveis Latentes.....	28
2.2.3	Processos Autocorrelacionados e Refino de Petróleo.....	40
2.2.4	Caracterização do óleo diesel	43
3	APLICAÇÃO.....	45
3.1	Contexto	45
3.2	Apresentação do Problema.....	46
3.3	Solução Proposta	46
3.4	Resultados Alcançados.....	49
3.4.1	Resultados para a propriedade temperatura de destilação (10% rec.) ...	58
3.4.2	Resultados para a propriedade temperatura de destilação (50% rec.) ...	63
3.4.3	Resultados para a propriedade temperatura de destilação (85% rec.) ...	67
3.4.4	Resultados para a propriedade temperatura de destilação (90% rec.) ...	71
3.4.5	Resultados para a propriedade índice de cetano	75
3.4.6	Resultados para a propriedade viscosidade cinemática	79
3.4.7	Resultados para a propriedade ponto de fulgor	83
3.4.8	Resultados para a propriedade destilação densidade	86
4	CONCLUSÕES	90
4.1	Verificação dos objetivos	90
4.2	Sugestões para continuidade do trabalho	91
	REFERÊNCIAS.....	93
	ANEXO A	101

1 INTRODUÇÃO

1.1 Contextualização do Problema

Segundo Manzano (2005), refinar petróleo é uma atividade estratégica para a economia dos países e seu segmento industrial está no centro do panorama macroeconômico e geopolítico internacional.

Ao mesmo tempo em que as demandas pelo refino de petróleo crescem (LIU et al., 2007), aumentam-se as exigências por qualidade e menores custos (MORO, 2003). Para tal, necessita-se de um maior conhecimento dos processos industriais e de uma abordagem mais realística das variáveis (COSTA; EPPRECHT; CARPINETTI, 2005; AGUILERA; ESCABIAS; VALDERRAMA, 2006).

Segundo Speight (1998), o petróleo é caracterizado por uma mistura complexa de compostos orgânicos, formado predominantemente por hidrocarbonetos (carbono e hidrogênio), além de outros compostos como nitrogenados, oxigenados, sulfurados e metais como níquel e vanádio. São quimicamente diferentes entre si em função da infinidade de fatores composicionais possíveis. De acordo com Cuellar (2009), em uma parafina com 43 átomos de carbono, por exemplo, pode haver aproximadamente 10^{15} compostos químicos diferentes - um dos fatores que inviabiliza a utilização direta da composição da mistura.

Os produtos derivados do petróleo têm especificações baseadas em suas propriedades físico-químicas (HIDAJAT; CHONG, 2000) que podem variar de maneira significativa com a modificação do elenco de petróleos durante seu processamento, mantendo-se as mesmas condições de controle de produção, o que compromete os padrões de qualidade dos seus produtos. Isso leva à necessidade de se determinar ou prever propriedades com a maior frequência possível em detrimento às análises pontuais tradicionais.

Segundo Dave et al. (2003), o uso de instrumentação avançada em conjunto com técnicas estatísticas multivariadas para a determinação das propriedades dos produtos, é uma das maneiras de maximizar o rendimento das operações de refino de petróleo.

1.2 Objetivos e Justificativas

1.2.1 Objetivo geral

Este trabalho tem por objetivo geral desenvolver modelos multivariados de previsão com a finalidade de monitorar em tempo real a qualidade no processamento do óleo diesel numa refinaria de petróleo.

1.2.2 Objetivos específicos

1. Avaliar o desempenho dos modelos de previsão por meio de validação.
2. Testar os modelos validados numa refinaria de petróleo brasileira - processo real de produção.

1.2.3 Justificativas

A contínua inovação dos produtos e processos e as constantes mudanças no ambiente interno e externo justificam uma demanda, sistemática e consistente, de transformação de dados brutos, coletados por meio de sistemas de informação, em conhecimento útil para tomada de decisões.

Devido ao número de variáveis empregadas e a sua natureza multicolinear, este trabalho foi conduzido segundo abordagens da Estatística Multivariada.

O monitoramento do refino de petróleo em tempo real apresenta uma série de vantagens quando comparado à abordagem tradicional de ensaios em laboratório, fornecendo informações essenciais para o controle do sistema de mistura, mediante previsão de diversos parâmetros de qualidade e controle de processo tais como:

1. Agilidade no processo de tomada de decisão;
2. Utiliza técnicas não destrutivas;
3. Custo de implementação relativamente baixo em relação ao retorno financeiro para o negócio;
4. Necessita de pequenas quantidades de amostra;
5. Não gera resíduos químicos;
6. Elimina reprocessamento de produtos, o que reduz o impacto sobre o meio ambiente;
7. Reduz a quantidade de ensaios físico-químicos no laboratório de qualidade.

1.3 Métodos de Pesquisa

O método de pesquisa utilizado em todas as etapas do trabalho foi a Pesquisa-Ação. De acordo com Miguel et al. (2010) este trabalho pode ser classificado como de natureza aplicada, com objetivos axiomáticos normativos e de abordagem quantitativa. As técnicas utilizadas foram a análise de experimentos e a modelagem computacional.

As fases da pesquisa estão ilustradas na Figura 1 e detalhadas a seguir:

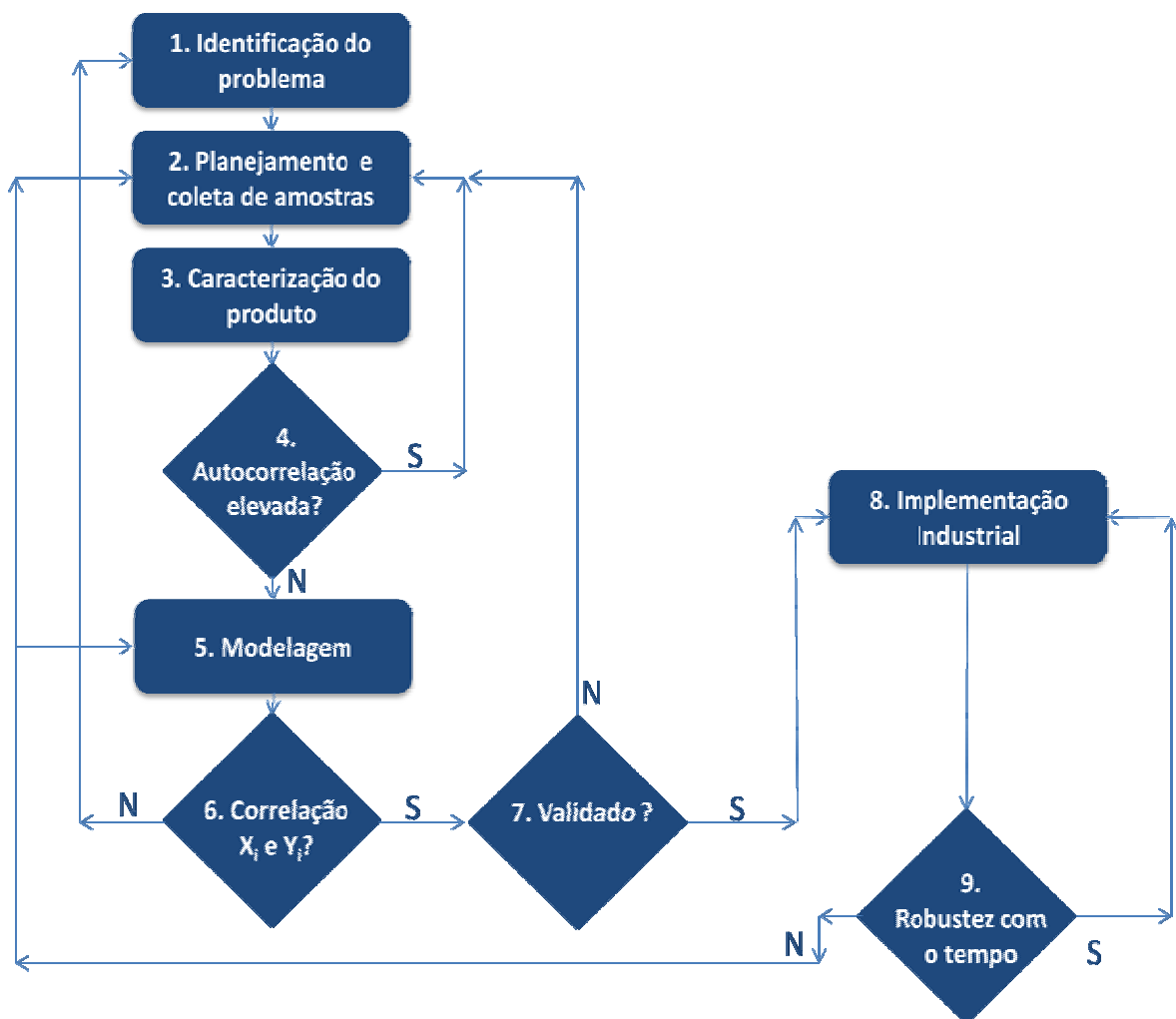


Figura 1. Etapas da Pesquisa

1. Identificação do problema: Em uma refinaria brasileira foram identificados os possíveis ganhos oriundos do monitoramento em tempo real da produção de óleo diesel utilizando instrumentação avançada e análise multivariada.

2. Planejamento e coleta de amostras: Devido à natureza intrínseca de um processo químico, observações sucessivas em intervalos curtos de tempo de uma mesma característica de qualidade costumam produzir dados autocorrelacionados (COSTA; EPPRECHT; CARPINETTI, 2010), por isso a coleta de amostras deste trabalho foi planejada ocorrer seis horas após troca de carga da refinaria. Com a modificação das misturas de petróleos ocorrem mudanças significativas em sua composição química, permitindo formar amostras independentes umas das outras. Isso se deve ao fato de que em média 20 petróleos diferentes e de origens diferentes fazem parte de uma mesma mistura a ser processada numa determinada campanha de produção na refinaria. Segundo Thomas (2004) petróleos obtidos de diferentes reservatórios possuem composições químicas diferentes com variação de componentes dentro das classes dos alcanos normais, isoalcanos, cicloalcanos, aromáticos, asfáltênicos e naftênicos. Além disso, alguns são bem escuros e opacos, densos e viscosos, ácidos ou não, contendo pouco ou nenhum gás, enquanto que outros são castanhos ou bastante claros, com baixa viscosidade e densidade e quantidade apreciável de gás em solução.
3. Caracterização do produto: Nesta etapa, as amostras de óleo diesel foram submetidas a ensaios de laboratório, objetivando determinar as variáveis de entrada, X_i , que são as absorções de radiação infravermelha da instrumentação avançada, cujo detalhamento será realizado no item 2.2.1 e, as variáveis respostas, Y_i , propriedades físico-químicas serem preditas por modelos matemáticos e que será detalhadas no item 2.2.4.
4. Autocorrelação elevada?: Caso não seja tomado o cuidado de coletar as amostras após troca da matéria-prima, a probabilidade de formar um banco de dados com baixa variabilidade é alta, principalmente se as amostras forem coletadas em intervalos curtos de tempo, devido, ao fenômeno da autocorrelação dentro de uma mesma campanha de produção. Neste caso, seria necessário voltar à etapa 2, caso contrário, seguir para a etapa 5.
5. Modelagem: Nesta etapa, os modelos matemáticos serão desenvolvidos utilizando o software *The Unscrambler*® via matriz de correlação. Para cálculos auxiliares foi utilizado o software Minitab®.
6. Há correlação entre X_i e Y_i ? Caso não haja correlação entre as variáveis preditoras X_i e as variáveis respostas Y_i , é necessário voltar à etapa 1 e verificar as falhas na

identificação do problema e nos métodos propostos. Caso contrário, seguir para a etapa 7.

7. Modelo Validado?: O desempenho dos modelos foi avaliado por meio de validação interna e externa, seguida de um teste estatístico para comparação dos resultados. Se o desempenho for considerado estatisticamente satisfatório, seguir para a etapa 8, caso contrário, será necessário voltar à etapa 2. Ainda que um modelo multivariado apresente correlação entre as variáveis preditoras X_i e variáveis respostas Y_i e é considerado não validado, isto ocorre porque o modelo apresenta baixa robustez na predição de amostras externas, que não fizeram parte do modelo, por isso, neste caso, seria necessário coletar novas amostras que representem esses novos perfis de produção, estudá-las e agregá-las no conjunto de modelagem.
8. Implementação industrial: Com intuito de testar os modelos em um processo real, os mesmos foram implementados numa refinaria brasileira para monitoramento em tempo real da produção de óleo diesel.
9. Robustez dos modelos com o tempo?: Uma vez implementados em escala industrial, os modelos devem passar de tempos em tempos por um procedimento análogo à validação inicial – validação contínua, pois ao longo do tempo novos perfis de produção ou diferentes misturas de petróleos podem surgir na refinaria e os modelos podem necessitar de atualizações. Neste caso, é necessário voltar à etapa 2 para coleta de novas amostras. Caso contrário, os modelos podem ser mantidos em uso na indústria.

1.4 Delimitações do Trabalho

Os resultados obtidos são válidos apenas para o produto óleo diesel, embora a abordagem utilizada possa ser estendida a qualquer produto de petróleo numa refinaria.

1.5 Estrutura do Trabalho

Esta dissertação está dividida em mais três capítulos, além deste introdutório.

O Capítulo 2 traz aplicações da Análise Multivariada, em seguida, trata da Ciência e Tecnologia envolvidas na mineração dos dados, e, equacionamentos. Posteriormente,

menção sobre processos autocorrelacionados e refino de petróleo. Por fim, a caracterização do óleo diesel.

No Capítulo 3, foi apresentado o problema: aumento da demanda por óleo diesel e os resultados dos modelos matemáticos multivariados. Os gráficos apresentados foram obtidos pelos softwares *The Unscrambler*® e Minitab®.

No Capítulo 4, estão as considerações finais da dissertação.

2 REFERENCIAL TEÓRICO

2.1 Aplicações da Análise Multivariada

Métodos baseados fortemente em estatística e inteligência artificial como os métodos multivariados ou quimiométricos têm sido amplamente utilizados na indústria petrolífera (KIM; LEE; KIM, 2009). Vários artigos foram escritos sobre aplicações da análise multivariada na predição de propriedades de derivados de petróleo (SANTOS JUNIOR et al., 2005; CHUNG, 2007).

Pasadakis, Sourligas e Foteinopoulos (2006) utilizaram os seis primeiros componentes da Análise por Componentes Principais (PCA – *Principal Components Analysis*) como variáveis de entrada na modelagem não linear de propriedades de derivados do petróleo.

Pasquini e Bueno (2007) propuseram uma nova abordagem para prever o ponto de ebulição verdadeiro do petróleo e seu grau API (*American Petroleum Institute*) – medida de densidade relativa de líquidos, por Mínimos Quadrados Parciais (PLS – *Partial Least Squares*) e Redes Neurais Artificiais (ANN – *Artificial Neural Network*). Amostras de misturas de petróleo foram obtidas de várias regiões produtoras do Brasil e do exterior. Nesta aplicação, os modelos obtidos pelo método PLS apresentaram desempenho superior às redes neurais. O curto intervalo de tempo necessário para previsão de propriedades justifica a proposta de caracterização do petróleo de maneira mais ágil para o monitoramento de processos de refino.

Teixeira et al. (2008) em trabalhos com gasolinas brasileiras utilizaram o algoritmo multivariado Modelagem Independente e Flexível por Analogia de Classe (SIMCA – *Soft Independent Modeling of Class Analogy*) para análise da formação de *clusters*. Com o objetivo de quantificar o teor de adulterações das gasolinas por outros hidrocarbonetos, aplicaram PLS. Por fim, os modelos foram validados internamente pelo algoritmo de validação cruzada (*cross validation*) e externamente com um conjunto independente de amostras.

Bao e Dai (2009) estudaram diferentes métodos multivariados, incluindo técnicas lineares e não lineares, buscando minimizar o erro de previsão de modelos desenvolvidos para controle de qualidade da gasolina.

Lira et al. (2010) aplicaram o método PLS para inferência dos parâmetros de qualidade: densidade, teor de enxofre e temperaturas de destilação da mistura diesel/biodiesel,

apresentando grande economia de tempo em comparação com os métodos tradicionais de caracterização por equipamentos de laboratório.

Aleme, Corgozinho e Barbeira (2010) realizaram um estudo de classificação de amostras de diesel utilizando PCA para a discriminação do tipo de óleo e para previsão de sua origem.

Em outra abordagem, vários trabalhos apresentam aplicações híbridas ou comparativas da Análise Multivariada com outros importantes métodos da Engenharia de Produção:

Paiva, Ferreira e Balestrassi (2007) combinaram o Método de Superfície de Resposta (RSM – *Response Surface Method*) do Projeto de Experimentos (DOE – *Design of Experiments*) com Análise por Componentes Principais na otimização de múltiplas respostas correlacionadas num processo de manufatura.

Huang, Hsu e Liu (2009) utilizaram Mahalanobis-Taguchi integrado com Redes Neurais Artificiais na mineração de dados para busca de padrões e modelagem na área de fabricação. Pal e Maiti (2010) utilizaram o algoritmo Mahalanobis-Taguchi na redução da dimensionalidade de dados multivariados para posterior otimização via Metaheurísticas.

Liu et al. (2007) realizaram inferências sobre parâmetros de qualidade do querosene de aviação utilizando Regressão Linear Múltipla (MLR – *Multiple Linear Regression*) e ANN. O trabalho mostrou que a performance da modelagem via ANN foi superior e atende às repetibilidades dos métodos de referência.

Em estudos de otimização de modelos multivariados, observam-se aplicações combinadas da Análise Multivariada com Metaheurísticas, tais como: cozimento simulado (SAUNIER, et al., 2009), algoritmo genético (GA – *Genetic Algorithm*) (ROY; ROY, 2009) busca tabu (QI; SHI; KONG, 2010), enxame de partículas (PAL; MAIT, 2010) e colônia de formigas (GOODARZI; FREITAS; JENSEN, 2009; ALLEGRIINI; OLIVERI, 2011).

Com o objetivo de otimizar a dimensionalidade de modelos multivariados e evitar o fenômeno de sobreajuste (*overfitting*) na determinação de componentes principais, Xu e Liang (2001) empregaram a Simulação de Monte Carlo em conjuntos de dados simulados e em dois casos reais. Gourvénec et al. (2003) compararam Monte Carlo *cross-validation* com o método tradicional de validação cruzada (*cross-validation*) do algoritmo PLS na determinação do número adequado de variáveis latentes.

Adler e Yazhemsky (2010) combinaram a Simulação de Monte Carlo, PCA e a Análise Envoltória de Dados (DEA - *Data Envelopment Analysis*) num contexto em que há um

número relativamente grande de variáveis com relação ao número de observações para auxílio à decisão.

Na modelagem por múltiplos critérios (MCDM – *Multiple Criteria Decision-Making*), Llobet et al. (2005) utilizaram Fuzzy na classificação de amostras de batatas fritas (*crispis*). Para previsão de propriedades oxidativas e hidrolíticas, foi empregado um nariz eletrônico baseado em modelos PLS, com seleção prévia de variáveis de entrada via metaheurística GA.

Wu, Feng e Wen (2011), em estudos relacionados à Botânica, compararam o desempenho do crescimento de uma espécie de árvore – *Carya Cathayensis Sarg* pelos métodos PCA e Análise Hierárquica (AHP – *Analytic Hierarchy Process*), elencando vantagens e desvantagens de cada método, embora os resultados obtidos por ambos foram basicamente idênticos.

Zhang et al. (2006) combinaram o método PROMETHEE (*Preference Ranking Organization Method for Enrichment Evaluations*), oriundo do ELECTRE (*Elimination et Choix Traduisant la Réalité*), e GAIA (*Geometrical Analysis for Interactive Assistance*) com os métodos quimiométricos PCA e PLS para classificar 67 óleos vegetais e determinar um indicador de qualidade dos produtos. Purcell, O'Shea e Kokot (2007) também combinaram PROMETHEE e GAIA com PCA e PLS em estudos relacionados à clonagem de cana-de-açúcar.

Em relação aos gráficos de controle destinados ao monitoramento do vetor de médias, Machado e Costa (2008) estudaram o desempenho dos gráficos T^2 de Hotelling baseados em componentes principais para o monitoramento de processos multivariados. Lourenço et al. (2011) utilizaram princípios da Tecnologia Analítica de Processos (PAT – *Process Analytical Technology*) na construção das cartas de controle a partir dos *scores* dos primeiros componentes principais de modelos PLS em função do tempo, para o monitoramento *on-line* de processos de granulação numa indústria farmacêutica.

Além disso, a Análise Multivariada é uma importante técnica de auxílio à decisão nas mais variadas áreas do conhecimento como: *Data Mining* (KETTANEH; BERGLUND; WOLD, 2005); Econometria (MACKAY, 2006); Marketing (AHN; CHOI; HAN, 2007) e *Supply Chain Management* (POZO et al., 2012).

2.2 REVISÃO SOBRE OS MÉTODOS UTILIZADOS

2.2.1 Ciência e tecnologia envolvidas na mineração dos dados

Na indústria de petróleo, os sinais de radiação infravermelha, gerados por sensores estão associados à previsão da qualidade de produtos destilados como nafta, gasolina, diesel e querosene de aviação (KIM; CHO; PARK, 2000).

Pasquini (2003) explica essa instrumentação: a radiação policromática emitida por uma fonte tem comprimentos de onda selecionados pelo interferômetro de Michelson, conforme ilustrado na Figura 2. O divisor de feixe possui um índice de refração tal que, aproximadamente metade da radiação é direcionada para o espelho 1, e, a outra metade é transmitida, atingindo o espelho 2. Em ambos, a radiação é novamente refletida e a resultante é direcionada para a amostra. As diferenças de caminho ótico ocorrem devido ao movimento relativo dos espelhos, o que promove a controlada interferência destrutiva das ondas, selecionando desta forma, apenas comprimentos de onda na região do infravermelho.

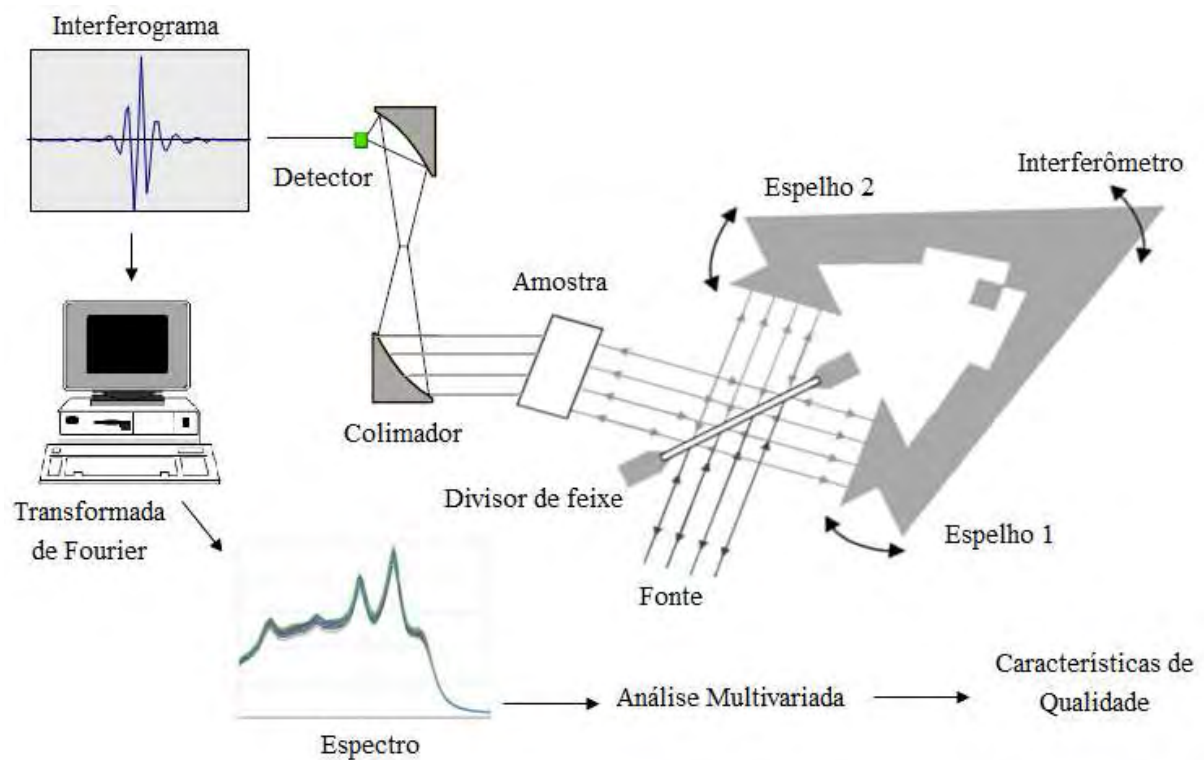


Figura 2. Esquema da tecnologia para aquisição do banco de dados (Adaptado de Pasquini, 2003)

Um interferograma é obtido como resultado da intensidade de radiação, que é transmitida e chega ao detector, em função da diferença de caminho percorrido pelas ondas (KHASHAN; NAGGAR; SHADDAD, 2000). Por meio do cálculo da Transformada de Fourier (FT – *Fourier Transform*), o interferograma, pode ser escrito como uma soma de senos

e cossenos (TARUMI et al, 2005) e neste caso, passa a ser chamado de espectro de transmitância, T (FORATO; FILHO; COLNAGO, 1997). Por fim, o espectro de transmitância, T , é convertido em espectro de absorbância, A , pelo cologaritmo de T (SUAREZ *et al.*, 2011). A absorbância pode ser interpretada como a quantidade de radiação que a amostra absorve e a transmitância, a fração da radiação que a amostra não absorve. Estes fenômenos ocorrem em função da sua composição química (KRAMER; SMALL, 2007).

As ligações químicas do tipo carbono- hidrogênio (C-H), oxigênio-hidrogênio (O-H) e nitrogênio-hidrogênio (N-H), presentes em produtos de petróleo (PASQUINI; BUENO, 2007), são as responsáveis pela absorção da radiação infravermelha, porém, não são muito intensas e sobrepõem-se, conforme espectros de óleo diesel utilizados neste trabalho e representados na Figura 3. As largas bandas espectrais formadas são de difícil interpretação (SKOOG; HOLLER; CROUCH, 2007) devido ao fenômeno da colinearidade (NAES; MARTENS, 1984). A origem deste fenômeno está associada à maneira pela qual a radiação infravermelha interage com a matéria e pode ser demonstrada pela Mecânica Quântica no trabalho de Pasquini (2003).

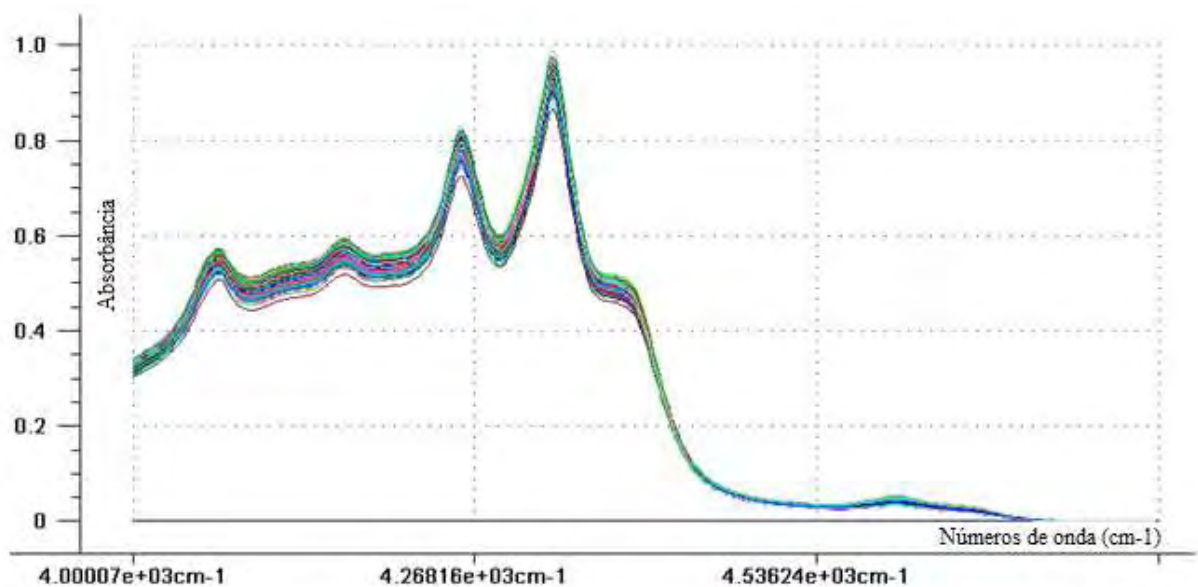


Figura 3. Conjunto de espectros de absorção do óleo diesel

Essas variáveis de absorção, X_i , variáveis de entrada, que compõe a matriz de processo, são correlacionadas entre si, portanto são ditas colineares ou multicolineares (NAES et al., 2002).

Para ilustrar a colinearidade, seja a matriz fictícia $\mathbf{X}$ com a_{ij} termos em i linhas e j colunas, onde a_{ij} representa a absorção da radiação de três amostras i ($i=1, 2, 3$) em dois comprimentos de onda j ($j=1,2$).

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix}$$

As colunas de $\mathbf{X}$ são linearmente dependentes, portanto, as variáveis das colunas j_1 e j_2 são colineares, ou seja, quando j_1 aumenta, j_2 aumenta na mesma proporção. Isso faz com que o determinante de $\mathbf{X}'\mathbf{X}$ seja zero, onde $\mathbf{X}'$ é a matriz transposta de $\mathbf{X}$.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 14 & 28 \\ 28 & 56 \end{bmatrix}$$

Então, o $\det(\mathbf{X}'\mathbf{X}) = (14.56) - (28.28) = 0$ e isto, segundo Naes et. al (2002) significa que há erro de matriz singular e que serão propagados erros na determinação de propriedades dependentes, Y , pelos métodos de regressão que não se baseiam em componentes principais, como a MLR.

No entanto, as abordagens multivariadas de Regressão por Componentes Principais (PCR – *Principal Components Regression*) e PLS têm se mostrado bastante adequadas graças à redução de dimensionalidade, que cria um novo conjunto de variáveis denominadas componentes principais (RAJALAHTI; KVALHEIM, 2011).

Esses métodos aplicados em sinais de radiação possibilitam manipular os dados de absorção, associados a mais de uma frequência ou comprimento de onda ao mesmo tempo (HELFER et al., 2006).

2.2.2 Análise Multivariada

2.2.2.1 Considerações iniciais

Os processos de fabricação em geral, podem apresentar muitas vezes variáveis correlacionadas em função da qualidade dos processos envolver, um grande número de características (PAIVA, 2006). Métodos de redução de dimensão auxiliam na eliminação da redundância dos dados e na geração de bases de dados com menos variáveis (MINGOTI, 2005).

A Análise Multivariada corresponde a um conjunto de métodos estatísticos no qual todas ou muitas variáveis de um conjunto de dados constituem informação para tomada de decisão (RAJALAHTI; KVALHEIM, 2011), tais como regressão linear múltipla (MLR), análise por componentes principais (PCA), regressão por componentes principais (PCR) e regressão por mínimos quadrados parciais (PLS) (OTTO, 2007; ESCANDAR et al., 2006).

2.2.2.2 Regressão Linear Múltipla

A MLR relaciona a propriedade de interesse com os dados observados por meio da equação 1, onde X_i são as variáveis explicativas independentes, β_i são os coeficientes de regressão e ε_i , o erro da estimativa (PIRES et al., 2008):

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \varepsilon_i \quad (1)$$

Segundo Montgomery (2004), o Método dos Mínimos Quadrados Ordinários (OLS – *Ordinary Least Squares*) permite determinar β_i , minimizando a soma dos quadrados dos erros, conforme equação 2.

$$L = \sum_{i=1}^n \varepsilon_i^2 \quad (2)$$

Substituindo a equação 1 em 2, tem-se:

$$L = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j X_{ji} \right)^2$$

A função L deve ser minimizada em função de $\beta_0, \beta_1, \dots, \beta_k$.

$$\frac{\partial L}{\partial \beta_j} = -2 \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \hat{\beta}_j X_{ji} \right) = 0$$

Simplificando, tem-se a igualdade:

$$\hat{\beta}_0 \sum_{i=1}^n X_{ik} + \hat{\beta}_1 \sum_{i=1}^n X_{ik} X_{i1} + \hat{\beta}_2 \sum_{i=1}^n X_{ik} X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{ik}^2 = \sum_{i=1}^n X_{ik} y_i$$

Em notação matricial, tem-se $\mathbf{Y} = \mathbf{\beta X} + \boldsymbol{\varepsilon}$, onde:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}, \mathbf{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Então:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$L = \mathbf{y}' \mathbf{y} - \boldsymbol{\beta}' \mathbf{X}' \mathbf{y} - \mathbf{y}' \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta}$$

$$L = \mathbf{y}' \mathbf{y} - 2\boldsymbol{\beta}' \mathbf{X}' \mathbf{y} + \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta}$$

Minimizando a função:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{X}' \mathbf{y} + 2\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = 0$$

Portanto, obtém-se a equação 3:

$$\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{y} \quad (3)$$

e, conseqüentemente, os coeficientes são determinados pela equação 4:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (4)$$

Além da questão da multicolinearidade, para o uso dessa técnica, o número de amostras deve ser maior que o número de variáveis (MUTIHAC; MUTIHAC, 2008).

No entanto, para minimizar essas dificuldades, a seleção de variáveis pode anteceder a MLR, conforme representada na Figura 4 mediante conhecimento prévio de quais variáveis se relacionam com a propriedade de interesse (não se aplica a este trabalho), ou ainda, através de métodos mais sofisticados como a Metaheurística GA (KONZEN et al., 2003) e o Algoritmo das Projeções Sucessivas (SPA – *Successive Projections Algorithm*) (OTERO et al., 2011).

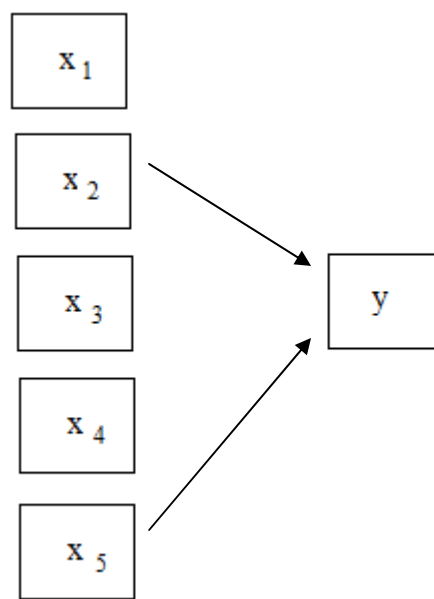


Figura 4. Seleção de variáveis para regressão por MLR

Segundo Naes et al. (2002), neste ambiente de muitas variáveis e colineares, a alternativa mais adequada à regressão clássica é a abordagem por componentes principais, que será exposta adiante e, especificamente, para tratar de dados oriundos de sinais de frequência o método PLS se mostra ser o mais adequado.

2.2.2.3 Componentes Principais e Variáveis Latentes

Segundo Morrison (2005), a técnica de Análise por Componentes Principais, PCA, foi originada por K. Pearson em 1901 e posteriormente publicada por Hotteling em 1933. Tem por objetivo explicar a estrutura da matriz variância-covariância $\Sigma_{p \times p}$ ou matriz de correlação

$P_{p \times p}$ de um vetor aleatório $\mathbf{X}$, composto de p -variáveis aleatórias, pela construção de k combinações lineares, ortogonais, das variáveis originais, com $k < p$.

A seguir, serão mostrados alguns conceitos introdutórios da Álgebra Linear e Estatística Multivariada, segundo Mingoti (2005), para que então as equações dos componentes principais possam ser formalmente apresentadas.

Seja $\mathbf{X}$ um vetor contendo p componentes, onde cada componente é uma variável aleatória, isto é, X_i é uma variável aleatória, $\forall i=1, 2, \dots, p$. Então, $\mathbf{X}$, denominado vetor aleatório é denotado por:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

Seja a matriz de covariâncias $\Sigma_{p \times p}$:

$$\Sigma_{p \times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

onde $\sigma_{ii} = \sigma_i^2$, covariância de X_i e X_j , $\text{Cov}(X_i, X_j) = \sigma_{ij}$ e é igual a esperança $E[(X_i - \mu_i)(X_j - \mu_j)]$, para vetor de médias μ .

No caso de um vetor aleatório de dimensão p , ou seja, $\mathbf{X}' = [X_1 \ X_2 \ \dots \ X_p]$ diz-se que este vetor tem uma distribuição normal p -variada, e denota-se $\mathbf{X} \sim N(\mu, \Sigma_{p \times p})$, se a função densidade de probabilidade de $\mathbf{X}$ for dada pela equação 5, embora a suposição de distribuição normal multivariada não seja requisito necessário para que a técnica de componentes principais possa ser utilizada:

$$f(x_1, x_2, \dots, x_p) = \frac{1}{2\pi^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right\} \quad (5)$$

A quantidade $(\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)$ é referida como distância de Mahalanobis do vetor $\mathbf{x}$ ao vetor de médias μ .

Se $\Sigma_{p \times p}$ é uma matriz de covariâncias, então existe uma matriz ortogonal $\mathbf{O}_{p \times p}$, isto é, $\mathbf{O}'\mathbf{O} = \mathbf{O}\mathbf{O}' = \mathbf{I}_{p \times p}$, tal que:

$$\mathbf{O}'\Sigma\mathbf{O} = \begin{bmatrix} \lambda_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \lambda_p \end{bmatrix} = \Lambda$$

onde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ são autovalores ordenados em ordem decrescente da matriz $\Sigma_{p \times p}$ e $\mathbf{I}_{p \times p}$ é a correspondente matriz identidade.

A *i-ésima* coluna da matriz $\mathbf{O}_{p \times p}$ é o autovetor normalizado $\mathbf{e}_i$ correspondente ao autovalor λ_i , $i=1,2,\dots,p$, denotado por:

$$\mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

com comprimento unitário:

$$\|\mathbf{e}_i\| = \sqrt{(e_{i1}^2 + e_{i2}^2 + \dots + e_{ip}^2)} = 1$$

Por fim, seja o sistema de combinações lineares de $\mathbf{X}$ constituído de i equações do tipo (MACHADO; COSTA, 2008; ANDERSON, 2003):

$$PC_1 = \mathbf{a}_1' \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$PC_2 = \mathbf{a}_2' \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

Generalizando, tem-se a equação 6:

$$PC_i = \mathbf{a}_i' \mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p \quad (6)$$

E deste modo, tem-se as equações 7 e 8:

$$\text{Var}(PC_i) = \mathbf{a}_i' \Sigma_{p \times p} \mathbf{a}_i \quad i = 1, 2, \dots, p \quad (7)$$

$$\text{Cov}(PC_i, PC_j) = a_i' \sum_{p \times p} a_j \quad i \neq j, i, j = 1, 2, \dots, p \quad (8)$$

De acordo com Johnson e Wichern (2002) e Mingoti (2005), o objetivo dos componentes principais é encontrar os valores dos coeficientes a_{ij} tais que $a_i' a_i = 1$, de modo que as combinações lineares $PC_1, PC_2, \dots, PC_i$ são não correlacionadas entre si e que tenham a máxima variância.

A Desigualdade de Cauchy-Schwarz assegura que a máxima variância de $PC_1 = a_1' X$, sob a restrição $a_1' a_1 = 1$ é igual a λ_1 e é obtida quando $a_1 = e_1$, ou seja, o autovetor normalizado correspondente a λ_1 , sendo PC_1 chamado de primeiro componente principal (PC – *Principal Component*). A variância máxima de $PC_2 = a_2' X$ sob as restrições $a_2' a_2 = 1$ e $\text{cov}(PC_1, PC_2) = 0$, é igual a λ_2 e é obtida quando $a_2 = e_2$, o autovetor normalizado correspondente a λ_2 sendo PC_2 chamado de segundo componente principal. A máxima variância de $PC_3 = a_3' X$, sob as restrições $a_3' a_3 = 1$ e $\text{Cov}(PC_1, PC_3) = \text{Cov}(PC_2, PC_3) = 0$ é igual a λ_3 , sendo PC_3 o terceiro componente principal.

Desta forma, a variância máxima de $PC_i = a_i' X$, sob as restrições $a_i' a_i = 1$ e $\text{Cov}(PC_i, PC_j) = 0$ para $j < i$, é obtida quando $a_i = e_i$, ou seja, o autovetor normalizado correspondente ao autovalor λ_i , sendo PC_i chamado de *i-ésimo* componente principal. Sob forma de função objetivo e restrições, sejam as equações 9 e 10:

$$\text{Maximizar } \text{Var}(e_i' X) \quad (9)$$

$$\text{Sujeito a: } \begin{cases} \text{Cov}(PC_i, PC_j) = 0 \text{ para } j < i \\ e_i' e_i = 1 \end{cases} \quad (10)$$

A proporção da variação total de X que é explicada pelo *j-ésimo* componente principal é definida conforme equação 11:

$$\frac{\lambda_j}{\sum_{i=1}^p \lambda_i} \quad (j \leq i) \quad (11)$$

A proporção da variação total que é explicada pelos k primeiros componentes principais é definida conforme equação 13:

$$\frac{\sum_{j=1}^k \lambda_j}{\sum_{i=1}^p \lambda_i} \quad (12)$$

Contudo, na prática, a matriz Σ_{pxp} é desconhecida e precisa ser estimada pela matriz de covariância amostral S_{pxp} .

$$S_{pxp} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Sendo $S_{ij} = S_{ji}$ e S_{ii} definidos pelas equações 13 e 14:

$$S_{ii} = \frac{\sum_{i=1}^n (x_{il} - \bar{X}_i)^2}{n-1} \quad (13)$$

$$S_{ij} = \frac{\sum_{l=1}^n (x_{il} - \bar{X}_i)(x_{jl} - \bar{X}_j)}{n-1} \quad (14)$$

Os componentes principais também podem ser obtidos via matriz de correlações $\mathbf{P}_{pxp}$ e, conseqüentemente matriz de correlação amostral $\mathbf{R}_{pxp}$ dadas por:

$$\mathbf{P}_{pxp} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{R}_{pxp} = \begin{bmatrix} 1 & R_{12} & \cdots & R_{1p} \\ R_{21} & 1 & \cdots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p1} & R_{p2} & \cdots & 1 \end{bmatrix}$$

onde ρ_{ij} é o coeficiente de correlação entre a i -ésima e j -ésima variáveis do vetor X definido pela equação 15:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \quad , \quad -1 \leq \rho_{ij} \leq 1, \quad i, j = 1, 2, \dots, p \quad (15)$$

Portanto, o coeficiente de correlação amostral R_{ij} será função de S_{ij} e S_{ii} .

Freitas et al. (2012) e Otto (2007) realizaram estudos sobre as projeções das amostras nos componentes principais no espaço tridimensional. Os objetos da Figura 5 representam amostras, centralizadas na média, compostas de três variáveis aleatórias X_1 , X_2 e X_3 .

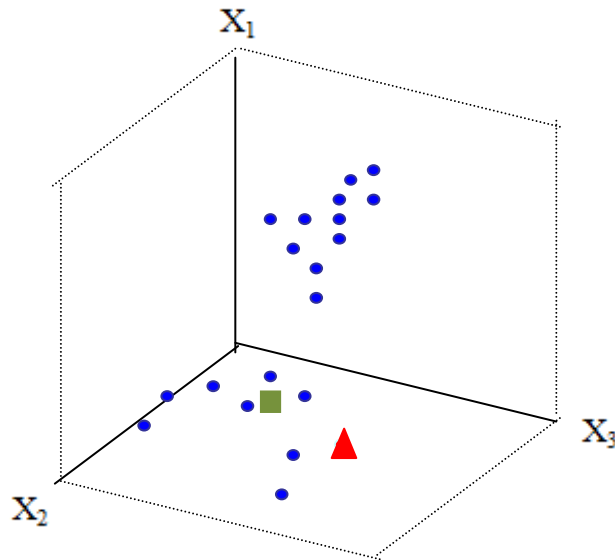


Figura 5. Conjunto de amostras no espaço tridimensional

O primeiro componente principal PC1 é uma combinação linear dos valores X_1 , X_2 e X_3 , que represente o máximo possível de variabilidade dos dados. O eixo do primeiro componente principal é representado, na Figura 6.

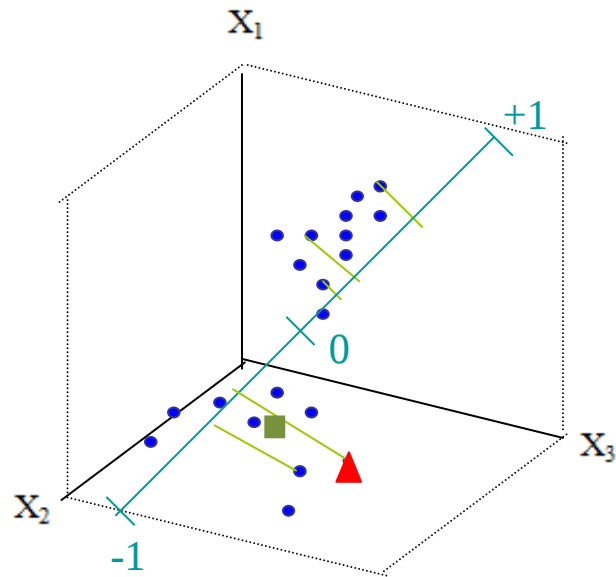


Figura 6. Primeiro componente principal plotado no espaço tridimensional

As projeções de cada amostra sobre esse novo eixo representam os valores denominados *scores* do primeiro componente principal PC1. Na prática, o *score* é o valor numérico do PC (DAVIS et al., 2000).

Para exemplificar, as amostras representadas por triângulo e quadrado, apresentam os mesmos valores de *scores* neste PC1, conforme ilustrado na Figura 7.

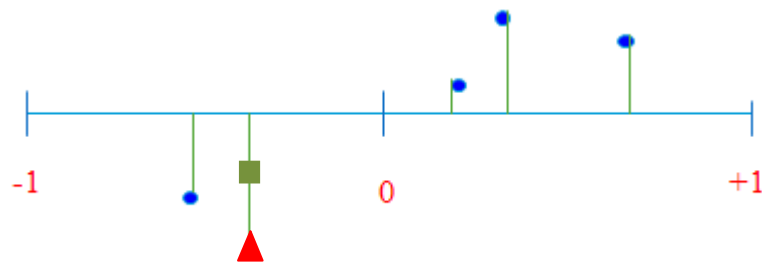


Figura 7. Projeção das amostras no primeiro componente principal

Se um segundo componente principal (PC2) for adicionado ao sistema, o novo eixo deve ser ortogonal ao eixo de PC1 e posicionado de forma a obter a máxima variabilidade residual possível dos dados. Novamente, os *scores* de PC2 podem ser obtidos pela projeção das amostras sobre este eixo, conforme ilustrado na Figura 8.

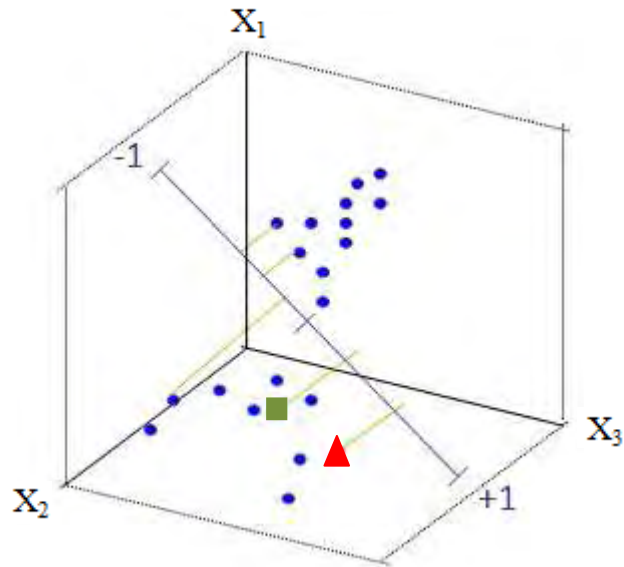


Figura 8. Segundo componente principal plotado no espaço tridimensional

As amostras representadas por triângulo e quadrado apresentam valores distintos de *scores* no PC2, conforme ilustrado na Figura 9, portanto, este foi o responsável por explicar a variabilidade existente entre elas.

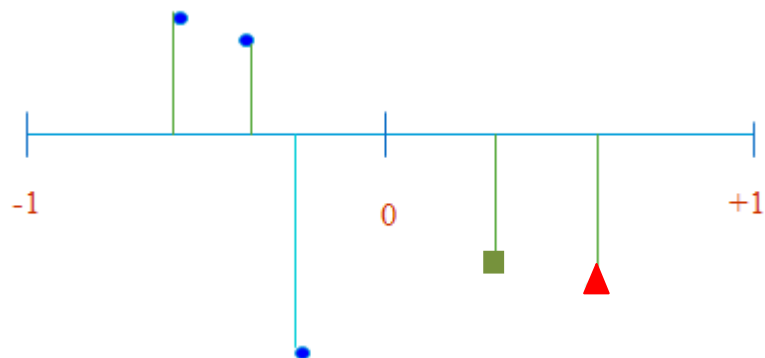


Figura 9. Projeção das amostras no segundo componente principal

Na interpretação geométrica, os *scores* relacionam os componentes principais com as amostras, enquanto os *loadings*, os relacionam com as variáveis (LORHO; WESTAD; BRO, 2006). Sua projeção é interpretada como o ângulo formado entre a variável e a PC, mais especificamente o cosseno do ângulo, representando o quanto aquela variável contribui na formação do PC (OTTO, 2007). Assim, ângulos pequenos fornecem elevados valores de cosseno e significa que essas variáveis contribuem mais para explicação da variabilidade do que aquelas formadas por ângulos maiores, conforme representado na Figura 10 (LI; QIN; ZHOU, 2010).

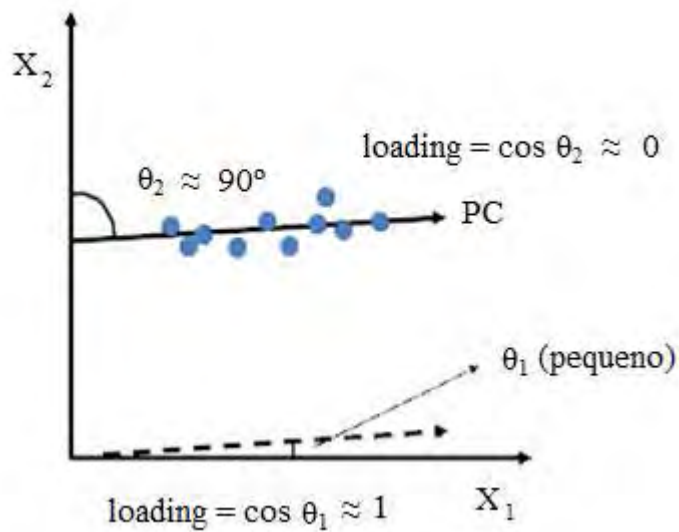


Figura 10. Interpretação geométrica dos *loadings*

Todo o desenvolvimento feito até agora não se trata de regressão por componentes principais, apenas redução de dimensionalidade dos dados, porém, é requisito para apresentação das equações por PCR e PLS, que empregam essas combinações lineares nesse processo.

Segundo Naes et al. (2002), a PCR já supera as limitações da MLR, pois a regressão é feita não com as variáveis originais, mas com os *scores* da matriz obtidos numa PCA.

Da mesma maneira, os componentes principais são combinações lineares determinadas pela habilidade em explicar a variabilidade em X . O primeiro componente principal $\hat{t}_1$ é dado pela combinação linear com a maior variância possível. O autovetor normalizado associado é denotado por $\hat{p}_1$. O segundo componente $\hat{t}_2$, é definido da mesma maneira e com a restrição de não ser correlacionado com $\hat{t}_1$. A segunda direção do vetor é denotada por $\hat{p}_2$ e assim por diante.

A matriz de *scores* dos componentes principais é denotada por $\hat{T}$ e a matriz de *loading* é denotada por $\hat{P}$ e pode ser expressa pela equação 16:

$$X = \hat{T} P' \quad (16)$$

O próximo passo é utilizar a matriz $\hat{T}$ na equação de regressão ao invés da matriz original X . O modelo de regressão pode ser escrito conforme equação 17 (NAES et al., 2002):

$$Y = \hat{T} q + f \quad (17)$$

e os coeficientes q são estimados por OLS, conforme equação 18.

$$\hat{q} = (T'T)^{-1}T'Y \quad (18)$$

A Figura 11 ilustra a compressão de dados, onde todas as quatro variáveis X_i são transformadas em apenas dois componentes principais t_1 e t_2 .

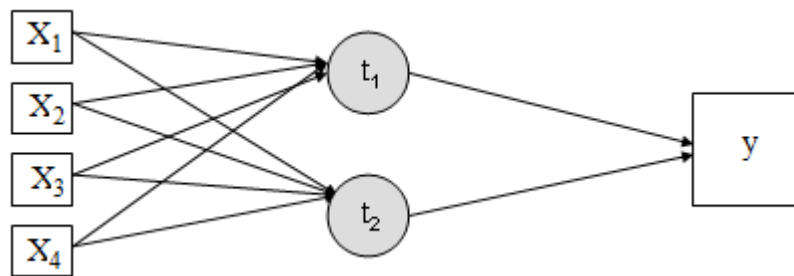


Figura 11. Variáveis transformadas em combinações lineares, relacionadas com Y por uma equação de regressão (Adaptado de NAES et al., (2002)

Uma vez que os componentes principais são ortogonais entre si, o problema de multicolinearidade é contornado (NAES; MARTENS, 1984).

São utilizados na PCR, os primeiros componentes principais que melhor explicam a variabilidade dos dados em X , portanto os k primeiros componentes principais da PCA. Este método, porém, não é capaz de garantir que a matriz T possua a melhor covariância com a matriz Y . Isso ocorre principalmente em função da decomposição gerada pela técnica PCA ser aplicada somente na matriz das variáveis explicativas X (NAES et al., 2002).

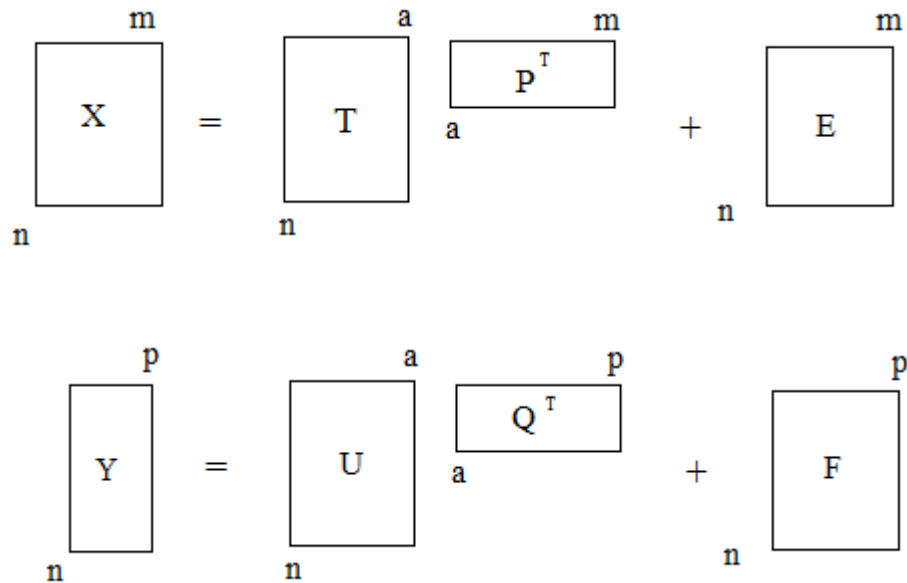
Na regressão por Mínimos Quadrados Parciais, PLS, os componentes principais são definidos levando-se em consideração os valores da variável Y . Nessa técnica, cada componente é obtido de forma a maximizar a covariância entre Y e todas as possíveis funções lineares das variáveis X . Com isso, são obtidos componentes que melhor se relacionam com a variável Y e, ao mesmo tempo, explicam da melhor maneira possível a variabilidade em X (JOLLIFFE, 2002).

Inicialmente, foi proposto um procedimento para obtenção dos blocos T e U , isto é, basicamente era realizada uma PCA para a matriz X e outra para a matriz Y . As matrizes X e Y são, então, decompostas em combinações lineares de h , componentes principais, conforme mostrado nas equações 19 e 20 (GELADI; KOWALSKI, 1986):

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \sum \mathbf{t}_h \mathbf{p}_h' + \mathbf{E} \quad (19)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} = \sum \mathbf{u}_h \mathbf{q}_h' + \mathbf{F} \quad (20)$$

Graficamente,



onde, $\mathbf{T}$ e $\mathbf{U}$ são as matrizes de *scores* das matrizes $\mathbf{X}$ e $\mathbf{Y}$, respectivamente; $\mathbf{P}$ e $\mathbf{Q}$ são as chamadas matrizes de *loadings* das matrizes $\mathbf{X}$ e $\mathbf{Y}$, respectivamente e $\mathbf{E}$ e $\mathbf{F}$ são os resíduos.

A correlação entre os dois blocos, $\mathbf{X}$ e $\mathbf{Y}$, é uma relação linear dada pelo coeficiente de regressão linear, conforme equação 21 (GELADI, KOWALSKI, 1986):

$$\hat{\mathbf{u}}_h = \mathbf{b}_h \mathbf{t}_h \quad (21)$$

Porém, a metodologia descrita não garante a melhor covariância entre os blocos $\mathbf{X}$ e $\mathbf{Y}$, pois $\mathbf{T}$ e $\mathbf{U}$ são obtidos separadamente (ZAMORA et al., 1997).

Com a evolução computacional, o PLS passa a empregar o algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*) no qual os componentes principais, agora chamados de variáveis latentes (LV – *Latent Variable*), produzem uma matriz de peso $\mathbf{W}$ que reflete uma estrutura de covariância entre os blocos $\mathbf{X}$ e $\mathbf{Y}$ (MARTINS; TEÓFILO; FERREIRA, 2010). A matriz $\mathbf{W}$ é obtida de forma a aumentar a covariância entre os vetores $\mathbf{t}$

e $\mathbf{u}$, isto é, ocorrem mudanças em $\mathbf{t}$ e $\mathbf{u}$ até que se encontre a melhor relação linear entre eles (ERGON, 2004).

As colunas $\mathbf{W}$ são vetores peso para $\mathbf{X}$, produzindo a matriz de *scores* $\mathbf{T}$ ($\mathbf{T} = \mathbf{XW}'$). Cálculos de OLS são realizados em $\mathbf{Y}$ e $\mathbf{T}$, para produzir os pesos $\mathbf{q}$ para $\mathbf{Y}$ ($\mathbf{Y} = \mathbf{TQ} + \mathbf{F}$) (BAFFI; MARTIN; MORRIS, 1999).

Com o aumento da Cov ($\mathbf{u}$, $\mathbf{t}$) pelo algoritmo de NIPALS, entretanto, os vetores $\mathbf{t}$ deixam de ser ortogonais. Para solucionar esse problema, após, verificada a convergência da relação entre $\mathbf{t}$ e $\mathbf{u}$, um bloco $\mathbf{P}$, matriz de *loading*, é calculado de modo a garantir a ortogonalidade entre os vetores da matriz $\mathbf{T}$ ($\mathbf{X} = \mathbf{TP} + \mathbf{E}$) (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

A segunda variável latente é encontrada da mesma maneira, porém usando as matrizes de resíduos $\mathbf{E}$ e $\mathbf{F}$, ao invés das matrizes de dados $\mathbf{X}$ e $\mathbf{Y}$, após subtração da primeira variável latente. A idéia é explicar a variabilidade residual a partir de outra LV pelo mecanismo de validação cruzada (CV – *Cross Validation*) (BASTIEN; VINZI; TENENHAUS, 2005). Este algoritmo é capaz de reduzir gradualmente, aleatoriamente, uma amostra de cada vez, (*leave-one-out*) do conjunto de dados (KIRSANOV et al., 2012) e um modelo construído a partir de amostras remanescentes é testado pela comparação com os valores verdadeiros das amostras excluídas (RONCHETTI; FIELD; BLANCHARD, 1997).

Vários métodos heurísticos e critérios estatísticos podem ser utilizados para a determinação do número de componentes principais ou variáveis latentes tais como: Critério de Kaiser, *Scree plot*, Validação Cruzada, Indicador de Malinowski. (OTTO, 2007; HENRY; PARK; SPIEGELMAN, 1999). A validação cruzada é um método prático para tal determinação e, com elevado grau de previsibilidade em análise quimiométrica (FERREIRA; MONTANARI; GAUDIO, 2002), pois permite verificar o poder preditivo para cada componente adicionado, e então parar a extração destes quando começarem a ocorrer componentes com baixo poder preditivo (MORELLATO, 2010).

A validação cruzada permite estabilizar o número de variáveis latentes entre dois extremos denominados *underfitting* e *overfitting*. (BARROS; RUTLEDGE, 2004). O primeiro fenômeno ocorre quando o modelo não capturou, suficientemente, importante variabilidade dos dados. Por outro lado, o segundo ocorre, quando se utiliza mais variáveis latentes que o necessário, ou seja, elevada redundância oriunda das variáveis $\mathbf{X}$. Em ambos os casos, os modelos resultarão em previsões de baixa qualidade.

A Figura 12 ilustra os dois fenômenos (NAES et al., 2002):

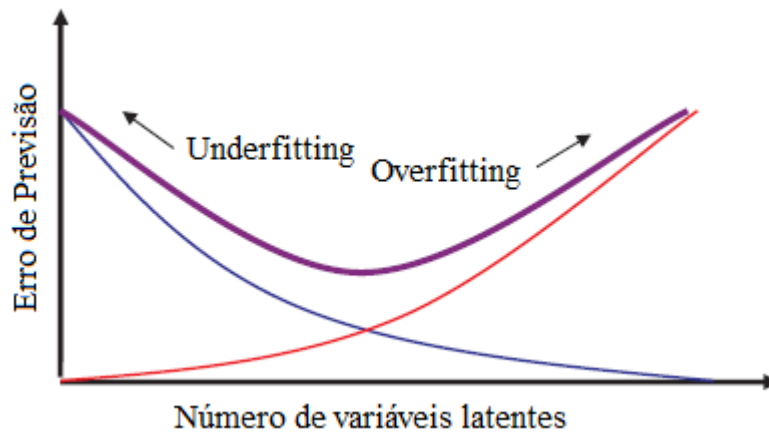


Figura 12. Fenômenos *underfitting* e *overfitting*

Os coeficientes de regressão são obtidos pela equação 22:

$$\hat{\mathbf{b}} = \hat{\mathbf{W}}(\hat{\mathbf{P}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{q}} \quad (22)$$

Durante o desenvolvimento dos modelos, pontos discrepantes, isto é, *outliers*, podem ser identificados e removidos, quando necessário, dentre outras técnicas estatísticas, através das análises de *leverage* e resíduos de *Student* (SHIROMA; SAONA, 2009). A grandeza *leverage*, h , pode ser interpretada como a distância de uma amostra ao centróide do conjunto de dados e pode ser calculada pela equação 23 (SILVA et al., 2012). Altos valores de *leverage* significam que a amostra está distante da média e apresenta grande influência no modelo. Os resíduos de *Student* podem ser interpretados como a diferença entre os valores reais e os valores preditos pelo modelo.

$$h_{ii} = \frac{1}{n} + (\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \quad (23)$$

Portanto, serão *outliers*, amostras com elevados valores de *leverage*. Dentre essas, aquelas que apresentarem elevados resíduos, devem ser eliminadas. Do contrário, são *outliers* que devem não somente ser mantidos no modelo, mas, estudados e interpretados.

2.2.3 Processos Autocorrelacionados e Refino de Petróleo

De acordo com Costa, Epprecht e Carpinetti (2005) atualmente, processos industriais contínuos, cada vez mais automatizados, como os químicos, costumam produzir dados autocorrelacionados, isto é, com um grau de correlação entre observações sucessivas X_i e X_{i-k} de uma mesma característica de qualidade, onde k indica a defasagem das observações.

O coeficiente de autocorrelação amostral pode ser calculado pela equação 24:

$$r_k = \frac{\sum_{i=k+1}^n (x_i - \bar{X})(x_{i-k} - \bar{X})}{\sum_{i=1}^n (x_i - \bar{X})^2} \quad (24)$$

No Capítulo 3 será ilustrado, através da propriedade densidade, que este fenômeno também ocorre nos processos de refino do petróleo e quais as implicações neste trabalho, do ponto de vista da modelagem matemática e planejamento experimental.

Camolesi (2009) ao modelar propriedades do querosene de aviação, ainda que no monitoramento *off-line*, identifica que a melhor oportunidade para coleta de amostras, visando compor um modelo, contornando a questão da autocorrelação, é após a troca de carga a refinaria, isto é, elenco de petróleos que serão refinados, considerando algumas horas, em torno de seis horas, para estabilização das unidades operacionais.

A seguir, de uma maneira simplificada, serão apresentados os principais esquemas de refino, com intuito de contextualizar e mostrar a complexidade e interligações dos macro-processos industriais autocorrelacionados, segundo a Agência Nacional do Petróleo.

O primeiro processo de uma refinaria é a destilação atmosférica ou destilação direta, onde os componentes do óleo cru são separados em diferentes cortes utilizando-se diferentes pontos de ebulição, conforme esquema de produção nível 1 representado pela Figura 13. Os principais produtos obtidos nesse processo são: gás liquefeito de petróleo (GLP), nafta – precursor da gasolina, querosene de aviação, diesel e óleo combustível.



Figura 13. Esquema de Produção nível 1

Adicionalmente, as refinarias utilizam uma segunda torre, a destilação a vácuo, para produzir cortes de gasóleos. Essas correntes intermediárias alimentam um processo químico denominado craqueamento catalítico fluido (FCC - *Fluid Catalytic Cracking*). Neste, duas correntes nobres são geradas: o GLP, e a gasolina. Trata-se de um esquema de refino bem mais flexível, embora, modernamente, possa, também, apresentar dificuldades para enquadramento de produtos em especificações mais rigorosas. A Figura 14 representa o esquema de produção nível 2.

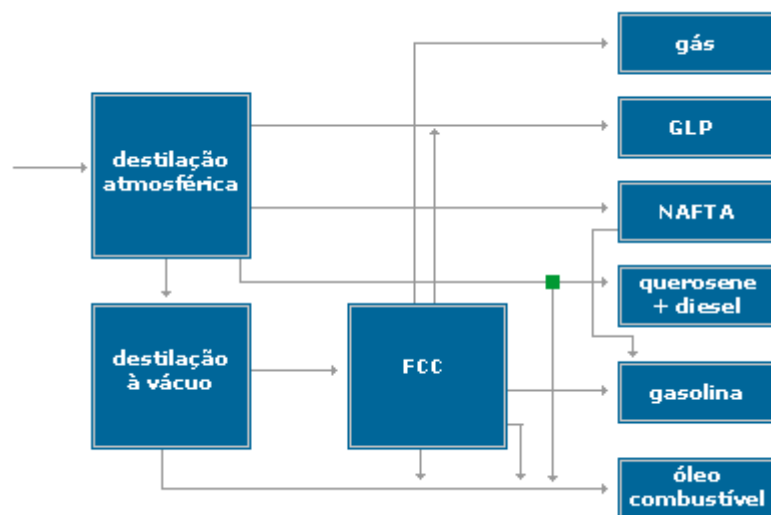


Figura 14. Esquema de Produção nível 2

O esquema de produção nível 3 é mais flexível e rentável que o anterior, pois utiliza o processo químico de coqueamento, que transforma uma fração de menor valor agregado - resíduo de vácuo das torres de destilação, em produtos mais nobres como GLP, gasolina, nafta e óleo diesel, conforme representado na Figura 15.

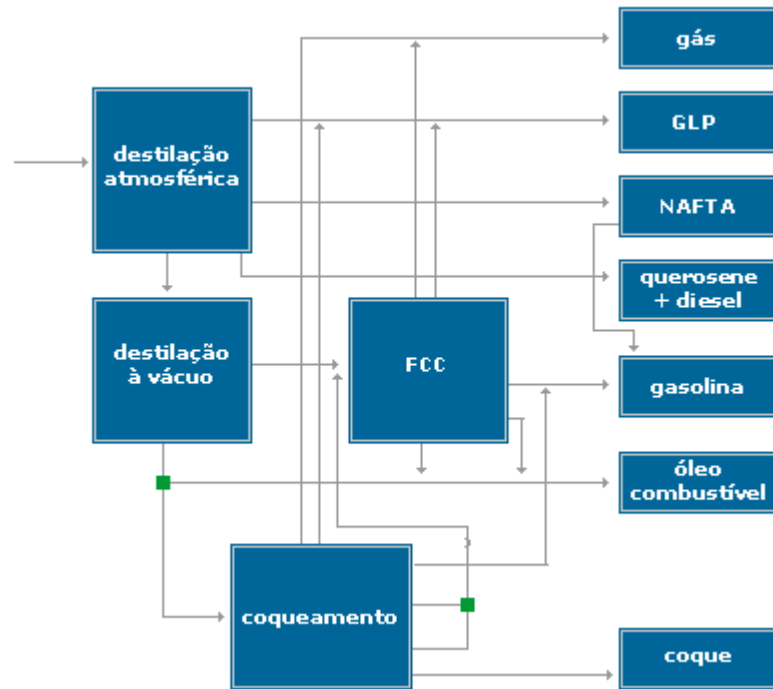


Figura 15. Esquema de Produção nível 3

O último esquema de refino representado incorpora o processo de hidrotratamento de frações médias geradas no coqueamento, possibilitando o aumento da oferta de óleo diesel de boa qualidade. Este esquema permite um maior equilíbrio na oferta de gasolina e de óleo diesel de uma refinaria, gerando, mais óleo diesel e menos gasolina que as configurações anteriores, conforme representado na Figura 16.

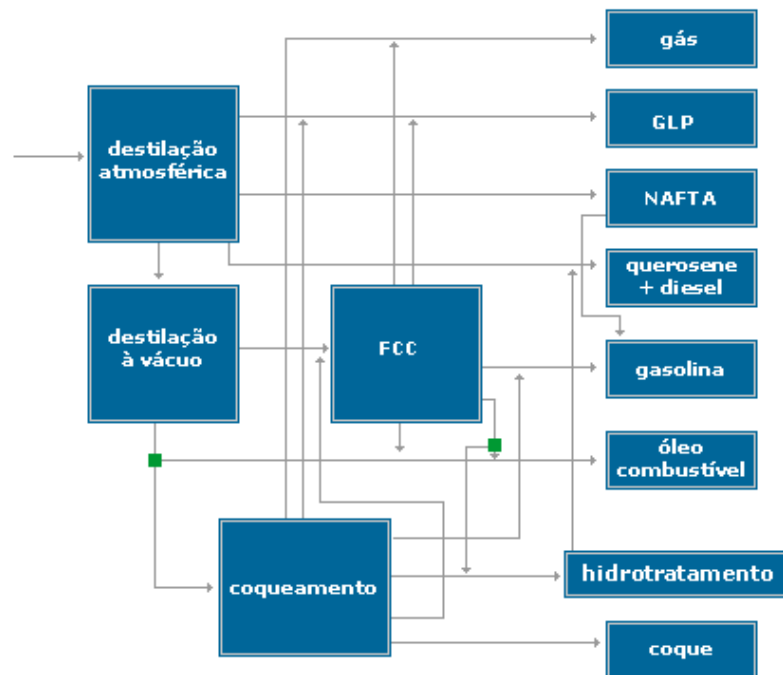


Figura 16. Esquema de Produção nível 4

Evidentemente, existem outros macro-processos e processos auxiliares, como estação de tratamento de água, despejo de efluentes, unidades de recuperação de enxofre, unidades de geração de hidrogênio e, conseqüentemente outras interligações, cujo detalhamento não é objeto deste trabalho.

2.2.4 Caracterização do óleo diesel

A composição química do óleo diesel influi diretamente no seu desempenho e está relacionada com o tipo de petróleo e com os processos utilizados para sua produção em refinarias.

As especificações técnicas criadas para a comercialização deste produto balizam para a área operacional o controle de qualidade a ser realizado através do atendimento a valores de um conjunto de propriedades verificadas por ensaios analíticos normatizados.

Dentre demais propriedades de especificação determinadas pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), as citadas a seguir, comportam-se como variáveis restrições à constante mudança da composição do petróleo (FERREIRA, 2008), por isso foram escolhidas para monitoramento via modelagem matemática.

De acordo com a ASTM D-4052 (2009), a densidade é uma propriedade utilizada como fator de transformação de volume em massa ou vice-versa. É utilizada para cálculos de quantidades (estocagem/suprimento) na comercialização do produto.

Segundo a ASTM D-86 (2010), as características de volatilidade do óleo diesel exercem grande influência sobre seu desempenho e são determinadas pelo ensaio de destilação. Os pontos percentuais recuperados (material condensado) permitem o levantamento da curva de destilação do produto que é utilizada para verificar se as proporções de frações leves e pesadas do combustível estão adequadas, visando obter um bom desempenho na combustão. Este ensaio leva 40 minutos para ser executado em laboratório.

A qualidade de ignição pode ser traduzida pelo número de cetano, conseqüentemente índice de cetano, que representa o atraso de ignição do combustível dentro da câmara de combustão e seu controle está associado ao desempenho global do motor, de acordo com a ASTM D-613 (2010). Este ensaio, que leva quatro horas, é executado em laboratório, por meio de um motor estacionário de combustão interna. Além da longa duração, essas máquinas apresentam altos custos de implementação e exigem mão-de-obra muito especializada para operá-las.

A menor temperatura em que os vapores despendidos do combustível aquecido sob determinadas condições inflam ao contato de uma chama, sem que a queima se mantenha denomina-se ponto de fulgor. O controle desta propriedade tem por objetivo minimizar riscos de incêndio e explosão associados ao transporte e manuseio do produto. O fulgor de um óleo diesel indica a temperatura abaixo da qual ele pode ser manuseado sem risco de fogo (ASTM D-93, 2005). Este ensaio leva 20 minutos para ser executado.

A viscosidade cinemática caracteriza-se pela medida da resistência do fluido ao escoamento a uma dada temperatura. Devido à grande influência da temperatura nos valores de viscosidade, é especificada uma faixa de trabalho, de modo que possíveis variações de temperatura possam ocorrer sem comprometimento da qualidade do produto no sistema de combustível (ASTM-D445, 2009). Este ensaio leva 40 minutos para ser executado.

3 APLICAÇÃO

3.1 Contexto

De uma maneira geral, quanto mais complexa uma refinaria de petróleo, maior a margem do refino e decisões de como operará-la e monitorá-la (WEIMIN; WENKAI; HUI, 2003), bem como quais produtos maximizar e quais as configurações de plantas construir, são fatores que proporcionam vantagens competitivas às indústrias petrolíferas (MANZANO, 2005).

A Figura 15 mostra um exemplo de ganho de margem de refino em função das etapas do processamento industrial:

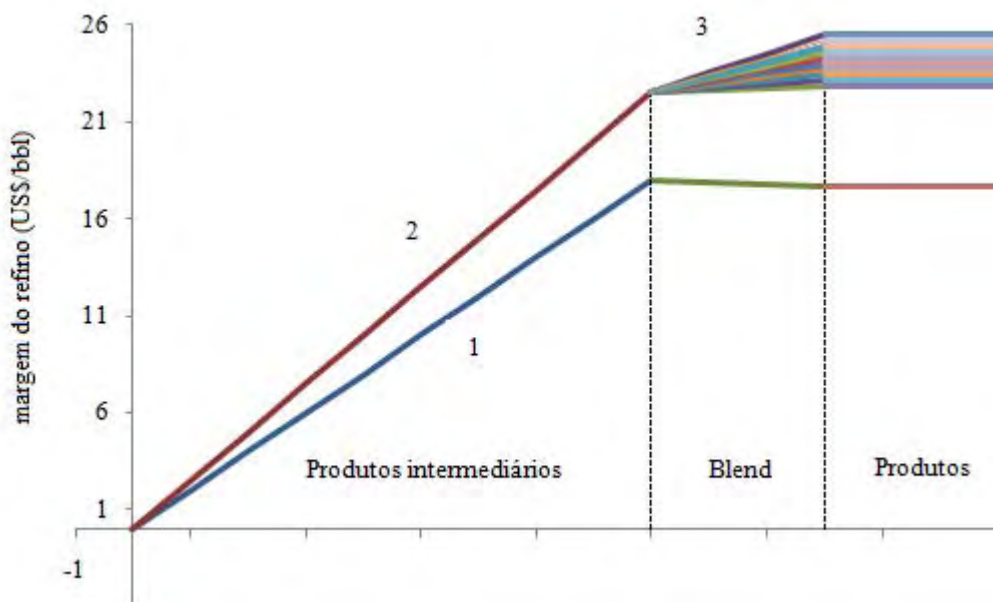


Figura 17. Exemplo de ganho de margem de refino (Adaptada de FERREIRA, 2008)

A etapa produtos intermediários, representa a agregação de valor pela passagem dos hidrocarbonetos nas unidades de processo, utilizando instrumentação básica (curva 1) ou controle avançado com instrumentação analítica (curva 2) (FERREIRA, 2008). A etapa *blend* visa aumentar a rentabilidade da produção considerando as restrições de qualidade dos produtos por meio de ajustes das unidades de processo que promoverão precisão para especificar os produtos com menor variabilidade, além de elevar a produção global dos derivados (curva 3).

3.2 Apresentação do Problema

O óleo diesel é o principal combustível comercializado no mercado brasileiro. Segundo Moreira (1996), seu consumo tem sido estudado, principalmente, com o objetivo de relacioná-lo com aspectos macroeconômicos e de avaliar características do setor de transporte.

De acordo com a projeção de Tolmasquim, Guerreiro e Gorini (2007), em 2030 o óleo diesel deve aumentar sua participação na matriz de consumo de derivados de petróleo para atendimento da demanda, conforme representação da Figura 16. Neste cenário, os derivados de petróleo manteriam sua posição de liderança com uma participação de 30% da matriz energética brasileira.

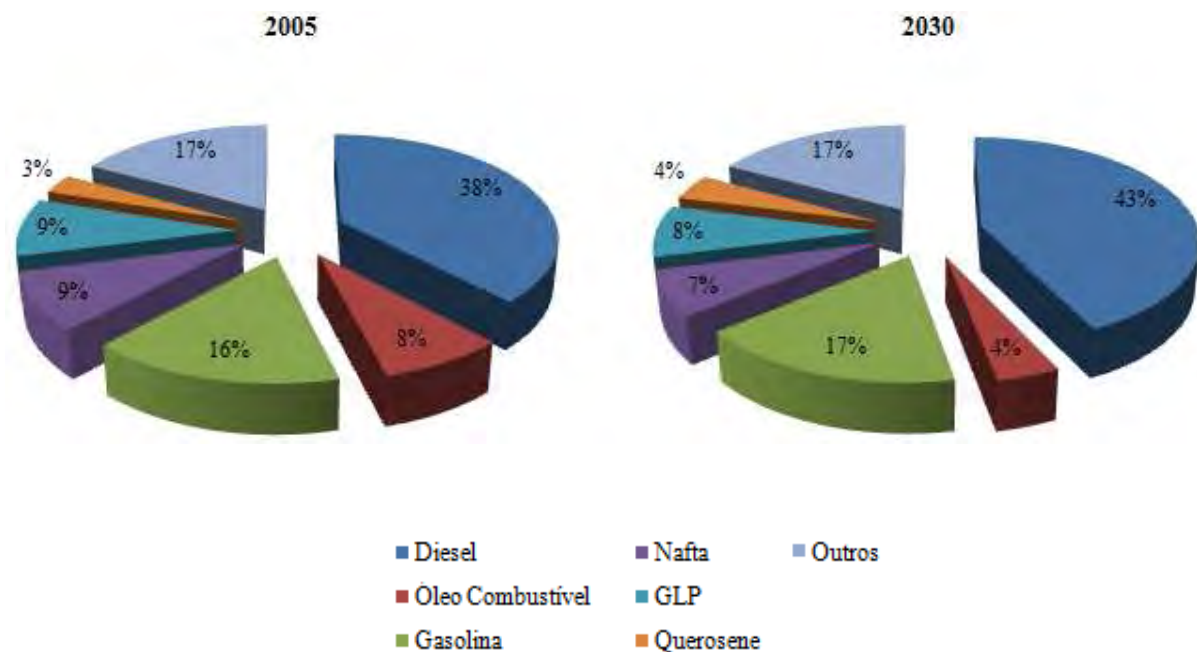


Figura 18. Estrutura do consumo de derivados no Brasil 2005-2030 (Adaptada de TOLMASQUIM; GUERREIRO; GORINE, 2007)

3.3 Solução Proposta

Dentro deste cenário, a solução proposta foi a implementação em escala industrial de modelos matemáticos desenvolvidos e validados para monitoramento *on-line* da produção de óleo diesel. Para tal, serão brevemente apresentados a seguir, alguns detalhes em termos de engenharia de processamento.

Como a matriz da Qualidade, bloco Y, formado pelas propriedades de especificação do produto, é função de sua composição química - matriz de Processo ou bloco X, e este é refletido pelo espectro de absorção de infravermelho, essa Ciência foi convertida em tecnologia no processamento de óleo diesel numa refinaria brasileira.

A Figura 17 representa um fluxograma simplificado da produção de diesel e o local da implementação industrial dos modelos matemáticos.

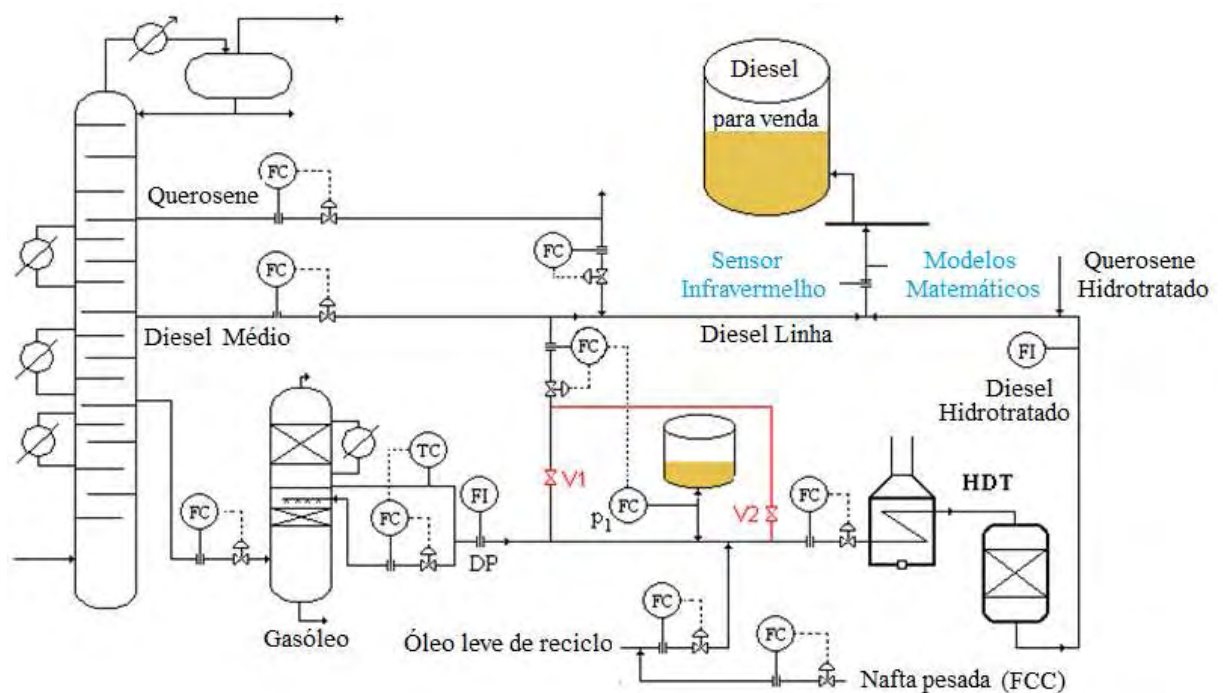


Figura 19. Fluxograma simplificado da produção de diesel numa refinaria de petróleo

Uma vez implementada a instrumentação avançada, foi possível a aplicação dos modelos matemáticos.

O arranjo da Figura 18 representa um esquema simplificado de como o projeto de amostragem foi capaz de viabilizar o monitoramento *on-line*:

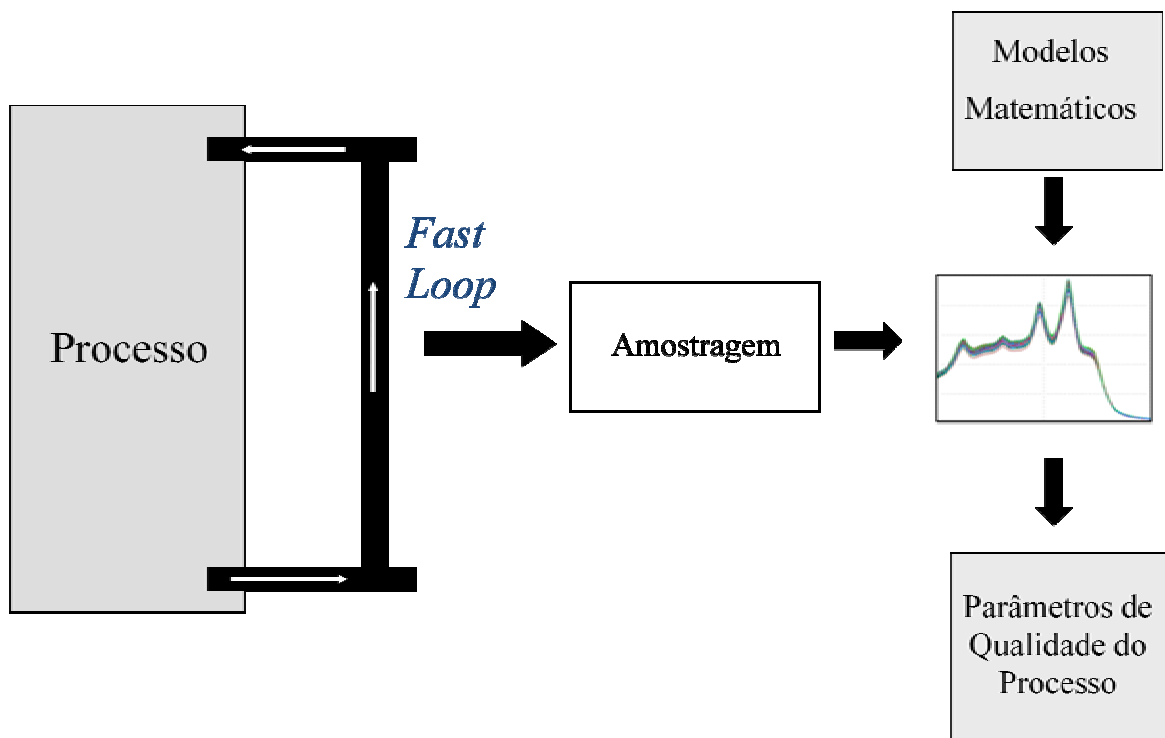


Figura 20. Arranjo simplificado para monitoramento *on-line*

Neste momento, foi determinado um ponto da tubulação do produto final em que a amostra representaria estatisticamente o processo. Nesta localização, foi derivada uma linha com rápida e constante circulação de produto (*fast loop*), com objetivo de garantir a homogeneidade e manutenção da representatividade da amostra. Em seguida, a mesma é bombeada para o chamado sistema de amostragem, com controle de pressão, temperatura e vazão até atingir o circuito ótico detalhado no Capítulo 2. Neste, a amostra é aprisionada por um sistema de válvulas e submetida à radiação infravermelha. Após um tempo de estabilização, são aplicados os modelos matemáticos para predição das características de qualidade do produto. A amostra é então liberada e retorna ao processo. Uma nova amostra é coletada e assim por diante. O tempo de ciclo, juntamente com o tempo computacional, não ultrapassa três minutos.

Os valores preditos pelos modelos são então enviados à central de controle da refinaria, por protocolos de comunicação, que utiliza essas informações para as sucessivas operações de *blending* do produto com flexibilidade e agilidade, maior rentabilidade, mantendo o tanque final dentro das especificações técnicas.

3.4 Resultados Alcançados

De acordo com o fluxograma representado pela Figura 1 do Capítulo 1, uma vez identificado o problema, iniciou-se a fase de planejamento e coleta de amostras.

As amostras foram planejadas a serem coletadas seis horas após troca de carga da refinaria, durante seis meses, para contornar o efeito do fenômeno da autocorrelação introduzido no Capítulo 2. Para demonstrar essa questão, foram coletadas 20 sucessivas amostras de óleo diesel (espaçadas de 3 minutos) e efetuadas medidas apenas da propriedade densidade, pois sua medição é totalmente automatizada e leva apenas alguns minutos e nesse sentido é a que menos onera mão-de-obra. A Tabela 1 apresenta o cálculo do coeficiente de autocorrelação para $k = 1$ ($0,0002832 / 0,00038 = 0,73691$). Esse valor representa forte correlação positiva.

Tabela 1. Cálculos intermediários para obtenção do coeficiente de autocorrelação (amostragens sucessivas)

i	X_i	X_{i+1}	$X_i - X_m$	$X_{i+1} - X_m$	$(X_i - X_m)(X_{i+1} - X_m)$	$(X_i - X_m)^2$
1	0,86916	0,86911	0,00412	0,00407	0,0000168	0,00002
2	0,86911	0,8693	0,00407	0,00426	0,0000174	0,00002
3	0,8693	0,87251	0,00426	0,00747	0,0000319	0,00002
4	0,87251	0,86962	0,00747	0,00458	0,0000343	0,00006
5	0,86962	0,86962	0,00458	0,00458	0,0000210	0,00002
6	0,86962	0,86833	0,00458	0,00329	0,0000151	0,00002
7	0,86833	0,86864	0,00329	0,00360	0,0000119	0,00001
8	0,86864	0,8673	0,00360	0,00226	0,0000082	0,00001
9	0,8673	0,86525	0,00226	0,00021	0,0000005	0,00001
10	0,86525	0,86657	0,00021	0,00153	0,0000003	0,00000
11	0,86657	0,85983	0,00153	-0,00521	-0,0000080	0,00000
12	0,85983	0,86017	0,00521	-0,00487	0,0000253	0,00003
13	0,86017	0,8601	0,00487	-0,00494	0,0000240	0,00002
14	0,8601	0,86042	0,00494	-0,00462	0,0000228	0,00002
15	0,86042	0,86369	0,00462	-0,00135	0,0000062	0,00002
16	0,86369	0,85839	0,00135	-0,00665	0,0000089	0,00000
17	0,85839	0,86223	0,00665	-0,00281	0,0000186	0,00004
18	0,86223	0,86208	0,00281	-0,00296	0,0000083	0,00001
19	0,86208	0,85839	0,00296	-0,00665	0,0000196	0,00001
20	0,85839	-	0,00665	-	-	0,00004
X_m	0,86504	-	-	Σ	0,0002832	0,00038

Contudo, o coeficiente de autocorrelação cai drasticamente de 0,73691 para 0,16175 ($0,0000063 / 0,000039 = 0,16175$) ao realizar amostragens após troca de carga da refinaria, conforme apresentado na Tabela 2.

Tabela 2. Cálculos intermediários para obtenção do coeficiente de autocorrelação (amostragens após troca de carga da refinaria)

i	X_i	X_{i+1}	$X_i - X_m$	$X_{i+1} - X_m$	$(X_i - X_m)(X_{i+1} - X_m)$	$(X_i - X_m)^2$
1	0,8691	0,8681	0,00006	-0,00094	-0,0000001	0,000000
2	0,8681	0,8678	0,00094	-0,00123	0,0000012	0,000001
3	0,8678	0,8674	0,00123	-0,00164	0,0000020	0,000002
4	0,8674	0,8685	0,00164	-0,00053	0,0000009	0,000003
5	0,8685	0,8685	0,00053	-0,00053	0,0000003	0,000000
6	0,8685	0,8693	0,00053	0,00027	-0,0000001	0,000000
7	0,8693	0,8691	0,00027	0,00006	0,0000000	0,000000
8	0,8691	0,869	0,00006	-0,00003	0,0000000	0,000000
9	0,869	0,8698	0,00003	0,00076	0,0000000	0,000000
10	0,8698	0,8702	0,00076	0,00116	0,0000009	0,000001
11	0,8702	0,871	0,00116	0,00196	0,0000023	0,000001
12	0,871	0,8708	0,00196	0,00177	0,0000035	0,000004
13	0,8708	0,8679	0,00177	-0,00113	-0,0000020	0,000003
14	0,8679	0,8681	0,00113	-0,00094	0,0000011	0,000001
15	0,8681	0,8656	0,00094	-0,00344	0,0000032	0,000001
16	0,8656	0,8717	0,00344	0,00266	-0,0000092	0,000012
17	0,8717	0,8705	0,00266	0,00147	0,0000039	0,000007
18	0,8705	0,8684	0,00147	-0,00064	-0,0000009	0,000002
19	0,8684	0,8699	0,00064	0,00086	-0,0000005	0,000000
20	0,8699	-	0,00086	-	-	0,000001
X_m	0,86903	-	-	Σ	0,0000063	0,000039

O principal problema numa situação de autocorrelação elevada é a dificuldade ou impossibilidade de desenvolvimento de modelos de regressão por PLS, daí a importância de se coletar amostras após troca da mistura de petróleos - situações nas quais se obtém a maior variabilidade possível entre os dados.

As Figuras 21 a 28 apresentam algumas medidas estatísticas para caracterizar as variáveis de resposta: temperatura de destilação correspondente aos pontos percentuais 10%, 50%, 85% e 90% de material condensado (recuperado), viscosidade cinemática, densidade, ponto de fulgor e índice de cetano.

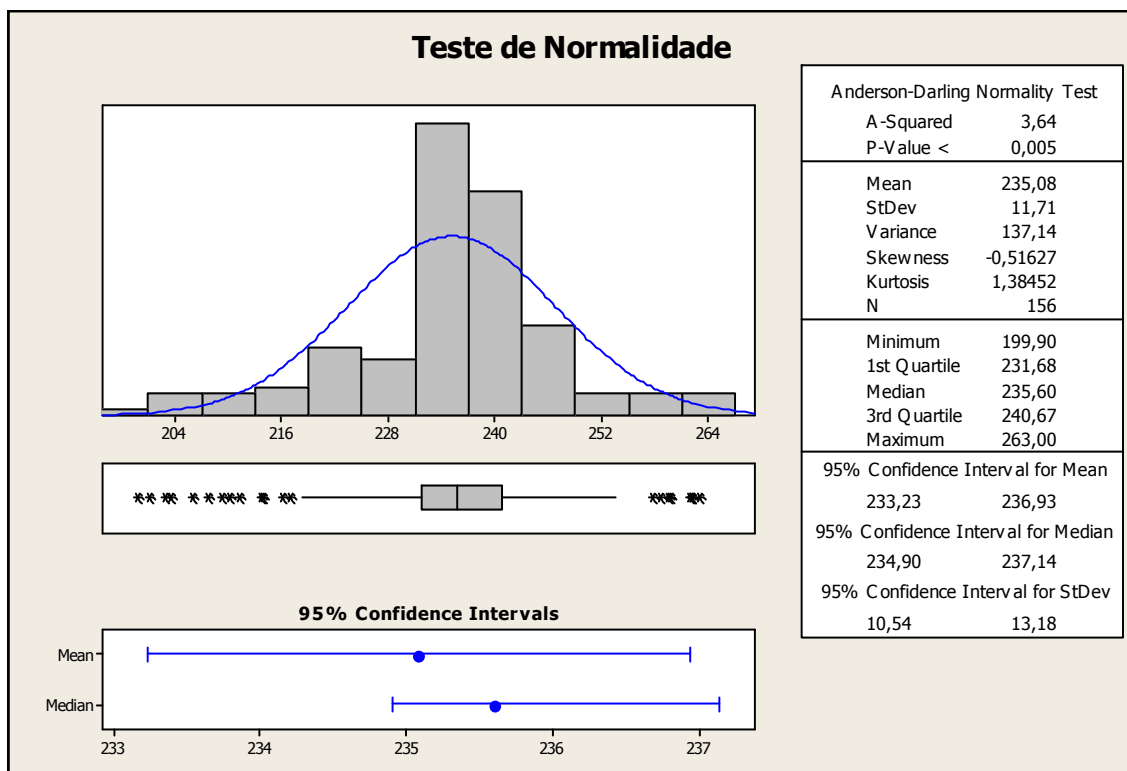


Figura 21. Histograma da propriedade temperatura de destilação 10% recuperados

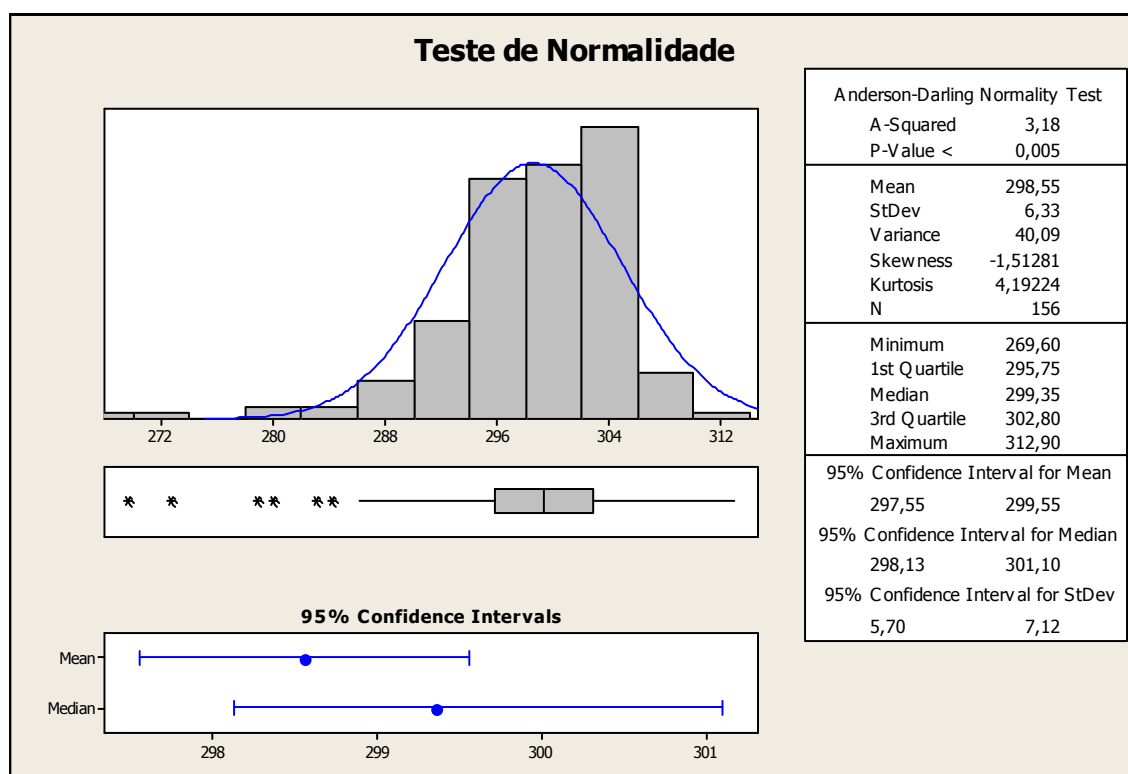


Figura 22. Histograma da propriedade temperatura de destilação 50% recuperados

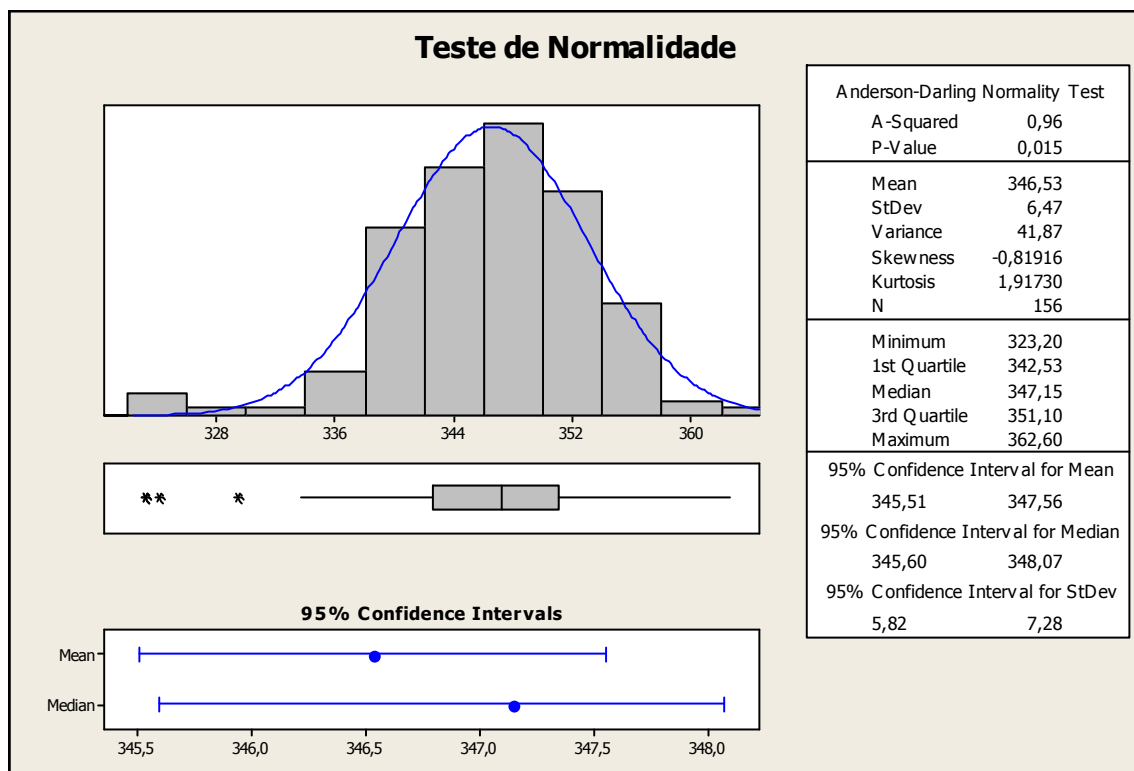


Figura 23. Histograma da propriedade temperatura de destilação 85% recuperados

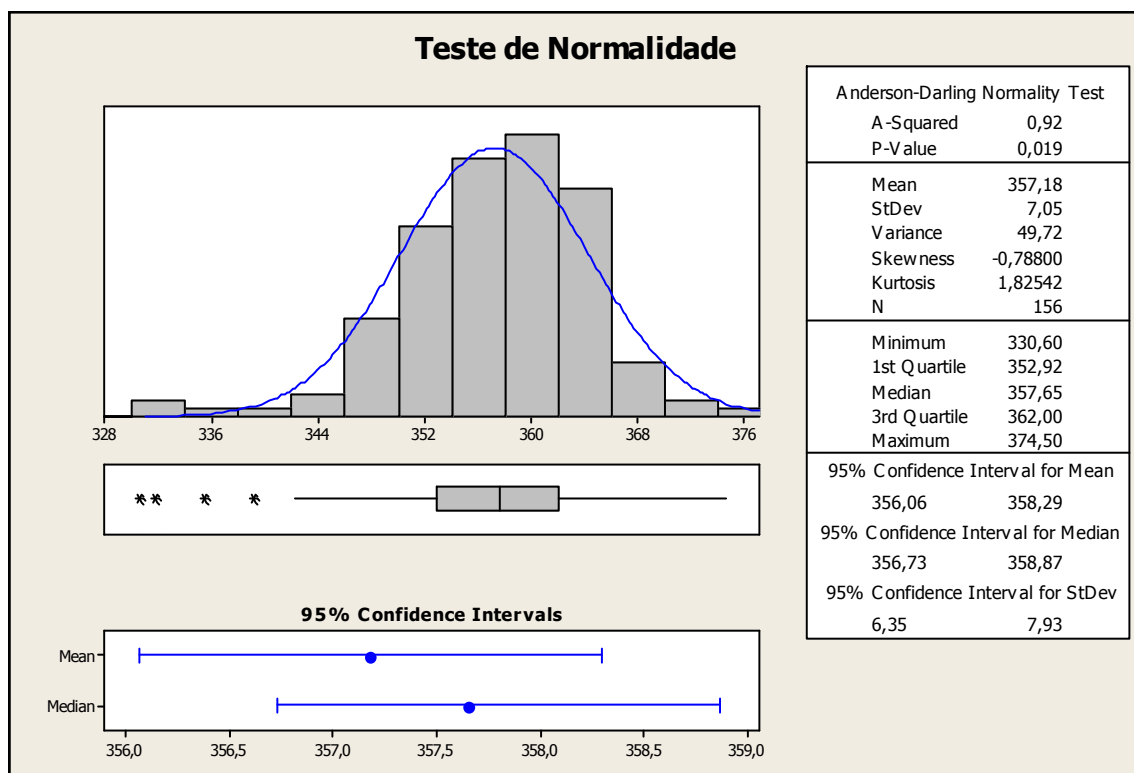


Figura 24. Histograma da propriedade temperatura de destilação 90% recuperados

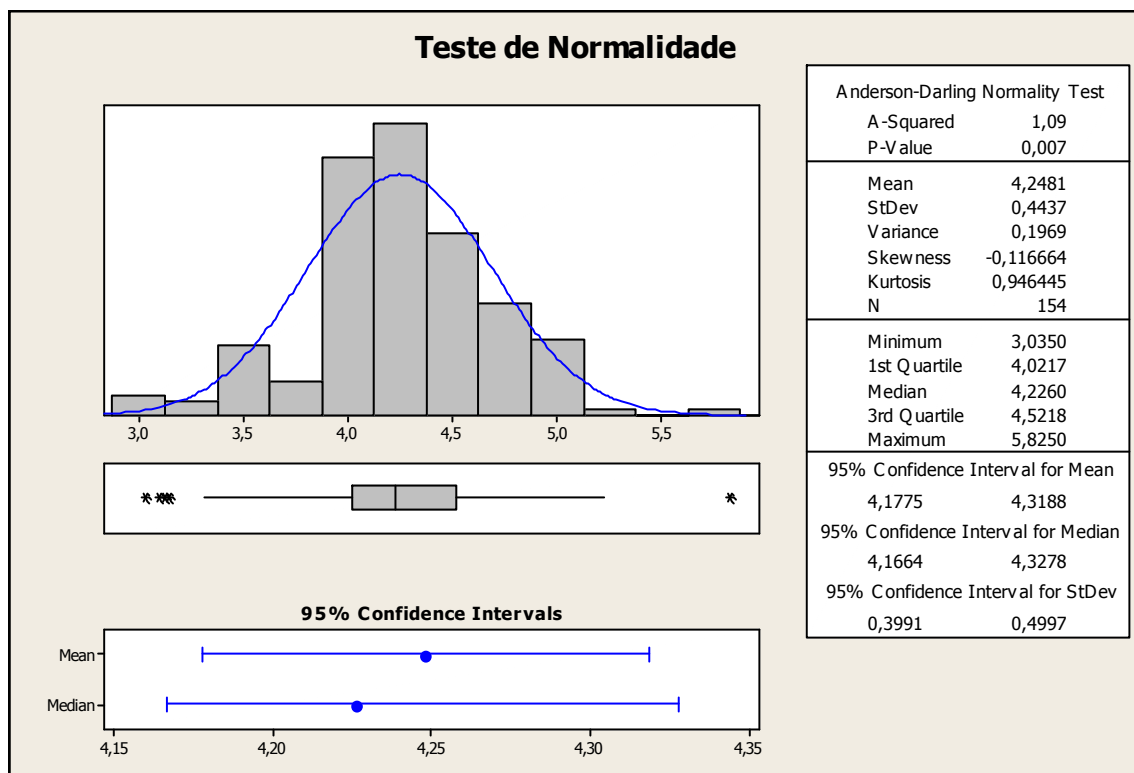


Figura 25. Histograma da propriedade viscosidade cinemática

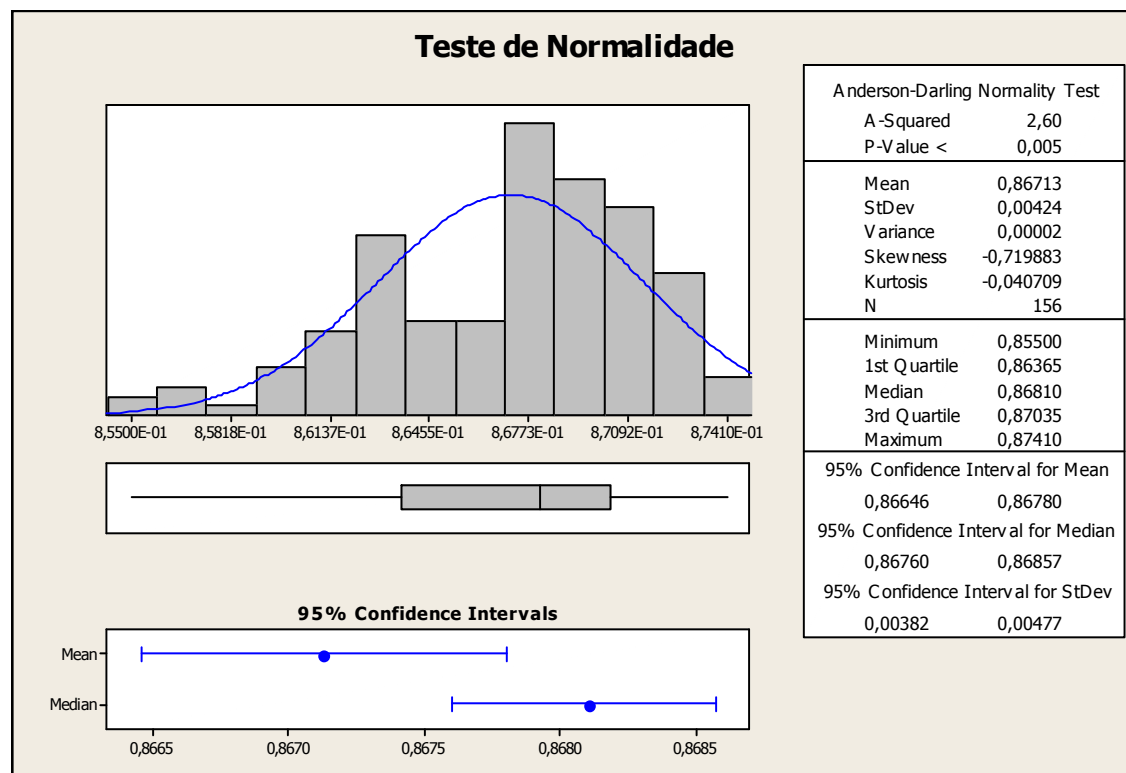


Figura 26. Histograma da propriedade densidade

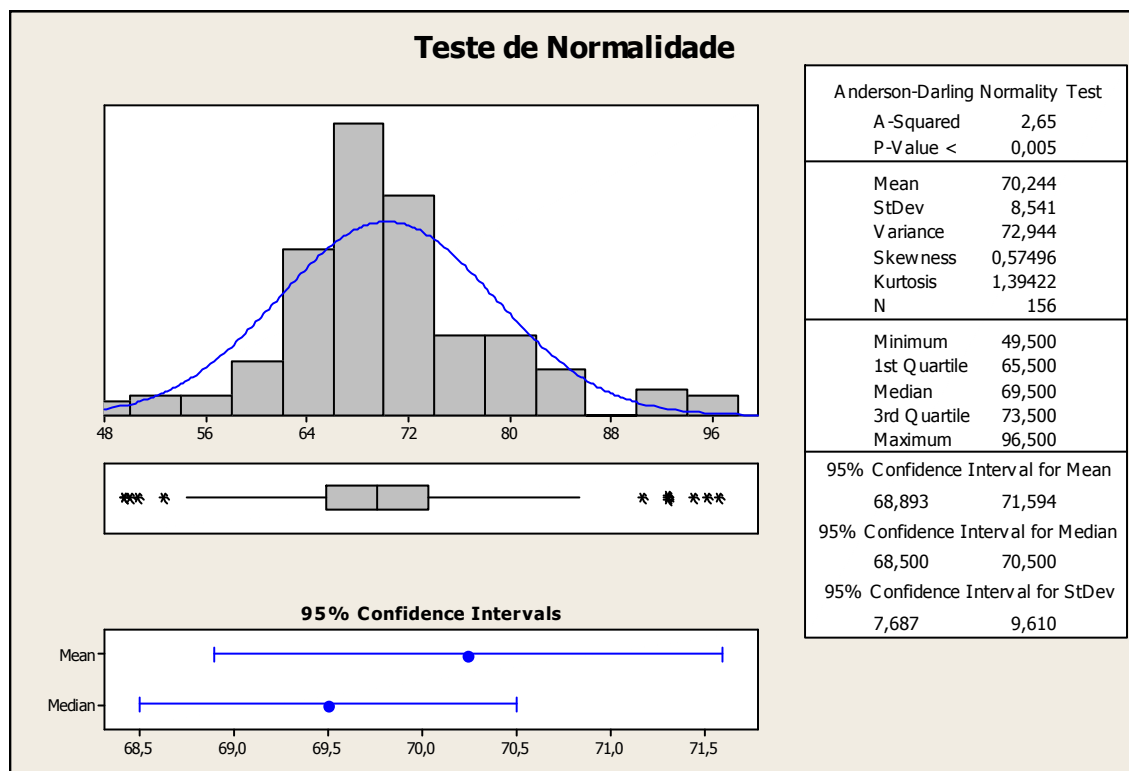


Figura 27. Histograma da propriedade ponto de fulgor

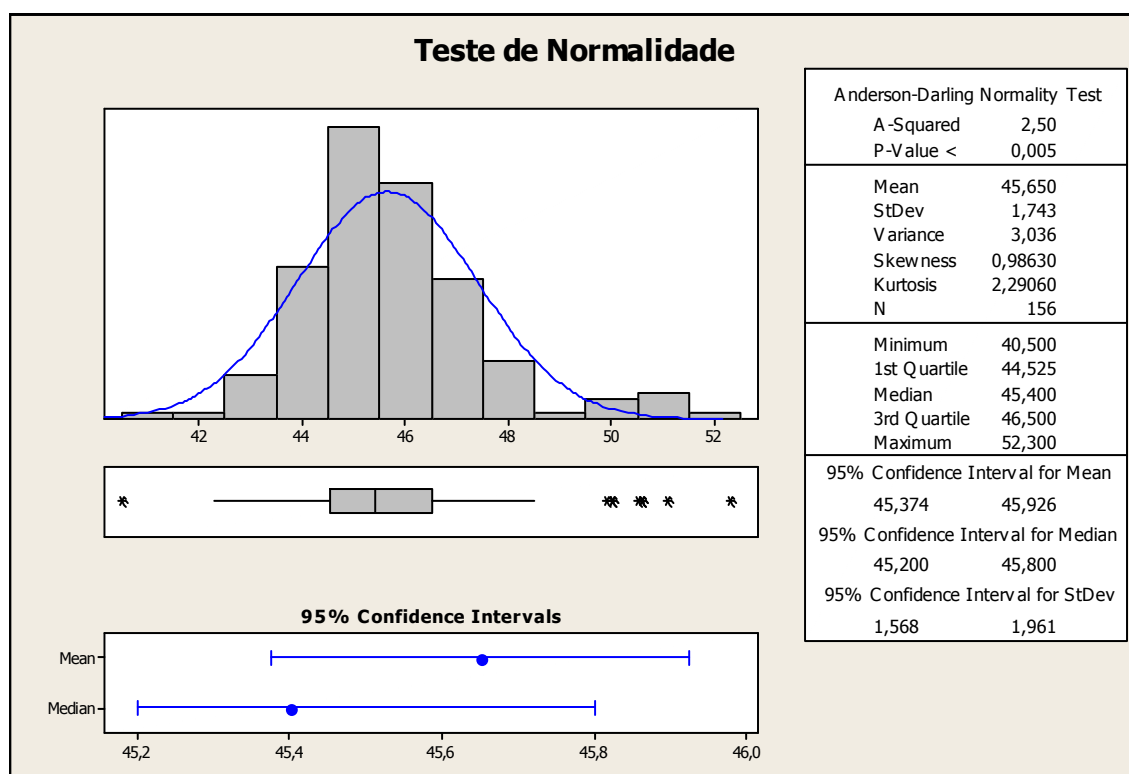


Figura 28. Histograma da propriedade índice de cetano

Com intuito de conhecer melhor o comportamento estatístico das variáveis X , foi realizada uma PCA no conjunto total de 156 amostras e um gráfico de *scores* dos dois primeiros componentes principais, PC1 e PC2, conforme Figura 29. O eixo principal do elipsóide está na direção correspondente à variável X_i de maior variabilidade, ou autovetor relacionado com o maior autovalor da matriz de covariâncias $\Sigma_{p \times p}$. O segundo eixo está relacionado com a variável de segunda maior variância.

Os dados encontram-se bem espaçados ao longo do elipsóide, indicando boa variabilidade na coleta dos dados. A elipsóide é conhecida como distribuição T^2 de Hotelling e matematicamente representa a mesma interpretação da distância de Mahalanobis introduzida na distribuição normal multivariada no Capítulo 2.

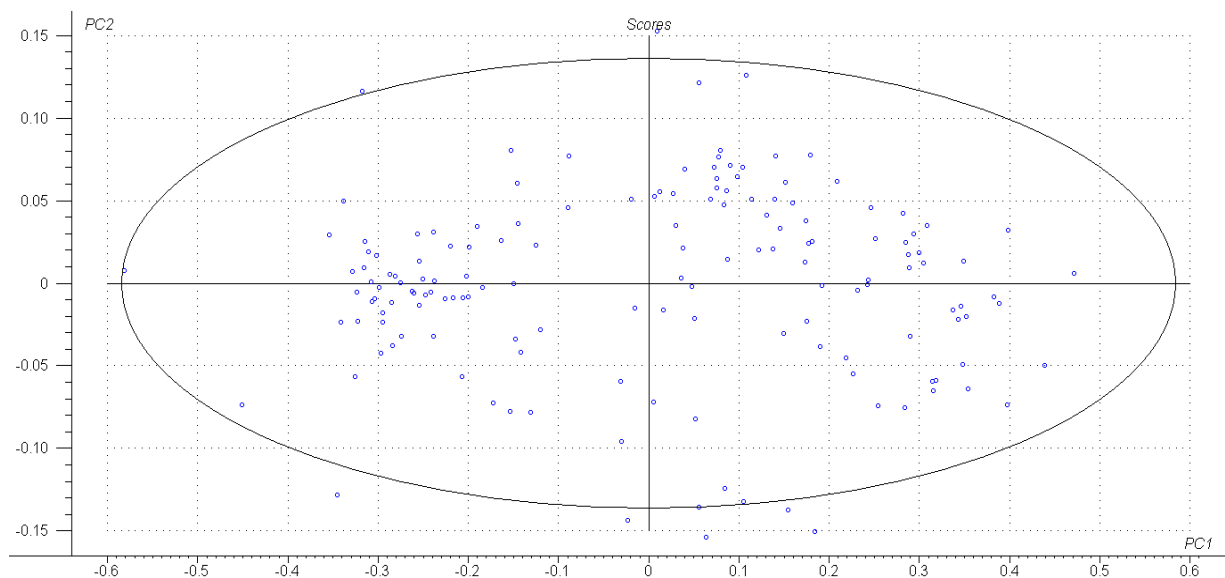


Figura 29. Scores dos dois primeiros componentes principais da PCA

Como os *loadings* representam o quanto uma variável contribui na formação do PC, os *x-loadings* apresentam um perfil gráfico que muito se assemelha ao próprio gráfico bidimensional dos espectros, pois quanto maior a absorção da radiação num dado comprimento de onda, maior a probabilidade dessas variáveis serem influentes para formação do PC, conforme ilustrados nas Figuras 31 e 31.

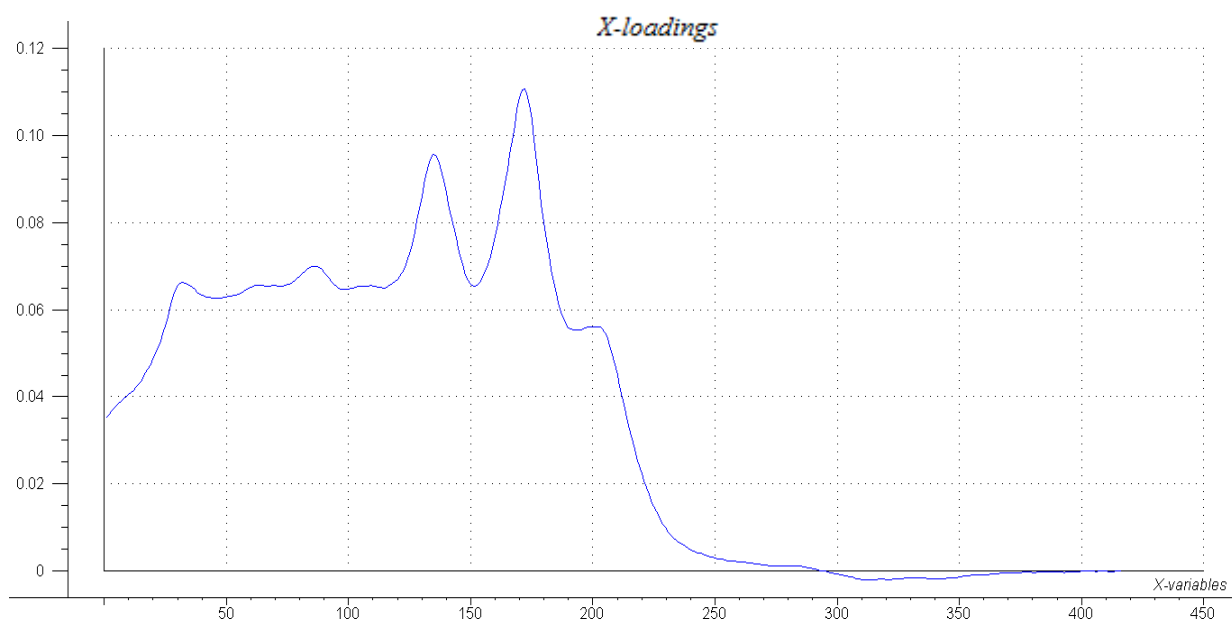


Figura 30. x-loadings da PCA

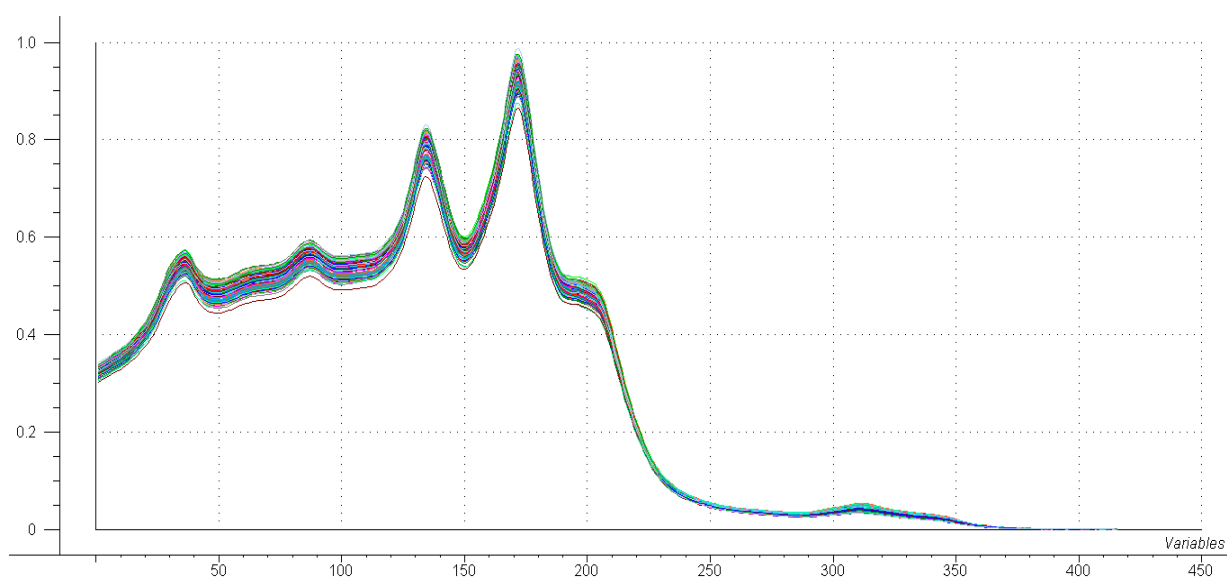


Figura 31. Espectros das amostras de óleo diesel

As absorções de 800 variáveis X do espectro foram comprimidas em apenas dois componentes principais para este estudo estatístico do comportamento das variáveis, com variância explicada acumulada de 99,11%, conforme representada na Figura 31.

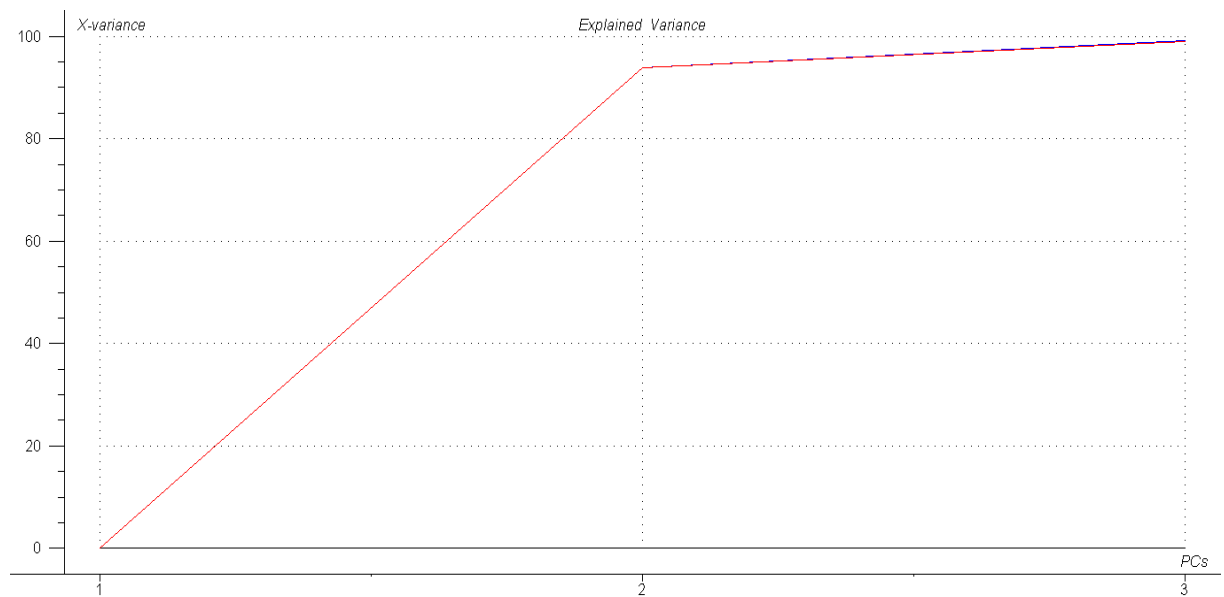


Figura 32. Variância explicada acumulada em função do número de componentes principais para a PCA

Seguindo o fluxograma 1 do Capítulo 1, uma vez realizado o planejamento, a coleta e a caracterização de amostras em laboratório, o próximo passo é a modelagem propriamente dita. Para cada propriedade, foi desenvolvido um modelo matemático PLS diferente e independente dos demais, a partir das variáveis X. Foram traçados gráficos exibindo os valores preditos versus valores reais juntamente aos coeficientes de correlação (R^2).

Para auxiliar a determinação do número de variáveis latentes entre os extremos *underfitting* e *overfitting*, foi utilizado o algoritmo *leave-one-out-cross-validation*. A validação cruzada, também funciona como um importante indicador da performance dos modelos, pois o RMSECV (*Root Mean Square Error of Cross Validation*) corresponde ao desvio padrão dos resíduos (diferença entre os valores preditos pelo modelo e os valores de referência) durante a fase de modelagem. O número de variáveis latentes foi determinado, portanto, quando o RMSECV é mínimo.

Durante o desenvolvimento dos modelos, os *outliers* com altos resíduos foram eliminados pela equação 24.

Uma vez desenvolvido o modelo de regressão, este precisou ser testado quanto ao seu desempenho. Para isso, utilizou-se um conjunto independente de amostras, chamado conjunto de validação externa. Neste trabalho, segundo norma ASTM-D6122, 20 amostras foram selecionadas aleatoriamente para compor este conjunto de validação. Os valores preditos foram comparados aos resultados verdadeiros da propriedade, conforme os métodos de

referência. Para o cálculo do erro de previsão, foi utilizado o parâmetro RMSEP (*Root Mean Square Error of Prediction*) que corresponde ao desvio-padrão dos resíduos para validação externa. O RMSECV foi comparado estatisticamente com RMSEP e não há evidências de diferenças significativas com 95% de confiança.

A Tabela 3 apresenta um resumo dos resultados de modelagem e validação para cada propriedade modelada.

Tabela 3. Resumo dos resultados de modelagem e validação

Propriedade	Número de Amostras	LV	R ²	RMSECV	RMSEP
Temperatura de destilação (10% rec.)	153	7	0,96630	3,0°C	3,9°C
Temperatura de destilação (50% rec.)	150	9	0,95324	1,8°C	2,1°C
Temperatura de destilação (85% rec.)	147	10	0,91527	2,4°C	2,7°C
Temperatura de destilação (90% rec.)	150	10	0,93919	2,2°C	2,9°C
Índice de Cetano	149	6	0,94625	0,5	0,7
Viscosidade Cinemática a 40°C	141	9	0,93082	0,15 cSt	0,16 cSt
Ponto de Fulgor	123	5	0,90057	3,3°C	3,4°C
Densidade	145	8	0,98630	0,00067	0,0012

A seguir serão apresentados detalhes gráficos sobre os resultados alcançados.

3.4.1 Resultados para propriedade temperatura de destilação (10% recuperados)

Para descrever a propriedade temperatura de destilação correspondente a 10% de material recuperado são necessárias sete variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 4 e Figura 33, acima de sete LV, não há ganhos percentuais significativos em variância explicada, complementar da variância residual. Ao contrário, a partir de 12 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui. Esse fenômeno ocorre porque essas últimas LV não estariam modelando sinal e sim ruído.

Tabela 4. Variância explicada acumulada (%) versus número de variáveis latentes - destilação (10% rec.)

Variável Latente	Variância Explicada Acumulada (%)
1	4,99
2	28,24
3	69,65
4	86,78
5	87,60
6	88,79
7	89,98
8	90,65
9	90,74
10	90,81
11	90,58
12	89,61

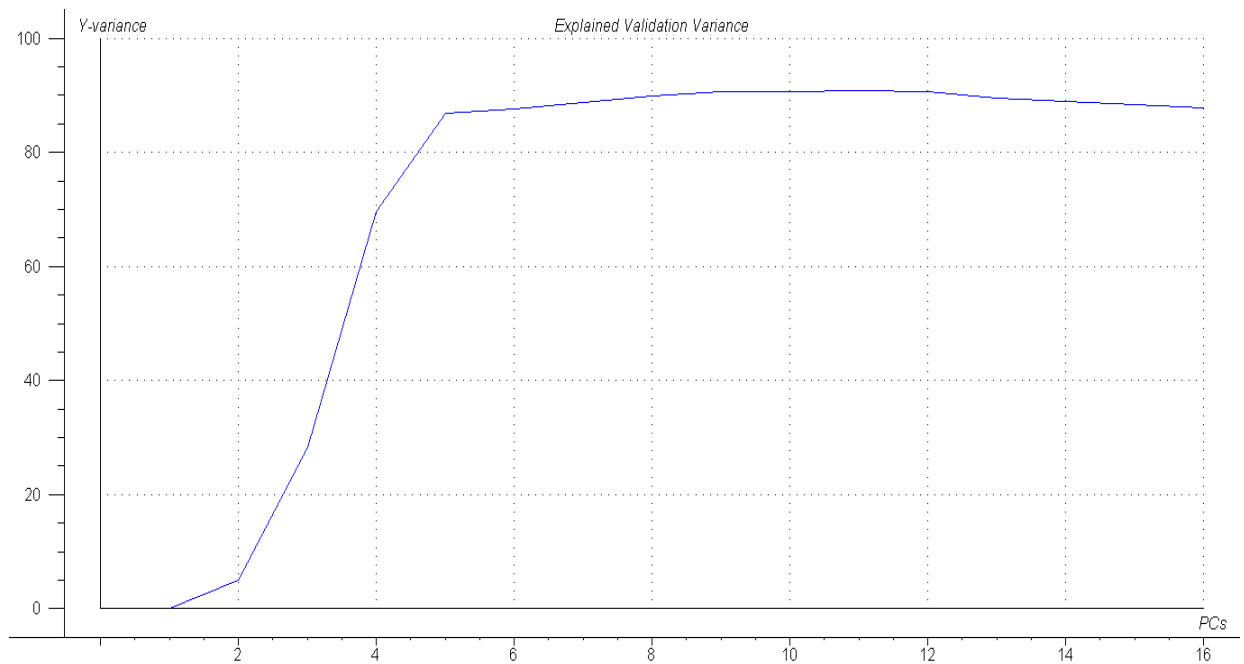


Figura 33. Variância explicada versus número de variáveis latentes para temperatura de destilação (10% rec.)

Caso não fosse utilizado o algoritmo de validação cruzada, sempre a variância explicada acumulada totalizaria 100%, conforme ilustrado na Figura 34:

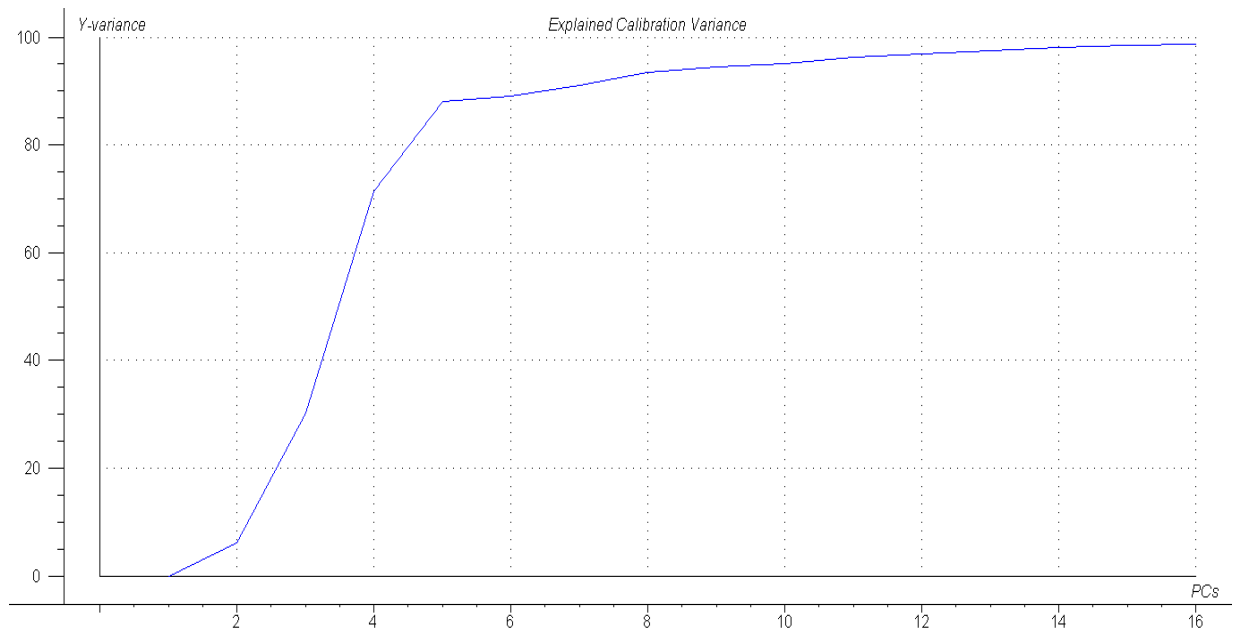


Figura 34. Variância explicada versus número de variáveis latentes sem utilizar validação cruzada

Assim, a determinação do número de variáveis latentes seria 16, isto é, sempre maior quando comparada à determinação que utiliza o algoritmo de validação cruzada. Isto ocorre, porque o CV simula situações reais nas quais os modelos serão submetidos, isto é simula previsões de novas amostras. Caso não seja utilizado, no mundo real, as previsões seriam de baixa robustez devido ao fenômeno *overfitting*.

A Figura 35 destaca um *outlier*, com elevado resíduo e alavancagem e que, portanto, foi removido do conjunto de dados.

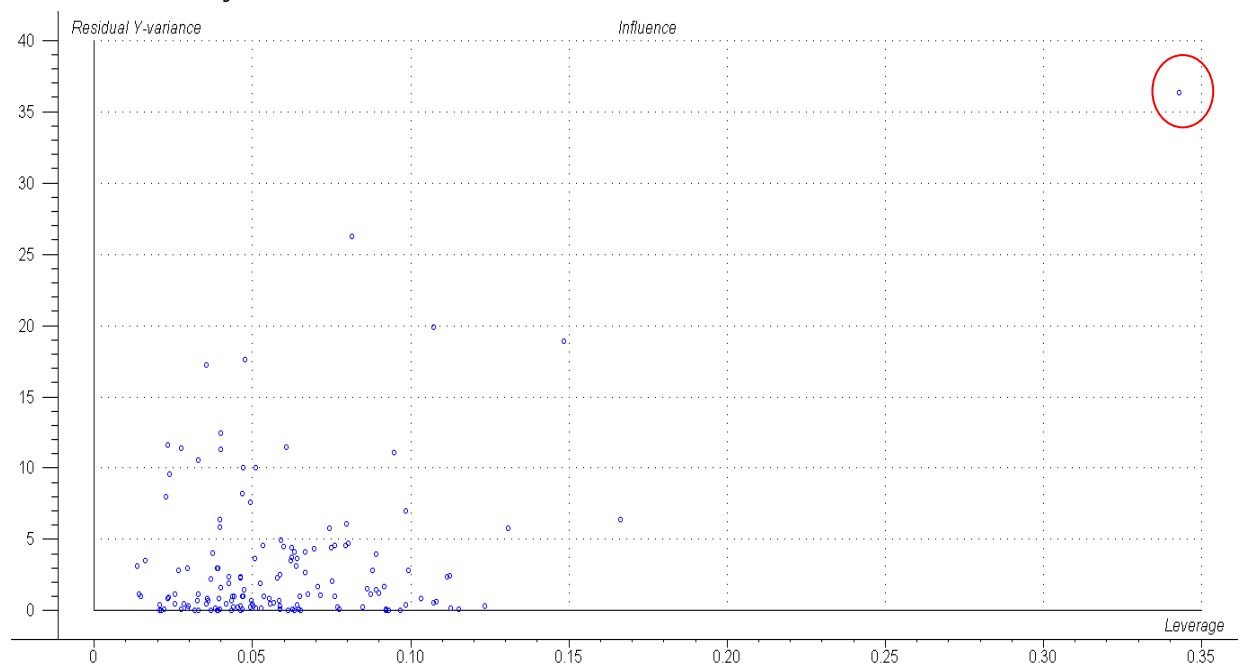


Figura 35. Resíduos de Student versus *Leverage*

A partir das próximas propriedades estas duas últimas análises gráficas serão suprimidas.

A Figura 36 mostra o gráfico de valores preditos versus valores reais da propriedade temperatura de destilação correspondente a 10% de material recuperado para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,96663 e RMSECV de 3,0°C.

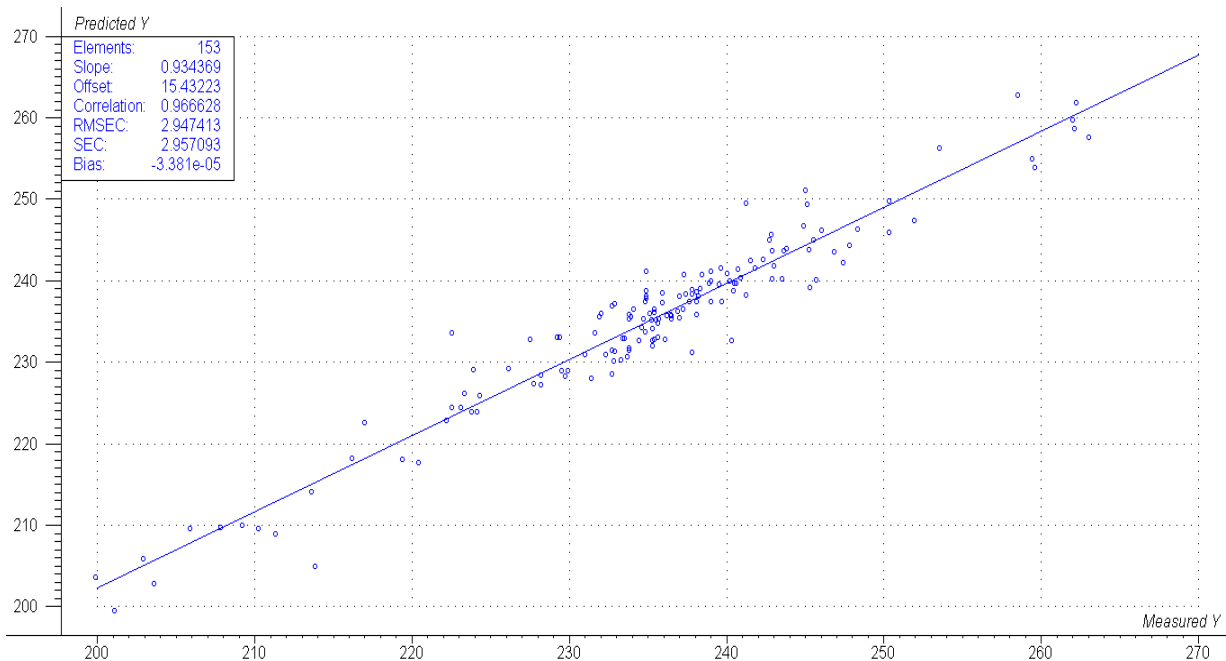


Figura 36. Valores preditos versus valores medidos para temperatura de destilação 10% rec.

A Tabela 5 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi de 3,9°C.

Tabela 5. Conjunto de validação externa – temperatura de destilação (10% rec.)

Amostra	Real	Predito	Resíduo
1	246,5	251,7	5,2
2	249,5	248,5	-1,0
3	245,0	253,1	8,1
4	250,0	258,1	8,1
5	209,5	208,9	-0,6
6	257,0	257,6	0,6
7	230,5	231,9	1,4
8	237,5	240,8	3,3
9	251,5	246,5	-5,0
10	259,0	260,8	1,8
11	256,0	254,0	-2,0
12	258,0	257,2	-0,8

Tabela 5. Conjunto de validação externa – temperatura de destilação (10% rec.)

13	252,5	251,1	-1,4
14	239,5	238,2	-1,3
15	252,0	249,6	-2,4
16	247,8	244,7	-3,1
17	249,2	243,7	-5,5
18	252,5	250,9	-1,6
19	235,6	242,1	6,5
20	245,1	246,2	1,1

As Figuras 37 e 38 ilustram que os resíduos deste modelo seguem distribuição normal com *p-value* de 0,139.

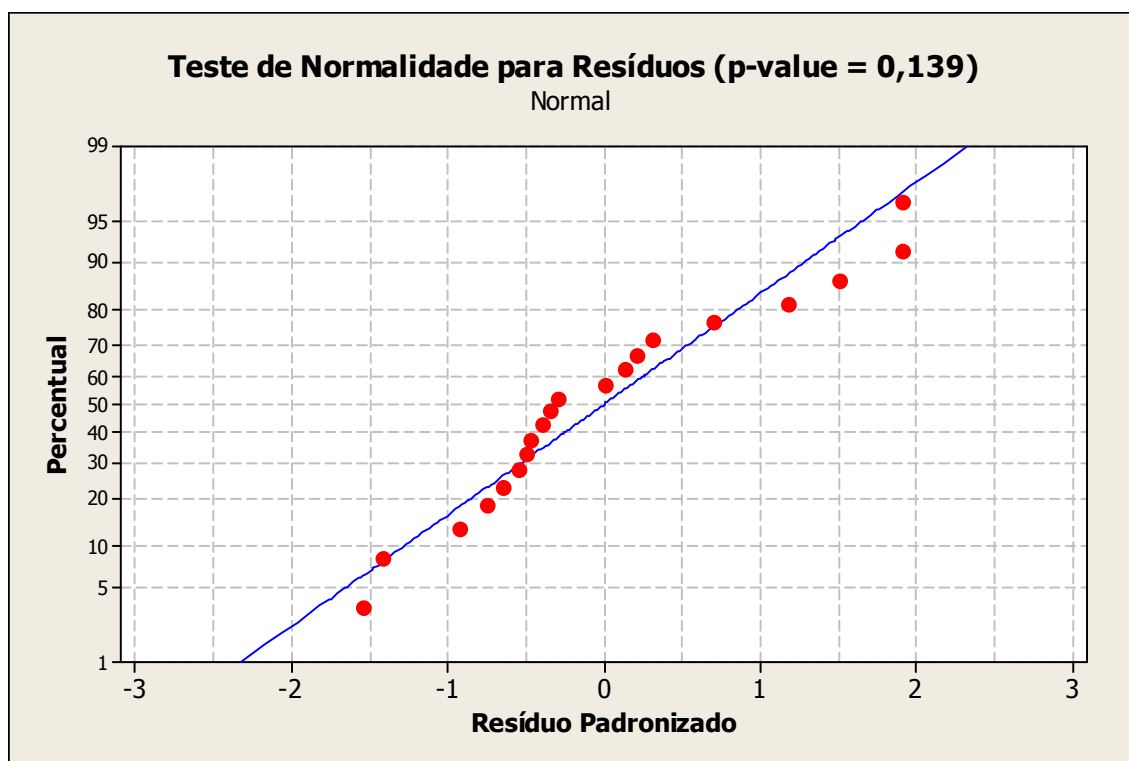


Figura 37. Teste de normalidade para propriedade temperatura de destilação 10% rec.

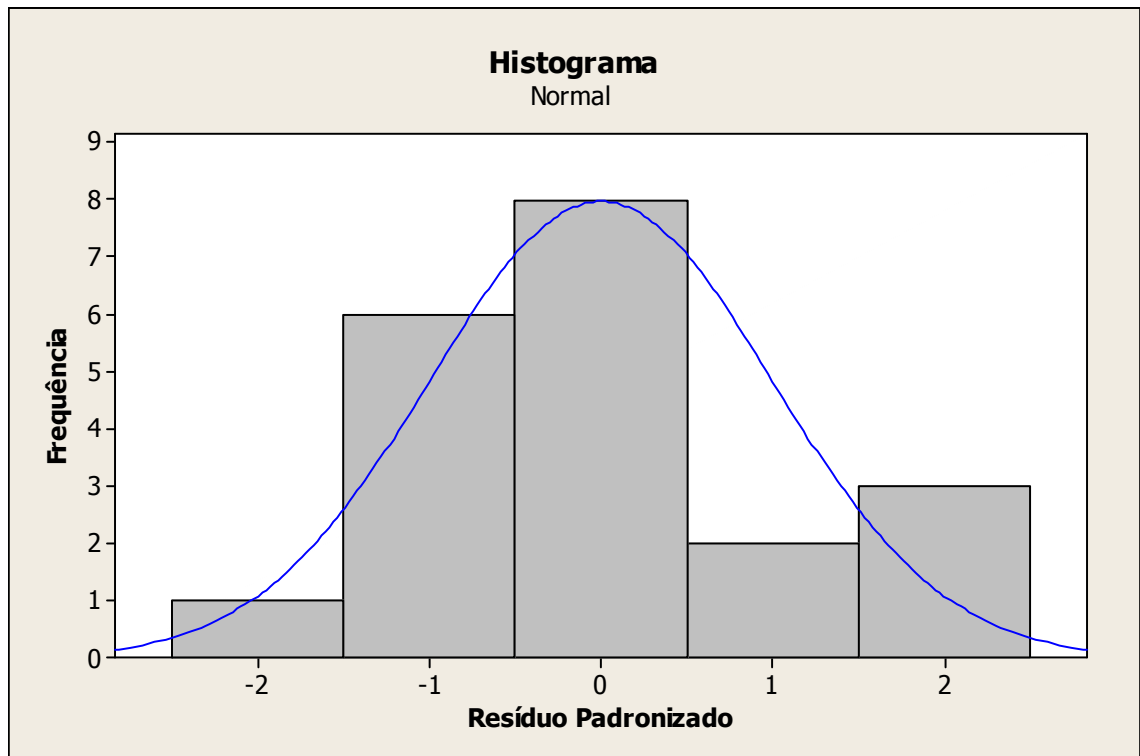


Figura 38. Histograma do resíduo padronizado para propriedade temperatura de destilação 10% rec.

3.4.2 Resultados para propriedade temperatura de destilação (50% recuperados)

Para descrever a propriedade temperatura de destilação correspondente a 50% de material recuperado são necessárias nove variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 6 e Figura 39, acima de nove LV, não há ganhos percentuais significativos em variância explicada. Ao contrário, a partir de 13 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui.

Tabela 6. Variância explicada acumulada (%) versus número de variáveis latentes – temperatura de destilação (50% rec.)

Variável Latente	Variância Explicada Acumulada (%)
1	2,77
2	17,81
3	50,25
4	70,47
5	73,07

Tabela 6. Variância explicada acumulada (%) versus número de variáveis latentes – temperatura de destilação (50% rec.)

6	79,57
7	82,43
8	83,78
9	84,44
10	83,53
11	83,47
12	83,90

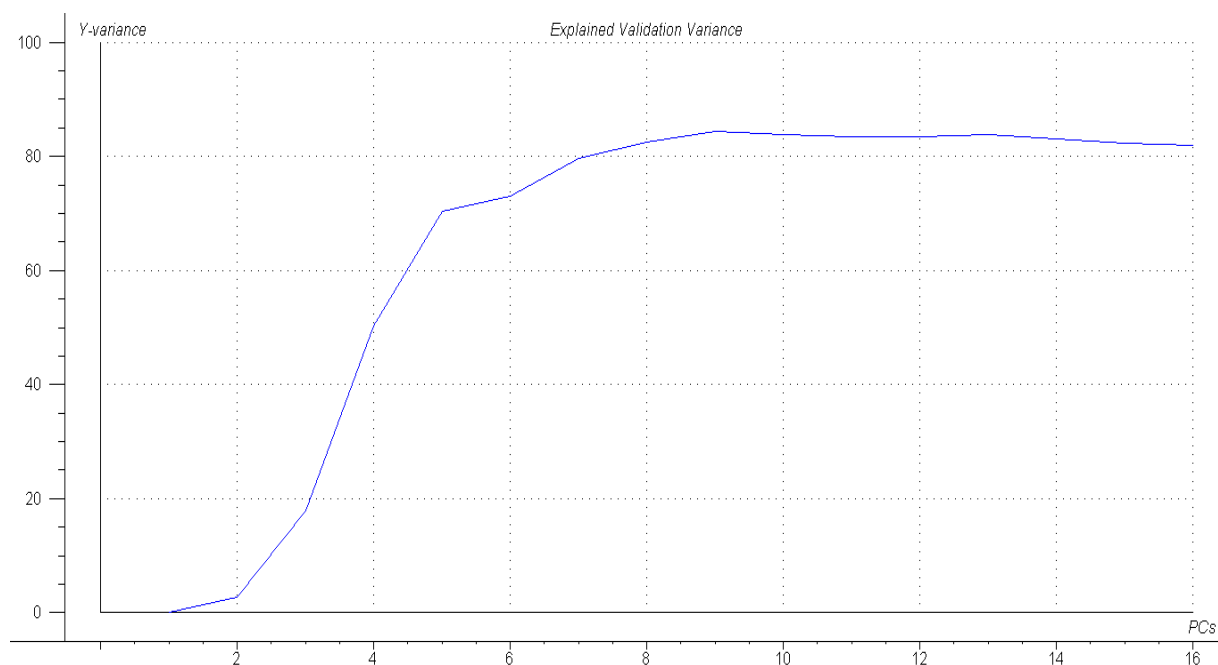


Figura 39. Variância explicada versus número de variáveis latentes para temperatura de destilação (50% rec.)

A Figura 40 mostra o gráfico de valores preditos versus valores reais da propriedade destilação 50% recuperados para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,95324 e RMSECV de 1,8°C.

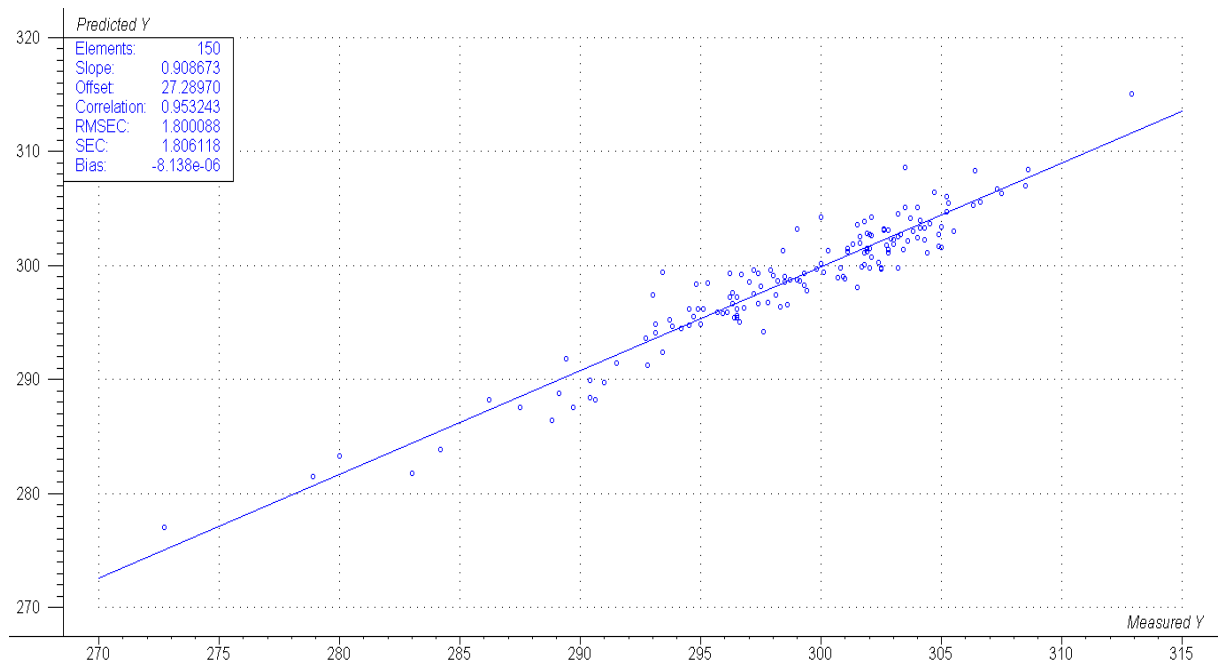


Figura 40. Valores preditos versus valores medidos para temperatura de destilação (50% rec.)

A Tabela 7 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 2,1°C.

Tabela 7. Conjunto de validação externa - destilação (50% rec.)

Amostra	Real	Predito	Resíduo
1	298,0	301,1	3,1
2	300,5	299,9	-0,6
3	296,5	297,6	1,1
4	303,5	305,8	2,3
5	301,5	305,5	4,0
6	287,0	289,7	2,7
7	293,0	294,6	1,6
8	305,0	303,7	-1,3
9	296,0	295,8	-0,2
10	299,5	300,0	0,5
11	299,5	299,3	-0,2
12	303,0	304,4	1,4
13	302,0	301,6	-0,4
14	293,0	293,6	0,6
15	307,5	304,0	-3,5
16	292,2	291,0	-1,2

Tabela 7. Conjunto de validação externa - destilação (50% rec.)

17	302,1	304,2	2,1
18	302,7	301,9	-0,8
19	286,0	288,8	2,8
20	294,9	300,0	5,1

As Figuras 41 e 42 ilustram que os resíduos deste modelo seguem distribuição normal com p -value de 0,929.

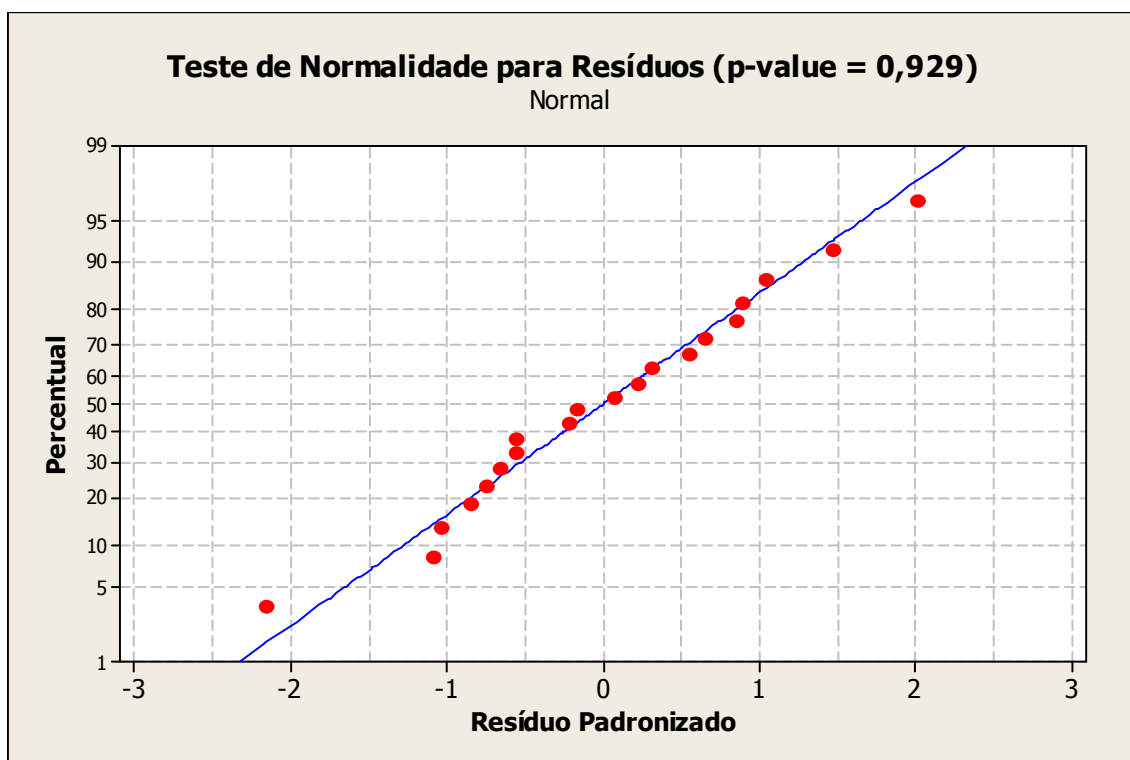


Figura 41. Teste de normalidade para propriedade temperatura de destilação 50% rec.

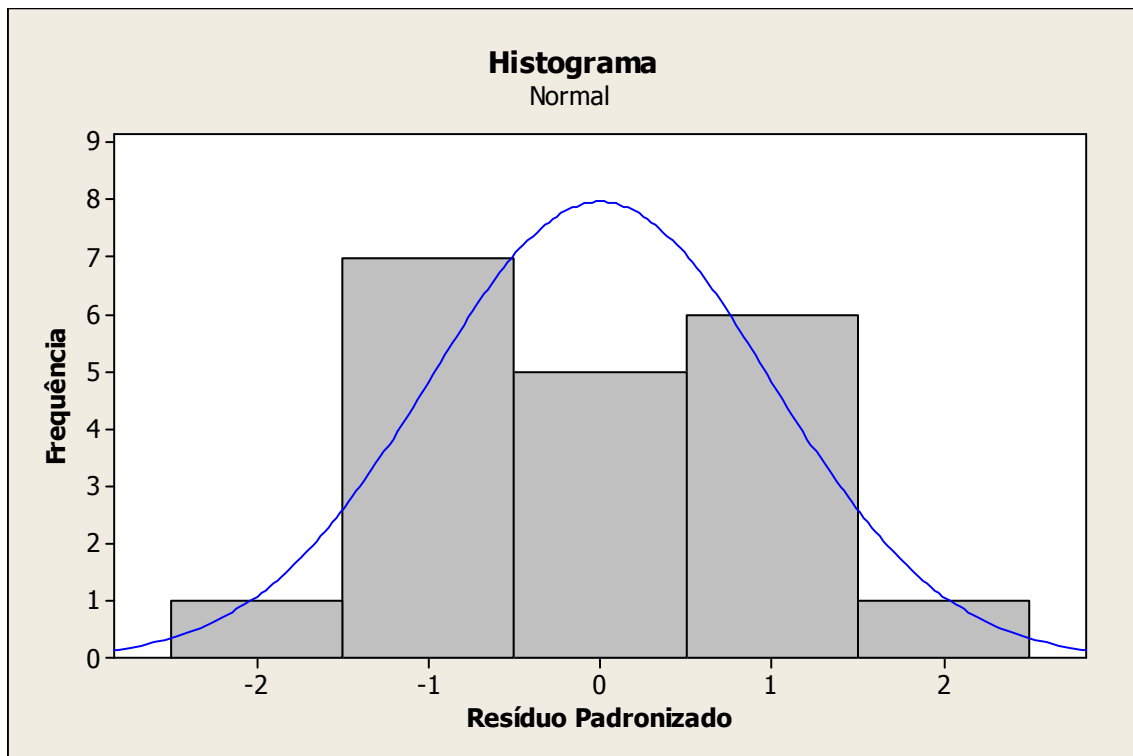


Figura 42. Histograma do resíduo padronizado para propriedade temperatura de destilação 50% rec.

3.4.3 Resultados para propriedade temperatura de destilação (85% recuperados)

Para descrever a propriedade temperatura de destilação correspondente a 85% de material recuperado são necessárias 10z variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 8 e Figura 43, acima de 10 LV, não há ganhos percentuais significativos em variância explicada.

Tabela 8. Variância explicada acumulada (%) versus número de variáveis latentes – temperatura de destilação (85% rec.)

Variável Latente	Variância Explicada Acumulada (%)
1	5,67
2	16,57
3	47,02
4	49,39
5	56,06
6	68,45
7	71,23
8	74,01

Tabela 8. Variância explicada acumulada (%) versus número de variáveis latentes – temperatura de destilação (85% rec.)

9	73,57
10	74,87
11	74,66
12	73,99
13	74,25
14	73,96
15	74,32

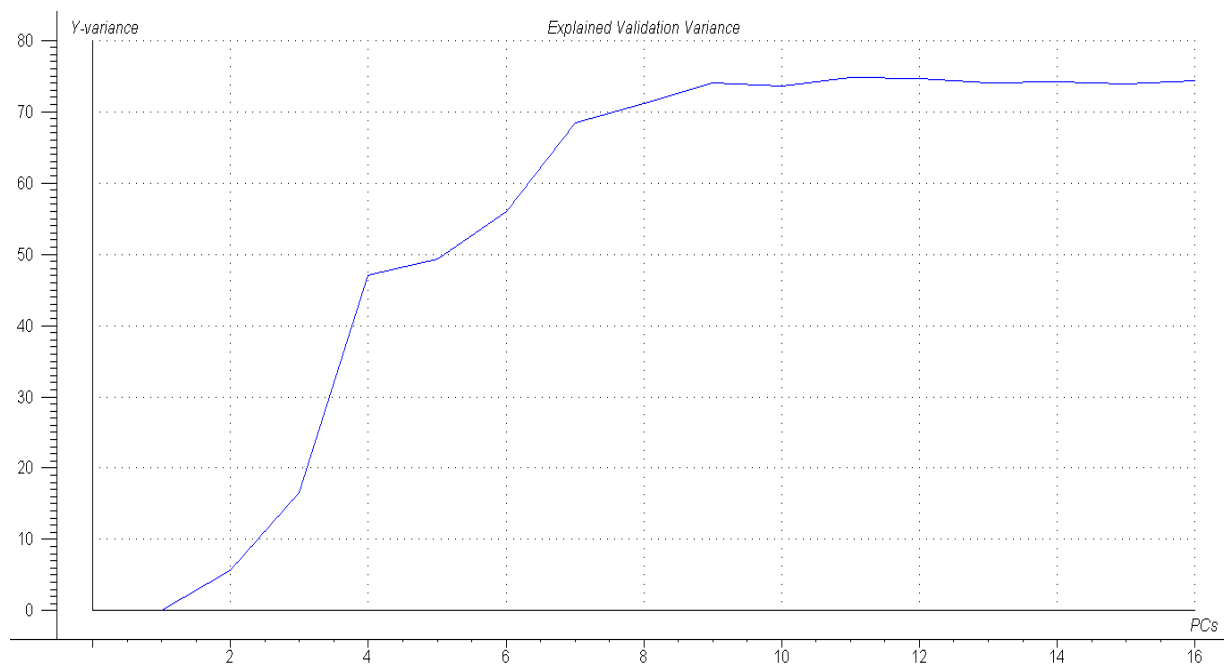


Figura 43. Variância explicada versus número de variáveis latentes para temperatura de destilação (85% rec.)

A Figura 44 mostra o gráfico de valores preditos versus valores reais da propriedade destilação 85% recuperados para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,91527 e RMSECV de 2,4°C.

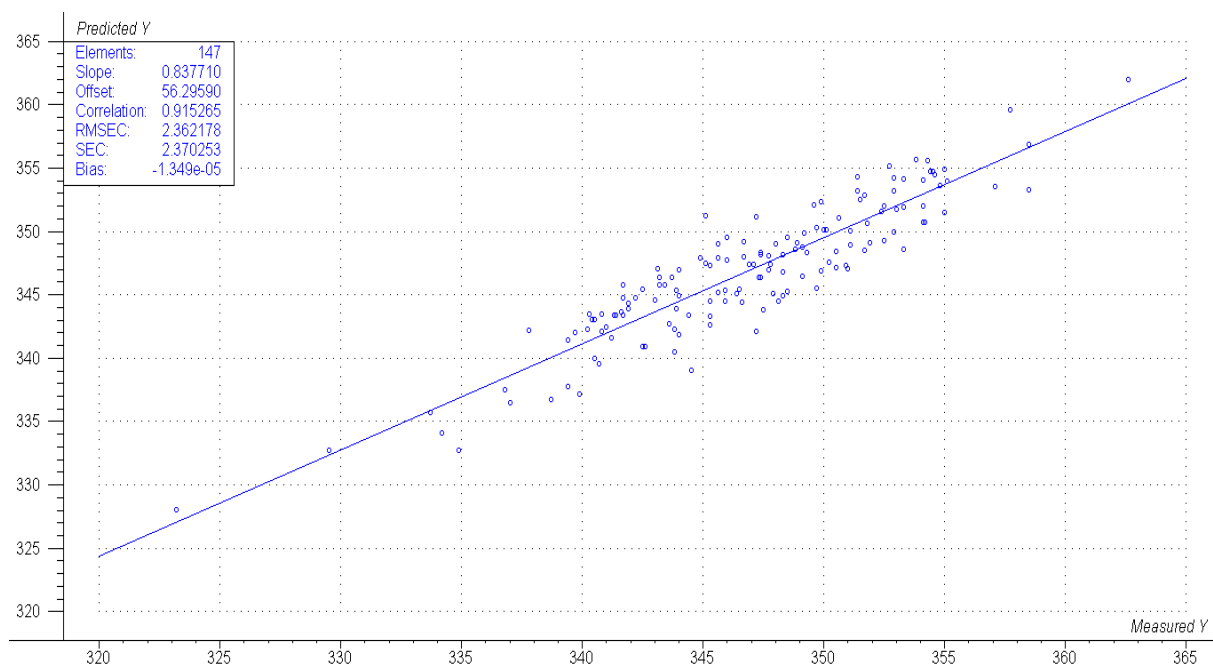


Figura 44. Valores preditos versus valores medidos para temperatura de destilação (85% rec.)

A Tabela 9 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 2,7°C.

Tabela 9. Conjunto de validação externa – temperatura de destilação (85% rec.)

Amostra	Real	Predito	Resíduo
1	354,0	357,8	3,8
2	356,0	358,2	2,2
3	365,0	361,3	-3,7
4	359,5	360,7	1,2
5	360,5	362,8	2,3
6	357,0	353,5	-3,5
7	359,5	358,1	-1,4
8	365,0	362,3	-2,7
9	358,0	355,0	-3,0
10	358,5	354,9	-3,6
11	360,0	359,0	-1,0
12	357,5	358,5	1,0
13	360,0	358,3	-1,7
14	357,5	355,5	-2,0
15	365,5	361,5	-4,0
16	336,9	342,4	5,5
17	336,9	334,0	-2,9

Tabela 9. Conjunto de validação externa – temperatura de destilação (85% rec.)

18	346,6	347,7	1,1
19	340,2	338,1	-2,1
20	358,9	359,0	0,1

As Figuras 45 e 46 ilustram que os resíduos deste modelo seguem distribuição normal com p -value de 0,124.

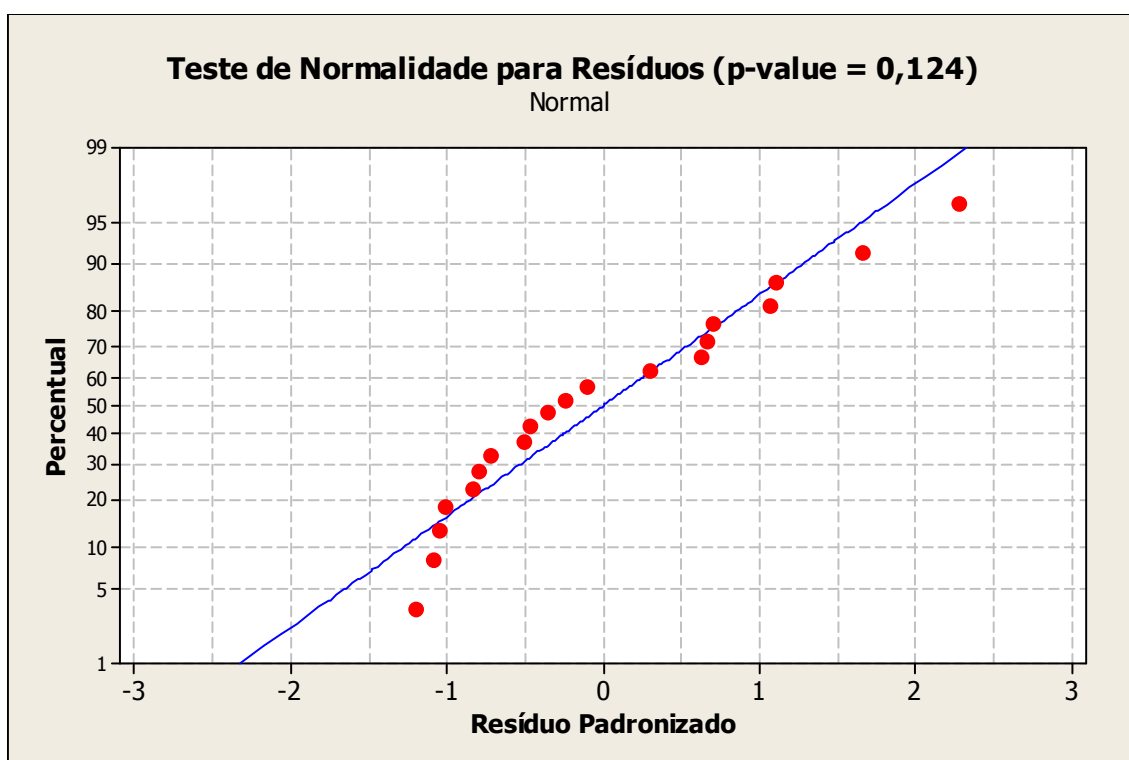


Figura 45. Teste de normalidade para propriedade temperatura de destilação 85% rec.

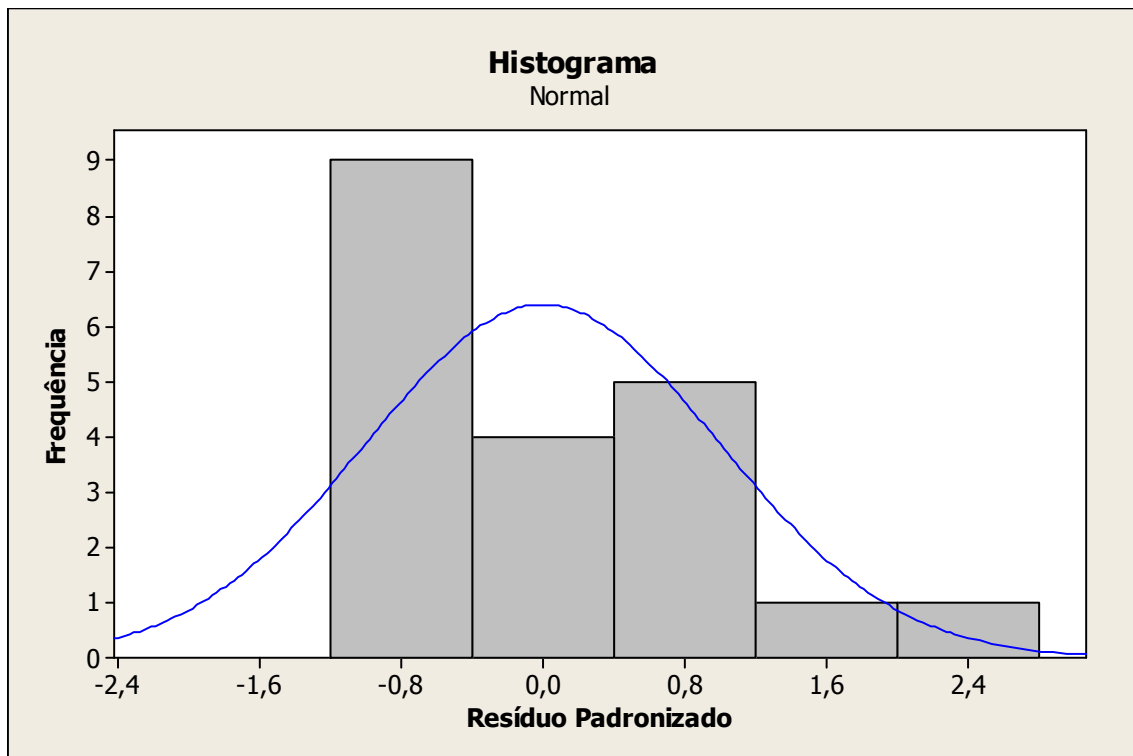


Figura 46. Histograma do resíduo padronizado para propriedade temperatura de destilação 85% rec.

3.4.4 Resultados para propriedade temperatura de destilação (90% recuperados)

Para descrever a propriedade temperatura de destilação correspondente a 90% de material recuperado são necessárias 10 variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 10 e Figura 47, acima de 10 LV, não há ganhos percentuais significativos em variância explicada. Ao contrário, a partir de 14 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui.

Tabela 10. Variância explicada acumulada (%) versus número de variáveis latentes - destilação (90% rec.)

Variável Latente	Variância Explicada Acumulada (%)
1	6,55
2	14,14
3	35,45
4	41,99
5	50,21

Tabela 10. Variância explicada acumulada (%) versus número de variáveis latentes - destilação (90% rec.)

6	62,25
7	64,95
8	67,55
9	66,42
10	67,37
11	67,23
12	67,33

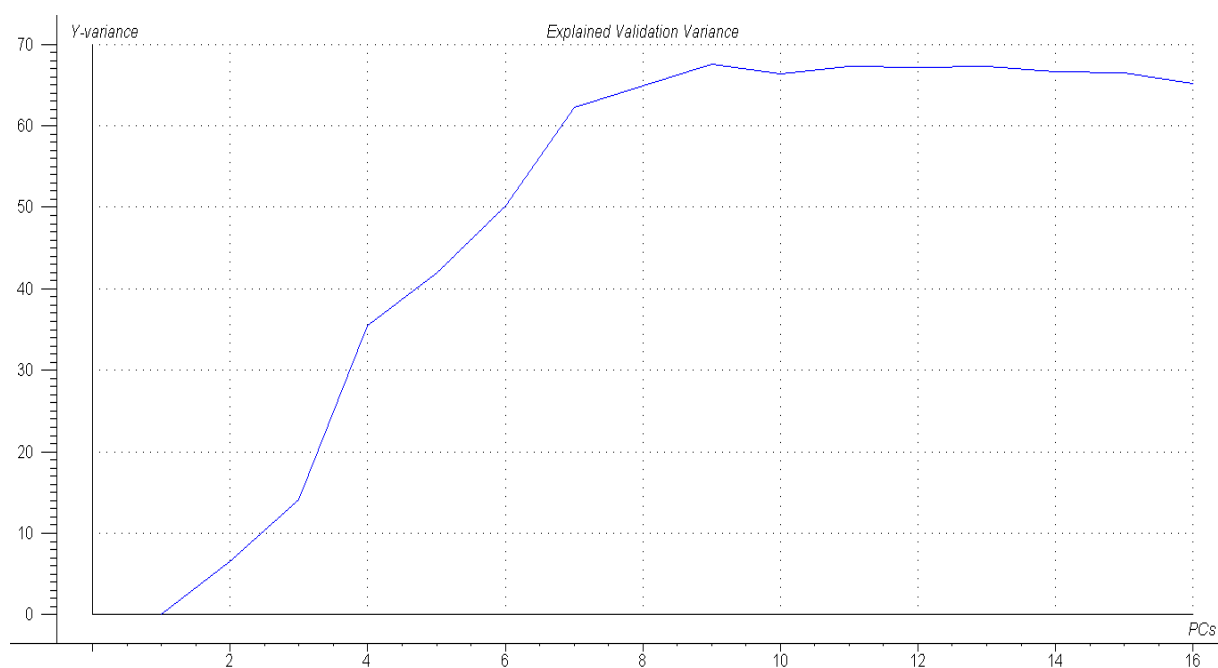


Figura 47. Variância explicada versus número de variáveis latentes para temperatura de destilação (90% rec.)

A Figura 48 mostra o gráfico de valores preditos versus valores reais da propriedade destilação 90% recuperados para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,93919 e RMSECV de 2,2°C.

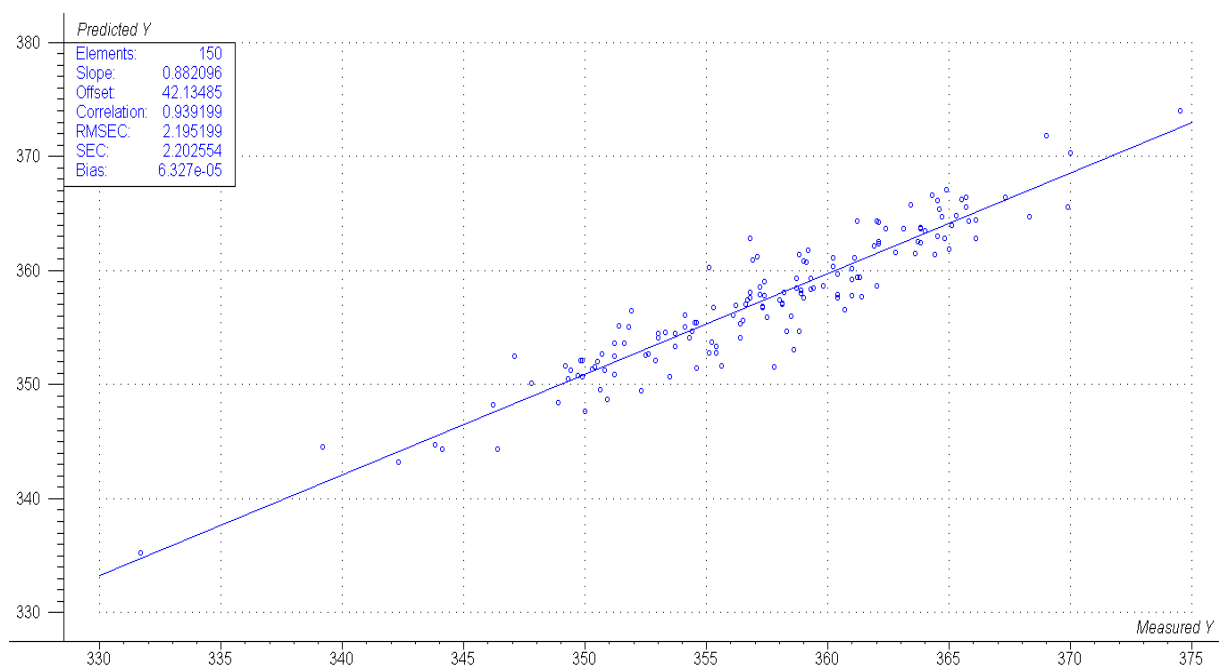


Figura 48. Valores preditos versus valores medidos para temperatura de destilação (90% rec.)

A Tabela 11 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 2,9°C.

Tabela 11. Conjunto de validação externa – temperatura de destilação (90% rec.)

Amostra	Real	Predito	Resíduo
1	358,3	356,0	-2,3
2	355,4	353,5	-1,9
3	360,4	358,6	-1,8
4	362,0	359,0	-3,0
5	355,2	358,2	3,0
6	356,4	359,0	2,6
7	357,3	358,0	0,7
8	356,6	356,0	-0,6
9	355,1	352,9	-2,2
10	342,2	345,1	2,9
11	346,2	347,7	1,5
12	343,8	346,0	2,2
13	350,0	348,7	-1,3
14	346,4	343,9	-2,5
15	342,3	345,3	3,0
16	350,0	350,1	0,1

Tabela 11. Conjunto de validação externa – temperatura de destilação (90% rec.)

17	344,6	348,8	4,2
18	371,7	369,9	-1,8
19	364,2	356,4	-7,8
20	365,1	364,5	-0,6

As Figuras 49 e 50 ilustram que os resíduos deste modelo seguem distribuição normal com p -value de 0,139.

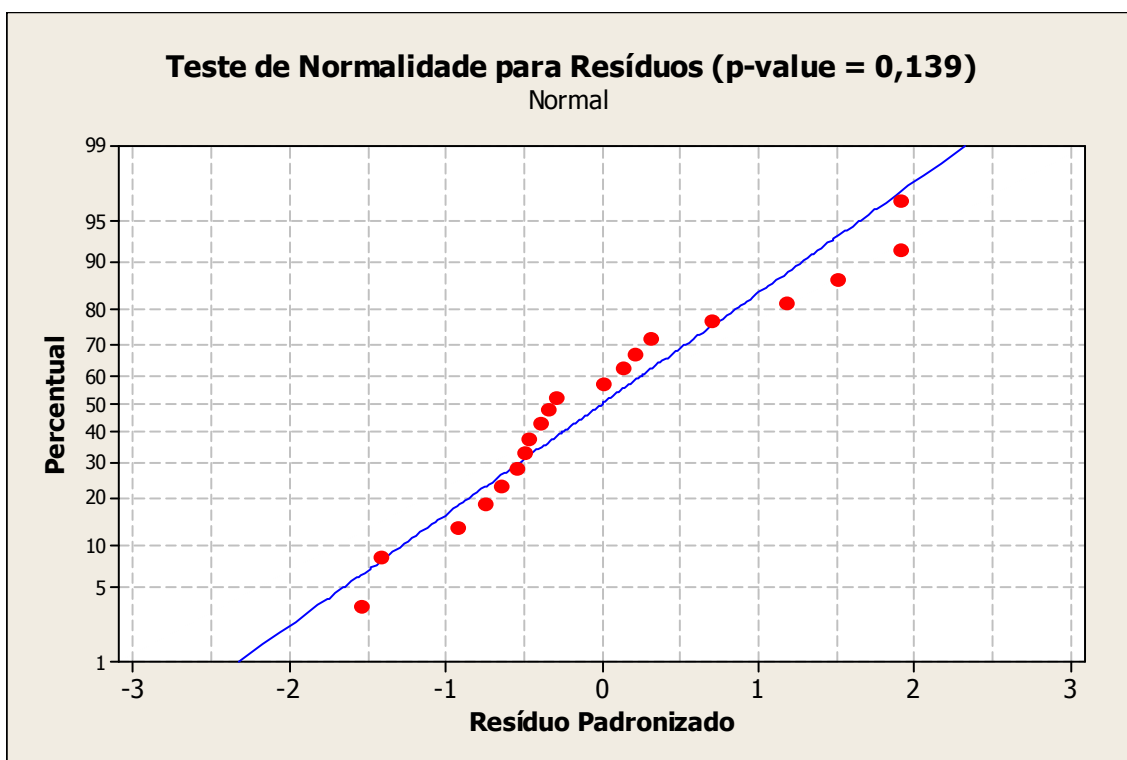


Figura 49. Teste de normalidade para propriedade temperatura de destilação 90% rec.

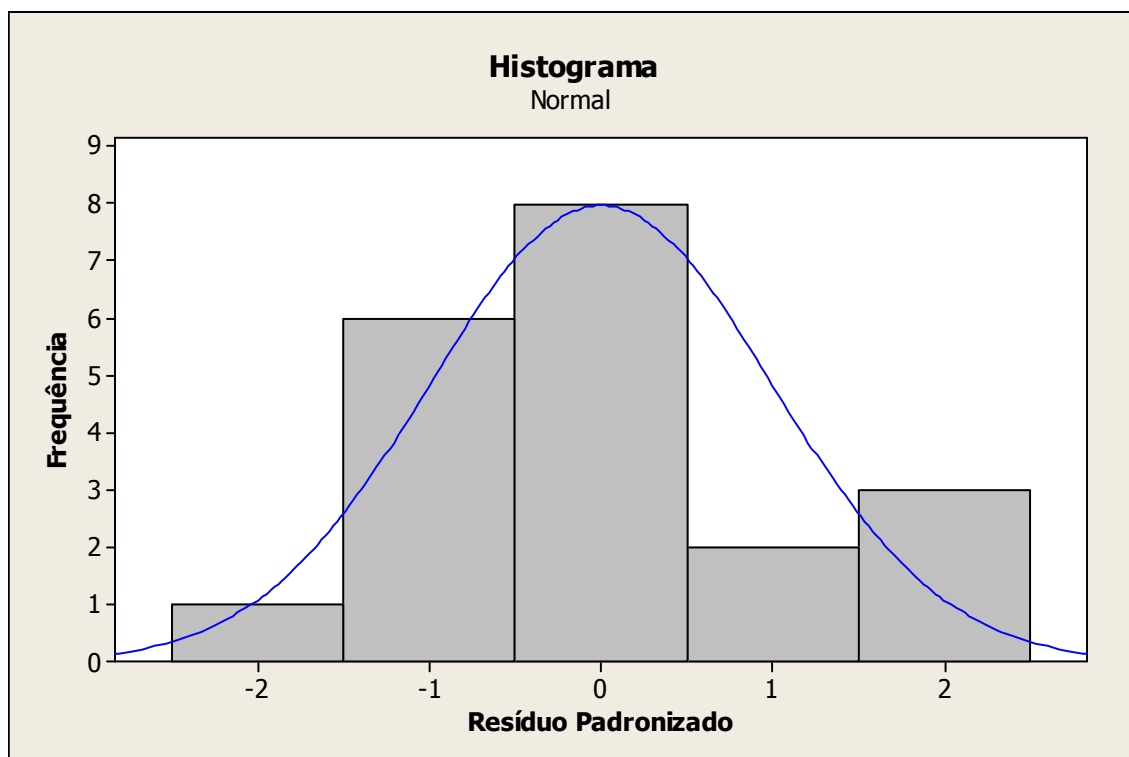


Figura 50. Histograma do resíduo padronizado para propriedade temperatura de destilação 90% rec.

3.4.5 Resultados para propriedade índice de cetano

Para descrever a propriedade índice de cetano são necessárias 06 variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 12 e Figura 51, acima de 06 LV, não há ganhos percentuais significativos em variância explicada. Ao contrário, a partir de 10 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui.

Tabela 12. Variância explicada acumulada (%) versus número de variáveis latentes - índice de cetano

Variável Latente	Variância Explicada Acumulada (%)
1	15,03
2	80,41
3	81,91
4	84,10
5	87,56
6	87,25
7	88,44

Tabela 12. Variância explicada acumulada (%) versus número de variáveis latentes - índice de cetano

8	88,68
9	88,54
10	87,70
11	87,06
12	86,58

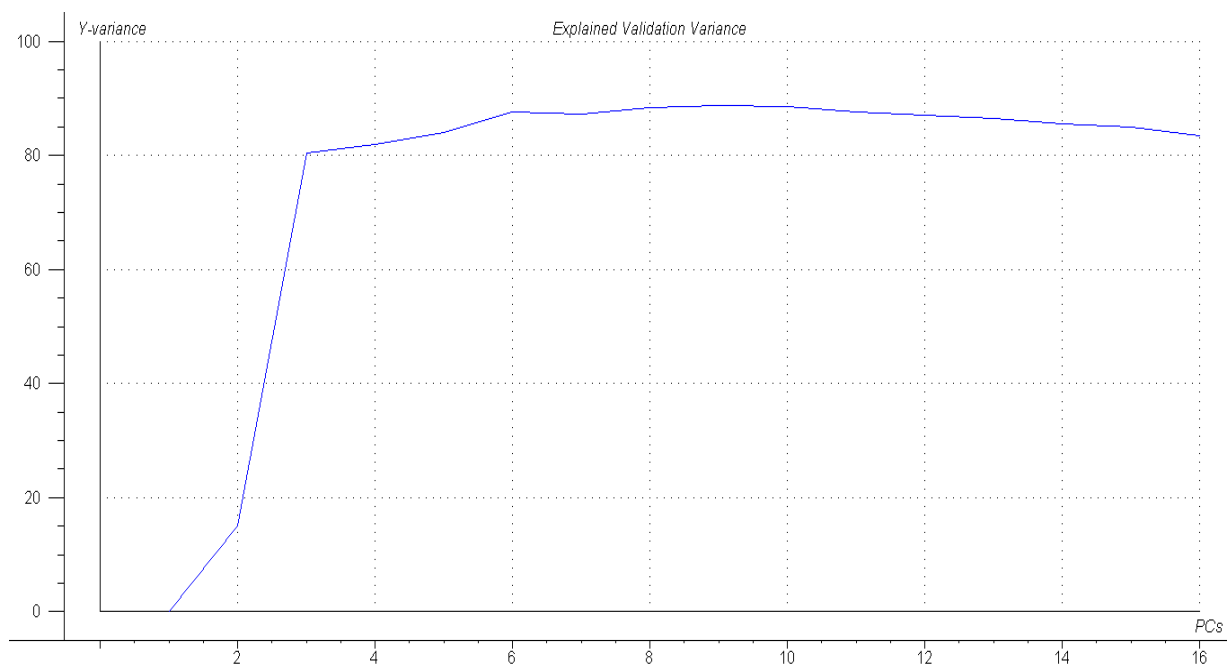


Figura 51. Variância explicada versus número de variáveis latentes - índice de cetano

A Figura 52 mostra o gráfico de valores preditos versus valores reais da propriedade índice de cetano para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,94625 e RMSECV de 0,5.

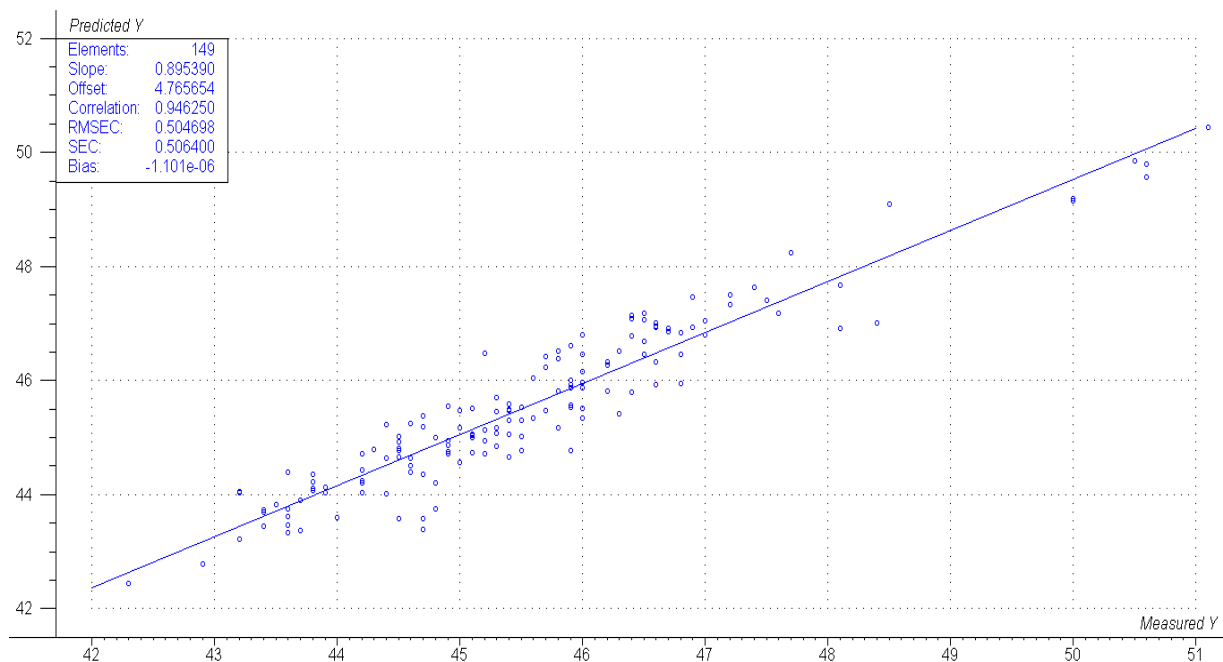


Figura 52. Valores preditos versus valores medidos para índice de cetano

A Tabela 13 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 0,7.

Tabela 13. Conjunto de validação externa – índice de cetano

Amostra	Real	Predito	Resíduo
1	45,3	44,8	-0,5
2	45,0	44,6	-0,4
3	44,6	44,4	-0,2
4	46,4	47,0	0,6
5	46,6	46,4	-0,2
6	47,5	47,5	0,0
7	46,4	47,0	0,6
8	46,8	46,5	-0,3
9	46,5	47,0	0,5
10	45,8	46,6	0,8
11	46,4	45,8	-0,6
12	45,8	46,5	0,7
13	46,0	46,6	0,6
14	46,6	46,0	-0,6
15	45,7	46,4	0,7
16	50,0	49,0	-1,0

Tabela 13. Conjunto de validação externa – índice de cetano

17	48,0	47,3	-0,7
18	47,2	45,7	-1,5
19	51,3	51,2	-0,1
20	50,9	50,5	-0,4

As Figuras 53 e 54 ilustram que os resíduos deste modelo seguem distribuição normal com *p-value* de 0,123.

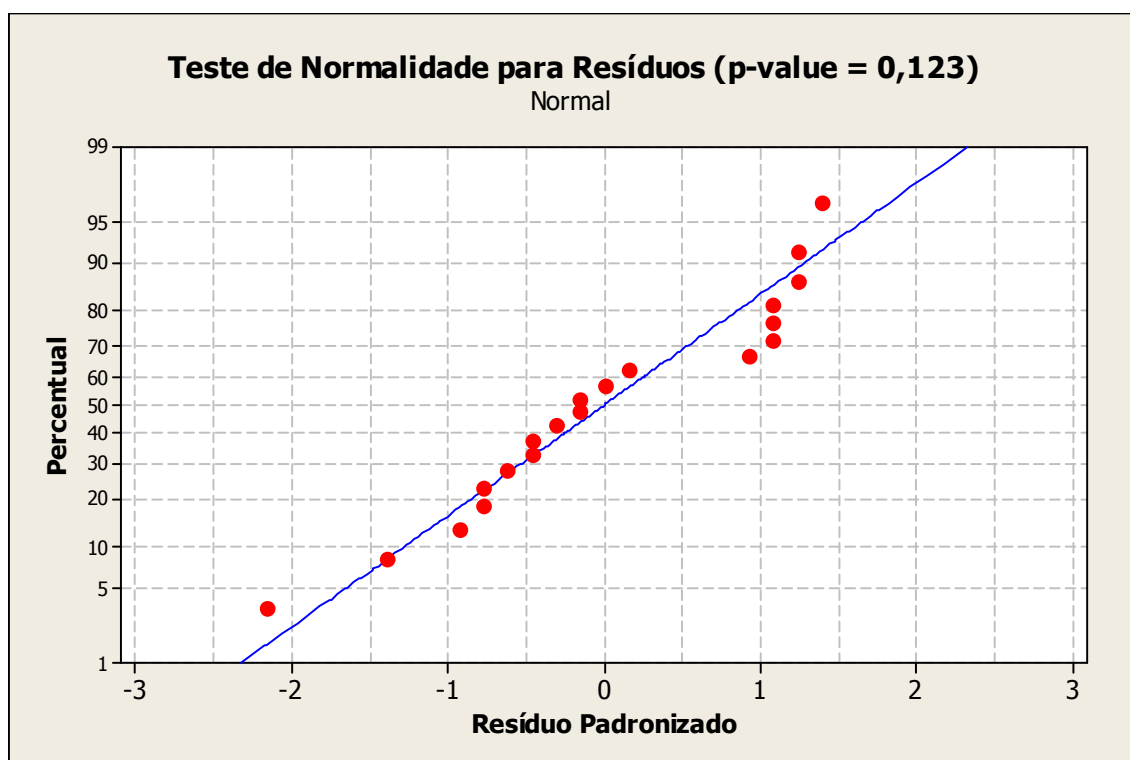


Figura 53. Teste de normalidade para propriedade índice de cetano

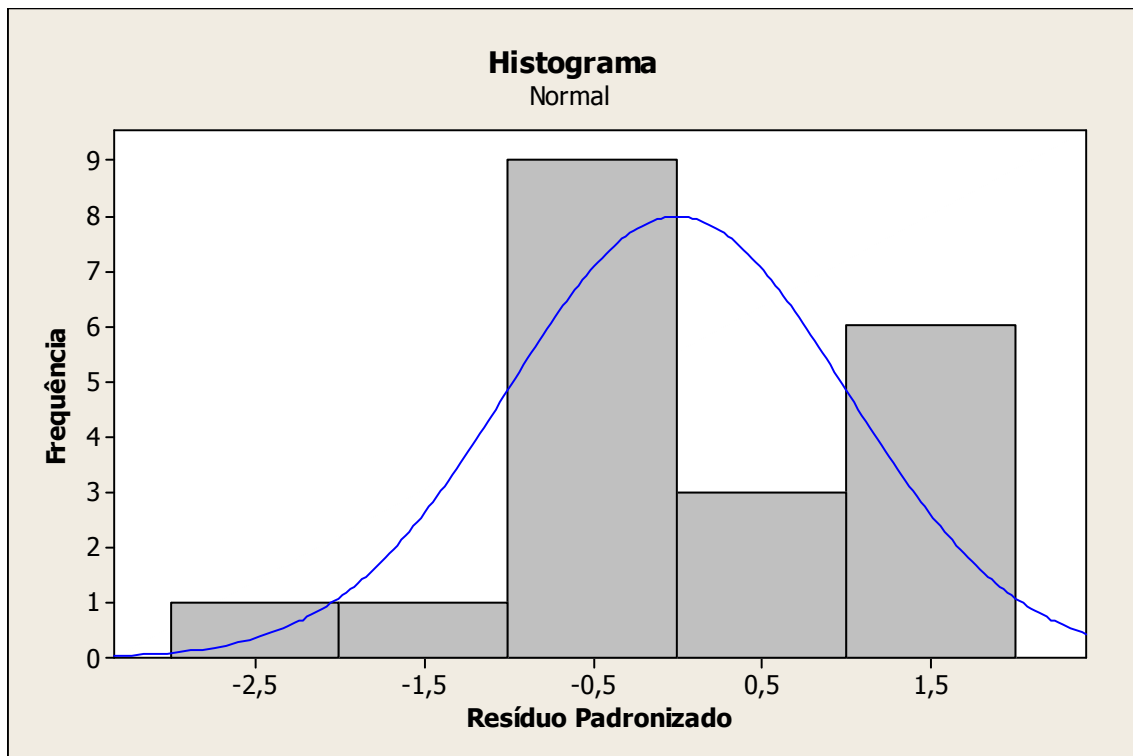


Figura 54. Histograma do resíduo padronizado para propriedade índice de cetano

3.4.6 Resultados para propriedade viscosidade cinemática

Para descrever a propriedade viscosidade cinemática são necessárias 09 variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 13 e Figura 55, acima de 09 LV, não há ganhos percentuais significativos em variância explicada. Ao contrário, a partir de 10 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui.

Tabela 13. Variância explicada acumulada (%) versus número de variáveis - viscosidade cinemática

Variável Latente	Variância Explicada Acumulada (%)
1	6,02
2	27,99
3	52,28
4	70,12
5	72,53
6	76,80
7	78,57

Tabela 13. Variância explicada acumulada (%) versus número de variáveis - viscosidade cinemática

8	80,04
9	80,10
10	77,77
11	76,31
12	74,75

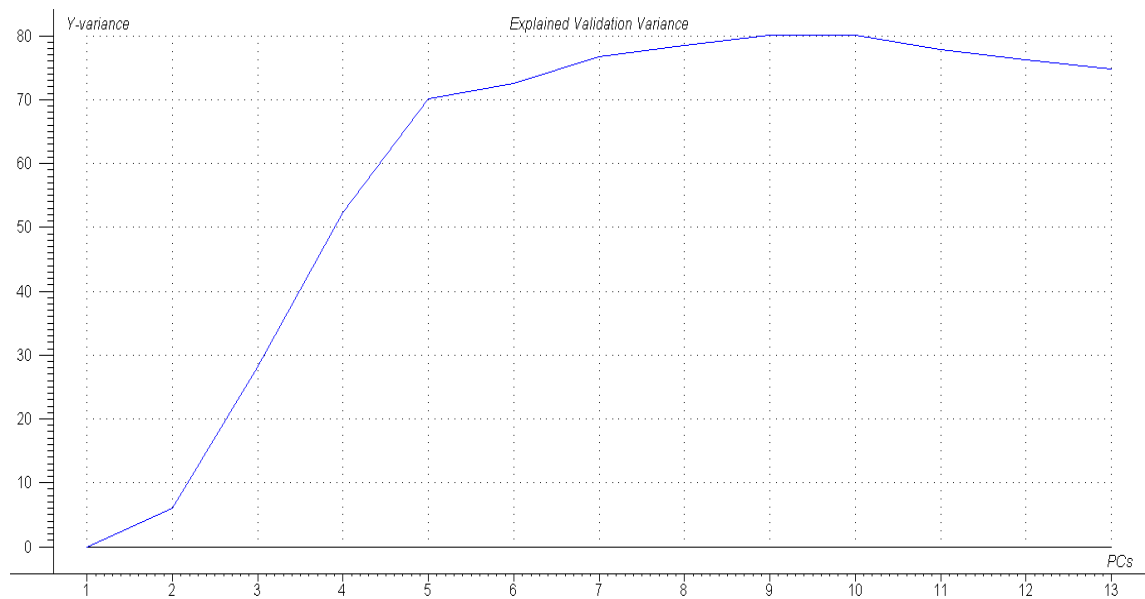


Figura 55. Variância explicada versus número de variáveis latentes para viscosidade cinemática

A Figura 56 mostra o gráfico de valores preditos versus valores reais da propriedade viscosidade cinemática para o modelo PLS desenvolvido com coeficiente de correlação R^2 0,93082 e RMSECV de 0,15 cSt.

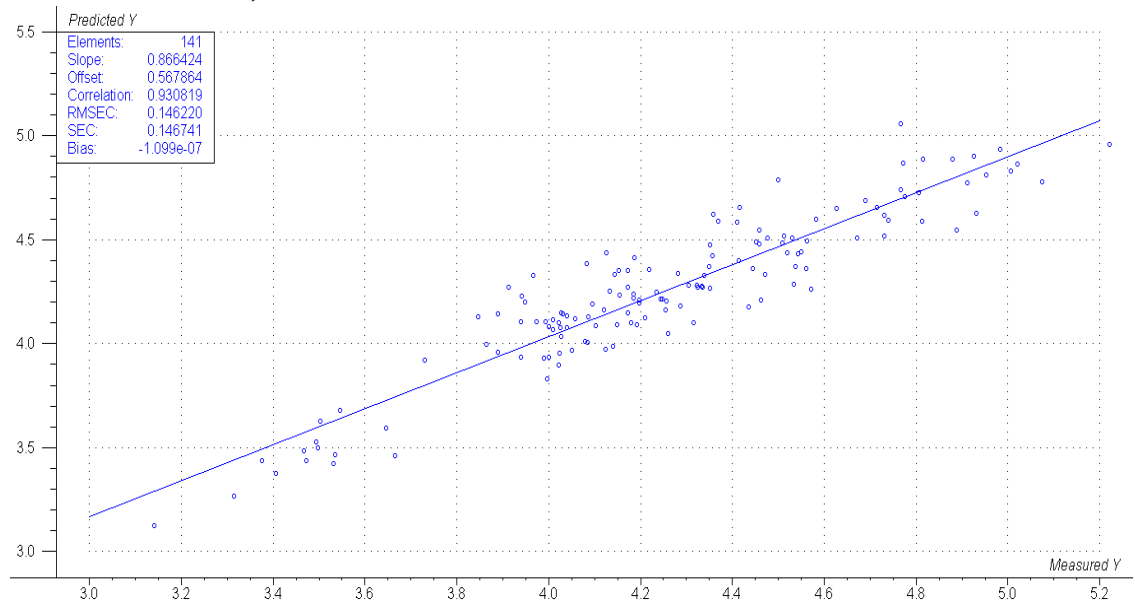


Figura 56. Valores preditos versus valores medidos para viscosidade cinemática

A Tabela 14 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi de 0,16 cSt.

Tabela 14. Conjunto de validação externa – viscosidade cinemática

Amostra	Real	Predito	Resíduo
1	4,578	4,867	0,289
2	4,647	4,707	0,060
3	4,521	4,351	-0,170
4	4,897	4,752	-0,145
5	4,936	5,105	0,169
6	3,597	3,559	-0,038
7	4,201	4,060	-0,141
8	5,081	5,234	0,153
9	4,083	4,215	0,132
10	4,864	4,600	-0,264
11	4,863	4,804	-0,059
12	4,977	4,899	-0,078
13	5,107	4,880	-0,227
14	3,922	3,903	-0,019
15	5,033	5,197	0,164
16	5,014	5,007	-0,008
17	4,606	4,807	0,201
18	4,460	4,613	0,153
19	4,721	4,631	-0,090
20	4,329	4,583	0,254

As Figuras 57 e 58 ilustram que os resíduos deste modelo seguem distribuição normal com *p-value* de 0,374.

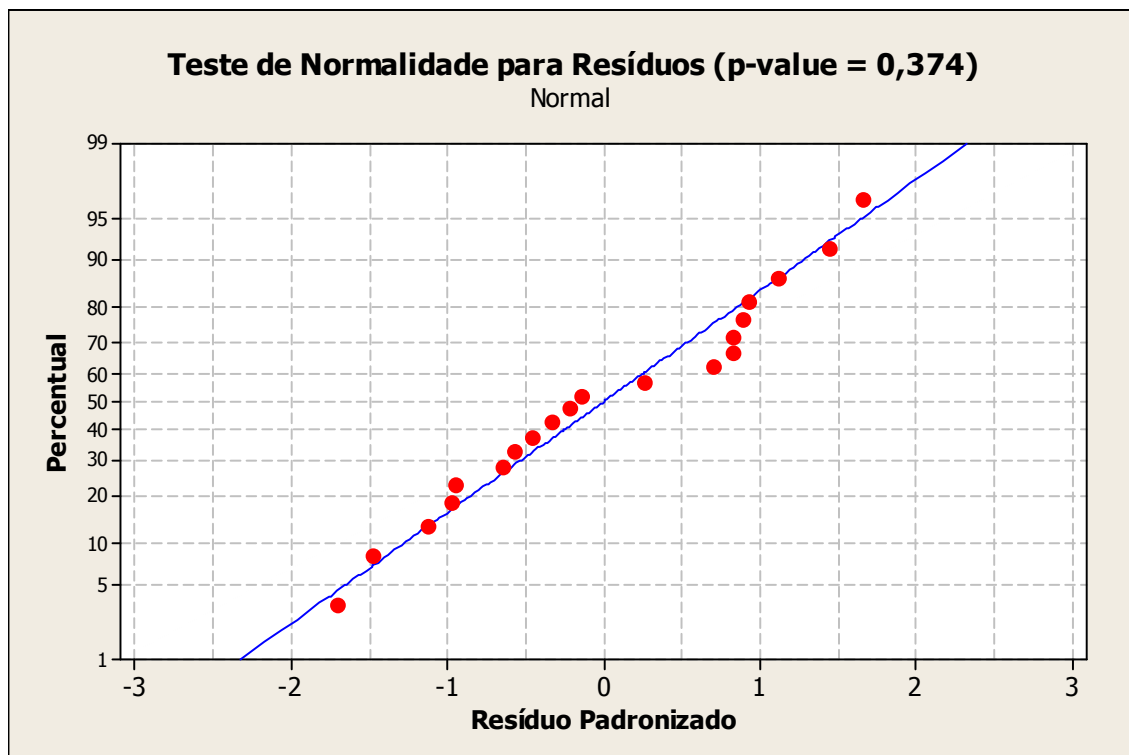


Figura 57. Teste de normalidade para propriedade viscosidade cinemática

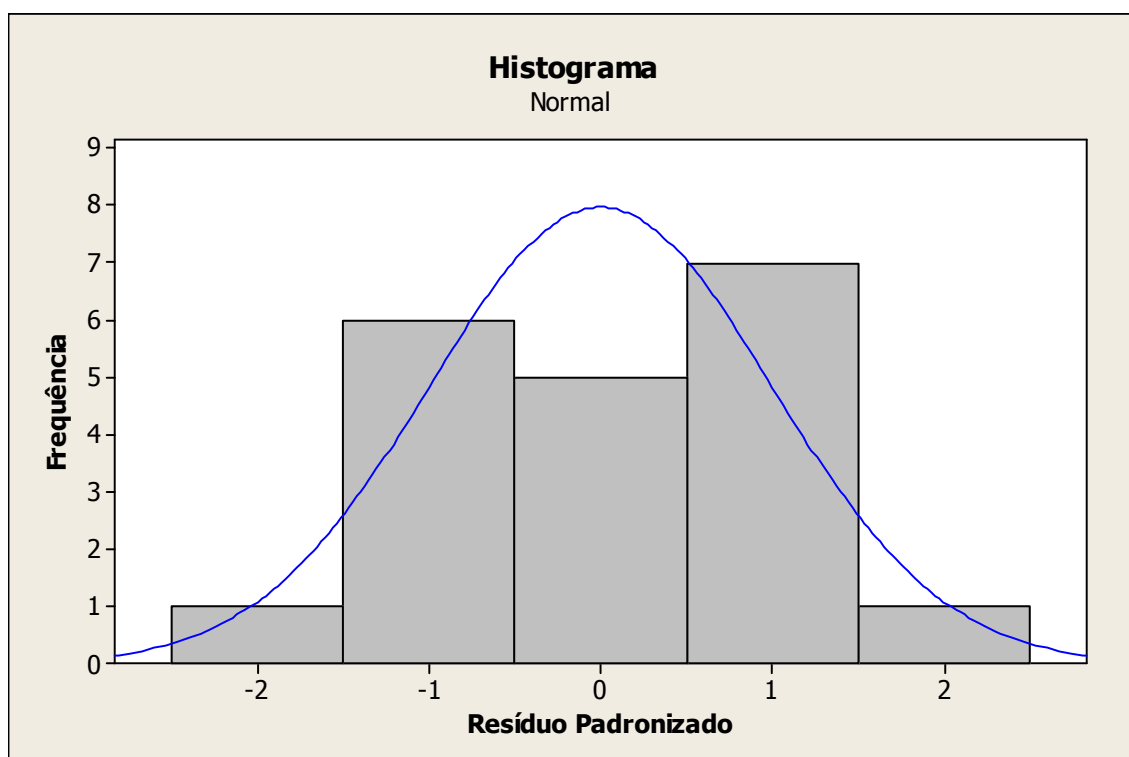


Figura 58. Histograma do resíduo padronizado para propriedade viscosidade cinemática

3.4.7 Resultados para propriedade ponto de fulgor

Para descrever a propriedade ponto de fulgor são necessárias 05 variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 15 e Figura 59, acima de 05 LV, não há ganhos percentuais de variância explicada. Ao contrário, a partir de 07 LV começa a ocorrer o fenômeno *overfitting*, pois com o aumento do número de LV, o percentual de variância explicada diminui.

Tabela 15. Variância explicada acumulada (%) versus número de variáveis latentes - ponto de fulgor

Variável Latente	Variância Explicada Acumulada (%)
1	15,89
2	54,77
3	67,93
4	78,81
5	79,19
6	78,97
7	77,34
8	74,05
9	76,14
10	70,17
11	68,31
12	66,92

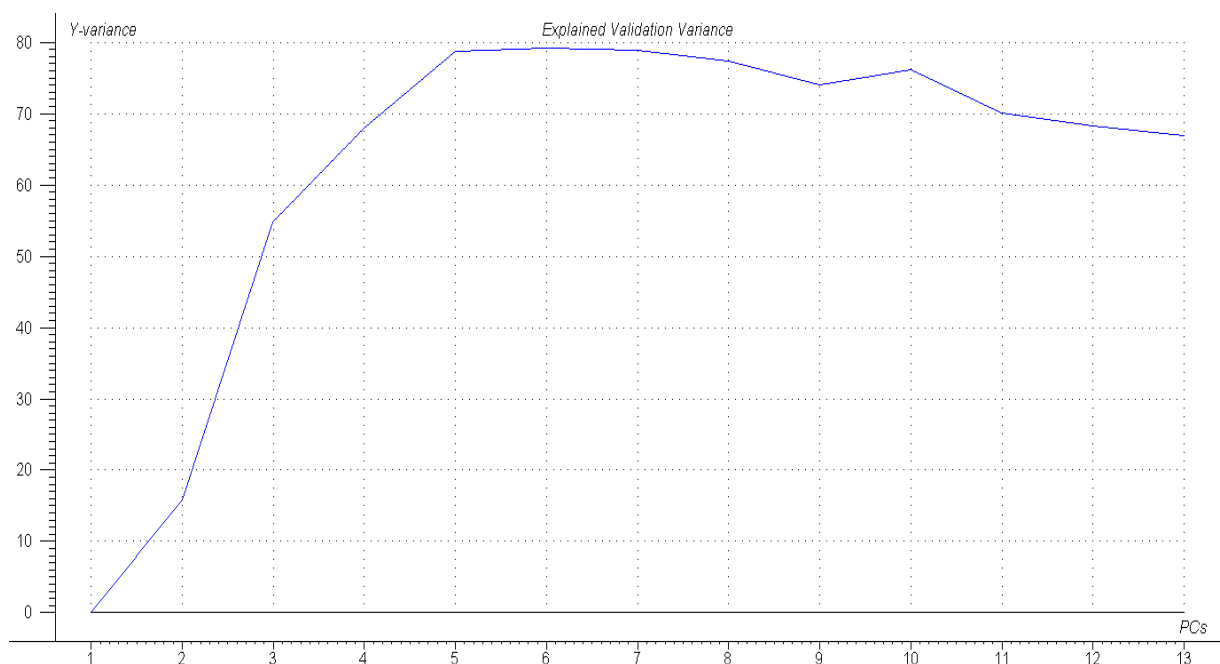


Figura 59. Variância explicada versus número de variáveis latentes para ponto de fulgor

A Figura 60 mostra o gráfico de valores preditos versus valores reais da propriedade ponto de fulgor para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,90057 e RMSECV de 3,3°C.

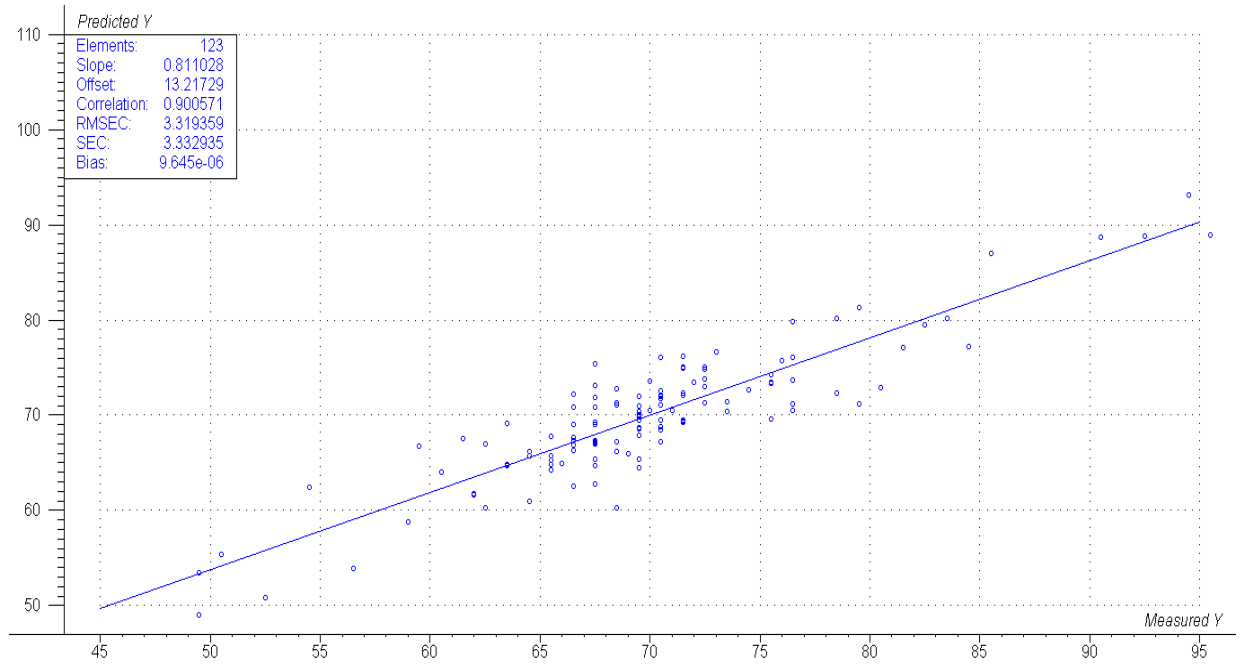


Figura 60. Valores preditos versus valores medidos para ponto de fulgor

A Tabela 16 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 3,4°C.

Tabela 16. Conjunto de validação externa – ponto de fulgor

Amostra	Real	Predito	Resíduo
1	63,5	70,4	6,9
2	65,5	67,1	1,6
3	70,5	73,0	2,5
4	70,5	72,7	2,2
5	69,5	66,2	-3,3
6	70,5	69,4	-1,1
7	64,5	67,1	2,6
8	70,5	68,7	-1,8
9	69,5	71,6	2,1
10	69,5	70,0	0,5
11	76,5	70,5	-6,0
12	76,5	71,3	-5,2
13	68,5	71,2	2,7
14	66,5	72,3	5,8

Tabela 16. Conjunto de validação externa – ponto de fulgor

15	70,5	71,0	0,5
16	83,4	84,0	0,6
17	92,5	89,0	-3,5
18	69,5	69,7	0,2
19	63,5	65,0	1,5
20	68,5	73,0	4,5

As Figuras 61 e 62 ilustram que os resíduos deste modelo seguem distribuição normal com *p-value* de 0,504.

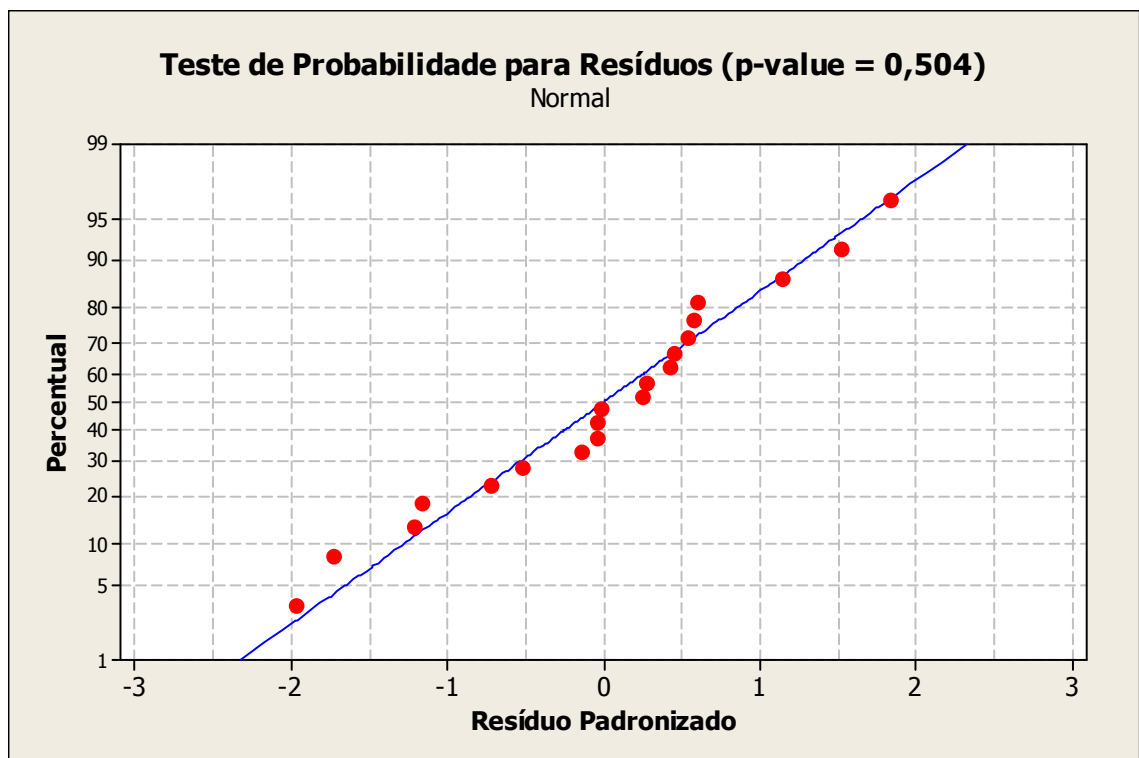


Figura 61. Teste de normalidade para propriedade ponto de fulgor

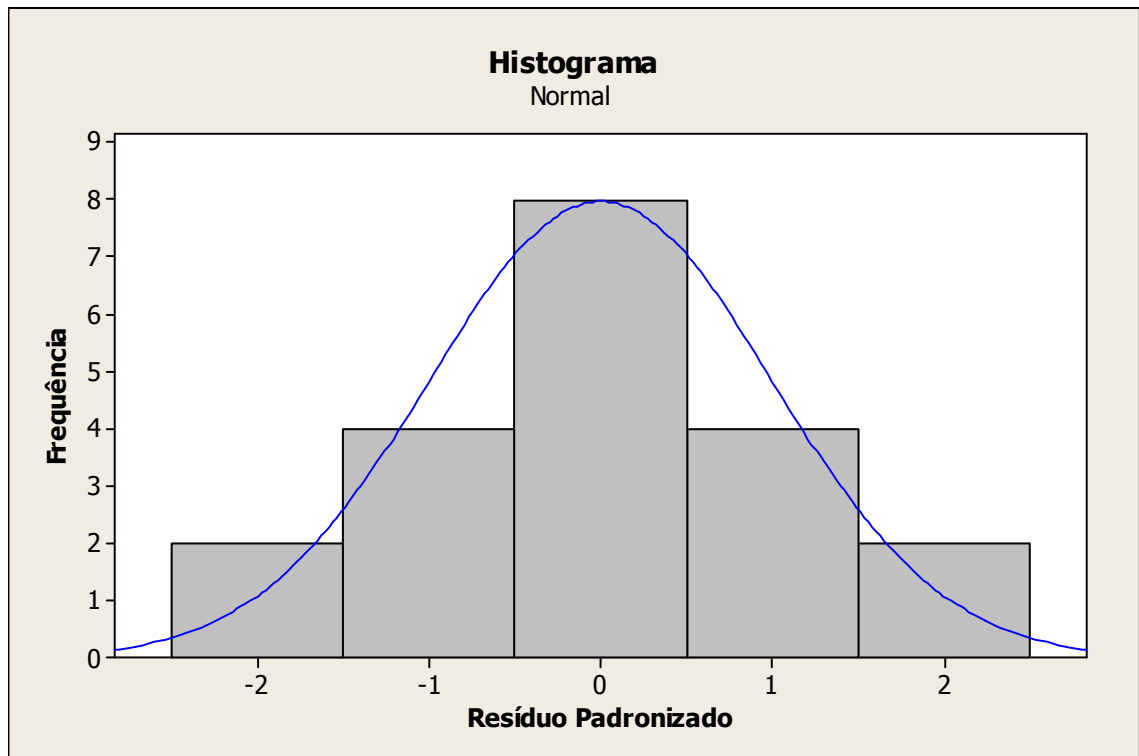


Figura 62. Histograma do resíduo padronizado para propriedade ponto de fulgor

3.4.8 Resultados para propriedade densidade

Para descrever a propriedade densidade são necessárias 08 variáveis latentes, segundo o algoritmo de validação cruzada. Conforme Tabela 17 e Figura 63, acima de 08 LV, não há ganhos percentuais significativos em variância explicada.

Tabela 17. Variância explicada acumulada (%) versus número de variáveis latentes - densidade

Variável Latente	Variância Explicada Acumulada (%)
1	8,61
2	26,38
3	64,66
4	92,41
5	93,64
6	96,28
7	97,07
8	97,29
9	97,27
10	97,27

Tabela 17. Variância explicada acumulada (%) versus número de variáveis latentes – densidade

11	97,31
12	97,39

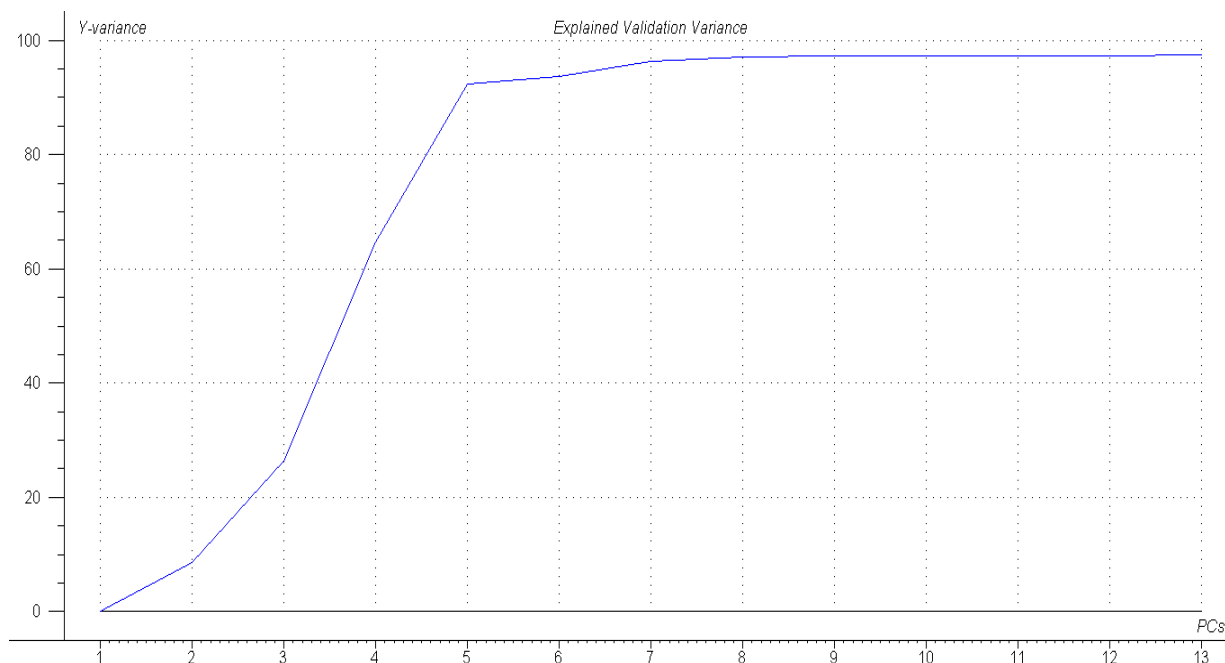


Figura 63. Variância explicada versus número de variáveis latentes para densidade

A Figura 64 mostra o gráfico de valores preditos versus valores reais para a propriedade densidade para o modelo PLS desenvolvido com coeficiente de correlação R^2 de 0,98630 e RMSECV de 0,00067.

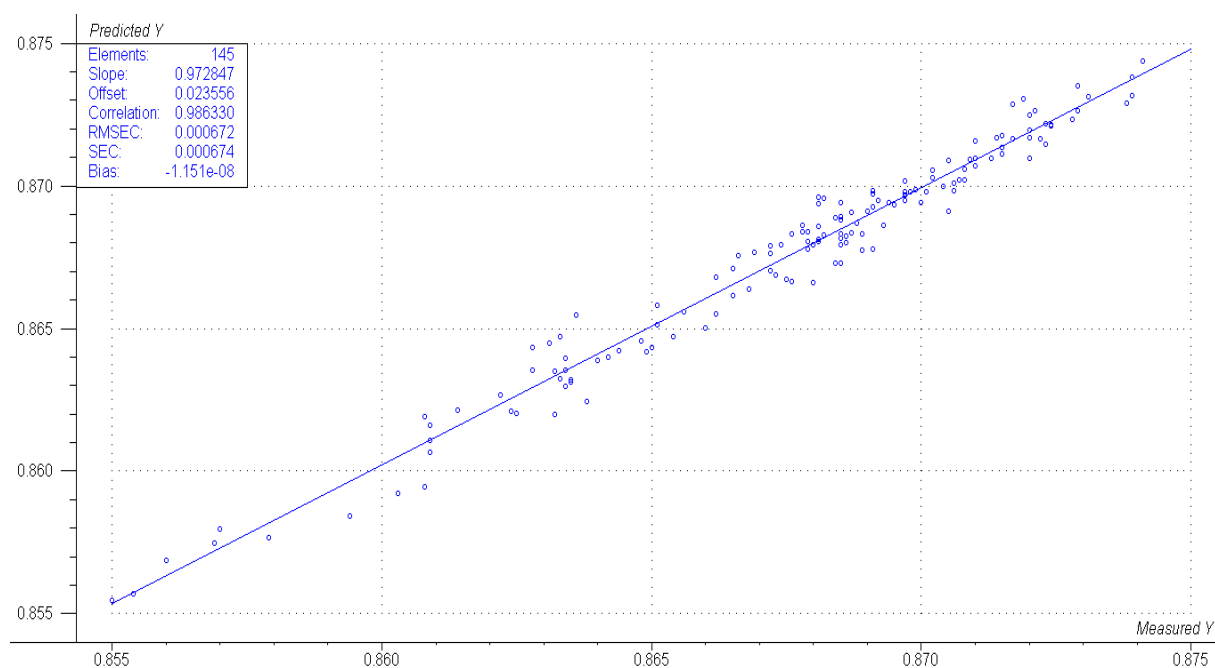


Figura 64. Valores preditos versus valores medidos para densidade

A Tabela 18 representa os resultados da validação externa. Para esta propriedade o RMSEP obtido foi 0,0012.

Tabela 18. Conjunto de validação externa – densidade

Amostra	Real	Predito	Resíduo
1	0,8675	0,8688	0,0013
2	0,8721	0,8724	0,0003
3	0,8713	0,8706	-0,0007
4	0,8734	0,8713	-0,0021
5	0,8740	0,8758	0,0018
6	0,8552	0,8559	0,0007
7	0,8625	0,8634	0,0009
8	0,8763	0,8767	0,0004
9	0,8620	0,8618	-0,0002
10	0,8643	0,8648	0,0005
11	0,8726	0,8720	-0,0006
12	0,8670	0,8673	0,0003
13	0,8765	0,8730	-0,0035
14	0,8630	0,8628	-0,0002
15	0,8789	0,8786	-0,0003
16	0,8640	0,8644	0,0004
17	0,8596	0,8589	-0,0007
18	0,8620	0,8613	-0,0007
19	0,8572	0,8572	0,0000
20	0,8669	0,8683	0,0013

As Figuras 65 e 66 ilustram que os resíduos deste modelo seguem distribuição normal com *p-value* de 0,139.

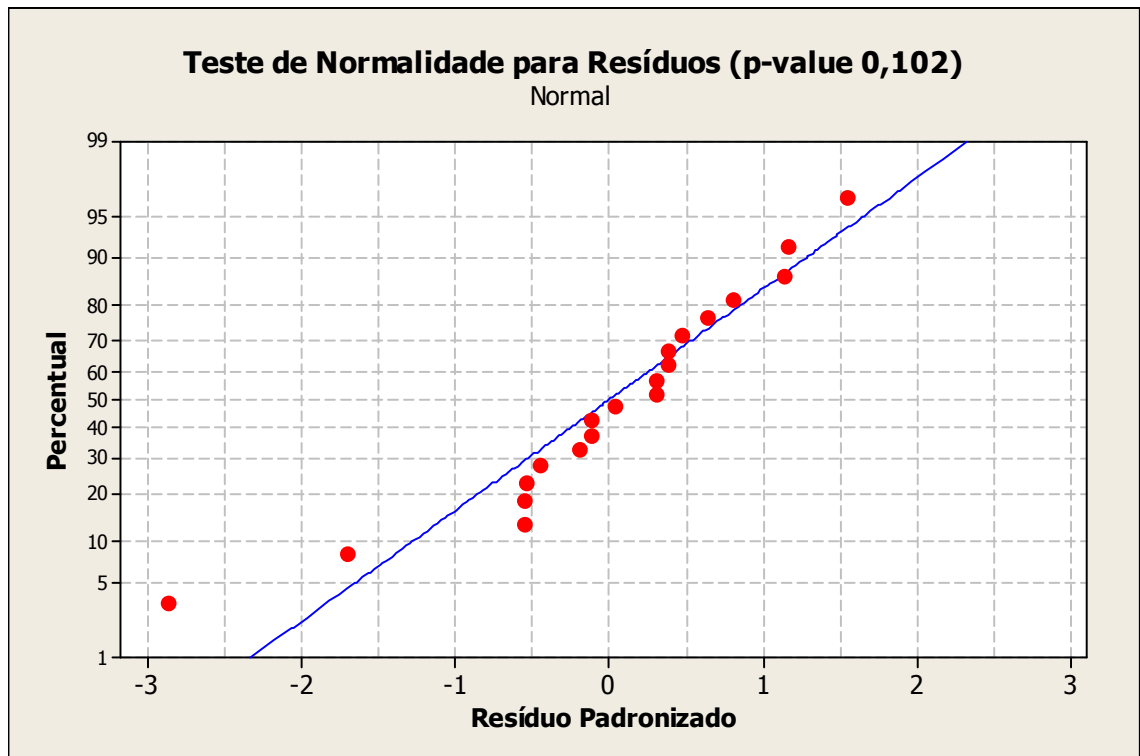


Figura 65. Teste de normalidade para propriedade densidade

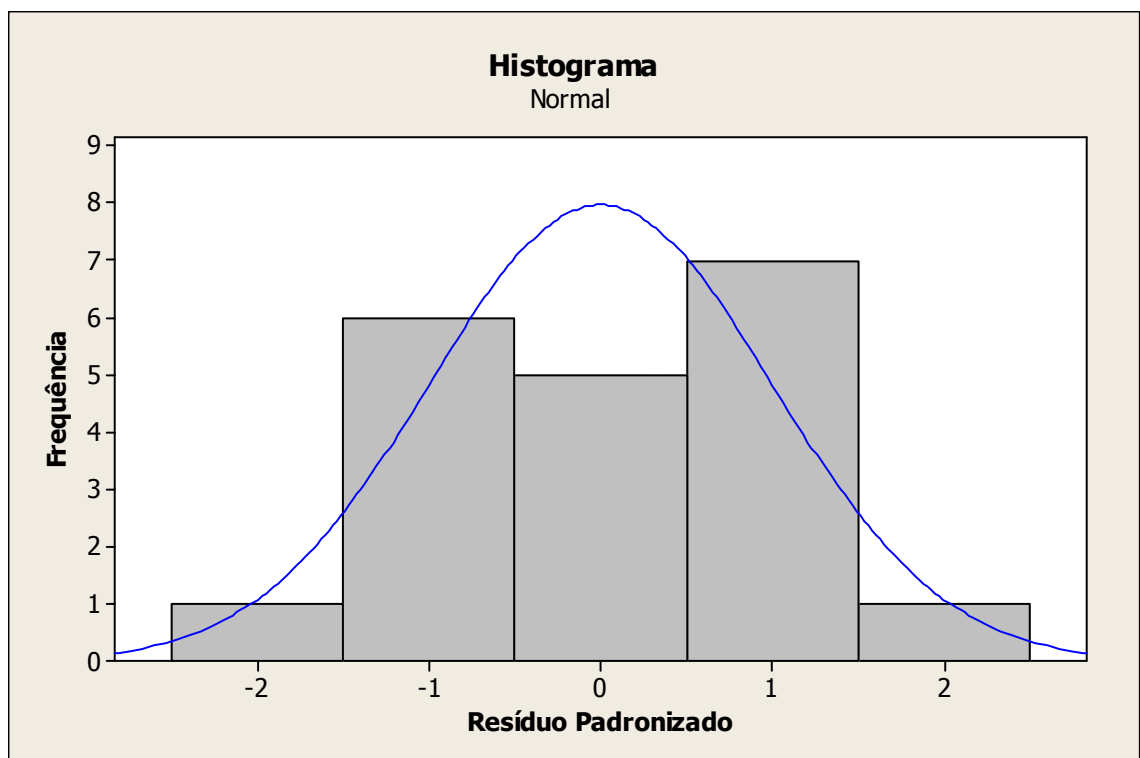


Figura 66. Histograma do resíduo padronizado para propriedade densidade

A equivalência estatística dos indicadores RMSECV e RMSEP revela que os modelos apresentam robustez na predição de amostras externas.

Os modelos foram implementados em escala industrial para monitoramento *on-line*. Apenas os ensaios de certificação de venda do produto estão sendo realizados em laboratório, por exigência da ASTM. Todos os tanques finais de diesel atendem às especificações técnicas do produto.

Por fim, foi determinado um Plano de Inspeção da Qualidade semanal de validação contínua dos modelos com o objetivo de avaliar e garantir a robustez ao longo do tempo.

4 CONCLUSÕES

4.1 Verificação dos objetivos

Foi possível modelar matematicamente as propriedades destilação, índice de cetano, ponto de fulgor, viscosidade cinemática e densidade do óleo diesel pela técnica multivariada PLS a partir da absorção da radiação infravermelha.

Os modelos desenvolvidos foram validados externamente segundo protocolo da norma ASTM D-6122 e suas previsões apresentam precisão equivalente aos métodos de referência conforme ASTM D-4052, D-86, D-613, D93 e D-445.

Os resultados obtidos foram utilizados em uma planta industrial numa refinaria brasileira e contribuiu para redução da variabilidade do processo, aumento da rentabilidade da produção e agilidade na tomada de decisão no sistema de *blending*, pois ensaios que levavam quatro horas para serem executados em laboratório, não levam mais de três minutos.

Embora se dispusesse de um banco de dados relativamente grande, a compressão dos dados em não mais de dez variáveis latentes foi suficiente para explicação de sua variabilidade.

O algoritmo de validação cruzada se mostrou eficiente na determinação da dimensão dos modelos, isto é, a determinação do número de variáveis latentes quando o RMSECV é mínimo foi robusta devido à proximidade estatística deste indicador com o RMSEP. Isso fez com que os modelos não operassem sob condições de *underfitting* e *overfitting*.

4.2 Sugestões para continuidade do trabalho

Este projeto de pesquisa pode ter continuidade, a ideia seria verificar possíveis não linearidades e comparar a modelagem via Análise Multivariada com Redes Neurais Artificiais, estendendo para os demais produtos da refinaria como petróleo, nafta, gasolina, reformado e querosene de aviação. Além disso, explorar abordagens híbridas na determinação do número de variáveis latentes, como Simulação de Monte Carlo ou Metaheurísticas, já que este é um ponto crucial nos métodos que utilizam redução de dimensão. Por fim, realizar monitoramento *on-line* de uma refinaria de petróleo.

REFERÊNCIAS

ADLER, N.; YAZHEMSKY, E. Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction. **European Journal of Operational Research**, v. 202, p. 273-284, 2010.

AGÊNCIA DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Esquemas de Produção no Refino de Petróleo**: Disponível em: <<http://www.anp.gov.br/?pg=7854&m=esquema+de+refino&t1=&t2=esquema+de+refino&t3=&t4=&ar=0&ps=1&cachebust=1331008874709>>. Acesso em: 06 03 2012.

AGUILERA, A. M.; ESCABIAS, M.; VALDERRAMA, M. J. Using principal components for estimating logistic regression with high-dimensional multicollinear data. **Computational Statistics & Data Analysis**, v. 50, p. 1905-1924, 2006.

AHN, H.; CHOI, E.; HAN, I. Extracting underlying meaningful features and canceling noise using independent component analysis for direct marketing. **Expert Systems with Applications**, v. 33, p. 181-191, 2007.

ALEME, H. G.; CORGOZINHO, C. N. C.; BARBEIRA, P. J. S. Diesel oil discrimination by origin and type using physicochemical properties and multivariate analysis. **Fuel**, v. 89, p. 3151-3156, 2010.

ALLEGRI, 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, p. 18-25, 2011.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard test method for calculate cetane index**. ASTM D-613 International, 2010, p.1-5.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard test method for density, relative density, and API gravity of liquids by digital density meter**. ASTM D-4052 International, 2009, p.1-8.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard test method for distillation of petroleum products at atmospheric pressure**. ASTM D-86 International, 2010, p.1-28.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard test method for flash point by tag closed cup tester**. ASTM D-93 International, 2005, p.1-12.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard test method for kinematic viscosity of transparent and opaque liquids**. ASTM D-445 International, 2009, p.1-10.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. **Standard Practice for Validation of the Performance of Multivariate Online, At-Line, and Laboratory**

Infrared Spectrophotometer Based Analyzer Systems. ASTM D-6122 International, 2010, p.1-5.

ANDERSON, T. W. **An Introduction to Multivariate Statistical**, 3 ed. Hoboken Wiley Series in Probability and Statistics, 2003. 742p.

BAFFI, G.; MARTIN, E. B.; MORRIS, A. J. Non-linear projection to latent structures revisited (the neural network PLS algorithm). **Computers and Chemical Engineering**, v. 23, p. 1283-1307, 1999.

BAO, X.; DAI, L. Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties. **Fuel**, v.88, p. 1216-1222, July 2009.

BARROS, A. S.; RUTLEDGE, D. N. Principal components transform-partial least squares: a novel method to accelerate cross-validation in PLS regression. **Chemometrics and Intelligent Laboratory Systems**, v. 73, p. 245-255, 2004

BASTIEN, P.; VINZI, V. E.; TENENHAUS, M. PLS generalised linear regression. **Computational Statistics & Data Analysis**, v. 48, p. 17-46, 2005.

CAMOLESI, V. J. **Caracterização do querosene através da espectroscopia de infravermelho próximo**. 2009. 50 f. Dissertação (Mestrado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

CHUNG, H. Applications of near-infrared spectroscopy in refineries and important issues to address. **Applied Spectroscopy Reviews**, v. 42, n.3 p. 251-285, 2007.

COSTA, A. F. B.; EPPRECHT, E. K.; CARPINETTI, L.C. R. Introdução e conceitos fundamentais. In: **Controle Estatístico de Qualidade**. São Paulo. Atlas, p15-20, 2010.

CUELLAR, F. D. R. **Representação do petróleo através de componentes reais para a simulação de processo de refino**. 2009. 54 f. Dissertação (Mestrado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

DAVE, D.; MURTUZA, Z.; DABHIYA, S. V. K.; GANGULY, S. Online tuning of a steady state crude distillation unit model for real time application. **Journal of Process Control**, Washington, v. 13, p-267-282, 2003.

DAVIS, J. F.; PIOVOSO, M. J.; HOO, K. A.; BAKSHI, B. R. Process Data Analysis and Interpretation. **Advances in Chemical Engineering**, v. 25, p. 1-103, 2000.

ERGON, R. Informative PLS score-loading plots for process understanding and monitoring. **Journal of Process Control**, v. 14, p. 889-897, 2004.

ESCANDAR, G. M.; DAMIANI, P.; GOICOEGHEA, H. C.; OLIVIERI, A. C. A review of multivariate calibration methods applied to biomedical analysis. **Process Microchemical Journal**, v. 82, p. 29-42, 2006.

FERREIRA, A. C. **Modelos de otimização na produção de óleo diesel: uma aplicação industrial**. 2008. 107 f. Tese (Doutorado em Engenharia Química) – Faculdade de Engenharia Química, Universidade Estadual Campinas, Campinas, 2008.

FERREIRA, M. M. C.; MONTANARI, C. A.; GAUDIO, A. C. Seleção de Variáveis em QSAR. **Química Nova**, v. 25, p. 439-448, 2002.

FORATO, L. A.; FILHO, R. B.; COLNAGO, L. A. Estudos de métodos de aumento de resolução de espectros de FTIR para análises de estruturas secundárias de proteínas. **Química Nova**, v. 20, p. 146-150, 1997.

FREITAS, L. V.; FREITAS, A. P. B. R.; MARINS, F. A. S.; LOURES, C. C. A.; SILVA, M. B. Multivariate modeling in quality control of viscosity in fuel: an application in oil industry. **Fuel Injection in Automotive Engineering**. 1ed. Rijeka: InTech, 2012, 144p.

GELADI, P.; KOWALSKI, B. Partial Least Square Regression: A Tutorial. **Analytica Chimica Acta**, v. 185, p. 1-17, 1986.

GOODARZI, M.; FREITAS, M.; JENSEN, R. Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions. **Chemometrics and Intelligent Laboratory Systems**, v. 98, p. 123-129, 2009.

GOURVÉNEC, S.; PIERNA, J. A. F.; MASSART, D. L.; RUTLEDGE. An evaluation of the PoLiSh smoothed regression and the Monte Carlo Cross-Validation for the determination of the complexity of a PLS model. **Chemometrics and Intelligent Laboratory Systems**, v. 68, p. 48-51, 2003.

HELPER, G. A.; FERRÃO, M. F.; FERREIRA, C. V.; HERMES, N. Aplicação de métodos de análise multivariada no controle qualitativo de essências alimentícias empregando espectroscopia no infravermelho médio. **Ciência e Tecnologia de Alimentos**, Campinas, p.779-786, 2006.

HENRY, R. C.; PARK, E. S.; SPIEGELMAN, C. H. Comparing a new algorithm with the classic methods for estimating the number of factors. **Chemometrics and Intelligent Laboratory Systems**, v. 48, p. 91-97, 1999.

HIDAJAT, K.; CHONG, S. Quality characterization of crude oils by partial least square calibration of NIR spectral profiles. **Journal of Near Infrared Spectroscopy**, Sussex, v.8, p. 53-59, 2000.

HUANG, C. L.; HSU, T. S.; LIU, C. M. The Mahalanobis–Taguchi system – Neural network algorithm for data-mining. **Expert Systems with Applications**, v. 36, p. 5475-5480, 2009.

JOLLIFFE, I. T. **Principal Component Analysis**, 2 ed. New York. Springer Series in Statistics, 2002. 350p.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**, 5 ed. Madinson: Prentice-Hall, 2002. 797p.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS with very large data sets. **Computational Statistics & Data Analysis**, v. 48, p. 69-85, 2005.

KHASHAN, M. A.; NAGGAR, A. M. E.; SHADDAD, E. A new method of determining the optical constants of a thin film from its reflectance and transmittance interferograms in a wide spectral range: 0.2–3 μm . **Optics Communications**, v. 178, p. 123-132, 2000.

KIM, K.; CHO, I.; PARK, J. Use of real-time NIR (near infrared) spectroscopy for the on-line optimization of a crude distillation unit. In: NPRA, **Computer Conference**. Chicago, 2000.

KIM, D.; LEE, J.; KIM, J. Application of near infrared diffuse reflectance spectroscopy for on-line measurement of coal properties. **Korean Journal of Chemical Engineering**, v. 26, p. 489-495, 2009.

KIRSANOV, D.; MEDNOVA, O.; VIETORIS V.; KILMARTIN P. A.; LEGIN, A. Towards reliable estimation of an “electronic tongue” predictive ability from PLS regression models in wine analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 90, p. 109-116, 2012.

KONZEN, P. H.; FURTADO, J. C.; CARVALHO, C. W.; FERRÃO, M. F.; MOLZ, R. F.; BASSANI, I. A.; HÜNING, S. L. Otimização de métodos de controle de qualidade de fármacos usando algoritmo genético e busca tabu. **Pesquisa Operacional**, v. 23, p. 189-207, 2003.

KRAMER, K. E.; SMALL, G. W. Robust absorbance computations in the analysis of glucose by near-infrared spectroscopy. **Vibrational Spectroscopy**, v. 43, p. 440-446, 2007.

LI, G.; QIN, S. J.; ZHOU, D. Geometric properties of partial least squares for process monitoring. **Automatica**, v. 46, p. 204-210, 2010.

LIRA, L. F. B.; VASCONCELOS, F. V. C.; PEREIRA, C. F.; PAIM, A. P. S.; STRAGEVITCH, L.; PIMENTEL M. F. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. **Fuel**, v. 89, p. 405-409, 2010.

LIU, H.; YU, J.; XU, J.; FAN, Y.; BAO, X. Identification of key oil refining technologies for China National Petroleum Co. (CNPC). **Energy Police**, v. 35, p. 2635 -2647, 2007.

LLOBET, M. V. E.; BREZMES J.; VILANOVA, X.; CORREIG, X. A fuzzy ARTMAP- and PLS-based MS e-nose for the qualitative and quantitative assessment of rancidity in crisps. **Sensors and Actuators B**, v. 106, p. 677-686, 2005.

LORHO, G.; WESTAD, F.; BRO, R. Generalized correlation loadings Extending correlation loadings to congruence and to multi-way models. **Chemometrics and Intelligent Laboratory Systems**, v. 84, p. 119-125, 2006.

LOURENÇO, V.; HERDLING, T.; REICH, G.; MENEZES, J. C.; LOCHMANN, D. Combining microwave resonance technology to multivariate data analysis as a novel PAT tool to improve process understanding in fluid bed granulation. **European Journal of Pharmaceutics and Biopharmaceutics**, v. 78, p. 513-521, 2011.

MACKAY, D. Chemometrics, econometrics, psychometrics - How best to handle hedonics? **Expert Systems with Applications**, v. 17, p. 529-535, 2006.

MACHADO, M. A. G.; COSTA, A. F. B. The use of principal components and univariate charts to control multivariate processes. **Pesquisa Operacional**, v. 28, p. 173-196, 2008.

MANZANO, F. S. **Supply chain practices in the petroleum downstream**. 2005. 30 f. Dissertation (Master of Engineering in Logistics) – Massachusetts Institute of Technology, Cambridge, 2005.

MARTINS, J. P. A.; TEÓFILO, R. F.; FERREIRA, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. **Journal of Chemometrics**, v. 24, p. 320-332, 2010.

MIGUEL, A. C. M; FLEURY, A; MELLO, C. H. P; NAKANO, D. N; TURRIONI, J. B; HO, L. L; MORABITO, R; MARTINS, R. A; PUREZA, V. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 1 ed. Rio de Janeiro. Editora Campus, 2010, 248p.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada**. 1. ed. Belo Horizonte: Editora UFMG, 2005. 297p.

MONTGOMERY, D. C. **Introduction to statistical quality control**. 5. ed. New York: John Wiley & Sons, 2004, 776p.

MOREIRA, A. R. B. Modelos para a projeção do consumo nacional e regional de óleo diesel. **Instituto de Pesquisa Econômica Aplicada**. Rio de Janeiro, 1996.

MORRISON, D. F. **Multivariate Statistical Methods**, 4 ed. Philadelphia. Duxbury Advanced Series, 2005. 498p.

MORO, L. F. L. Process technology in the petroleum refining industry – current situation and future trends. **Computers & Chemical Engineering**, v. 27, p. 1303 - 1305, 2003.

MUTIHAC, L.; MUTIHAC, R. Mining in chemometrics. **Analytica Chimica ACTA**, v. 612, p. 1-18, 2008.

NAES, T.; ISAKSSON, T.; FEARN, T.; DAVIES, T. Partial Least Squares. In: **Multivariate calibration and classification**. Chichester. NIR Publications, 2002.

NAES, T.; MARTENS, H. Multivariate calibration II. Chemometric methods. **Trends in analytical chemistry**, v. 3, p. 266-271, 1984.

NEIRO, S. M. S.; PINTO, J. M. A general modeling framework for the operational planning of petroleum supply chains. **Computers and Chemical Engineering**, v. 28, p. 871-896, 2004.

OTERO, R. L. S.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; CAVALHEIRO, E. T. G. Thermogravimetric determination of l-ascorbic acid in non-effervescent formulations using

multiple linear regression with temperature selection by the successive projections algorithm. **Thermochimica Acta**, v. 526, p. 200-204, 2011.

OTTO, M. Modeling. In: **Chemometrics: Statistics and computer application in analytical chemistry**. New York. Wiley, 2007.

PAIVA, A. P. **Metodologia de superfície de resposta e análise de componentes principais em otimização de processos de manufatura com múltiplas respostas correlacionadas**. 2006. 129 f. Tese (Doutorado em Engenharia Mecânica) – Instituto de Engenharia Mecânica, Universidade Federal de Itajubá, Itajubá, 2006.

PAIVA, A. P.; FERREIRA, J. R.; BALESTRASSI, P. P. A multivariate hybrid approach applied to AISI 52100 hardened steel turning optimization. **Journal of Materials Processing Technology**, v. 189, p. 26-35, 2007.

PAL, A.; MAITI, J. Development of a hybrid methodology for dimensionality reduction in Mahalanobis–Taguchi system using Mahalanobis distance and binary particle swarm optimization. **Expert Systems with Applications**, v. 37, p. 1286-1293, 2010.

PASADAKIS, N.; SOURLIGAS, S.; FOTEINOPOULOS, Ch. Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks. **Fuel**, v. 85, p. 1131-1137, 2006.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. **Journal of Chemical Brazilian Society**, v. 14, p. 198-219, 2003.

PASQUINI, C.; BUENO, A. F. Characterization of petroleum using near-infrared spectroscopy: Quantitative modeling for the true boiling point curve and specific gravity. **Fuel**, v. 86, p. 1927-1934, 2007.

PIRES, J. C. M.; MARTINS, F. G.; SOUSA, S. I. V.; FERRAZ, M. C. M. A.; PEREIRA, M. C. Selection and validation of parameters in multiple linear and principal component regressions. **Environmental Modelling & Software**, v. 23, p. 50-55, 2008.

POZO, C.; FERMENIA, R. R.; CABALLERO, J.; GOSA, G. G. JIMÉNEZ, L. On the use of Principal Component Analysis for reducing the number of environmental objectives in multi-objective optimization: Application to the design of chemical supply chains. **Chemical Engineering Science**, v. 69, p. 146-158, 2012.

PURCELL, D. E.; O'SHEA, M. G.; KOKOT, S. Role of chemometrics for at-field application of NIR spectroscopy to predict sugarcane clonal performance. **Chemometrics and Intelligent Laboratory Systems**, v. 87, p. 113-124, 2007.

QI, S.; SHI, W. M.; KONG, W. Modified tabu search approach for variable selection in quantitative structure–activity relationship studies of toxicity of aromatic compounds. **Artificial Intelligence in Medicine**, v. 49, p. 61-66, 2010.

RAJALAHTI, T.; KVALHEIM, O. M. Multivariate data analysis in pharmaceuticals: A tutorial review. **International Journal of Pharmaceutics**, v. 417, p. 280-290, 2011.

RONCHETTI, E.; FIELD, C.; BLANCHARD, W. Robust linear model selection by cross-validation. **Journal of the American Statistical Association**, v. 92, p. 1017-10231, 1997.

ROY, K.; ROY, P. P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. **European Journal of Medicinal Chemistry**, v. 44, p. 2913-2922, 2009.

SANTOS Jr V. O.; Oliveira F. C. C.; LIMA, D. G.; PETRY, A. C.; GARCIA, E.; SUAREZ, P. A. Z.; RUBIM, J. C. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. **Analytica Chimica ACTA**, v. 547, p. 188-196, 2005.

SAUNIER, O.; BOCQUET, M.; MATHIEU, A.; ISNARD, O. Model reduction via principal component truncation for the optimal design of atmospheric monitoring networks. **Atmospheric Environment**, v. 43, p. 4940-4950, 2009.

SHIROMA C.; SAONA, L. R. Application of NIR and MIR spectroscopy in quality control of potato chips. **Journal of Food Composition and Analysis**, v. 22, p. 596-605, 2009.

SILVA, M. A. M.; FERREIRA, M. H.; BRAGA, J. W. B.; SENA, M. M. Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy. **Talanta**, v. 89, p. 342-351, 2012.

SKOOG, D. A.; HOLLER, F. J.; CROUCH, S. R. **Princípios de Análise Instrumental**, 6 ed. Porto Alegre: Editora Bookman, 2007. 1055 p.

SPEIGHT, J. Chemical Industries In: **The chemistry and technology of petroleum**. New York. 1998.

SUAREZ, J. R. C.; LONDOÑO, L. C. P.; REYES, M. V.; DIEM, M.; TAGUE, T. J.; RIVERA, S. P. H. **Fourier Transforms - New Analytical Approaches and FTIR Strategies**: Open-Path FTIR Detection of Explosives on Metallic Surfaces. Rijeka: InTech, 2011. 520p.

TARUMI, T.; SMALL, G. W.; COMBS, R. J.; KROUTIL, R. T. Infinite impulse response filters for direct analysis of interferogram data from airborne passive Fourier transform infrared spectrometry. **Vibrational Spectroscopy**, v. 37, p. 39-52, 2005.

TEIXEIRA, L. S. G.; OLIVEIRA, F. S.; SANTOS, H. C. S.; CORDEIRO, P. W. L.; ALMEIDA, S. Q. Multivariate calibration in Fourier transform infrared spectrometry as a tool to detect adulterations in Brazilian gasoline. **Fuel**, v. 87, p. 346-352, 2008.

THOMAS, J. E. **Fundamentos de Engenharia de Petróleo**. 2. ed. Rio de Janeiro: Editora Interciência, 2004. 269p.

TOLMASQUIM, M. T.; GUERREIRO, A.; GORINI, R. Matriz energética brasileira: Um prospectiva. **Novos Estudos**, p.47-69, 2007.

XU, Q. S.; LIANG, Y. Z. Monte Carlo cross validation. **Chemometrics and Intelligent Laboratory Systems**, v. 56, p. 1-11, 2001.

WEIMIN, L.; WENKAI, L.; HUI, C. W. Integrating neural network models for refinery planning. **Process Systems Engineering**, v. 15, p. 1304-1309, 2003.

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

WU, D. S.; FENG, X.; WEN, Q.Q. The Research of Evaluation for Growth Suitability of *Carya Cathayensis* Sarg. Based on PCA and AHP. **Procedia Engineering**, v. 15, p. 1879-1883, 2011.

ZAMORA, P. P.; PONCE, L. C.; NAGATA, N.; POPPI, R. J. Alternativas quimiométricas para a resolução de problemas analíticos clássicos. Determinação espectrofotométrica de misturas de zircônio e háfnio. **Química Nova**, v. 20, p. 469-474, 1997.

ZHANG, G.; NI, Y.; CHURCHILL, J.; KOKOT, S. Authentication of vegetable oils on the basis of their physico-chemical properties with the aid of chemometrics. **Talanta**, v. 70, p. 293-300, 2006.

ANEXO A – Autorização para utilização dos dados da Petrobras



DECLARAÇÃO

São José dos Campos, 14 de dezembro de 2010

Para: Faculdade de Engenharia de Guaratinguetá – FEG/UNESP

Empregado: Leandro Valim de Freitas
Matricula: 971952-0

Declaro para os devidos fins que o referido empregado tem autorização para utilizar os dados gerados no Laboratório da Refinaria Henrique Lage, constantes em sua monografia do curso de Especialização em Gestão da Produção e em sua dissertação de Mestrado em Engenharia Mecânica (Gestão e Otimização).

Atenciosamente

Pedro Luiz Corrêa Garcia
Gerente Setorial de Desenvolvimento e Qualidade de Produtos
Refinaria Henrique Lage